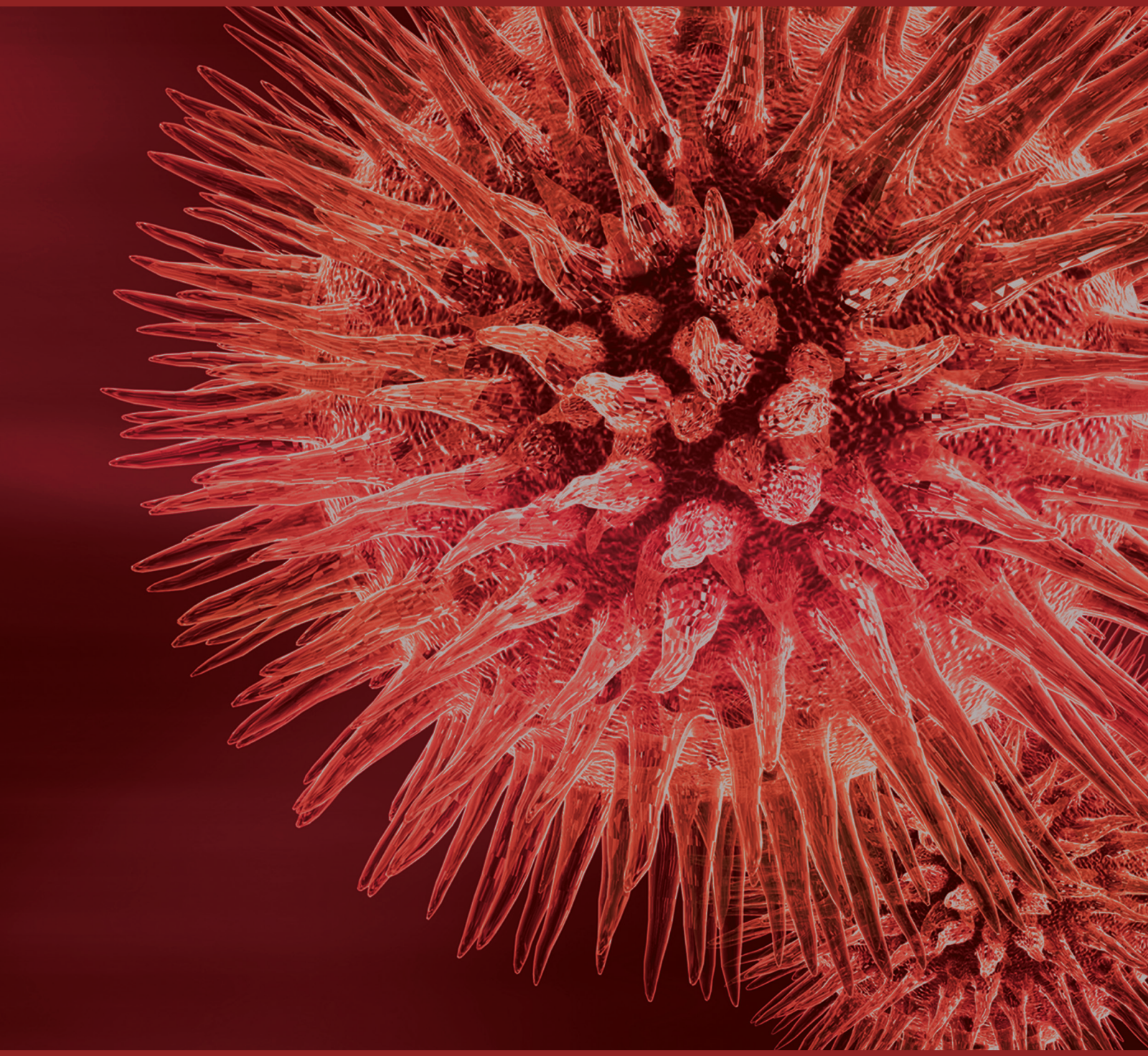
BioMed Research International

Big Data and Network Biology

Guest Editors: Shigehiko Kanaya, Md. Altaf-Ul-Amin, Samuel Kuria Kiboi, and Farit Mochamad Afendi



Big Data and Network Biology

BioMed Research International

Big Data and Network Biology

Guest Editors: Shigehiko Kanaya, Md. Altaf-Ul-Amin, Samuel Kuria Kiboi, and Farit Mochamad Afendi

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 836708, 2 pages
http://dx.doi.org/10.1155/2014/836708

Editorial

Big Data and Network Biology

Shigehiko Kanaya,[1] Md. Altaf-Ul-Amin,[1] Samuel Kuria Kiboi,[2] and Farit Mochamad Afendi[3]

[1] *Nara Institute of Science and Technology (NAIST), Nara 630-0192, Japan*
[2] *University of Nairobi, Nairobi 00100, Kenya*
[3] *Bogor Agricultural University, Java Barat 16680, Indonesia*

Correspondence should be addressed to Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 19 May 2014; Accepted 19 May 2014; Published 15 June 2014

With the data deluge caused by the recent high throughput experiments in molecular biology emerged the popular topics such as big data biology and network biology aiming at understanding life as a system by integrating and applying knowledge and facilities of different branches of science including mathematics, physics, statistics, chemistry, computer science, and information technology. Naturally, the spectrum of topics under big data and network biology is widespread and the present special issue is not an exhaustive representation of the subject. Nonetheless the articles selected for this special issue represent recent trends and versatile knowledge concerning the title topic, that we have the pleasure of sharing with the readers. Data-intensive sciences like contemporary biology consist of three basic activities: capture, curation, and analysis. Being in the bioinformatics domain, this special issue mainly focuses on analysis; that is, it contains articles about novel tools and methodologies for data analysis and mining and review articles describing databases, tools, and algorithms useful for curation and analysis of biological data.

This special issue contains fifteen papers. Three papers discuss software tools for analyzing different "omics" data. Three other papers are review papers discussing versatile aspects of systems biology. Two of the papers present biomedical text mining approaches. The other seven papers are methodology articles related to genomics, transcriptomics, proteomics, metabolomics, and herbal medicines.

The paper "*Systems biology in the context of big data and networks,*" which is a review paper, gives an overview of the progress in big data biology and data handling and also introduces some applications of networks and multivariate analysis in systems biology.

The paper "*AmalgamScope: merging annotations data across the human genome*" presents a new interactive software tool developed to assist scientists with annotation of the human genome and in particular the integration of the annotations from multiple data types, using gene identifiers and genomic coordinates. Supported platforms include next-generation sequencing and microarray technologies.

The paper "*Integration of residue attributes for sequence diversity characterization of terpenoid enzymes*" first determined important metrics describing the biochemical and physical attributes of amino acids. It utilized random forest algorithm to reduce redundancies in the amino acid index. This research contributes a different mechanism toward handling protein sequences. It especially quantifies the sequence information to numerical scale and thus facilitates the application of computational algorithms.

The paper entitled "*OWL reasoning framework over big biological knowledge network*" presents a general OWL (web ontology language) reasoning framework to systematically study and reveal the implicit relationships among biological entities from big biological networks. In their experiment, the authors focused on association between traditional Chinese medicine (TCM) and Western medicine (WM). The derived associations are useful for biologists to promote the development of novel drugs and also for modernization of TCM.

The paper entitled "*A knowledge-driven approach to extract disease-related biomarkers from the literature*" developed a text mining approach to extract a dataset of biomarkers related to diseases covering all therapeutic areas by

exploiting a large literature database. Additionally, this work presents a bibliometric analysis of the journals reporting biomarker related information during the last 40 years.

The paper "*Integrated analysis of gene network in childhood leukemia from microarray and pathway databases*" delineates differential responses of acute lymphoblastic leukemia (ALL) subtypes, B-ALL and T-ALL, to glucocorticoids (GCs) treatment by identifying the differences among biological processes, molecular pathways, and interaction networks that emerge from the action of GCs.

The paper "*Tools and databases of the KOMICS web portal for preprocessing, mining, and dissemination of metabolomics data*" presents a metabolomics web portal which includes the tools and databases for preprocessing, mining, visualization, and publication of metabolomics data. Metabolomics research is increasingly utilized for applications such as disease diagnostics, biomarker discovery, and assessment of food quality.

The paper entitled "*Supervised clustering based on DPClusO: prediction of plant-disease relations using Jamu formulas of KNApSAcK database*" proposes a new approach to predict the relation between effective therapeutic plant and disease using network analysis and supervised clustering based on the ingredient data of Indonesian herbal medicines called Jamu. Scientific analysis of traditional medicines is important because such medicines have been developed through hundreds of years of human experience.

The paper entitled "*A novel feature selection strategy for enhanced biomedical event extraction using the Turku system*" proposed a method to enhance the performance of Turku Event Extraction System (TEES). This work developed and applied an accumulated effect evaluation (AEE) algorithm to identify important features for text-mining classifiers.

The paper entitled "*A novel bioinformatics method for efficient knowledge discovery by BLSOM from big genomic sequence data*" constructed oligonucleotide BLSOM corresponding to a wide range of vertebrate genomes and detected differences between human and mouse genomes. Due to its high classification and visualization power, BLSOM recognized the species-specific key combination of oligonucleotide frequencies in each genome, described as "genome signature."

The paper titled "*Applied graph-mining algorithms to study biomolecular interaction networks*" is a review paper. Graph comparison and module detection are the two most commonly used strategies for analyzing PPI or other biological networks. This paper summarizes the current literature on graph kernel and graph alignment methods for graph comparison, as well as a variety of module detection approaches including seed-and-extend, hierarchical clustering, optimization-based, probabilistic, and frequent subgraph methods.

The paper "*An unsupervised approach to predict functional relations between genes based on expression data*" first digitizes the log-ratio type gene expression data to a matrix consisting of 1, 0, and −1 indicating highly expressed, no major change, and highly suppressed conditions, respectively, for genes. For each pair of genes, a probability density mass function table is constructed indicating nine joint probabilities and those probability values were intelligently utilized to find functional relations between genes.

The paper entitled "*Survey of network-based approaches to research of cardiovascular diseases*" is a review article on cardiovascular diseases (CVDs) which are the leading health problems worldwide. Biomolecular interaction networks generated from available data are excellent platforms for understanding linkage of all processes within a living cell, including processes that underlie diseases. This work reviewed approaches that explore and use relationships between topological properties of biological networks and mechanisms underlying CVDs.

The paper "*Essential functional modules for pathogenic and defensive mechanisms in Candida albicans infections*" adopted a systems biology approach to construct the early-stage and late-stage protein-protein interaction (PPI) networks for both *C. albicans* and zebrafish and, by comparing those PPI networks, identified several critical functional modules in both pathogenic and defensive mechanisms.

The paper entitled "*Visualization of genome signatures of eukaryote genomes by batch-learning self-organizing map with a special emphasis on Drosophila genomes*" generated the BLSOMs for tetra- or pentanucleotide composition in approximately one million sequence fragments derived from 101 eukaryote genomes. BLSOM recognized phylotype-specific characteristics (e.g., key combinations of oligonucleotide frequencies) in the genomic sequences, by clustering the sequences without adding any prior information regarding the species.

Acknowledgments

We heartily thank the authors for their excellent and fundamental contributions and their patience in communicating with us. Finally we acknowledge the dedicated works of all reviewers of these papers for their critical and helpful comments.

Shigehiko Kanaya
Md. Altaf-Ul-Amin
Samuel Kuria Kiboi
Farit Mochamad Afendi

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 428570, 11 pages
http://dx.doi.org/10.1155/2014/428570

Review Article

Systems Biology in the Context of Big Data and Networks

Md. Altaf-Ul-Amin,[1] Farit Mochamad Afendi,[2] Samuel Kuria Kiboi,[3] and Shigehiko Kanaya[1]

[1] *Nara Institute of Science and Technology, Japan*
[2] *Bogor Agricultural University, Indonesia*
[3] *University of Nairobi, Kenya*

Correspondence should be addressed to Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 17 January 2014; Revised 8 April 2014; Accepted 1 May 2014; Published 27 May 2014

Academic Editor: Aparup Das

Science is going through two rapidly changing phenomena: one is the increasing capabilities of the computers and software tools from terabytes to petabytes and beyond, and the other is the advancement in high-throughput molecular biology producing piles of data related to genomes, transcriptomes, proteomes, metabolomes, interactomes, and so on. Biology has become a data intensive science and as a consequence biology and computer science have become complementary to each other bridged by other branches of science such as statistics, mathematics, physics, and chemistry. The combination of versatile knowledge has caused the advent of big-data biology, network biology, and other new branches of biology. Network biology for instance facilitates the system-level understanding of the cell or cellular components and subprocesses. It is often also referred to as systems biology. The purpose of this field is to understand organisms or cells as a whole at various levels of functions and mechanisms. Systems biology is now facing the challenges of analyzing big molecular biological data and huge biological networks. This review gives an overview of the progress in big-data biology, and data handling and also introduces some applications of networks and multivariate analysis in systems biology.

1. Introduction

Biology has recently become a "big-data science" mainly supported by the advances in high-throughput experimental technologies. Data-intensive science consists of three basic activities: capture, curation, and analysis [1]. All three of these phases of handling big data raise many new research challenges to pursue in systems biology. The big data challenges are not only their size but also their increasing complexity. The emergence of big data biological sciences, such as systems biology, and their growing impact on health, nutrition, ecosystems, and other societal issues have only recently become the focus of scholars in social studies, science, and information studies [2]. Jim Gray proposed the fourth data paradigm and farming of the "data deluge;" that is, the capacity to measure, store, analyze, and visualize data is the new reality to which science must adapt. The heart of the fourth paradigm is data and it sits alongside empiricism (1st paradigm), theory (2nd paradigm), and simulation (3rd paradigm), which together form the continuum we think of as the modern scientific method [1]. Systems biology is one of the several other subjects including astronomy, ecology, and meteorology where challenges of the fourth data paradigm have become relevant. The basic purpose of systems biology is the system-level understanding of a cell or an organism, which can be summarized in the context of molecular networks as (1) an understanding of the structure of all the components of a cell/organism up to molecular level, (2) the ability to predict the future state of the cell/organism under a normal environment, (3) the ability to predict the output responses for a given input stimulus, and (4) the ability to estimate the changes in system behavior upon perturbation of the components or the environment. In a cell or organism the primary-level components, for example, the molecules, are of numerous types and numbers and hence system-level understanding of a cell/organism is still a very difficult

task. However along the way to achieve the theoretical goal of systems biology, that is, to understand life scientifically, many other practical applications will be invented. Practical applications will include development of new generation medical tests, drugs, foods, fuel, materials, sensors, and other applications. Systems biology now faces the challenges of analyzing large amounts of molecular biological data and huge biological networks.

2. Big Picture of Hierarchy and Networks in Systems Biology

The hierarchy shown in Figure 1(a) summarizes the major types of molecules being studied in systems biology, which aims to determine the functions of the molecules of each layer and how these molecules interact with each other within individual layers and between layers to perform biological tasks. Figure 1(b) shows the overview of the accumulated data in the KNApSAcK database which has been developed to facilitate the knowledge discovery regarding plants and plant-human omics [3]. The upper part of Figure 1(a) can be regarded as an example of a big picture of networks in systems biology. This figure implies the existence and abstraction of networks in individual species and across species. Numerous studies constructed suitable networks for understanding systems or subsystems within species. Networks representing systems or subsystems can also be compared or linked between species (Figure 1(b)). This world is cohabitated by humans and many other species and the understanding of the interactions at the molecular level among all the species is important for healthy and sustainable living for humans and other organisms.

3. Data Types in Systems Biology

Many experiments are conducted in systems biology like many other branches of science; these experiments produce various types of data. Currently in systems biology some of the popularly-used data types are as follows.

3.1. Sequences. The DNA is a molecule of double helix structure that consisted of two complementary strands of sequences of four nucleotide bases—adenine, thymine, guanine, and cytosine, represented as A, T, G, and C, respectively [4]. DNA contains all the necessary information preserved in the order of the nucleotide sequences. Hence, it is important to determine the sequences accurately. A gene is usually a continuous part of one of the DNA strands and contains codes for one or a few different proteins. The proteins are essential molecules that consisted of amino acid sequences. From the starting site of a gene, every three nucleotides are called a codon and a codon corresponds to an amino acid. It is in this way that a gene preserves the code of a protein. For example ATGAAGCTACTGTCTTCTATCGAACAAGCATGCGAT is the sequence of the first 30 nucleotides of GAL4 gene of yeast and KLLSSIEQAC is the sequence of the first 10 amino acids of the corresponding protein. There is variation in codon usage by different organisms and links can be established between codon usage and the biological characteristics of an organism [5, 6]. Development of DNA sequencing techniques started in the 1970s and since then various methods have been developed [7–9]. The sequence of individual genes, group of genes, parts of chromosomes, full chromosomes, or entire genomes are determined for different purposes. Recent developments in next generation sequencing techniques have greatly reduced the time and cost of sequencing [10–12].

3.2. Molecular Structure. Determination and prediction of the three-dimensional structures of omics molecules are also very important. DNA is packed into protein-DNA structures referred to as chromatin mainly to fit the long DNA chain inside the cell. The primary protein components of chromatin are histones. The DNA packaging protects DNA from damage and plays important roles in gene regulation by allowing or blocking the binding of transcription factors and other molecules to DNA. Usually, proteins also work by forming complexes with other proteins. In general it can be stated that DNA, RNA, and protein molecules usually bind with one another dynamically to perform different cellular functions. Therefore, not only the sequences but also the three-dimensional structures of the omics molecules are important for predicting the possibility of binding between molecules and thus predict the functions of uncharacterized molecules. X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy are the experimental procedures used for determining the 3D structures of proteins. A number of methods for the computational prediction of protein structure from its sequence have been developed [13, 14]. Also, there are computational methods for the prediction of RNA structures [15–17]. There are numerous software tools for predicting and visualizing 3D structures of proteins and RNAs. A comprehensive list of these tools can be found in scientific literature. Molecular structure data are therefore three-dimensional geometrical figures of versatile shapes or related information that can be easily converted to three-dimensional structures usually with the aid of computer software.

3.3. Gene Expression. Gene expression is the process of extracting information of a gene and is the initial step of producing gene products such as mRNAs which are usually translated to proteins and functional RNAs such as rRNA or tRNA. Gene expression is known to take place in all life forms, that is, eukaryotes (unicellular and multicellular), prokaryotes (bacteria and Archaea), and viruses—to generate the macromolecular machinery and building blocks for life. Though most cells in an organism contain the same genes, not all of the genes are used in each cell. Some genes are turned on, or "expressed," when needed in particular types of cells. Microarray technology [18, 19] allows us to look at many genes at once and determine which are expressed in a particular cell type and to what extent. Next generation sequencers are also currently used to determine the gene expression [20]. To say that "a gene is highly expressed" means many copies of mRNA corresponding to that gene are produced in the cell. The extent of expression of genes is

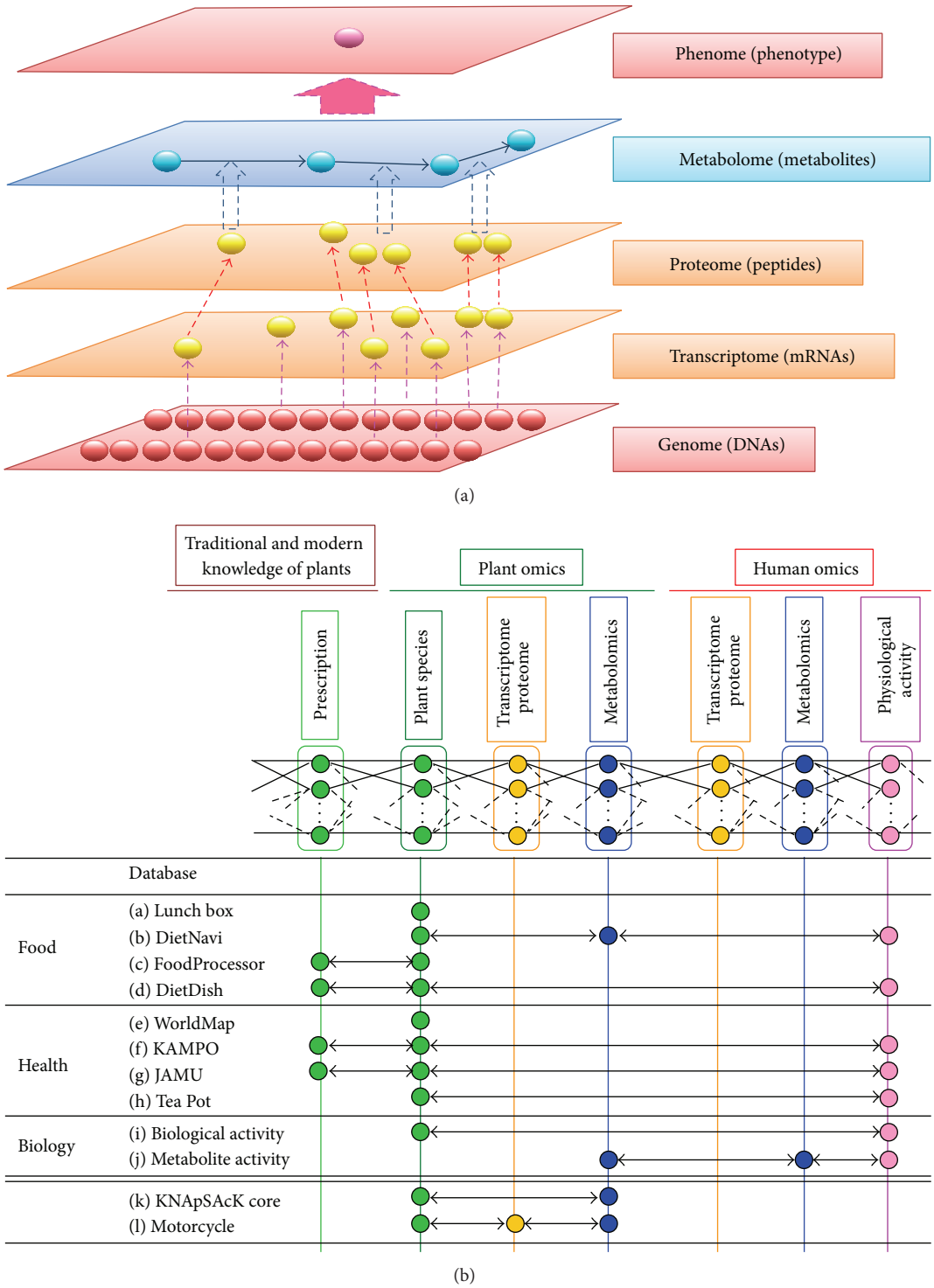


Figure 1: (a) Hierarchy of the omics molecules and (b) integrated platform of knowledge in KNApSAcK database for plants and plant and human omics.

usually measured for comparison by using samples collected under different experimental conditions, for example, sick and healthy tissues, normal cells or cells put under certain stress or starving. Gene expression data is usually represented as a matrix where the rows represent genes and the columns represent experimental conditions; that is, gene expression data are multivariate data.

3.4. Binding Sites and Domains. Many important cell processes such as RNA transcription, DNA packing, DNA

replication, DNA recombination, and DNA repair are initiated and regulated by binding of proteins to selected DNA sequences. A position weight matrix (PWM) is a commonly-used representation of motifs (patterns) in biological sequences [21]. A PWM, also called position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), is a matrix of score values that gives a weighted match to any given substring of fixed length. A DNA-binding domain (DBD) of a protein is an independently folded domain that contains at least one motif that identifies and binds double- or single-stranded DNA. A DBD can recognize a specific DNA sequence usually known as a recognition sequence or have a general affinity to DNA [22]. The domains of proteins and the binding sites at DNA are therefore part of the sequences of the corresponding proteins and the DNA, respectively.

3.5. Protein-Protein Interaction (PPI). In cells, thousands of different types of proteins act as enzymes-catalysts to chemical reactions of the metabolism, components of cellular machinery (e.g., ribosomes), regulators of gene expression, and so on. Some proteins play specific roles in special cellular compartments, whereas others move from one compartment to another carrying mass or information. A protein may work as an individual entity, but usually two or more proteins bind together and form a complex to carry out their biological functions. The RNA polymerase, a large molecular machine that copies information from DNA to produce mRNA, is indeed a big protein complex that consisted of many proteins. Proteins are generally bound together in a complex not by chemical bonds but by other forces. Usually PPI data are represented as binary relation between two proteins whether they are part of two-protein complex or multi-protein complex. All or a number of the PPIs of an organism can be represented as a network where a protein represents a node and an interaction represents an edge. Experiments that are used to determine PPIs are yeast two hybrid system (Y2H) [23, 24], affinity purification coupled to MS (AP-MS) [25], and so forth.

3.6. Mass Spectrometry. Mass spectrometry (MS) is an analytical technique that produces spectra (singular spectrum) of the masses of the molecules comprising a sample. The spectra are used to determine the elemental composition of a sample, the masses of particles and of molecules, and to elucidate the chemical structures of molecules, such as peptides, metabolites, and other chemical compounds. Mass spectrometry works by ionizing chemical compounds to generate charged molecules or molecule fragments and measuring their mass-to-charge ratios [26]. Mass spectrometry data can be represented as 2- (molecular weight versus magnitude) or 3- (molecular weight versus magnitude versus time) dimensional arrays; that is, they can be treated as multivariate data.

3.7. Metabolic Pathways. Living cells generate energy and produce building material for cell components and replenishing enzymes by the process of metabolism. All organisms live and grow by receiving food and nutrients from the environment. The foods are processed through thousands of reactions. In cells chemical reactions take place around-the-clock, constantly breaking and making chemical molecules and transferring ions and electrons. These reactions are called metabolic pathways. All or a group of known metabolic reactions of an organism can be represented as a network where metabolites are considered as nodes and a reaction between them is represented as edges. The edges in metabolic pathways correspond to one or more enzymes. Metabolic reactions follow the laws of physics and chemistry and thus modeling of metabolic reactions requires considering many physicochemical constraints [27]. In summary, it can be said that in terms of structure, extensively-used data in systems biology consist of four types: sequence data, 3D-structure data, multivariate data, and network data. However, the present challenge is that the amount of data is expanding rapidly requiring new tools and algorithms for handling big data. One type of data can be converted to another type for convenience of analysis. In the following section we discuss how networks can be generated from multivariate data and sequence data.

4. Network Generation from Different Data Types

In multivariate data, entities are represented by multiple variables and each entity can be regarded as a point in a multidimensional space or as a profile wave sketched according to the data values. Therefore, to convert multivariate data to a network, it is necessary to use a metric or some kind of measure that can assess distance or similarity between two multivariate entities. Widely-used distance or similarity measures are Euclidean distance, Manhattan distance, Mahalanobis distance, Correlation, and so forth [28–30]. The value of correlation ranges from −1 to +1 and the higher the value between two multivariate entities the more similar the entities. The opposite of distance can be used as a measure of similarity. Usually similarity between each pair of entities is calculated and then a threshold similarity is decided based on statistical analysis or some other important criteria, for example, to ensure scale-free degree distribution of the generated network or something like that. After selecting the threshold, all entities of the multivariate data are considered as the nodes of a network and an edge is inserted between the pair of entities for which the similarity is more than the threshold. A weighted network can be constructed by considering the similarity values as the weights of the edges. Sometimes one type of network is converted to another type for the convenience of applying algorithms or for some other purposes. In [31] the metabolic pathways are converted to a simple network of enzymes/genes. After that, graph spectral clustering was applied to the converted networks corresponding to *M. tuberculosis*, *M. leprae,* and *E. coli*. It was observed that reactions belonging to fatty acid biosynthesis and the FAS-II cycle of the mycolic acid pathway in *M. tuberculosis* form distinct, tightly connected subclusters. Also, based on degree centrality and eigenvector centrality the important

genes in the networks were determined and their functions were analyzed. In [32] a PPI network was converted to the corresponding line graph for the convenience of applying a clustering algorithm. The conversion to line graph helped to place the related proteins to densely connected regions or clusters and thus paved the way to obtain useful results by the application of a graph clustering algorithm.

5. Big Biological Databases

Curation and analysis become important after capturing data from various experiments. Curation includes storage, retrieval, spreading around the world, filtering and integrating the data. The engineering techniques for these jobs are already known, but when that data is in the petabyte scale, it becomes complicated. Algorithms and software tools developed for the analysis of biological data also face the problems of scalability when data becomes very big. However, many big databases have been created around the world for curation and analysis of biological data and their data volume and performance are gradually improving. DNA Data Bank of Japan [33] and GenBank [34] are big databases of primary nucleotide sequences of many organisms which are related to the bottom level (Genome) of the hierarchy shown in Figure 1(a). PGDBj is a portal website for the integration of plant genome-related databases [35]. Gene expression omnibus (GEO) from NCBI is a data repository of array- or sequence-based gene expression profiles. ATTED-II is a database of coexpressed genes [36]. Information about noncoding RNA (ncRNA) families and other structured RNA elements can be found in Rfam database [37]. For the sequences and annotations of microRNAs, a useful database is miRBase [38]. GEO, ATTED-II, Rfam, and miRBase are related to the transcriptome level of Figure 1(a). UniProt is a comprehensive and freely accessible database of protein sequences and functional information of proteins [39]. The PROSITE database [40] consists of entries describing the protein families, domains, and functional sites as well as amino acid patterns and profiles of them. BIND [41] and BioGRID [42] are databases of protein-protein interactions. UniProt, PROSITE, BIND, and BioGRID are related to the proteome level of Figure 1(a). A central archive of macromolecular structural data is wwPDB [43]. The data accumulated in wwPDB is freely and publicly available to the global community. There are four member sites of wwPDB as follows: RCSB PDB (USA), PDBe (Europe), PDBj (Japan), and BMRB (USA). NetPath [44] is a manually curated database of signal transduction pathways in human. For metabolic pathways KEGG is a rich and well known database. KNApSAcK is a metabolomics database which was initially developed as a species metabolite relational database [45] and afterwards extended to KNApSAcK family databases containing information about herbal medicines [46, 47] and metabolite activities [3]. KEGG and KNApSAcK are mainly associated with metabolome level of Figure 1(a). A comprehensive list of the omics databases can be found by searching the internet with the term "list of biological databases."

6. Multivariate Analysis in Systems Biology

After capture and curation of data, the next step is analysis. Algorithms for analyzing multivariate data developed for other applications are currently used extensively in systems biology. The well-known methods for handling multivariate data are related to dimension reduction, clustering, classification, and regression. Often, dimension reduction is done before applying other methods. Principal component analysis (PCA) is the popular algorithm for dimension reduction [48]. PCA is a mathematical process that converts the values of a set of possibly-correlated variables into a set of values of uncorrelated variables which are called principal components. This transformation assigns the largest possible variance to the first principal component and usually the sum of variance of first few components approaches the total variance of all the variables in the original data. Therefore, variable reduction is performed by replacing all the original variables by the first few components obtained from PCA analysis.

Regression analysis is a process for estimating the relationships between dependent variables (response variable) and independent (predictor) variables. Most regression analysis techniques estimate coefficients to establish a linear relation between dependent and independent variables. Least squares regression [49] and partial least squares (PLS) regression [50] are popular regression techniques. In multivariate data analysis, classification is the problem of identifying the category of a new observation from among a set of categories. Support vector machine (SVM) is a popular algorithm for classifying multivariate entities into two categories [51]. A multivariate entity can be regarded as a point in a multidimensional space. Usually an optimum hyperplane is determined based on training data so that multivariate entities of one category fall on one side of the hyperplane, while the entities of the other category fall on the other side. The concept of SVM can be extended to classify multivariate entities into multiple categories. Another type of classifier is the neural network [52], which is a naïve way of electronically simulating the function of the human brain. It is difficult to make a single formal definition of all the methods considered neural networks in the scientific literature. Usually, a neural network consists of a layer of input nodes and a layer of output nodes and several hidden layers of nodes. A neural network can be trained to use it as a classifier of multivariate entities. A multivariate data vector can be applied to the input nodes and after mathematically processing values applied at the input nodes by functions associated to the hidden nodes some values are propagated to the output nodes, which are utilized to determine the class of the input multivariate entity. The functions associated to the hidden nodes are determined or optimized based on the training data. The naïve Bayes classifier [53] is another popular supervised classification technique applicable to multivariate data. This classification algorithm is named after Thomas Bayes (1702–1761), who proposed the Bayes theorem. However it is called naïve Bayes because it naively assumes that the features or variables that describe a multivariate entity are mutually independent. Naïve Bayes classifier usually computes the probability that

a multivariate entity belongs to a certain class given its features. Usually a set of training data or well-defined probability density functions are used to estimate different probabilities required to classify a multivariate entity. Random forest [54] is another classification method. The random forest is an ensemble classifier which constructs multiple decision trees. Each tree is constructed using a subset of training data and a subset of variables. Class assignment is made by the number of votes from all of the trees. Random forests can also be used to rank the importance of the variables in a regression or classification problem. Some other classification algorithms are partial least squares discriminant analysis (PLS-DA) [55] and soft independent modeling of class analogy (SIMCA) [56].

Another multivariate technique common in systems biology is clustering. This is the task of dividing a set of entities or objects into several groups or clusters in such a way that the objects in the same cluster are more similar in some sense to each other than to those in other clusters. Clustering and classification are related concepts, but in the case of classification, the categories are known beforehand, whereas in case of clustering, usually the categories are understood after applying a clustering algorithm. Hierarchical clustering [57, 58] is the widely used algorithm for clustering of multivariate data. Hierarchical clustering is subdivided into 2 types: agglomerative methods and divisive methods. Agglomerative methods proceed by a series of fusions of the objects into groups eventually encompassing all objects in a single group. On the other hand, the divisive method separates the objects successively into finer groupings, eventually keeping each object in a single group. Hierarchical clustering is a technique that organizes elements into a tree. *K*-mean clustering [59] and self-organizing mapping (SOM) [60, 61] are also important clustering algorithms applicable to multivariate data. *K*-mean is one of the simplest unsupervised clustering methods. One disadvantage of *K*-mean clustering is that it is necessary to guess and set the number of clusters in the targeted dataset before applying the algorithm. In case of SOM, multidimensional data/input vectors are mapped onto a two-dimensional array of nodes. Data points assigned to a node or nearby nodes are considered as a cluster.

Data assimilation can be referred to as state estimation which is the process of combining a model with observational data to estimate the state of a system. By data assimilation, a quantity of interest is estimated by combining observational data with the underlying dynamical principles governing the system under investigation. There are applications of data assimilations in systems biology. The data assimilation technique was applied to elucidate the dynamics of time-lagged gene-to-metabolite networks of *Escherichia coli* [62]. State transitions in the transcriptome of *Bacillus subtilis* and in both transcriptome and metabolome of *Arabidopsis thaliana* were predicted using a data assimilation technique called linear dynamical system model [63].

Numerous researches have been conducted in systems biology based on multivariate data analysis. We briefly discuss a few examples below.

6.1. Application of BL-SOM. Batch learning self-organizing map (BL-SOM) is a novel neural-network algorithm that has been applied to efficiently and comprehensively analyze codon usage in approximately 60,000 genes from 29 bacterial species simultaneously [61]. In the original SOM method [60], the initial weight vectors are set by random values, but in BL-SOM the vectors are initialized by PCA, which is a statistical method that performs linear mapping to extract optimal features from an input distribution in the mean squared error sense. This technique allows the resulting SOM to be independent of input vectors. BL-SOM makes it possible to cluster and visualize the genes of individual species separately at a much higher resolution than can be obtained with PCA because PCA works based on linear mechanism while SOM can be trained to adapt non-linear mechanisms. The organization of the SOM can be explained by the genome G + C and tRNA compositions of the individual species. This work further used SOM to examine codon usage heterogeneity in the *E. coli* O157 genome, which contains "O157-unique segments" (O-islands), and showed that SOM is a powerful tool for characterization of horizontally transferred genes. Another example of the application of BL-SOM is the investigation of the enzyme sequence diversity related to secondary metabolism [64]. Initially, a map was constructed by using a big data matrix that consisted of the frequencies of all possible dipeptides in the protein sequence segments of plants and bacteria. The enzyme sequence diversity of the secondary metabolic pathways was examined by identifying clusters of segments associated with certain enzyme groups in the resulting map. The extent of diversity of fifteen secondary metabolic enzyme groups was discussed. On the resulting map, six clusters were rich with fragments of monoterpene, sesquiterpene, diterpene, and triterpene synthases. Nine clusters are corresponding to eight types of phenylpropanoids which are flavonoid and isoflavonoid synthases. Five clusters were associated to acetyl-, O-methyl-, and N-methyl transferases. As a whole these results show sequence similarities between specific types of enzymes related to secondary metabolic pathways.

6.2. Application of PLS-DA Model. PLS-DA is an example of a multivariate model that has been applied in systems biology a case study being our previous work on Indonesian herbal medicines, popularly known as Jamu. These medicines are prepared from a mixture of several plants. The plants are chosen so that the Jamu has the desired efficacy. Thus, the composition of the plants used in a Jamu formula determines its efficacy. A model using partial least square discriminant analysis (PLS-DA) has been developed to predict the efficacy of Jamu based on the information of plants used in Jamu formula [55]. In this analysis, among 3,138 Jamu medicines, the efficacies of 2,248 Jamu medicines (71.6) were correctly predicted. Hence, the efficacy in most Jamu medicines can be predicted on the basis of the ingredient medicinal plants. In addition, the regression coefficients of the PLS-DA model, which relates plant usage in Jamu as predictors and Jamu efficacy as response, can be helpful in determining which plants in the ingredients of Jamu are used as main ingredients, which contribute primarily to the medicines' efficacies, and

which plants are used as supporting ingredients. Plants that act as main ingredients will have a significant effect on the developed model. Due to the absence of parametric testing for the PLS-DA coefficients, the evaluation for significance is performed using permutation testing, in which the distribution of coefficients under the null hypothesis is generated via resampling of the existing data. The resampling is performed by permuting the order of the responses (in this case, Jamu efficacies) while maintaining the order of the predictors (in this case, plant utilization as Jamu ingredients) so that the existing relationship between the predictors and the response is destroyed and a new data set is generated under the null hypothesis; that is, plant utilization in Jamu does not affect Jamu efficacy. If such resampling is performed many times and the PLS-DA model is applied on the new data generated from the resampling, the accumulation of the PLS-DA coefficients obtained from this process generates a distribution, against which a P value can be calculated and subsequently evaluated for significance. From the testing, it was observed that 234 plants (50.3 among all 465 plants) showed no significant status for all 9 efficacies; whereas the other 231 plants have a significant status of which 189 plants (40.6) are significant only for 1 efficacy, 38 plants (8.2) are significant for 2 efficacies, and the other 4 plants (0.9) are significant for 3 efficacies.

7. Network Analysis in Systems Biology

For system-level understanding, initially the elements of a system are connected based on their mutual relation and a network is formed. Global network properties such as average path length, clustering coefficient, and degree distribution [65] are determined to assess the overall characteristics of the network such as how they are formed, what model they fit, how robust they are, and how tightly the elements are connected. There are numerous algorithms for finding clusters in a network. As a flexible notion the densely connected regions of a network are called clusters. Also, there are precise definitions of network clusters such as k-core, k-plex, and n-clan [66–68]. In recent years network theory has been substantially applied in systems biology. Construction and analysis of biological networks have become highly popular among omics researchers. In the following section we discuss some of the applications of networks and network algorithms in systems biology.

7.1. Function Prediction. Functions of many omics molecules or entities, for example, genes, mRNAs, proteins, and also metabolites, are still unknown. A system-level approach of predicting functions of an unknown entity is performed by constructing a network of that entity and other known and unknown entities. Usually, after constructing a network, some suitable clustering method is applied. There are versatile graph clustering methods such as based on density and periphery [69], random walk [70], betweenness centrality [71], and so on. Usually the entities belonging to the same cluster are considered to have similar function based on the hypothesis "guilt by association" and therefore if the majority of the members of a cluster have some known function, then the unknown members are also assumed to have that function.

7.2. Protein Complex Detection. Protein molecules may act individually, but in most cases to perform a biological task they form complexes by binding with one or more other protein molecules. High throughput experiments such as yeast two hybrid system (Y2H) [23, 24] and affinity purification coupled to MS (AP-MS) [25] are used to determine the global set of interacting protein pairs. Such protein pairs can be represented as a network which is known as a PPI network. Usually it is assumed that a set of proteins in a densely connected region in a PPI network correspond to a protein complex. A good number of researches have been conducted to computationally detect protein complexes by applying clustering algorithms to PPI networks [72–76]. In those studies it was shown that real protein complexes of yeast substantially matched with computationally detected protein complexes.

7.3. Prediction of Interaction. The presence of statistically significant complementary domain pairs in interacting protein pairs determined in the context of a PPI network indicates that certain domains facilitate protein binding [77, 78]. Thus the presence of complementary domains in two new proteins implies the possibility that they might interact inside the cell. Thus, PPI networks of one or more species can be used to first determine complementary domain pairs and then to predict interactions between new protein pairs corresponding to a species.

7.4. Analyzing Evolution. PathBLAST [79] is a network alignment and search tool for comparing protein interaction networks across species to identify protein pathways and complexes that have been conserved by evolution. The basic method searches for high-scoring alignments between pairs of protein interaction paths, for which proteins of the first path are paired with putative orthologs occurring in the same order in the second path.

7.5. Information Integration. Networks can be constructed by combining different types of information, thus being helpful for integrated analysis of different omics molecules based on their relations. An integrative network of *C. elegans* embryogenesis genes based on three types of data (protein-protein interaction, expression profiling similarity, and phenotyping profiling similarity) was studied in [80]. This study showed that gene pairs connected by interactions supported by multiple data are more likely to belong to the same GO category. For example in [81] gene expression profiles and mass spectrometry profiles are merged by using appropriate normalization of the data and a combined network of genes and metabolites has been constructed which helped find related genes and metabolites. A very large network of more than 60,000 interactions was reported [82] by integrating transcription factor binding, PPI, and protein phosphorylation data of yeast. This network was found to contain 7 types of 3-molecule motifs involving kinases out of which 5 types were overrepresented.

7.6. Determination of Important Entities. It is easy to realize that in the context of a network all nodes are not equally important. For example, a node with very high degree is obviously more important compared to a node having degree 1 or 2. There is an important relation between vertex degree and functional importance of the vertices in biological networks [83]. It has been reported that in PPI networks the removal of highly connected proteins is more likely to have more lethal effect [84]. The importance of a node in a network is precisely and mathematically determined by the centrality measures, for example, degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, and so forth. In [85] a list and definitions of 17 types of different centrality measures are presented.

7.7. Disease Diagnosis. Biological networks can be utilized to identify biomarkers for disease diagnosis. Even a subnetwork also might be a biomarker. Protein network and mRNA profiles can be integrated to identify subnetwork biomarkers, that is, highly connected genes of a subnetwork whose sum of expression can be a marker of a disease state. There are several network-based approaches for identifying disease genes and protein interaction subnetworks which are disease signatures [86–88]. The application of a network analysis to metabolic PET (positron emission tomography) data obtained from patients with Parkinson's disease resulted in the identification and validation of two distinct spatial covariance patterns associated with the motor and cognitive manifestations of the disease [89].

7.8. Drug Development. Complicated diseases such as cancer, Alzheimer, mental disorder, and heart diseases are very complex and caused by multiple molecular abnormalities. The drug discovery process of these diseases needs to target not a single molecule but entire molecular pathways of various cellular omics networks. Recently biological networks, for example, PPI networks and gene expression networks, are extensively used to find drug targets [90–92]. In [93], a method for drug target identification was proposed by combining information about drug therapeutic similarity, chemical similarity, and protein-protein interaction network using linear regression.

7.9. Prediction of Drug-Drug Interactions. Understanding drug-drug interaction is important for drug development and drug administration. A drug interaction is a situation in which a substance (usually a drug) affects the activity of another drug when both are administered together. Drug-drug interaction is a significant cause of adverse drug reaction, especially in population on multiple medications. Drug-drug interaction can be categorized into three types: pharmaceutical, pharmacokinetic (PK), and pharmacodynamic (PD). A prediction method of pharmacodynamic drug-drug interaction through protein-protein interaction networks is proposed in [94]. This work introduced a metric called "S-score" that measures the strength of network connection between drug targets. Thus drug-drug interaction was determined by assessing the interaction between the drug targets in the context of the whole PPI network.

7.10. Comparison of Biological Mechanisms. Different types of biological networks, for example, PPI networks, gene regulatory networks, and metabolic pathways, and so forth, are system-level representations of biological mechanisms. Interesting results were obtained by comparing biological networks with random networks of the same size [69, 95] or biological networks derived under different contexts [96]. Usually such comparisons are performed in the context of global network properties like degree distribution, average path length, and clustering coefficient, and so forth. Though global level degree distribution of PPI networks of many species follows power law, subtle differences between PPI networks of different species can be found by using other concepts. Not only PPI network but also other types of biological networks of different species can be compared to decipher the differences in mechanisms to explain phenotypes and other useful matters. A distance measure called relative graphlet frequency distance is presented in [97] which is based on the frequency of undirected induced subgraphs of size three to five. This measure was used to compare PPI networks of *E. coli* and yeast with different artificial networks [98]. Another concept of comparing two networks especially regulatory networks is on the basis of network motifs which are reoccurring patterns in complex networks and thus in some sense similar to the motifs in gene or protein sequences. It is shown in [99] that three highly overrepresented network motifs are present in the transcriptional interaction network of *E. coli*.

8. Conclusions

To understand a living organism as a system we first need to understand a cell as a system. This means we need to comprehensively understand the functions of each gene/protein/metabolite and how they work as an individual or in a group. The advancement in molecular biological experiments is producing huge piles of data related to genome and RNA sequence, protein and metabolite abundance, protein-protein interaction, gene expression, and so on. It is important to handle these huge data efficiently and scientifically to understand the cell as a system and to develop new applications in biotechnology and biomedical fields. This, in turn, necessitates the usage of high speed computers and integrating knowledge from other branches of science, for example, statistics, mathematics, physics, chemistry, and so on. The data we need to handle is of old formats, but the present challenge is that it has grown very big and needs the integration of different data types. This can be done by developing efficient scaling techniques for the current software tools and statistical and mathematical models for data handling. The application of network theory and algorithms can facilitate analyzing and integrating big data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] G. Bell, T. Hey, and A. Szalay, "Computer science: beyond the data deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009.

[2] W. Callebaut, "Scientific perspectivism: a philosopher of science's response to the challenge of big data biology," *Studies in History and Philosophy of Science C :Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, no. 1, pp. 69–80, 2012.

[3] Y. Nakamura, F. M. Afendi, A. K. Parvin et al., "KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities," *Plant and Cell Physiology*, vol. 55, no. 1, p. e7, 2014.

[4] E. R. Kandel, J. H. Schwartz, T. M. Jessell et al., *Principles of Neural Science*, vol. 4, McGraw-Hill, New York, NY, USA, 2000.

[5] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, "Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis," *Journal of Molecular Evolution*, vol. 53, no. 4-5, pp. 290–298, 2001.

[6] Y. Xu, P. Ma, P. Shah, A. Rokas, Y. Liu, and C. H. Johnson, "Non-optimal codon usage is a mechanism to achieve circadian clock conditionality," *Nature*, vol. 494, no. 7439, pp. 116–120, 2013.

[7] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.

[8] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 2, pp. 560–564, 1977.

[9] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[10] N. Hall, "Advanced sequencing technologies and their wider impact in microbiology," *Journal of Experimental Biology*, vol. 210, no. 9, pp. 1518–1525, 2007.

[11] T. Tucker, M. Marra, and J. M. Friedman, "Massively parallel sequencing: the next big thing in genetic medicine," *The American Journal of Human Genetics*, vol. 85, no. 2, pp. 142–154, 2009.

[12] J. R. Ten Bosch and W. W. Grody, "Keeping up with the next generation: massively parallel sequencing in clinical diagnostics," *Journal of Molecular Diagnostics*, vol. 10, no. 6, pp. 484–492, 2008.

[13] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *Journal of Molecular Biology*, vol. 300, no. 4, pp. 1005–1016, 2000.

[14] Y. Zhang, "Progress and challenges in protein structure prediction," *Current Opinion in Structural Biology*, vol. 18, no. 3, pp. 342–348, 2008.

[15] V. Reinharz, F. Major, and J. Waldispühl, "Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure," *Bioinformatics*, vol. 28, no. 12, pp. i207–i214, 2012.

[16] Z. Wang and J. Xu, "A conditional random fields method for RNA sequence-structure relationship modeling and conformation sampling," *Bioinformatics*, vol. 27, no. 13, pp. i102–i110, 2011.

[17] C. Laing and T. Schlick, "Computational approaches to 3D modeling of RNA," *Journal of Physics Condensed Matter*, vol. 22, no. 28, Article ID 283101, 2010.

[18] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[19] D. A. Lashkari, J. L. Derisi, J. H. Mccusker et al., "Yeast microarrays for genome wide parallel genetic and gene expression analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 24, pp. 13057–13062, 1997.

[20] T. T. Torres, M. Metta, B. Ottenwälder, and C. Schlötterer, "Gene expression profiling by massively parallel sequencing," *Genome Research*, vol. 18, no. 1, pp. 172–177, 2008.

[21] I. Ben-Gal, A. Shani, A. Gohr et al., "Identification of transcription factor binding sites with variable-order Bayesian networks," *Bioinformatics*, vol. 21, no. 11, pp. 2657–2666, 2005.

[22] D. M. Lilley, *DNA-Protein: Structural Interactions*, IRL Press, 1995.

[23] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[24] S. V. Rajagopala, P. Sikorski, J. H. Caufield, A. Tovchigrechko, and P. Uetz, "Studying protein complexes by the yeast two-hybrid system," *Methods*, vol. 58, no. 4, pp. 392–399, 2012.

[25] A. C. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[26] O. D. Sparkman, "Mass spectrometry desk reference," *Journal of the American Society for Mass Spectrometry*, vol. 11, no. 12, p. 1144, 2000.

[27] B. O. Palsson, *Systems Biology*, Cambridge University Press, New York, NY, USA, 2006.

[28] R. Gentleman, V. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, Springer, New York, NY, USA, 2005.

[29] S. K. Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, Radius Press, 1991.

[30] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, pp. 59–66, 1988.

[31] K. D. Verkhedkar, K. Raman, N. R. Chandra, and S. Vishveshwara, "Metabolome based reaction graphs of *M. tuberculosis* and *M. leprae*: a comparative network analysis," *PLoS ONE*, vol. 2, no. 9, article e881, 2007.

[32] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks," *Proteins: Structure, Function and Genetics*, vol. 54, no. 1, pp. 49–57, 2004.

[33] E. Kaminuma, T. Kosuge, Y. Kodama et al., "DDBJ progress report," *Nucleic Acids Research*, vol. 39, no. 1, pp. D22–D27, 2011.

[34] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Research*, vol. 36, supplement 1, pp. D25–D30, 2008.

[35] E. Asamizu, H. Ichihara, A. Nakaya et al., "Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases," *Plant and Cell Physiology*, vol. 55, no. 1, p. e8, 2014.

[36] T. Obayashi, Y. Okamura, S. Ito et al., "ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants," *Plant and Cell Physiology*, vol. 55, no. 1, p. e6, 2014.

[37] S. W. Burge, J. Daub, R. Eberhardt et al., "Rfam 11.0: 10 years of RNA families," *Nucleic Acids Research*, vol. 41, no. 1, pp. D226–D232, 2013.

[38] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D68–D73, 2014.

[39] The UniProt Consortium, "Update on activities at the universal protein resource (UniProt) in 2013," *Nucleic Acids Research*, vol. 41, no. 1, pp. D43–D47, 2013.

[40] C. J. A. Sigrist, E. de Castro, L. Cerutti et al., "New and continuing developments at PROSITE," *Nucleic Acids Research*, vol. 41, no. 1, pp. D344–D347, 2013.

[41] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.

[42] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D535–D539, 2006.

[43] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, "The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D301–D303, 2007.

[44] K. Kandasamy, S. S. Mohan, R. Raju et al., "NetPath: a public resource of curated signal transduction pathways," *Genome Biology*, vol. 11, no. 1, article R3, 2010.

[45] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin et al., "KNApSAcK: a comprehensive species-metabolite relationship database," in *Plant Metabolomics*, pp. 165–181, Springer, New York, NY, USA, 2006.

[46] F. M. Afendi, T. Okada, M. Yamazaki et al., "KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research," *Plant and Cell Physiology*, vol. 53, no. 2, p. e1, 2012.

[47] F. M. Afendi, N. Ono, Y. Nakamura et al., "Data mining methods for omics and knowledge of crude medicinal plants toward big data biology," *Computational and Structural Biotechnology Journal*, 2013.

[48] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[49] J. Aldrich, "Doing least squares: perspectives from Gauss and Yule," *International Statistical Review*, vol. 66, no. 1, pp. 61–81, 1998.

[50] B. Wilson, "Using PLS to investigate interaction effects between higher order branding constructs," in *Handbook of Partial Least Squares*, pp. 621–652, Springer, New York, NY, USA, 2010.

[51] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[52] I. Aleksander and H. Morton, *An Introduction to Neural Computing*, vol. 240, Chapman & Hall, London, UK, 1990.

[53] T. M. Mitchell, *Machine Learning*, McGraw-Hill, Burr Ridge, Ill, USA, 1997.

[54] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[55] F. M. Afendi, L. K. Darusman, M. Fukuyama, M. Altaf-Ul-Amin, and S. Kanaya, "A Bootstrapping approach for investigating the consistency of assignment of plants to Jamu efficacy by PLS-DA Model," *Malaysian Journal of Mathematical Sciences*, vol. 6, no. 2, pp. 147–164, 2012.

[56] S. Wold and M. Sjöström, "SIMCA: a method for analyzing chemical data in terms of similarity and analogy," in *Chemometrics: Theory and Application*, vol. 52 of *ACS Symposium Series*, pp. 243–282, 1977.

[57] D. Defays, "An efficient algorithm for a complete link method," *The Computer Journal*, vol. 20, no. 4, pp. 364–366, 1977.

[58] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, pp. 30–34, 1973.

[59] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, p. 14, Berkeley, Calif, USA, 1967.

[60] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[61] S. Kanaya, M. Kinouchi, T. Abe et al., "Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome," *Gene*, vol. 276, no. 1-2, pp. 89–99, 2001.

[62] H. Takahashi, R. Morioka, R. Ito et al., "Dynamics of time-lagged gene-to-metabolite networks of *Escherichia coli* elucidated by integrative omics approach," *OMICS: A Journal of Integrative Biology*, vol. 15, no. 1-2, pp. 15–23, 2011.

[63] R. Morioka, S. Kanaya, M. Y. Hirai, M. Yano, N. Ogasawara, and K. Saito, "Predicting state transitions in the transcriptome and metabolome using a linear dynamical system model," *BMC Bioinformatics*, vol. 8, article 343, 2007.

[64] S. Ikeda, T. Abe, Y. Nakamura et al., "Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSAcK motorcycle database," *Plant and Cell Physiology*, vol. 54, no. 5, pp. 711–727, 2013.

[65] B. H. Junker and F. Schreiber, *Analysis of Biological Networks*, vol. 2, John Wiley & Sons, New York, NY, USA, 2008.

[66] S. B. Seidman, "Network structure and minimum degree," *Social Networks*, vol. 5, no. 3, pp. 269–287, 1983.

[67] J. Edachery, A. Sen, and F. J. Brandenburg, "Graph clustering using distance-k cliques," in *Graph Drawing*, pp. 98–106, Springer, New York, NY, USA, 1999.

[68] D. W. Matula, "k-Components, clusters and slicings in graphs," *SIAM Journal on Applied Mathematics*, vol. 22, no. 3, pp. 459–480, 1972.

[69] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinformatics*, vol. 7, article 207, 2006.

[70] S. M. van Dongen, *Graph Clustering by Flow Simulation*, 2000.

[71] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[72] M. Altaf-Ul-Amin, M. Wada, and S. Kanaya, "Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking," *ISRN Biomathematics*, vol. 2012, Article ID 726429, 11 pages, 2012.

[73] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, article 2, 2003.

[74] M. Wu, X. Li, C. K. Kwoh, and S. K. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinformatics*, vol. 10, article 169, 2009.

[75] H. C. M. Leung, Q. Xiang, S. M. Yiu, and F. Y. L. Chin, "Predicting protein complexes from PPI data: a core-attachment approach," *Journal of Computational Biology*, vol. 16, no. 2, pp. 133–144, 2009.

[76] K. Ning, *Refining Markov Clustering for Protein Complex Prediction by Incorporating Core-Attachment Structure*, World Scientific, 2009.

[77] K. Nishikata, M. Wada, H. Takahashi, K. Nakamura, S. Kanaya, and M. Altaf-Ul-Amin, "Predicting conformation of protein complexes by determining statistically significant domain-domain interactions," *Plant Biotechnology*, vol. 26, no. 5, pp. 495–501, 2009.

[78] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain-domain interactions from protein-protein interactions," *Genome Research*, vol. 12, no. 10, pp. 1540–1548, 2002.

[79] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, "PathBLAST: a tool for alignment of protein interaction networks," *Nucleic Acids Research*, vol. 32, supplement 2, pp. W83–W88, 2004.

[80] K. C. Gunsalus, H. Ge, A. J. Schetter et al., "Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis," *Nature*, vol. 436, no. 7052, pp. 861–865, 2005.

[81] F. Matsuda and K. Saito, "Integrative analysis of secondary metabolism and transcript regulation in *Arabidopsis thaliana*," in *The Handbook of Plant Metabolomics*, pp. 175–195.

[82] J. Ptacek, G. Devgan, G. Michaud et al., "Global analysis of protein phosphorylation in yeast," *Nature*, vol. 438, no. 7068, pp. 679–684, 2005.

[83] R. Albert, H. Jeong, and A. L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[84] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[85] B. H. Junker, D. Koschützki, and F. Schreiber, "Exploration of biological network centralities with CentiBiN," *BMC Bioinformatics*, vol. 7, article 219, 2006.

[86] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics*, vol. 10, article 73, 2009.

[87] R. K. Nibbe, S. Markowitz, L. Myeroff, R. Ewing, and M. R. Chance, "Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer," *Molecular & Cellular Proteomics*, vol. 8, no. 4, pp. 827–845, 2009.

[88] R. K. Nibbe, M. Koyutü, and M. R. Chance, "An integrative-omics approach to identify functional sub-networks in human colorectal cancer," *PLoS Computational Biology*, vol. 6, no. 1, Article ID 1000639, 2010.

[89] T. Eckert, C. Tang, and D. Eidelberg, "Assessment of the progression of Parkinson's disease: a metabolic network approach," *The Lancet Neurology*, vol. 6, no. 10, pp. 926–932, 2007.

[90] H. S. Lee, T. Bae, J. H. Lee et al., "Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug," *BMC Systems Biology*, vol. 6, article 80, 2012.

[91] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PLoS ONE*, vol. 4, no. 8, Article ID e6536, 2009.

[92] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, article 496, 2011.

[93] S. Zhao and S. Li, "Network-based relating pharmacological and genomic spaces for drug target identification," *PLoS ONE*, vol. 5, no. 7, Article ID e11764, 2010.

[94] J. Huang, C. Niu, C. D. Green, L. Yang, H. Mei, and J. D. J. Han, "Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network," *PLoS Computational Biology*, vol. 9, no. 3, Article ID e1002998, 2013.

[95] S. Bansal, S. Khandelwal, and L. A. Meyers, "Exploring biological network structure with clustered random networks," *BMC Bioinformatics*, vol. 10, article 405, 2009.

[96] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, no. 7006, pp. 308–312, 2004.

[97] O. Kuchaiev, A. Stevanović, W. Hayes, and N. Pržulj, "GraphCrunch 2: software tool for network modeling, alignment and clustering," *BMC Bioinformatics*, vol. 12, article 24, 2011.

[98] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[99] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 893501, 5 pages
http://dx.doi.org/10.1155/2014/893501

Research Article

AmalgamScope: Merging Annotations Data across the Human Genome

Georgia Tsiliki,[1] Konstantinos Tsaramirsis,[1,2] and Sophia Kossida[1]

[1] *Bioinformatics and Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, 115 27 Athens, Greece*
[2] *Henley Business School, Business Informatics, University of Reading, Whiteknights, Reading RG6 6UD, Uk*

Correspondence should be addressed to Georgia Tsiliki; gtsiliki@bioacademy.gr

Received 1 October 2013; Revised 21 February 2014; Accepted 18 April 2014; Published 20 May 2014

Academic Editor: Shigehiko Kanaya

The past years have shown an enormous advancement in sequencing and array-based technologies, producing supplementary or alternative views of the genome stored in various formats and databases. Their sheer volume and different data scope pose a challenge to jointly visualize and integrate diverse data types. We present AmalgamScope a new interactive software tool focusing on assisting scientists with the annotation of the human genome and particularly the integration of the annotation files from multiple data types, using gene identifiers and genomic coordinates. Supported platforms include next-generation sequencing and microarray technologies. The available features of AmalgamScope range from the annotation of diverse data types across the human genome to integration of the data based on the annotational information and visualization of the merged files within chromosomal regions or the whole genome. Additionally, users can define custom transcriptome library files for any species and use the file exchanging distant server options of the tool.

1. Background

A major advancement in the field of biomedical research is that currently researchers analyze multiple types of data, such as expression profiling, whole genome sequencing and other high-throughput experiments, which correspond to complementary views of a single organism [1]. Those diverse data types improve our ability to detect gene sets associated with a phenotype of interest. For instance, cancer genomes contain point mutations, methylation abnormalities, copy number, and expression changes not seen in normal tissues [2]. To fully comprehend those data, often called "omics" data, one needs to consult publicly available databases provided by major bioinformatics organizations, such as the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (EBI; http://www.ebi.ac.uk/). Their sheer volume accompanied with auxiliary information, justify for novel tools that are able to flexibly scale, integrate, and jointly visualize them.

Along these lines, a number of efforts have been established which differ on the volume, type, and scope of data considered. Well-known visualization tools, such as the Ensembl genome browser [3], the University of California Santa Cruz (UCSC) genome browser [4], and NCBI's MapViewer [5], are built on top of the corresponding online public databases and for that reason provide access to a broad range of annotational information for a wide variety of organisms and different genome assemblies. Many stand alone software exist which retrieve information from the above and other databases but specialize in producing summary and visualization results for specific species or data types. Recently produced visualization software mostly explore next-generation sequencing (NGS) data alongside extra annotation from a reference genome, for instance, the MapView [6], the interactive GenomeView [7], AnyExpress [8], and the MGAViewer [9] software which offers visualization options in metagenomics studies. Representative examples of tools that focus on merging information from different data studies are the generic genome browser (GBrowse) [10] and its web server WebGBrowse [11] which allows users to upload their own genomic data for display. AnyExpress [8] accepts as input aligned data, removes undesirable probes, such as sequence repeats, and generates a target-by-sample

text file. OmicsBrowse [12] connects different genomes and evolutional correspondences among multiple species derived from multiple data-servers; it displays both genetic maps and genomic annotations within wide chromosomal intervals and assists the user to select candidate genes by filtering their annotations or associated documents against user-specified keywords or ontology terms. Nevertheless, the above software accepts raw data or concentrates on analyzing user-defined chromosomal intervals and merges information from various data sources, rather than exploring data from whole array data.

The integrative genomics viewer (IGV) [13] is well suited for genome-wide exploration of NGS data. Particularly, the IGV tool has the ability to display data and dynamically group samples based on metadata supplied, giving particular emphasis on the data scaling option provided to the user. Additionally, GeneWeaver [14] is a curated repository of genomic experiments which allows the user to perform integrative functional genomics in combination with the incorporated data. When it comes to merging the data in an annotational level, in order to computationaly analyze them, it is crucial to be able to extract the data mapping and the summary information that is graphically displayed. Existing tools are not flexible enough to support custom changes, or updated versions of platform databases, whereas such options are available through less user-friendly programs. For example, the Galaxy platform [15] and also the statistical computing environment R (http://www.r-project.org/) via Bioconductor (http://www.bioconductor.org/) provide extensive annotation resources (e.g., AnnotationDBi) as well as merging capabilities (e.g., GenomicRanges, GenomicFeatures packages).

To address these issues, we present a processs which automates the matching of NGS and microarray data by scaling them to genes and also allows users to customize their database files in order to include only gene sets capturing established knowledge about biological processes and pathways. AmalgamScope (Amalgam) is an online stand alone tool with a user-friendly graphical interface which allows merging of large diverse datasets on an annotational level, as well as offering an interactive graphical user interface.

Amalgam aims to create a map of the human genome based on the annotation files supplied by the user, primarily focusing on the genomic context, whilst augmenting the available information from well-known databases. The user follows a stepwise procedure to derive the merged data, namely, a knowledgable summary of the human data uploaded, which can then be used for complementing computational analysis. The suggested pipeline greatly reduces the time, expertise, and error involved in assembling a common vocabulary for diverse data, offering the baseline for the development of analytical integration methodologies and also a common platform for reproducible exploration of the transcriptome. An important advantage of the tool is the option to upload customized input files for merging and placing equal weight on each dataset, and in a similar way, the integration of files can be based on customized library files of any species. Data supported are derived by sequencing and microarray technologies and currently include RNA sequencing (RNA-Seq), microarray gene expression, copy number variants (CNVs), and DNA methylation data, though it is possible to import any data files with entities that can be assigned to either genes or chromosomal locations in the human genome. Additionally, a special feature of Amalgam is the option to manage input or output from distant hosting servers.

2. Implementation

To merge the data uploaded, we consider a common denominator, that is, an entity which can serve as an intermediate link between the various data types. Since gene names and chromosomal locations are reported in most of the data produced by microarray and sequencing technologies, we have considered both entities as the default "regional units" where annotations from all datasets would be translated into. The user can choose a different regional unit as long as that is present in all loaded datasets and define it via the "Settings" option. Amalgam aims to identify the wider, in terms of base pairs, non-overlapping regional units across datasets which include at least one of the entities supplied. For that reason, it identifies a list of unique genomic features across datasets and by that constructs a vocabulary of regional units, where the merging scheme is based on and all visualization options are later displayed in. The output is an integrated view of the data mapped onto the human genome, along with genomic annotations from public databases.

2.1. Translating and Merging. By uploading the data files, the user is prompted to choose between local and web annotation retrievals. The local download retrieves data from three local repositories downloaded from NCBI, Ensembl (http://www.ensembl.org/index.html), and UCSC Browser (http://genome.ucsc.edu/) databases. The local library files include gene names and synonyms, gene identifiers, and chromosomal locations in base pairs. The web annotation retrieval is comparatively time-consuming, as it directly connects with the above databases to retrieve up-to-date and possibly missing information. Additionally, the user can choose to upload customized library files and proceed with merging the data.

Upon completion of the merging procedure, the user can browse and download the merged tabulated files as formated text or HTML file format. The latter includes active links to the above mentioned three databases. The results are organized based on the identified vocabulary of regional units and the data types considered. If the input data files include extra information, such as metadata or data values, then those will be also available in the merged files. Compressed formats can be emailed to a user specified address, enabling a quick data exchange (backup) of information which could be reused as input to avoid repetition of the same analysis. In Figure 1 we show a schematic representation of how K different datasets (Dataset $1 \cdots K$) are processed using Amalgam to produce a merged annotation file mapped onto the human genome. The user can download the merged file or alternatively launch the graphical interface options.

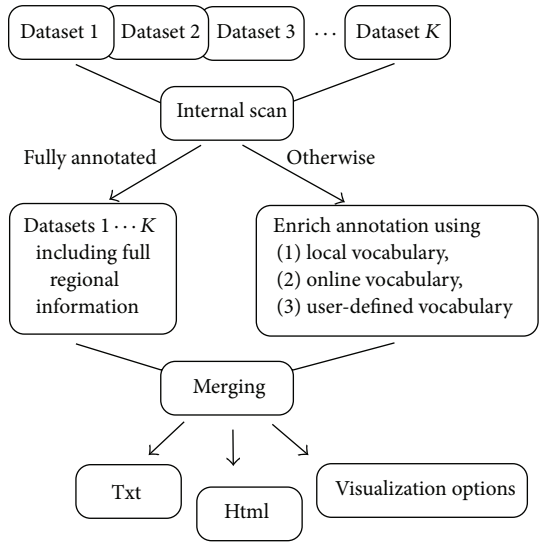


Figure 1: Schematic view of the procedure followed by Amalgam. Input files, Datasets $1 \cdots K$, are scanned to ensure that all entries include regional information. If that is not the case a reverse annotation procedure is followed to produce the unique annotational vocabulary. After merging the files, annotational information is available in Txt and Html formats, along with visualization options.

2.2. Visualization Options. Amalgam offers two visualization options for an intuitive real-time exploration of the merged data, called "Region Browser" and "Midnight Browser," respectively. The first visualization tab page displays the identified list of regional units and their summaries; upon clicking a region, region-specific details appear via the above-mentioned databases. The second visualization window provides chromosome and human genome maps of the integrated data, where the user can launch a genome-wide exploration to see the identified regional units together with their individual annotational information. Navigation through the merged data ranges from whole genome to base pair level. Figure 2 shows the processing windows of Amalgam. Particularly, in Figure 2(a) the window for uploading and translating the data, either locally or remotely, is shown, whereas Figure 2(b) shows the "Integrator" tab page as it appears when the mapping procedure is completed and the output file is available in Txt and Html formats. At the bottom of Figure 2, the Amalgam's integrated data visualization options are shown. In Figure 2(c) the merged files are presented in a list, given the regional units identified. For each regional unit, relative annotational information are supplied, whereas the user also has the option to further explore the information derived by NCBI, Ensembl, and UCSC databases. Figure 2(d) shows the "Midnight Viewer" tab window, where we can observe the regional units vocabulary derived by diverse datasets (i.e., microarray gene expression, RNA gene expression, and CNV). A search by term option is offered, as well as the FASTA files of the genes found in the identified regional units.

2.3. Use Distant Server. A special feature of Amalgam is the option to manage input or output files from local exchange email clients or file hosting services. Amalgam uses an internal timer to receive input from distant email servers and cloud storage servers, "E-mail Listener" and "Cloud Listener" options, respectively, as shown in Figure 2(a). Every few seconds, feedback is requested from the server. If a new email has been received and it is qualified for processing, Amalgam downloads the attached files, processes them locally, and resends them using the given email server. Similarly, if new files have been uploaded on a given directory, Amalgam processes them and saves the output to the given output directory. The reason for using internet hosting or emailing services is to allow the users to access their results from any network device.

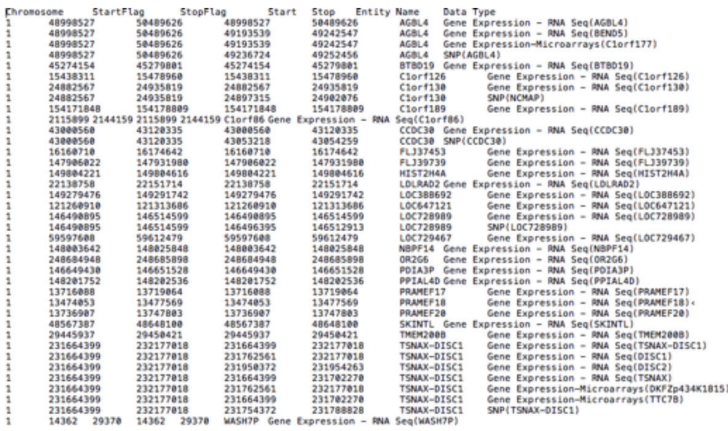
3. Results and Discussion

Amalgam was tested with publicly available data downloaded from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) data portal, but can be easily applied to any datasets given that gene name or gene identification is included at the uploaded files or alternatively chromosome name along with chromosomal locations given in base pairs. Datasets can be loaded from local or remote sources, enabling users to merge their data with other publicly available data of interest. As a case study, we applied Amalgam to merge three publicly available human breast cancer samples downloaded from the TCGA data portal, that is, an Agilent gene expression microarray (G4502A) sample, an Illumina HiSeq RNA-Seq (V2), and a copy number variant data sample from Affymetrix Genome wide SNP 6.0 array. Amalgam finds the wider, in terms of base pairs, regional units for all data samples to constitute a unique genome vocabulary of non-overlapping intervals. Each regional unit originates from one of the three data samples and may include subregions, that is, entries from the remaining two data samples. Diverse data types are merged, placing equal weights to each dataset.

(a) (b)

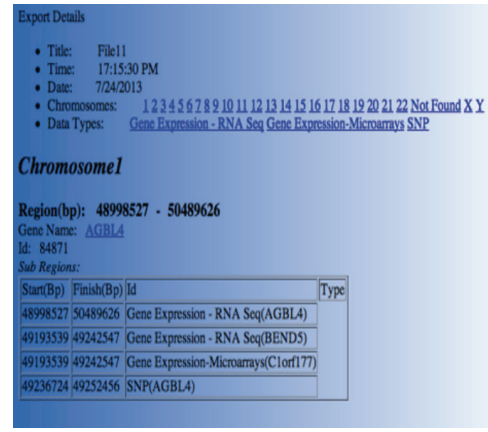
(c) (d)

FIGURE 2: Screenshots of Amalgam outlining the key features of the application. The "Translator" window (a) awaits for input files allowing users to upload text annotation or data files. When merging is complete, the output files can be downloaded by the "Integrator" tab page (b). The output merged files can be displayed in the two browsing windows. Particularly, genomic features, such as gene names and chromosomal positions, are extracted from the text files and are displayed in the "RegionBrowser" tab page (c). For each identified region, gene related information can be displayed from well-known public databases. In the "Midnight Viewer" tab page (d), the merged data are displayed on chromosomal or genome maps.



Chromosome	StartFlag	StopFlag	Start	Stop	Entity Name	Data Type
1	48998527	50489626	48998527	50489626	AGBL4	Gene Expression - RNA Seq(AGBL4)
1	48998527	50489626	49193539	49242547	AGBL4	Gene Expression - RNA Seq(BEND5)
1	48998527	50489626	49193539	49242547	AGBL4	Gene Expression-Microarrays(C1orf177)
1	48998527	50489626	49236724	49252456	AGBL4	SNP(AGBL4)
1	45274154	45279801	45274154	45279801	BTBD19	Gene Expression - RNA Seq(BTBD19)
1	15438311	15478960	15438311	15478960	C1orf126	Gene Expression - RNA Seq(C1orf126)
1	24882567	24935819	24882567	24935819	C1orf130	Gene Expression - RNA Seq(C1orf130)
1	24882567	24935819	24897315	24902076	C1orf130	SNP(NCMAP)
1	154171848	154178809	154171848	154178809	C1orf189	Gene Expression - RNA Seq(C1orf189)
1	2115099	2144159	2115099	2144159	C1orf86	Gene Expression - RNA Seq(C1orf86)
1	43000560	43120335	43000560	43120335	CCDC30	Gene Expression - RNA Seq(CCDC30)
1	43000560	43120335	43053218	43054259	CCDC30	SNP(CCDC30)
1	16160710	16174642	16160710	16174642	FLJ37453	Gene Expression - RNA Seq(FLJ37453)
1	147906022	147931980	147906022	147931980	FLJ39739	Gene Expression - RNA Seq(FLJ39739)
1	149804221	149804616	149804221	149804616	HIST2H4A	Gene Expression - RNA Seq(HIST2H4A)
1	22138758	22151714	22138758	22151714	LDLRAD2	Gene Expression - RNA Seq(LDLRAD2)
1	149279476	149291742	149279476	149291742	LOC388692	Gene Expression - RNA Seq(LOC388692)
1	121260910	121313686	121260910	121313686	LOC647121	Gene Expression - RNA Seq(LOC647121)
1	146490895	146514599	146490895	146514599	LOC728989	Gene Expression - RNA Seq(LOC728989)
1	146490895	146514599	146496395	146512913	LOC728989	SNP(LOC728989)
1	59597608	59612479	59597608	59612479	LOC729467	Gene Expression - RNA Seq(LOC729467)
1	148003642	148025848	148003642	148025848	NBPF14	Gene Expression - RNA Seq(NBPF14)
1	248684948	248685898	248684948	248685898	OR2G6	Gene Expression - RNA Seq(OR2G6)
1	146649430	146651528	146649430	146651528	PDIA3P	Gene Expression - RNA Seq(PDIA3P)
1	148201752	148202536	148201752	148202536	PPIAL4D	Gene Expression - RNA Seq(PPIAL4D)
1	13716088	13719064	13716088	13719064	PRAMEF17	Gene Expression - RNA Seq(PRAMEF17)
1	13474053	13477569	13474053	13477569	PRAMEF18	Gene Expression - RNA Seq(PRAMEF18)
1	13736907	13747803	13747803	13747803	PRAMEF20	Gene Expression - RNA Seq(PRAMEF20)
1	48567387	48648100	48567387	48648100	SKINTL	Gene Expression - RNA Seq(SKINTL)
1	29445937	29450421	29445937	29450421	TMEM2080	Gene Expression - RNA Seq(TMEM2080)
1	231664399	232177018	231664399	232177018	TSNAX-DISC1	Gene Expression - RNA Seq(TSNAX-DISC1)
1	231664399	232177018	231762561	232177018	TSNAX-DISC1	Gene Expression - RNA Seq(DISC1)
1	231664399	232177018	231954372	231954263	TSNAX-DISC1	Gene Expression - RNA Seq(DISC2)
1	231664399	232177018	231664399	231702270	TSNAX-DISC1	Gene Expression - RNA Seq(TSNAX)
1	231664399	232177018	231762561	232177018	TSNAX-DISC1	Gene Expression-Microarrays(DKFZp434K1815)
1	231664399	232177018	231664399	231702270	TSNAX-DISC1	Gene Expression-Microarrays(TTC7B)
1	231664399	232177018	231754372	231780828	TSNAX-DISC1	SNP(TSNAX-DISC1)
1	14362	29370	14362	29370	WASH7P	Gene Expression - RNA Seq(WASH7P)

(a)



(b)

FIGURE 3: Txt and Html formats of the merged file are produced. Three breast cancer TCGA samples were uploaded and merged. Each sample is derived by a different platform, namely, Agilent gene expression microarray, Illumina RNA-Seq, and Affymetrix copy number variant data. (a) The tabulated merged file includes chromosomal id, start/end positions of the identified regional units (StartFlag, StopFlag), and the subregions included in the identified regions (Start, Stop), the region name (Entity name) and the data type that the region belongs to. (b) The same information as in (a) is displayed in Html format with active links to the NCBI database.

In Figure 3 we show the merged files produced from the three data samples uploaded. As can be seen from the Txt (Figure 3(a)) and Html (Figure 3(b)) exports produced, information for the regional units identified is given, as well as the subregions included in each regional unit. In the example presented 52,932 unique regions were identified, from which 343 entries did not map to gene or other regional information based on our local databases. Using the "Web Annotation" option, we identified 293 additional entries. This information can be also displayed using Amalgam's visualization browsers as shown in Figures 2(c) and 2(d). Reusability is ensured since the user can relaunch Amalgam and merge an output from a previous Amalgam run with additional input sample files.

Additionally, custom files can be uploaded, either as input or library files, enabling users to augment, merge, and visualize their data with publicly available data. This, also ensures that the tool is up-to-date at all times, since apart from hosting custom-made libraries, it is also able to house a compiler when running.

4. Conclusions

Integration and analysis of large diverse datasets is a promising field of ongoing research towards the understanding of the genome and its relation to human disease [2, 16]. However, merging diverse data types in the genome often becomes a time-consuming task due to the continuous improvement of the underling technologies and certain format incompatibilities. This calls for intuitive tools able to flexibly integrate multiple data types and produce a common vocabulary map of the human genome given the supplied data. We have developed AmalgamScope, a user-friendly stand-alone tool for merging annotation files of diverse genomic data types, including data values. Our software provides a flexible framework to summarize fragmented annotation information across data types, relative to information extracted from web-based annotation resources, providing a common platform for reproducible exploration of the transcriptome. The integrated annotation files can considerably assist further analysis for data integration or visualization.

To improve the management and storage of input/output files, an emailing exchange procedure is included together with a cloud utility; however, in the future we aim to further explore Amalgam's cloud capabilities, given the complexity and volume of the data. Currently, we are extending Amalgam in a distributed computing environment to accommodate larger studies.

Availability and Requirements

Project Name. AmalgamScope.
Project Home Page. http://www.bioacademy.gr/bioinformatics/Amalgam/index.html.
Operating System. Microsoft Windows. Programming language: C sharp. License: Freeware.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

Georgia Tsiliki and Sophia Kossida were supported by the EU DICODE (Mastering Data-Intensive Collaboration and Decision Making) Collaborative Project (FP7, ICT-2009.4.3, Contract no. 257184) and EU COST Action SeqAhead (BM1006).

References

[1] D. E. Sullivan, J. L. Gabbard Jr., M. Shukla, and B. Sobrala, "Data integration for dynamic and sustainable systems biology resources: challenges and lessons learned," *Chemistry and Biodiversity*, vol. 12, pp. 1599–1610, 2002.

[2] S. Tyekucheva, L. Marchionni, R. Karchin, and G. Parmigiani, "Integrating diverse genomic data using gene sets," *Genome Biology*, vol. 12, no. 10, article R105, 2011.

[3] J. Stalker, B. Gibbins, P. Meidl et al., "The Ensembl web site: mechanics of a genome browser," *Genome Research*, vol. 14, no. 5, pp. 951–955, 2004.

[4] L. Meyer, A. Zweig, A. Hinrichs et al., "The UCSC genome browser database: extensions and updates 2013," *Nucleic Acids Research*, vol. 41, no. 1, pp. D64–D69, 2013.

[5] S. Dombrowski and D. Maglott, *Using the Map Viewer to Explore Genomes*, NCBI Handbook, 2003.

[6] H. Bao, H. Guo, J. Wang, R. Zhou, X. Lu, and S. Shi, "MapView: visualization of short reads alignment on a desktop computer," *Bioinformatics*, vol. 25, no. 12, pp. 1554–1555, 2009.

[7] T. Abeel, T. van Parys, Y. Saeys, J. Galagan, and Y. van de Peer, "GenomeView: a next-generation genome browser," *Nucleic Acids Research*, vol. 40, no. 2, article e12, 2011.

[8] J. Kim, K. Patel, H. Jung, W. P. Kuo, and L. Ohno-Machado, "AnyExpress: integrated toolkit for analysis of cross-platform gene expression data using a fast interval matching algorithm," *BMC Bioinformatics*, vol. 12, article 75, 2011.

[9] Z. Zhu, B. Niu, J. Chen, S. Wu, S. Sun, and W. Li, "MGAviewer: a desktop visualisation tool for analysis of metagenomics alignment data," *Bioinformatics*, vol. 29, no. 1, pp. 122–123, 2013.

[10] L. D. Stein, C. Mungall, S. Shu et al., "The generic genome browser: a building block for a model organism system database," *Genome Research*, vol. 12, no. 10, pp. 1599–1610, 2002.

[11] R. Podicheti, R. Gollapudi, and Q. Dong, "WebGBrowse—a web server for GBrowse," *Bioinformatics*, vol. 25, no. 12, pp. 1550–1551, 2009.

[12] T. Toyoda, Y. Mochizuki, K. Player, N. Heida, N. Kobayashi, and Y. Sakaki, "OmicBrowse: a browser of multidimensional omics annotations," *Bioinformatics*, vol. 23, no. 4, pp. 524–526, 2007.

[13] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler et al., "Integrative genomics viewer," *Nature Biotechnology*, vol. 29, no. 1, pp. 24–26, 2011.

[14] E. Baker, J. Jay, J. Bubier, M. Langston, and E. Chesler, "GeneWeaver: a web-based system for integrative functional genomics," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1067–D1076, 2012.

[15] J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, article R86, 2010.

[16] Q. Mo, S. Wang, V. Seshan et al., "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 11, pp. 4245–4250, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 753428, 10 pages
http://dx.doi.org/10.1155/2014/753428

Research Article

Integration of Residue Attributes for Sequence Diversity Characterization of Terpenoid Enzymes

Nelson Kibinge, Shun Ikeda, Naoaki Ono, Md. Altaf-Ul-Amin, and Shigehiko Kanaya

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan

Correspondence should be addressed to Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 1 November 2013; Accepted 21 February 2014; Published 11 May 2014

Academic Editor: Samuel Kuria Kiboi

Progress in the "omics" fields such as genomics, transcriptomics, proteomics, and metabolomics has engendered a need for innovative analytical techniques to derive meaningful information from the ever increasing molecular data. KNApSAcK motorcycle DB is a popular database for enzymes related to secondary metabolic pathways in plants. One of the challenges in analyses of protein sequence data in such repositories is the standard notation of sequences as strings of alphabetical characters. This has created lack of a natural underlying metric that eases amenability to computation. In view of this requirement, we applied novel integration of selected biochemical and physical attributes of amino acids derived from the amino acid index and quantified in numerical scale, to examine diversity of peptide sequences of terpenoid synthases accumulated in KNApSAcK motorcycle DB. We initially generated a reduced amino acid index table. This is a set of biochemical and physical properties obtained by random forest feature selection of important indices from the amino acid index. Principal component analysis was then applied for characterization of enzymes involved in synthesis of terpenoids. The variance explained was increased by incorporation of residue attributes for analyses.

1. Background

Biology and other modern sciences have become data intensive and in fact data-driven biology is now a full-fledged domain of specialization among the life sciences. Due to the amorphous nature of the accruing biological data, various databases have continually been developed to allow systematization [1]. A lot however still needs to be done to characterize these compilations into meaningful information. KNApSAcK database describes species-metabolite relationships, and within the KNApSAcK family we have developed an enzyme-reaction database called KNApSAcK motorcycle DB containing reactions and enzyme peptide sequences based on experimental evidence focusing on secondary metabolic reactions in plants.

Increasingly, the need to analyse data in such repositories has made advanced mathematical and statistical tools a mainstay of bioinformatics in recent years, more so in sequence-based analyses [2]. For molecular sequence data especially proteins, standard alphabetical notation of sequence information may not explicitly capture aspects such as their biochemical and physicochemical properties (BPPs) and may to some extent limit tractability to mathematical analyses. Ideally, computational analyses of the often heterogenous datasets require theoretical representations in forms suitable for various data processing models. This formal representation has been defined as sequence feature coding [3]. There has been no standard method of directly encoding quantifiable BPP of protein sequences hitherto. A key research question has thus been how to quantitatively characterize such data for computation whilst considering these aspects and other sequence metadata [4]. In the present study, we introduce a BPP subset for encoding amino acid residue properties into protein sequences during analyses. We found that this increases the flexibility of computational analyses focusing on facets of biochemical, physical, and evolutionary attributes of sequence data. Integration of BPP is employed in examination of diversity in enzymes related to secondary metabolite pathways, specifically those involved in terpenoid synthesis.

Researchers have proposed schemes to ensure amenability of sequences to computation, but it remains difficult to achieve computational objectivity while maintaining biological interpretability. 5-bit and 3-bit binary feature coding

of amino acids in peptide sequences acids have been used in studies such as [5]. White and Seffens also used 20-bit transformation for neural network application for translation of proteins [6]. A limitation of binary feature coding is the minimal biological information with respect to amino acid diversity. This is because bit coding does not account for relative similarities or differences between amino acids and neither is it flexible to integration of their BPP; besides, binary notation of highly conserved protein sets may also pose numerical difficulties to probability-based models.

Information theory has also been exploited as an alternative, where mutual information and entropy are estimated by the Shannon-Wiener indexing of amino acid properties [7, 8]. This way, distance and variation between amino acid units, is estimable and therefore is an improvement over the binary coding method. It, however, does not directly represent characteristic attributes such as polarity, molecular size, and other features of residues.

More recently, the amino acid index (AAindex) database has accumulated published data of amino acid properties [9]. Each index has a set of 20 numerical values of a BPP quantified and published in research literature. Currently, AAindex contains 544 indices describing quantifiable amino acids residue properties. This provides a foundation for BPP feature-coding protein by assigning "scale-measured" attributes of amino acids.

AAindex in its entirety (raw form) is not merited for feature coding since it is highly redundant and has some missing values. Atchley and colleagues proposed index reduction using multivariate factor analysis reducing it to five compressed factors [10]. This methodology is a useful solution for AAindex-based metrication of amino acid residues but factor analysis (FA) reduction complicates biological interpretability in downstream analyses. This is because FA just like principal component analysis (PCA) assumes an underlying linear independence of variables whose coefficients, also called "factors," are a proxy interpretation of AAindex variables [11]. This means that "factors" derived are pseudovariables of the actual original properties and in a way add to complexity of biological interpretation in subsequent downstream steps of sequence analyses.

In light of these challenges in analytical systematization of protein sequences, the present work applies a slightly different variable selection criterion from AAindex for the purpose of encoding BPP information into sequence data. This was achieved by use of random forest (RF) algorithm [12] to reduce redundancy and to maximize amino acid metadata captured in the AAindex. Eight BPP indices describing variability of amino acids were selected based on our experimental results. The derived reduced AAindex (rAAindex) is a subset of the original AAindex after elimination of redundancies. We further integrated the rAAindex in characterization of protein sequence diversity in the KNApSAcK motorcycle DB. The enzymes characterized are involved in secondary metabolic pathways of terpenoids and include monoterpenoid synthases, diterpenoid synthases, triterpenoid synthases, and sesquiterpenoid synthases.

2. Materials and Methods

2.1. Amino Acid Index and Random Forest Selection of Biochemical and Physical Properties. Amino acid index (AAindex) is a database of numerical indices describing biochemical and physical attributes of the 20 amino acids [9]. It provides a plausible starting point for interpreting peptide BPPs numerically through its "building blocks": the amino acids. We selected a set of important indices that broadly characterize amino acid BPP variation, where RF algorithm [12] was used for index selection.

We denote by X the set of amino acid indices as the explanatory variables. N denotes the set of amino acids (AA). The categorical predictor variable that best defines the AA population is its qualitative attribute of water interaction; that is, every amino acid is described as either hydrophobic or hydrophilic. We therefore denote Y as describing hydrophobicity or hydrophilicity of an amino acid. The ith amino acid, n_i, is described by a vector $(x_1, \ldots, x_m, y_i)$.

Raw AAindex is highly redundant and multicollinear. We initially processed it by removal of indices that had missing values for any amino acid. Redundant indices were eliminated by backward elimination of variables whose correlation coefficient was above a threshold of 0.85. 283 indices were retained for RF variable selection.

Random forest (RF) [12] is a popular algorithm in statistics and bioinformatics for two reasons:

(i) it is a powerful classification and regression tree (CART) tool that generates ensembles of decision trees. RF and other decision tree-based classifiers are nonparametric; that is, they do not assume underlying structure in a dataset and are useful for classification and regression modeling of complex biological data;

(ii) RF implements a mechanism of calculating variable importance scores (VIM) by permutation testing. These measures are useful in feature selection and provide an advantage which we explore in this work.

Besides these two advantages, further application to biological research has been documented in [14]. Detailed mechanism of the RF algorithm is described in [12, 15] although a generalized outlook of its concepts is illustrated in Figure 1. RF was implemented for selection of reduced AAindex (rAAindex) consisting of indices describing BPPs explaining the variation of amino acids.

In RF, sampling by bootstrap creates an "out-of-bag" (OOB) sample which is an important feature due to its usefulness in estimation of VIM. These scores are derived by permutation testing of the OOB data in the error estimation step. VIM scores of index x_i are described as the mean error rate over all trees in the RF ensemble. Detailed information on VIM calculation is described elsewhere [12], but for descriptive purposes, we simplify the formal representation of this measure as

$$\mathrm{VIM}\left(x_i\right) = \frac{1}{\mathrm{ntree}} \sum_{1}^{\mathrm{ntree}} \left(\mathrm{err}\cdot\widetilde{\mathrm{OOB}}_{\mathrm{t}} - \mathrm{err}\cdot\mathrm{OOB}_{\mathrm{t}}\right), \quad (1)$$

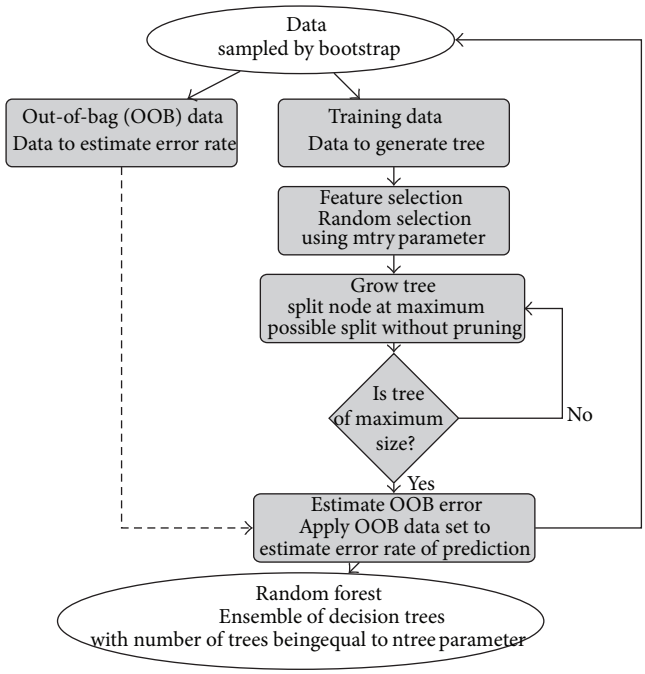


Figure 1: Random forest algorithm. Mechanism of the random forest (RF) algorithm starting from the data selection by bootstrapping up to variable importance calculation. The amino acid index data containing physicochemical metrics of amino acids was subjected to RF for index selection.

where VIM(x_i) is a function estimating the VIM score for variable x_i and ntree is the number of trees in the RF ensemble, whereas err $\cdot\widetilde{\text{OOB}}_t$ is the number of misclassifications tested on a tree t where the input was permuted values of variable x_i. Conversely, err $\cdot\text{OOB}_t$ is the number of misclassifications tested on a tree t whose input was the nonpermuted values of variable x_i.

Validity of permutation test derivation of VIM in the RF algorithm operates on the premise that if a variable is "important," then permuting its values (realistically) leads to reduced accuracy of class prediction. Variables were selected using the method described in [16]. The first step involved stochastically running 1000 RF classification trials of AAindex and each time recording the mean decrease in accuracy (VIM score). Indices were then ranked on a decreasing score order. The variation of these VIM scores was obtained and the point of minimum variance was initialized as a threshold, from which 93 amino acid indices were retained for the further index reduction by nested RF feature selection approach described in detail by Genuer et al. [16]. The threshold of significant deviation in the increasing error rates from the nested RF modeling was set to the number of variables above which the error rates significantly increase above the threshold of 0.02 percent meaning that at most a single amino acid misclassification could be accepted from the nested RF. A reduced amino acid index (rAAindex) was thus derived and its usefulness as a representation of amino acid information was tested on data from our KNApSAcK motorcycle database.

2.2. KNApSAcK Motorcycle DB: Peptide Sequence-Metabolic Reaction Relationship DB. It is necessary to extend the species-metabolite relationship DB by incorporating a secondary metabolite pathway DB that includes pathways with detected enzymatic reactions and other actual or predicted peptide sequences that may be involved in these pathways. We surveyed reactions of secondary metabolites in scientific literature, and amino acid sequences involved in secondary metabolism were obtained from public databases in PubMed.gov (http://www.ncbi.nlm.nih.gov/). All the data comprising 2,881 secondary metabolic reactions was accumulated in the KNApSAcK motorcycle DB (http://kanaya.naist.jp/motorcycle/top2.html) as shown in the main window of the KNApSAcK motorcycle DB (Figure 2); enzyme reactions can be retrieved using keywords of enzymes, species, genes, metabolites, and peptide sequences obtained from a BLASTP search. For metabolite search using its keywords, we obtain information on enzyme name, reaction involved, compound class (C-class in Figure 2) and subclass (C-subclass) of metabolic reactions, and reaction mechanisms. For BLASTP search, we can predict reaction equations for a targeted peptide sequence using information on the class and subclass of metabolic pathways (Figure 1(c)). Thus, the motorcycle DB makes it possible to predict enzyme reactions based on the class and subclass of metabolic reactions evidenced by experiments mentioned in scientific literature. This differentiates it from KEGG [17] and BioCyc [18]. We have thus far obtained 596,974 protein sequences of 59,165 plant species and 124,292 protein sequences of 66 bacterial species from the nonredundant protein sequences of PlantGDB.

For analytical purposes of the presently developed method, we narrowed our test dataset to terpenoid synthase peptide sequences with >200 amino acid residues. Terpenoids are organic metabolites of plants that have been shown to have insect-pesticide properties among other roles. Terpenoid synthases sequence from the KNApSAcK database [19] was examined for patterns in diversity. Enzymes annotated to four families, namely, monoterpenoid synthases, diterpenoid synthases, triterpenoid synthases, and sesquiterpenoid synthases, were examined by PCA. Understanding the data structure of these terpenoid enzyme subfamilies is important for annotation of similar organic compounds [20]. Multiple sequence alignment and gap removal were carried out to extract homologous regions of sequences from the four subfamilies of terpenoid synthases. These domains had a length of 28 residues for the 283 sequences. Binary and rAAindex (BPP) feature coding of amino acid residues in these sequences were compared.

2.3. Sequence Diversity Characterization Based on Principal Component Analyses (PCA). The present work attempts to examine diversity of secondary metabolic enzyme groups using datasets from the KNApSAcK motorcycle DB by

FIGURE 2: KNApSAcK motorcycle database. Enzyme-reaction database. (a) The main window of motorcycle. (b) An example of a keyword search. (c) An example of BLASTP search.

integrating amino acid attributes represented in the rAAindex where PCA was used to analyse variation. PCA is a technique that enables efficient interpretation of variation and relationship between variables in a huge dataset represented by higher dimensional vectors [21]. It is widely applied in bioinformatics as exemplified by Tatusov et al. who phylogenetically classified genomes by protein function [22]. For comparative purposes, a BPP integrated dataset was analysed in comparison to 8-bit binary-coded enzyme set. We initially generated lattices representing individual sequences and encoded by both rAAindex and the commonly used 8-bit binary feature coding. We hypothesized that the BPP-encoded set explains more variance of sequences and thus reflects the diversity of proteins based on BPP integration.

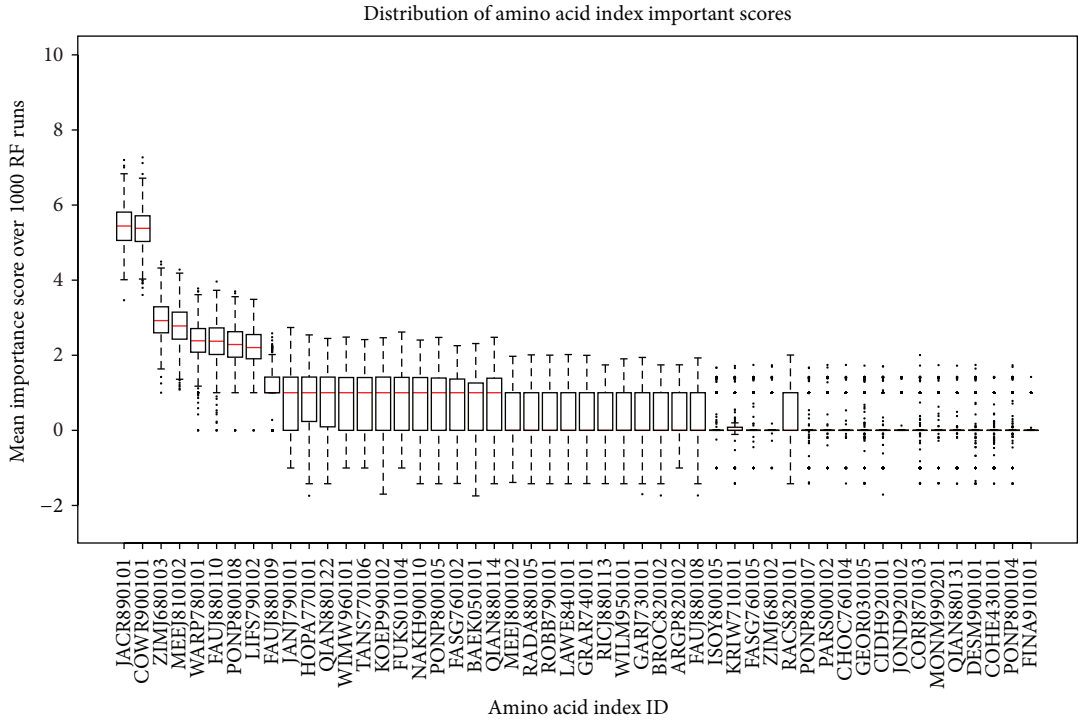


Figure 3: Variable importance scores (1000 RF trials). Variable importance score distribution for 1000 runs of random forest classification on the amino acid index. Each boxplot represents distribution of each property (also called variable) represented on the horizontal axis. The properties have been ordered in decreasing order of the median score (red line in boxplot). For easier visualization, the set has been truncated to show the top 50 properties. The corresponding properties are shown in Table 1.

3. Results and Discussion

3.1. Reduced Amino Acid Index. Physicochemical properties of amino acids quantitatively describe the overall biochemical behaviour of peptide and protein sequences [13]. The amino acid index database [9] has collected properties of amino acids measured by various researchers since the 1970s using scientific instruments and quantifiable metrics. It is essential to consider these properties in objective analyses of sequence data. Numeric quantification is also pivotal because it gives a flexible way of integrating this information in a mathematically and statistically amenable form different from the alphabetical string representation.

The AAindex database is highly redundant and has some missing values for certain properties. In its raw form, the AAindex is unsuitable for direct BPP encoding. Ideally, a reduced set would work for most sequence analyses. Researchers have utilized various compression techniques to reduce the AAindex. Atchley and colleagues used a multivariate factor analysis to propose a compressed variable set of five vectors describing amino acids in a multidimensional space [10]. More recently, fuzzy c-means algorithm has been applied in clustering the AAindex indices and the resultant clusters incorporated in a support vector machine modeling experiment to predict DNA-binding domains [23].

Factor analysis (FA) is a useful approach for the same purpose of defining a minimal set of "factors" that simplify interpretation of protein sequence characteristics. Random forest (RF) variable reduction differs from FA since RF selects important variables without compressing the whole variable set into fewer descriptive factors. RF has been proven to be a useful tool for biological data as described in [24–26]. Here, BPP selection entails minimizing the original variables in the AAindex by elimination of redundancy, high collinearity, and less informative variables whilst maintaining a sufficiently parsimonious set of the original BPP properties represented in the AAindex. Compression (as in FA) on the other hand is redefinition of original AAindex variables into new components by multivariate techniques such as PCA [21] and factor analyses [27]. We argue against compression in the context of AAindex data that while a minimized descriptive set is achieved, there results a complexity of biological interpretation that arises if the redefined variables are applied in subsequent downstream mathematical or statistical analyses.

From the initial 544 properties contained in the amino acid index database, 13 which had missing values were dropped. Initially, the redundancy was further reduced by dropping variables with a correlation coefficient greater than 0.85, further trimming the set to 283 amino acid indices. The retained indices were then subjected to the RF algorithm (1000 trials) and variable reduction was done using the technique in [16]. Variable importance scores (VIM) were ranked in decreasing order (Figure 3). The "importance" score of a BPP illustrates its significance with regard to amino acid classification. It shows the VIM score distribution for 1000 runs of random forest classification of BPP in the amino acid

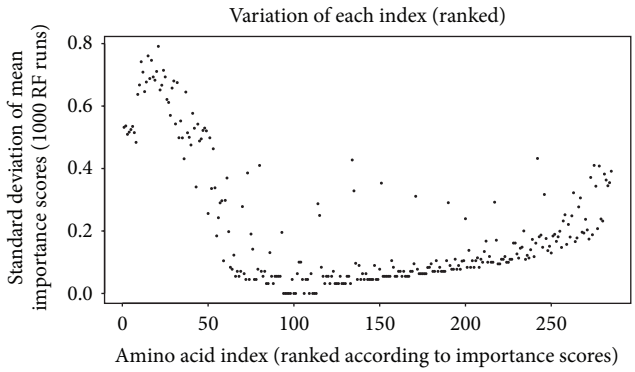


Figure 4: Variation of the BPP importance. Standard deviation of the importance scores of the properties (y-axis); models contribution of each property to the performance of the RF algorithm. Those variables with a close to zero variation are less "important." At the tail end, variance is more than zero due to chance.

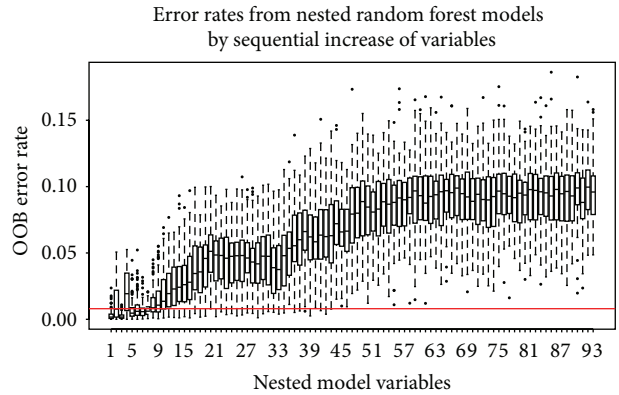


Figure 5: Nested random forest error rates. Nested random forest variable selection: variables have been ordered by their importance scores, new RF models are built by single variable addition in the nested RF setup, and RF error rates are measured. In this experiment, each box and whisker plot represents the distribution of error rates from 100 trials at each nested RF step. In total, there were 93 steps corresponding to the 93 top indices from previous step. The y-axis shows the error rates as a percentage. The threshold of acceptable mean error rate was set at 2 percent shown by the red horizontal line.

index. Each boxplot in the figure represents distribution of each BPP (also called variable) represented on the horizontal axis. These properties have been ordered in decreasing order of the median score (red line in boxplot). For easier visualization, the set in the figure has been truncated to show the top 50 properties by mean VIM ranking. The corresponding properties are shown in Table 1. Variation (standard deviation) of the ranked scores was observed as shown in Figure 4. Standard deviation of the importance scores of the properties (y-axis) models contribution of each property towards performance of the RF algorithm. Those variables with a close to zero variation are less "important." At the tail end variance is higher than zero but is large due to chance (P value > 0.05). Variables with a mean VIM score of 0 or less were dropped, lowering the retained variables to 93 indices. Nested RF modeling was done using the ranked indices, by starting with the highest ranked variable and subsequent addition remaining after each step. Error rates were estimated for each step of the nested modeling. Details of nested RF are explained in [16]. The threshold of acceptable error rate was set to 2 percent. Figure 5 shows that when the first 8 indices are used for classification, the error rate remains under the 2 percent error rate (horizontal red line in the figure) threshold, whereas it significantly rises with the addition of subsequent indices.

An RF-reduced subset of the amino acid index, rAAindex, (Table 2) with these 8 most important BPPs, is proposed for use in BPP encoding especially for statistical learning and other mathematical tasks involving protein sequences. The properties retained are shown in Table 3.

3.2. Characterization of Terpene Synthase Sequence Diversity. Terpenes are the largest group of plant natural products with a variety of core chemical structures comprising at least 30,000 compounds and synthesized by terpenoid synthases [28]. Terpene diversity is caused by the large number of different terpene synthases used in the first step of terpene synthesis, and some terpene synthases produce multiple products [29]. Terpene synthases are generally classified according to the number of carbons in their substrates, that is, geranyl diphosphate (C10, GPP) for monoterpene synthases, farnesyl diphosphate (C15, FPP) for sesquiterpene synthases, geranylgeranyl diphosphate (C20, GGPP) for diterpene synthases, and squalene for triterpene synthases (C30). The rather limited similarity of plant terpenoids [20] complicates annotation of their enzymes. Clustering algorithms improve the resolution to some extent. PCA was used to analyse variation among 4 terpene synthase subfamilies. We examined the performance of PCA classification when rAAindex BPP encoding is implemented, relative to the commonly used 8-bit binary encoding. The combined variance explained by the first two principal components is 30.02 (Figure 6 left) percent, whereas the variance explained by the first two components in binary encoded set is 14.84 percent (Figure 6 right). Triterpenoid synthases can be clearly distinguished from the other categories, which was also consistent with our previous findings [30].

We compared the performance of rAAindex to 8-bit binary encoding of amino acids [31] in an actual dataset (described in the data and methods section). Binary coding is the most popular representation scheme for machine learning tasks of protein data [5, 6] and was utilized as a benchmark of comparison to rAAindex encoding. Figure 6 (left) shows the fragment distribution of terpene synthases coded by 8-bit binary notation mainly clustered into five regions. Similarly, Figure 6 (right) shows the distribution of terpene synthases coded by rAAindex into the same 5 groups. Principal component analysis was performed on terpenoid synthase sequence subfamilies, where amino acid residues were encoded in 8-bit binary code. PC1 and PC2 show a combined variance of 14.84 percent explained variance. The second part

Table 1: Top 50 ranked indices. Descriptions of the top 50 indices by ranking of the VIM scores prior to nested RF. The first column represents the AAindex access ID, while the second is the corresponding BPP. Please refer to the amino acid index database in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/753428 for the references in column 2.

	ID	Property
1	RADA880101	Information value for accessibility; average fraction 35% (Biou et al., 1988)
2	ROSM880101	Information value for accessibility; average fraction 23% (Biou et al., 1988)
3	KIDA850101	Retention coefficient in TFA (Browne et al., 1982)
4	EISD840101	Normalized hydrophobicity scales for alpha + beta-proteins (Cid et al., 1992)
5	JACR890101	Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al., 1992)
6	COWR900101	Consensus normalized hydrophobicity scale (Eisenberg, 1984)
7	BLAS910101	Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)
8	MEEJ810101	Hydrophobic parameter pi (Fauchere-Pliska, 1983)
9	CIDH920104	Number of hydrogen bond donors (Fauchere et al., 1988)
10	GRAR740102	Number of full nonbonding orbitals (Fauchere et al., 1988)
11	ZIMJ680103	Polarity (Grantham, 1974) [13]
12	MEEJ810102	Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)
13	RADA880104	Hydropathy index (Kyte-Doolittle, 1982)
14	KUHL950101	Hydrophobic parameter (Levitt, 1976)
15	FAUJ880110	Conformational preference for parallel beta-strands (Lifson-Sander, 1979)
16	RADA880107	Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)
17	WARP780101	Retention coefficient in NaClO4 (Meek-Rossetti, 1981)
18	BIOV880102	Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)
19	BIOV880101	8 A contact number (Nishikawa-Ooi, 1980)
20	FASG890101	Partition coefficient (Pliska et al., 1981)
21	PONP800108	Average number of surrounding residues (Ponnuswamy et al., 1980)
22	FAUJ830101	Hydrophobicity (Prabhakaran, 1990)
23	CORJ870101	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)
24	MANP780101	Transfer-free energy from chx to wat (Radzicka-Wolfenden, 1988)
25	LIFS790102	Transfer-free energy from chx to oct (Radzicka-Wolfenden, 1988)
26	GUOD860101	Energy transfer from out to in (95% buried) (Radzicka-Wolfenden, 1988)
27	WOLS870101	Mean polarity (Radzicka-Wolfenden, 1988)
28	WOLR790101	Side chain hydropathy, uncorrected for solvation (Roseman, 1988)
29	LEVM760101	Side chain hydropathy, corrected for solvation (Roseman, 1988)
30	WOLR810101	Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)
31	ROSM880102	Average interactions per side chain atom (Warme-Morgan, 1978)
32	WOEC730101	Polar requirement (Woese, 1973)
33	PLIV810101	Hydration potential (Wolfenden et al., 1981)
34	RADA880108	Principal property value z1 (Wold et al., 1987)
35	FAUJ880109	Polarity (Zimmerman et al., 1968)
36	HOPA770101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)
37	CIDH920103	Hydropathy scale based on self-information values in the two-state model (9% accessibility) (Naderi-Manesh et al., 2001)
38	TANS770106	Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)
39	PRAM900101	Hydrophilicity scale (Kuhn et al., 1995)
40	ENGD860101	Retention coefficient at pH 2 (Guo et al., 1986)
41	BROC820101	Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)
42	NADH010103	Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases (Punta-Maritan, 2003)
43	NADH010102	Hydrophobicity index (Wolfenden et al., 1979)

Table 1: Continued.

	ID	Property
44	KYTJ820101	Hydrophobicity-related index (Kidera et al., 1985)
45	EISD860103	Weights from the IFH scale (Jacobs-White, 1989)
46	NISK800101	Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)
47	JURD980101	Scaled side chain hydrophobicity values (Black-Mould, 1991)
48	WIMW960101	NNEIG index (Cornette et al., 1987)
49	QIAN880122	Hydrophobicity index (Engelman et al., 1986)
50	PUNT030101	Hydrophobicity index (Fasman, 1989)

Table 2: Reduced Amino Acid Index (rAAindex). The first column contains the names of the 20 amino acids that make up protein sequences. Each of the 8 subsequent columns is a BPP attribute selected based on its importance in explaining variation of the amino acid. Each row is a vector describing an amino acid in 8 dimensions, each of which represents a physical property tabulated in Table 3.

Amino acid	JACR890101	COWR900101	ZIMJ680103	MEEJ810102	FAUJ880110	WARP780101	PONP800108	LIFS790102
Ala	0.18	0.42	0.00	1.00	0.00	10.04	6.05	1.00
Arg	−5.40	−1.56	52.00	−2.00	3.00	6.18	5.70	0.68
Asn	−1.30	−1.03	3.38	−3.00	3.00	5.63	5.04	0.54
Asp	−2.36	−0.51	49.70	−0.50	4.00	5.76	4.95	0.50
Cys	0.27	0.84	1.48	4.60	0.00	8.89	7.86	0.91
Gln	−1.22	−0.96	3.53	−2.00	3.00	5.41	5.45	0.28
Glu	−2.10	−0.37	49.90	1.10	4.00	5.37	5.10	0.59
Gly	0.09	0.00	0.00	0.20	0.00	7.99	6.16	0.79
His	−1.48	−2.28	51.60	−2.20	1.00	7.49	5.80	0.38
Ile	0.37	1.81	0.13	7.00	0.00	8.72	7.51	2.60
Leu	0.41	1.80	0.13	9.60	0.00	8.79	7.37	1.42
Lys	−2.53	−2.03	49.50	−3.00	1.00	4.40	4.88	0.59
Met	0.44	1.18	1.43	4.00	0.00	9.15	6.39	1.49
Phe	0.50	1.74	0.35	12.60	0.00	7.98	6.62	1.30
Pro	−0.20	0.86	1.58	3.10	0.00	7.79	5.65	0.35
Ser	−0.40	−0.64	1.67	−2.90	2.00	7.08	5.53	0.70
Thr	−0.34	−0.26	1.66	−0.60	2.00	7.00	5.81	0.59
Trp	−0.01	1.46	2.10	15.10	0.00	8.07	6.98	0.89
Tyr	−0.08	0.51	1.61	6.70	2.00	6.90	6.73	1.08
Val	0.32	1.34	0.13	4.60	0.00	8.88	7.62	2.63

Table 3: Annotation of the selected properties. Descriptions of the 8 indices selected after the nested random forest variable selection. The first column represents the AAindex access ID, while the second is the corresponding BPP. Please refer to the amino acid index database for the references in column 2.

	ID	Property
1	JACR890101	Number of full nonbonding orbitals (Fauchere et al., 1988)
2	COWR900101	Conformational preference for parallel beta-strands (Lifson-Sander, 1979)
3	ZIMJ680103	Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)
4	MEEJ810102	Average number of surrounding residues (Ponnuswamy et al., 1980)
5	WARP780101	Average interactions per side chain atom (Warme-Morgan, 1978)
6	FAUJ880110	Polarity (Zimmerman et al., 1968)
7	PONP800108	Weights from the IFH scale (Jacobs-White, 1989)
8	LIFS790102	Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)

of the figure illustrates PCA of the same data set encoding the biochemical and physical properties of amino acid residues described above. PC1 and PC2 in this case explain 30.02 percent variance showing that more sequence information is described by the BPP subset. The four subfamilies clustered are monoterpenoid, diterpenoid, triterpenoid, and sesquiterpenoid synthases. Triterpenoid synthases and subgroups of terpenoid synthases are distinctly different in structure compared to the other synthases. It is noted that three types of terpene synthases except for the diterpene synthases are

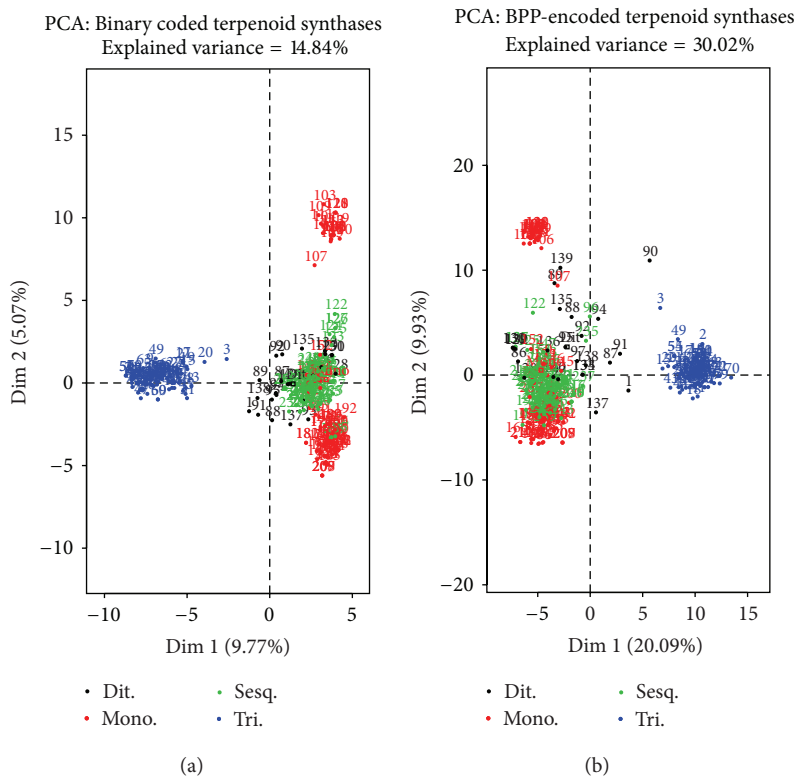


FIGURE 6: Principal component analysis of terpenoid synthases. Left: principal component analysis of the terpenoid synthase subfamilies where amino acid residues are encoded in 8-bit binary method. PC1 and PC2 show a combined variance of 14.84 percent explained variance. Right: principal component analysis of the same data set encoded using the biochemical and physical properties of amino acid residues. PC1 and PC2 in this case explain 30.02 percent variance. The four subfamilies clustered are monoterpenoid, diterpenoid, triterpenoid, and sesquiterpenoid synthases. Triterpenoid synthases and subgroups of terpenoid synthases are distinctly different in structure compared to the other synthases.

less divergent at the peptide sequence level; that is, small changes in peptide sequences of the terpene synthase make it possible to synthesize many different terpenoid compounds. The orders of fragments from the N- to the C-terminus in enzymes are arranged in two clusters for monoterpene synthases and are a single cluster for the other categories. Thus, monoterpene and sesquiterpene synthases are very similar in arrangement of peptide fragments, which is consistent with the similarity of the 3D structures in monoterpene and sesquiterpene synthases [32, 33] and with the fact that several bifunctional enzymes possess both sesquiterpene and monoterpene synthase activities [33, 34].

Diterpene and triterpene synthases have inherent structures specified by a single cluster for diterpene synthases and two clusters from the N- to C-terminus for triterpene synthases.

4. Conclusion

This paper has introduced a subset of eight biochemical and physical attributes of amino acids that can be encoded in protein sequences for sequence-based analysis. These features quantify attributes of individual residues in numerical metrics that improve amenability to mathematical and statistical tasks and also enhance biological interpretability of such tasks. Terpenoid synthases protein set was used to evaluate the encoding of these attributes by PCA. The terpenoid synthase subfamilies established that more variance is explained when BPPs are encoded compared to the commonly used binary encoding which does not integrate physicochemical aspects of protein sequences.

Conflict of Interests

The authors declare that there is no financial interest or conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Bioscience Database Center, Japan, and the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Grant-in-Aid for Scientific Research on Innovation Areas "Biosynthetic Machinery: Deciphering and Regulating the System for Creating Structural Diversity of Bioactivity Metabolites (2007)").

References

[1] L. D. Stein, "Integrating biological databases," *Nature Reviews Genetics*, vol. 4, no. 5, pp. 337–345, 2003.

[2] M. He and S. Petoukhov, *Mathematics of Bioinformatics: Theory, Methods and Applications*, vol. 19, John Wiley & Sons, 2011.

[3] L. Kong, Y. Zhang, Z.-Q. Ye et al., "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine," *Nucleic acids research*, vol. 35, supplement 2, pp. W345–W349, 2007.

[4] M. Vendruscolo and G. G. Tartaglia, "Towards quantitative predictions in cell biology using chemical properties of proteins," *Molecular BioSystems*, vol. 4, no. 12, pp. 1170–1175, 2008.

[5] A. Coghlan, D. A. Mac Dónaill, and N. H. Buttimore, "Representation of amino acids as five-bit or three-bit patterns for filtering protein databases," *Bioinformatics*, vol. 17, no. 8, pp. 676–685, 2001.

[6] G. White and W. Seffens, "Using a neural network to backtranslate amino acid sequences," *Electronic Journal of Biotechnology*, vol. 1, no. 3, pp. 17–18, 1998.

[7] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992.

[8] O. Weiss, M. A. Jiménez-Montaño, and H. Herzel, "Information content of protein sequences," *Journal of Theoretical Biology*, vol. 206, no. 3, pp. 379–386, 2000.

[9] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 28, no. 1, p. 374, 2000.

[10] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drüke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, pp. 6395–6400, 2005.

[11] W. J. Krzanowski, *Principles of Multivariate Analysis*, Oxford University Press, 2000.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[14] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. Konig, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[15] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[16] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[17] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[18] R. Caspi, H. Foerster, C. A. Fulcher et al., "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 36, no. 1, pp. D623–D631, 2008.

[19] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin et al., "KNApSAcK: a comprehensive species-metabolite relationship database," in *Plant Metabolomics*, pp. 165–181, Springer, 2006.

[20] J. Bohlmann, G. Meyer-Gauen, and R. Croteau, "Plant terpenoid synthases: molecular biology and phylogenetic analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 8, pp. 4126–4133, 1998.

[21] I. Jollie, *Principal Component Analysis*, Wiley Online Library, 2005.

[22] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev et al., "The COG database: new developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Research*, vol. 29, no. 1, pp. 22–28, 2001.

[23] H.-L. Huang, I.-C. Lin, Y.-F. Liou et al., "Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties," *BMC Bioinformatics*, vol. 12, no. 1, article S47, 2011.

[24] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[25] M. Sandri and P. Zuccolotto, "Variable selection using random forests," in *Data Analysis, Classification and the Forward Search*, pp. 263–270, Springer, 2006.

[26] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, article 307, 2008.

[27] P. Kline, *An Easy Guide to Factor Analysis*, Routledge, 1994.

[28] J. D. Connolly and R. A. Hill, *Dictionary of Terpenoids. 1. Mono- and Sesquiterpenoids*, vol. 1, CRC Press, 1991.

[29] J. Degenhardt, T. G. Köllner, and J. Gershenzon, "Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants," *Phytochemistry*, vol. 70, no. 15-16, pp. 1621–1637, 2009.

[30] S. Ikeda, T. Abe, Y. Nakamura et al., "Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSAcK Motorcycle database," *Plant and Cell Physiology*, vol. 54, no. 5, pp. 711–727, 2013.

[31] R. Staden, "Sequence data handling by computer," *Nucleic Acids Research*, vol. 4, no. 11, pp. 4037–4051, 1977.

[32] D. C. Hyatt, B. Youn, Y. Zhao et al., "Structure of limonene synthase, a simple model for terpenoid cyclase catalysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 13, pp. 5360–5365, 2007.

[33] D. A. Nagegowda, M. Gutensohn, C. G. Wilkerson, and N. Dudareva, "Two nearly identical terpene synthases catalyze the formation of nerolidol and linalool in snapdragon flowers," *Plant Journal*, vol. 55, no. 2, pp. 224–239, 2008.

[34] N. J. Nieuwenhuizen, M. Y. Wang, A. J. Matich et al., "Two terpene synthases are responsible for the major sesquiterpenes emitted from the flowers of kiwifruit (Actinidia deliciosa)," *Journal of Experimental Botany*, vol. 60, no. 11, pp. 3203–3219, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 272915, 16 pages
http://dx.doi.org/10.1155/2014/272915

Research Article

OWL Reasoning Framework over Big Biological Knowledge Network

Huajun Chen,[1] Xi Chen,[1] Peiqin Gu,[1] Zhaohui Wu,[1] and Tong Yu[2]

[1] *Department of Computer Science, Zhejiang University, Hangzhou 310027, China*
[2] *Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China*

Correspondence should be addressed to Huajun Chen; huajunsir@zju.edu.cn

Received 30 October 2013; Revised 25 February 2014; Accepted 19 March 2014; Published 27 April 2014

Academic Editor: Md. Altaf-Ul-Amin

Recently, huge amounts of data are generated in the domain of biology. Embedded with domain knowledge from different disciplines, the isolated biological resources are implicitly connected. Thus it has shaped a big network of versatile biological knowledge. Faced with such massive, disparate, and interlinked biological data, providing an efficient way to model, integrate, and analyze the big biological network becomes a challenge. In this paper, we present a general OWL (web ontology language) reasoning framework to study the implicit relationships among biological entities. A comprehensive biological ontology across traditional Chinese medicine (TCM) and western medicine (WM) is used to create a conceptual model for the biological network. Then corresponding biological data is integrated into a biological knowledge network as the data model. Based on the conceptual model and data model, a scalable OWL reasoning method is utilized to infer the potential associations between biological entities from the biological network. In our experiment, we focus on the association discovery between TCM and WM. The derived associations are quite useful for biologists to promote the development of novel drugs and TCM modernization. The experimental results show that the system achieves high efficiency, accuracy, scalability, and effectivity.

1. Introduction

With the explosive growth of biological data on the web, large volume data sets are generated rapidly in the field of biology. Up to February 2014, linked life data (LLD), a data integration platform in the biological domain (http://linkedlifedata.com/sources.html), contains 10,192,505,364 statements and 1,553,620,636 entitlements. Entrez Gene has more than 100 million gene records (http://www.ncbi.nlm.nih.gov/gene/). Bioportal contains 24, 828,631,205 annotations (http://www.bioportal.bioontology.org). UniProt [1] knowledge base (UniProtKB/Swiss-Prot) contains 53,249,714 sequence entries, comprising about 10 billion amino acids (ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt). Besides the obvious scalability issues, heterogeneities from different resources are another major challenge for big biological data integration and analysis. Biological data covers a quite wide range, including proteins, pathways, diseases, targets, genes, Chinese medical herbs, symptoms, and syndromes, which usually come from multiple isolated sources and have different formats and taxonomies.

Based on domain knowledge from different disciplines all regarding human biological systems, the decentralized data repositories are implicitly connected (such as Figure 1). Thus, without regard to the formatting issue, we can logically regard the large-scale, heterogeneous, and complex-associated biological data as a big biological knowledge network. Biologists will benefit a lot by mining and discovering the hidden association information from the network. For example, the implicit associations between TCM and WM can help biologists have a better understanding of the complex biological system from the two perspectives of TCM and modern biology. Besides, they can also greatly promote the combination of TCM and WM, which will be useful in explaining the science of TCM and developing novel drugs.

However, faced with such large-scale, heterogeneous, and linked biological data, how to provide an efficient approach to model, integrate, and analyze the big biological network becomes a challenge. To support challenging these efforts, a

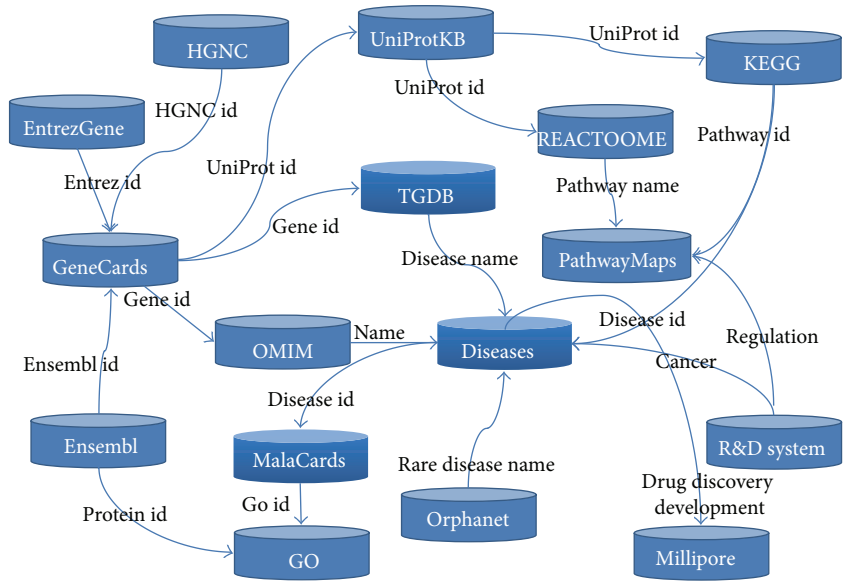


Figure 1: An implicitly linked biological knowledge network.

computational framework should meet the following several basic requirements:

(i) a biological conceptual network to model the concepts and corresponding relationships of modern biology and TCM;

(ii) a unified data model to integrate data across disparate data sources;

(iii) a collection of efficient and scalable computational services to analyze and discover new associations in the integrated biological knowledge network.

Semantic web technologies [2], most especially the OWL [3], are widely used in the life science and healthcare and provide us with an efficient way to create a conceptual model for the biological network by defining a specific ontology [4–6]. An ontology represents the formal and explicit concepts within a domain and the relationships between those concepts. In OWL, resources are identified with triple pattern $\langle \mathbf{s}, \mathbf{p}, \mathbf{o} \rangle$, representing a property **p** between subject **s** and object **o** [7]. It provides a simple graph data model for encoding networked data on the web using concepts and semantic relations. Every concept in the biological ontology maps a class of the biological network (e.g., a gene, herb, protein, drug, disease, etc.). The connections (e.g., treatment, possibleDrug, and encode) between biological classes are expressed as certain semantic rules (relations). For example, triple (Drug, treatment, Disease) represents a statement or a fact that drug class can link to disease class by the rule "treatment." The semantic rule "treatment" from the example can combine drug database and disease database. So semantic web technologies are able to help us construct a conceptual model to logistically organize and unify the versatile biological data by defining a unified biological ontology. Then based on the shared conceptual model, corresponding large-scale heterogeneous biological data sources can be mapped and merged into a big biological knowledge network.

A biological conceptual network can be divided into multiple chains. Every chain is composed of multiple classes of biological entities which are linked by several semantic rules (Figure 4). Reasoners are able to derive the implicit associations along the semantic rule chains. Thus, it becomes quite natural to make full use of reasoning method to accomplish the association discovery for the biological network. OWL reasoning technology is quite applicable to data analysis problems especially knowledge discovery problems involving complex semantic associations because it is able to infer logical consequences based on a set of asserted rules or axioms [8]. For rule-based reasoners, the OWL ontology definitions are first compiled into a set of rules. This rule set is then applied on the presented data set to generate the new inferred triples.

However, existing reasoners on single machine including Pellet [9], Fact++ [10], and Racer [11] work only on small or simple knowledge network because the reasoning algorithms are not scalable and usually are main memory oriented. As for the large biological data analysis, we have to devise an efficient and scalable reasoning algorithm. MapReduce is a simple and effective parallel programming model for big data processing on commodity computer cluster [12]. Users can implement a distributed program by simply specifying a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key. The computing framework is designed for batch-oriented work load, so it is quite effective in processing data/text intensive tasks. It is capable of processing the massive input data that is much larger than the total memories of these physical computing nodes. Developers also can add or delete computing nodes flexibly based on their

needs. These characteristics of MapReduce make it an ideal choice for big biological network reasoning. Figure 2 shows the basic workflow of MapReduce.

In this paper, we present a general OWL reasoning framework for modeling, integration, and analysis of the big biological network. Specifically speaking, our works are as follows.

(i) We design a unified biological ontology to model the complex biological conceptual network including TCM and WM. It provides an explicit specification of the conceptualization of the abstract view of the integrated biological network.

(ii) Based on the biological ontology, corresponding massive biological instance entities are integrated into a big linked biological knowledge network, which acts as the data model of the reasoning framework.

(iii) We propose several MapReduce-based property chain reasoning algorithms to discover the implicit associations between entities from the big biological knowledge network.

(iv) We present an implementation based on our prototype system and real biological data sets. The results show that the system achieves high efficiency, accuracy, scalability, and effectivity.

The remaining of this paper is organized as follows. In Section 2, we give the overall OWL reasoning framework over big biological network and related modules. Section 3 presents the detailed implementation of the distributed reasoning system. Section 4 introduces the experiment and the result analysis. Section 5 describes the related work, including OWL reasoning over biological data, massive biological data integration and search platforms, and large-scale semantic data reasoning systems. Section 6 gives conclusion.

2. OWL Reasoning Architecture and Modules

Three main modules have oriented our software development: ontology modeling module, data integration module, and distributed reasoning module. Ontology modeling module is used to construct a biological ontology to model the big biological conceptual network. Data integration module is responsible for creating a big linked biological knowledge network as the data model. Distributed reasoning module aims at deriving the implicit associations between different biological entities.

Figure 3 shows the schematic description of our OWL reasoning architecture. The unified biological ontology provides integration principles and reasoning rules to data integration module and distributed reasoning module, respectively. Data integration module outputs unified RDF triples to form the big biological linked knowledge graph as data model. Based on the conceptual model and data model, the distributed reasoning module implements the reasoning algorithm on a Hadoop cluster.

In the first subsection, we first introduce the method to build the unified biological ontology. The second subsection shows the process of data integration. The last part gives the brief introduction of the distributed reasoning process. The detailed implementation of the distributed reasoning module will be presented in the next section.

2.1. Unified BioTCM Ontology. To capture and model the complex biological network including modern biology and TCM, we construct a standard and sharable conceptual model by defining a unified biological ontology called unified BioTCM ontology with the help of some TCM and WM experts. It is an important component of the reasoning framework, playing a fundamental role in integrating disparate data sources and extracting reasoning rules, in that (1) it is a unique ontology, which captures the fundamental concepts, classes, and properties that help build the biological conceptual network including modern biology and TCM; (2) it defines the explicit semantic relations between different biological entities, which will act as the reasoning rules for cross-domain associated knowledge discovery.

Fundamentally, the unified BioTCM ontology provides a common generalized terminological and assertional base for mapping from multiple sources to a unified mapping schema. It is mainly a terminology box (TBOX) which consists of class hierarchies and class restrictions defined with object properties.

Figure 4 gives a brief introduction of the associated conceptual model for TCM and WM network. In the unified BioTCM ontology model, there are many key concepts: disease, drug, gene, protein, syndrome, symptom, target, TCM herb, TCM symptom, TCM syndrome, and so on. Mainly, specific disorders of certain genes can affect the encoding proteins, which cause diseases to appear. Proteins also can affect the gene expression. Drugs are used to treat diseases by interacting with the sequential proteins through possible targets and involved pathways. A pathway can trigger the assembly of new protein molecules. The herbs are the constituents making up drugs. The major link between modern biology and Chinese medicine is based on the fact that some western diseases are similar to TCM diseases, and it has been found that certain genes are responsible for some TCM diseases and that certain remedies (e.g., herbs) might cure the genetic disease by possible biological targets [13–18].

In Figure 4, the big biological conceptual network can be divided into multiple reasoning property chains. For the associated network of unified BioTCM ontology, we identify several property chains. Every property chain, consisting of several sequential semantic rules, can capture the implicit associations between every two specific biological classes by modeling the potential interactions of intermediate biological entities. This association information is useful in understanding the mechanisms of action of biological entities as a whole, especially those entities biological researchers are not familiar with.

2.2. Biological Data Integration. Since we have designed a well-defined comprehensive biological ontology, the TBOX from the ontology tells us which data needs to be collected and how its schemas should be. Thus, a big linked biological knowledge graph (also called assertion component (ABOX))

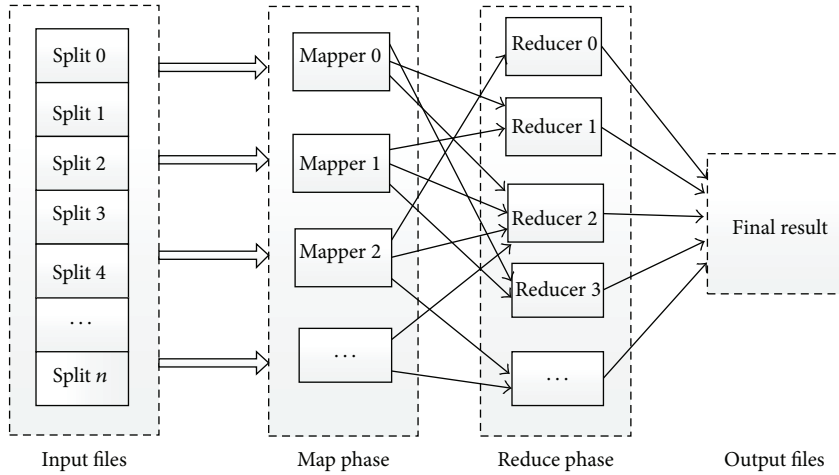


Figure 2: MapReduce workflow: map and reduce.

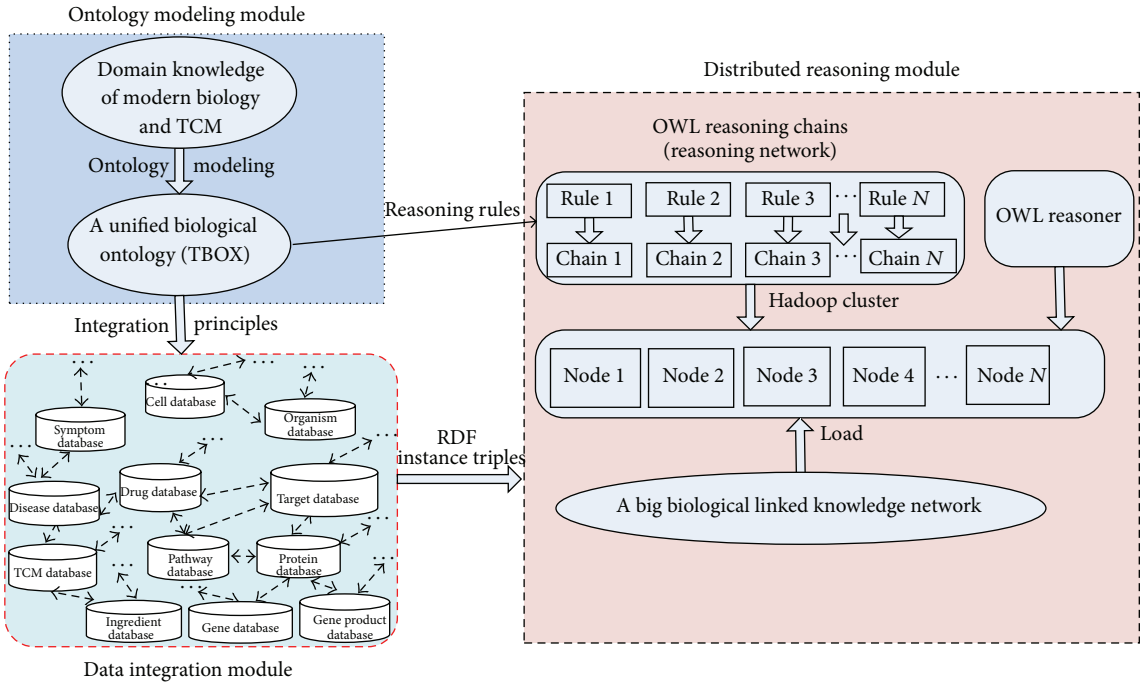


Figure 3: OWL reasoning framework and three composition modules.

can be created based on the TBOX. Another challenge in the integration of biological data lies in the format. Although there are numerous bioinformatics databases available, most of them do not share the uniform format. We utilize many different ways to transform these data into a standard RDF format.

For some text data, we utilize simple text mining method to extract required instance triples. For relational data, we use RDB2RDF tools such as D2R to implement transforming [19]. We also get some online gene data by web service, such as the NCBI efetch service (http://www.ncbi.nlm.nih.gov/books/NBK43082/). As a result, a big and comprehensive linked biological knowledge network is formed.

2.3. Distributed Reasoning. The distributed reasoning module is the core of our reasoning framework. It is composed of three parts: reasoning rules, reasoning objects, and distributed reasoning algorithm. Reasoning rules depict some basic association relationships between biological classes, which can be extracted from the unified biological ontology. Reasoning objects represent the biological entities that we want to discover the implicit associations between them. It can be formed by constructing a linked knowledge network.

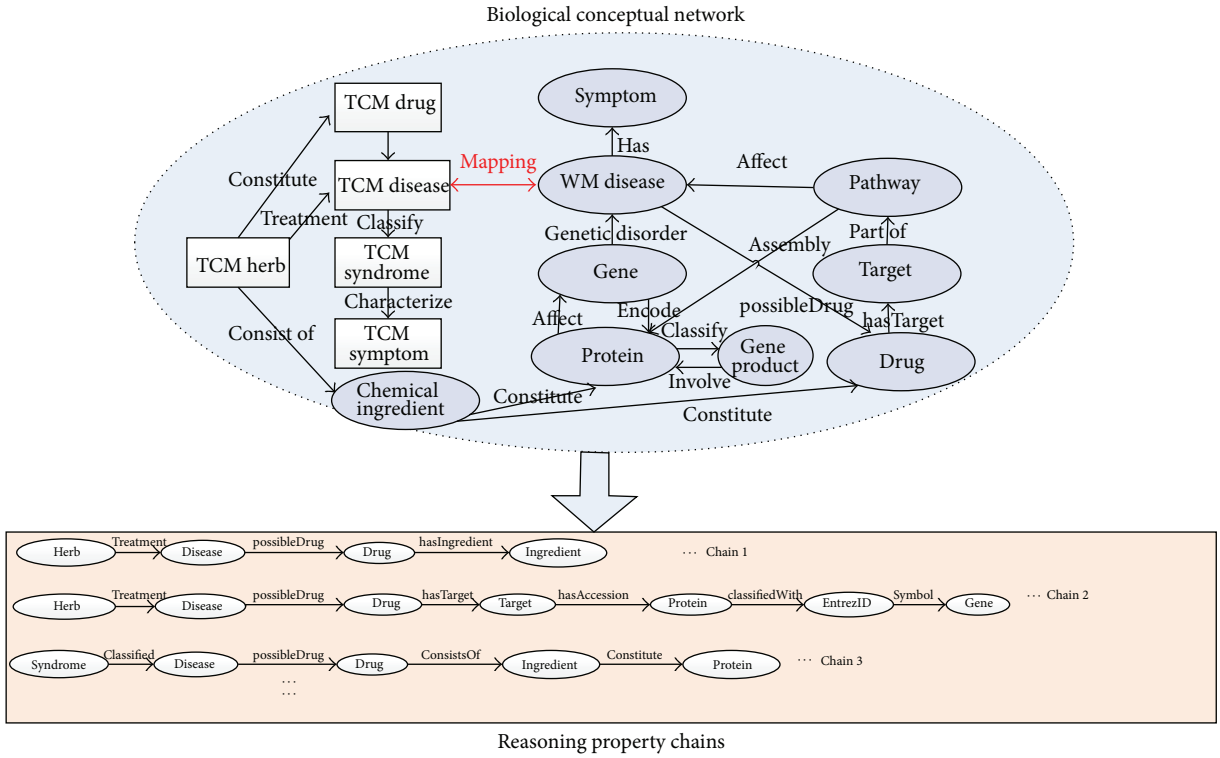


Figure 4: Biological conceptual network and corresponding reasoning property chains.

Distributed reasoning algorithm is dedicated to deploying an efficient and scalable reasoner over big biological network based on reasoning rules. The first two parts have been described. We will show the detailed realization of the distributed reasoning algorithm in the next section.

3. OWL Reasoning Algorithms Based on MapReduce

In the section, we first describe a typical biological reasoning problem and redefine it formally. Then we present a general reasoning algorithm framework and subsequently introduce a naïve OWL reasoning algorithm based on MapReduce. We call this implementation naïve because it is easy to understand but performs poorly. Therefore, in the next part, an improved algorithm is presented to deal with the conflict between the parallel mechanism of MapReduce and the sequential demands of a reasoning rule set. At last, to enhance the parallel capability and efficiency of reasoning system, a multichains reasoning algorithm is presented to accomplish multiple property chains reasoning processes in an iterative MapReduce job.

3.1. Biological Reasoning Example. Traditional Chinese medicine, which has existed for thousands of years in China, is yet to become an integral part of the standard healthcare system in western countries due to a lack of scientific evidence for its efficacy and safety [20]. Meanwhile, TCM is also gaining increasing attention from western healthcare practitioners because it is making favorable contributions to the development of novel drugs that are made of natural herbs. So it will become quite useful to reveal some implicit relationships between TCM and WM. Problem 1 describes a typical biological reasoning example.

Problem 1. In recent years, several herbs were found to exhibit a variety of effects through regulating a wide range of gene expressions or protein activities [17, 18]. To discover the implicit mappings between Chinese herbs and genes is a problem for biological researchers to solve for understanding the possible therapeutic mechanisms of TCMs via gene regulations.

We are able to get associations between herb and gene based on the corresponding OWL transitive property chain in the biological network (Chain 2 in Figure 4). The transitive relationship can be derived through the shared intermediates. In our reasoning system, relationships between two kinds of biological entities are expressed as reasoning rules. Typically, as is shown in Figure 4, some basic reasoning rules have been given directly by the biological ontology, such as "treatment" and "possibleDrug." But there does not exist a direct association rule between herb and gene. On this occasion, we need to create a reasoning rule set based on existing basic reasoning rules that can link them implicitly.

3.2. Formal Definition of Reasoning Problem. To address the problem efficiently, we define the following concepts.

Table 1: Variables table.

Variable symbol	Definition	Example
RRC	Reasoning rule chain	$\mathbf{RRC_0}$
OPC	OWL property chain	$\mathbf{OPC_0}$
PCS	Property chain set	$\mathbf{PCS_0}$
PID	Property ID	1
ORN	OWL reasoning network	$\mathbf{ORN_0}$
ARS	Associated result set	$\mathbf{ARS_0}$
G	Instance triple graph	$\mathbf{G_0}$
Class_k	An entity belonging to certain class	$\mathbf{Herb_0}$
$\mathbf{T_k}$	An instance triple	$\mathbf{T_0}$
R_k	A rule triple	$\mathbf{R_0}$
P_k	Property of the kth rule triple	$\mathbf{P_0}$

Definition 2 (reasoning rule chain (RRC)). A reasoning rule chain is a set of sequential basic reasoning rules. Every basic reasoning rule is given in advance which is formalized as a rule triple such as (Herb, treatment, Disease). The reasoning rule chain of Problem 1 can be described as $\mathbf{RCC_0}$ = {(Herb, treatment, Disease), (Disease, possibleDrug, Drug), (Drug, hasTarget, Target), (Target, hasAccession, Protein), (Protein, classifiedWith, EntrezID), (EntrezID, symbol, Gene)}.

Definition 3 (OWL property chain (OPC)). A OPC is made up of one or more sequential properties from the reasoning rule chain. Given a reasoning rule chain such as $\mathbf{RCC_0}$, $\mathbf{P_k}$ refers to the property of the kth rule triple. Initially, $\mathbf{OPC_k}$ equals $\mathbf{P_k}$. Therefore, we can get the following results: $\mathbf{OPC_0}$ = treatment, $\mathbf{OPC_1}$ = possibleDrug, ..., $\mathbf{OPC_5}$ = symbol. Then several consecutive sequential OPCs will form a new OPC with operation ⊗ if they meet merging condition. For example, if there exist some triples, ($\mathbf{Herb_0}$, treatment, $\mathbf{Disease_0}$), ($\mathbf{Disease_0}$, possibleDrug, $\mathbf{Drug_0}$), ..., ($\mathbf{EntrezID_0}$, symbol, $\mathbf{Gene_0}$), then we can derive a new triple ($\mathbf{Drug_0}$, P, $\mathbf{Gene_0}$) where P is expressed as ($\mathbf{OPC_0} \otimes \mathbf{OPC_1} \otimes \mathbf{OPC_2} \ldots \otimes \mathbf{OPC_5}$). To some extent, the reasoning process can be regarded as the iterated merging operations of OPCs.

Definition 4 (property chain set (PCS)). As the name suggests, the PCS is a set of sequential OPCs in a given triple graph. For $\mathbf{RCC_0}$, the initial PCS is expressed as $\mathbf{PCS_0}$ = {treatment, possibleDrug, hasTarget, hasAccession, classifiedWith, symbol}. In the process of reasoning, the PCS will vary with OPCs.

Definition 5 (property ID (PID)). We allocate an ID called PID to every OPC in the PCS. Initially, the PID of the first OPC in the $\mathbf{PCS_0}$ is set as 0, the second is 1,. . ., and the PID of the last OPC is 5 (the length of $\mathbf{PCS_0}$ is 6). Correspondingly, every instance triple also owns a PID because its predicate maps some OPC. For those triples whose OPCs are not included in the PCS, the PID is assigned as −1. These triples should be ignored in the process of reasoning.

Algorithm 1 is a specific example for these definitions (take chain 2 in Figure 4 for example). Table 1 shows the variable symbols and related definitions used by the paper. Based on the above definitions, Problem 1 can be redefined formally as Problem 6.

Problem 6. Input a quad (G, PCS_0, Herb, Gene); we are required to solve the problem: compute the triple collection $S = \{(O_0, \text{OPC}, O_5) \mid O_0 \in \text{Herb}, \text{OPC} = (\text{treatment} \otimes \text{possibleDrug} \otimes \text{hasTarget}. \ldots \otimes \text{symbol}), O_5 \in \text{Gene}\}$. G is the instance triple graph. The PCS_0 is the property chain set of G. Herb and Gene represent the two classes needed to explore implicit mappings.

Consider the following instance triple graph: $\mathbf{G_0}$ = {$\mathbf{T_0}$($\mathbf{Herb_0}$, treatment, $\mathbf{Disease_0}$), $\mathbf{T_1}$($\mathbf{Disease_0}$, possible Drug, $\mathbf{Drug_0}$), $\mathbf{T_2}$($\mathbf{Drug_0}$, hasTarget, $\mathbf{Target_0}$), $\mathbf{T_3}$($\mathbf{Target_0}$, hasAccession, $\mathbf{Protein_0}$), $\mathbf{T_4}$($\mathbf{Protein_0}$, classifiedWith, $\mathbf{EntrezID_0}$), $\mathbf{T_5}$ ($\mathbf{EntrezID_0}$, symbol, $\mathbf{Gene_0}$), $\mathbf{T_6}$($\mathbf{Herb_1}$, treatment, $\mathbf{Disease_0}$), $\mathbf{T_7}$($\mathbf{Target_0}$, geneSequence, $\mathbf{Sequence_0}$)}. According to the above three definitions, we can calculate the PID for every instance triple. For example, $\mathbf{T_0}$'s PID is 0 because its predicate "treatment" is the first OPC in $\mathbf{PCS_0}$. $\mathbf{T_7}$'s PID is −1 because its predicate "geneSequence" is not included in $\mathbf{PCS_0}$.

3.3. Framework of OWL Reasoning Algorithm. Given an input Quad0 = ($\mathbf{G_0}$, $\mathbf{PCS_0}$, Herb, Gene), to compute solution domain, we need to keep applying the rules to reason until we finish deriving the desired triples (fixpoint). It will involve multiple iterations. The number of iterations depends on the complexity of the input and efficiency of the algorithm.

In the workflow of the algorithm as shown in Algorithm 2, we firstly complete initialization by inputting a quad and setting a global variable to check fixpoint condition. Then the algorithm comes into the procedure of iterating. In every iteration, we load the triple graph and PCS. Then we perform a join with a MapReduce job. At last, new input triple graph and PCS are calculated for the next iteration.

3.4. Naïve OWL Reasoning Algorithm. To derive a new triple, we need another two triples as the sources. It is quite natural and direct to connect Herb with Drug through intermediate Disease based on the rule chain in Figure 4. That is to say, we firstly process the instance triples whose PID is 0 or 1 in every iteration. Based on the idea, we can specify the join condition: the objects of triples whose PID equals 0 must match the subjects of other triples whose PID is 1. For the sake of description, we define the concept of Join Candidate Set.

Definition 7 (Join Candidate Set). Join Candidate Set is a binary set of the instance triples that meet join condition. Once there exist two instance triples satisfied with the above join condition, such as $\mathbf{T_0}$ ($\mathbf{Herb_0}$, $\mathbf{P_0}$, $\mathbf{Disease_0}$), $\mathbf{T_1}$ ($\mathbf{Disease_0}$, $\mathbf{P_1}$, $\mathbf{Drug_0}$), we should add the element ($\mathbf{T_0}$, $\mathbf{T_1}$) to the Join Candidate Set. In every iteration, we firstly compute the Join Candidate Set, and then we can perform joins to derive some new triples over elements in the Join Candidate Set.

Reasoning Rule Chain: {(herb, treatment, Disease), (Disease, possibleDrug, Drug), (Drug, hasTarget, Target), (Target, hasAccession, Protein), (Protein, classifiedWith, EntrezID), (EntrezID, symbol, Gene)}
OWL Property Chain: OPC_0=treatment, OPC_1=possibleDrug, OPC_2=hasTarget, OPC_3=hasAccession, OPC_4=classifiedWith, OPC_5=symbol
Property Chain Set: {OPC_0, OPC_1, OPC_2, OPC_3, OPC_4, OPC_5}
Property ID: {OPC_0=0, OPC_1=1, OPC_2=2, OPC_3=3, OPC_4=4, OPC_5=5}

Algorithm 1: Formalized definitions for a specific reasoning example.

Initialization: instance triple graph, G_0; Property Chain Set, PCS_0; two classes required to explore implicit semantic associations, Herb and Gene; number that has been iterated, $I = 0$; number needed to be iterated, M;
Iteration:
while $I < M$ **do**
Step 1. Load triple graph and PCS on the current iteration, G_I, PCS_I;
Step 2. Group instance triples based on join key;
Step 3. Derive new instance triples;
Step 4. Update input instance triple graph, G_{I+1};
Step 5. Update PCS, PCS_{I+1};
Step 6. $I \leftarrow I + 1$;
end while

Algorithm 2: Framework of OWL reasoning algorithm.

After an iteration, the first two OPCs ($\mathbf{P_0}$ and $\mathbf{P_1}$) in the PCS will merge to a new OPC ($\mathbf{P_0} \otimes \mathbf{P_1}$) whose PID is set to 0. Meanwhile, the PID of all other OPCs reduces by 1. Obviously, the length of the PCS will also reduce by 1. When length of the PCS becomes 1, the algorithm ends. So for a PCS whose initial length is n, we need $n - 1$ iterations to finish reasoning.

Let us consider the same input Quad0 as above. In the first iteration, we derive two triples by computing the Join Candidate Set $\{(\mathbf{T_0}, \mathbf{T_1}), (\mathbf{T_6}, \mathbf{T_1})\}$: $\mathbf{T_8}(\mathbf{Herb_0}, \mathbf{P_0} \otimes \mathbf{P_1}, \mathbf{Drug_0})$, and $\mathbf{T_9}(\mathbf{Herb_1}, \mathbf{P_0} \otimes \mathbf{P_1}, \mathbf{Drug_0})$. Then $\{\mathbf{T_0}, \mathbf{T_1}, \mathbf{T_6}\}$ will be deleted from the input data. $\mathbf{T_7}$ is also removed because its OPC (GeneSequence) is not included in the $\mathbf{PCS_0}$. So the new input quad becomes $QUAD_1$ ($\mathbf{G_1}$, $\mathbf{PCS_1}$, Herb, Gene). Consider $\mathbf{G_1} = \{\mathbf{T_8}(\mathbf{Herb_0}, \mathbf{P_0} \otimes \mathbf{P_1}, \mathbf{Drug_0})$, $\mathbf{T_9}(\mathbf{Herb_1}, \mathbf{P_0} \otimes \mathbf{P_1}, \mathbf{Drug_0})$, $\mathbf{T_2}(\mathbf{Drug_0}$, hasTarget, $\mathbf{Target_0})$, $\mathbf{T_3}(\mathbf{Target_0}$, hasAccession, $\mathbf{Protein_0})$, $\mathbf{T_4}(\mathbf{Protein_0}$, classifiedWith, $\mathbf{EntrezID_0})$, $\mathbf{T_5}(\mathbf{EntrezID_0}$, symbol, $\mathbf{Gene_0})\}$. $\mathbf{PCS_1} = \{\mathbf{P_0} \otimes \mathbf{P_1}, \mathbf{P_2}, \mathbf{P_3}, \mathbf{P_4}\}$. Then we continue to apply the same method to perform joins until we get the final results: ($\mathbf{Herb_0}$, $\mathbf{P_0} \otimes \mathbf{P_1} \otimes \mathbf{P_2} \otimes \mathbf{P_3} \otimes \mathbf{P_4}$, $\mathbf{Gene_0}$) and ($\mathbf{Herb_1}$, $\mathbf{P_0} \otimes \mathbf{P_1} \otimes \mathbf{P_2} \otimes \mathbf{P_3} \otimes \mathbf{P_4}$, $\mathbf{Gene_0}$). As the length of $\mathbf{PCS_0}$ is 5, the total number of iterations is 4. The first iteration process is shown in Figure 5.

When deployed in MapReduce, every MapReduce job corresponds to an iteration procedure which performs a join. Mapper is used to separate all input triples into three groups based on PID: triples needed to be joined immediately, triples needed to be processed later, and irrelevant triples. Reducer is responsible for implementing joins to recalculate new input triple graph for the next iteration. At last, PCS is updated. Another similar MapReduce job continues to be executed until the length of PCS becomes 1.

3.5. Efficient OWL Reasoning Algorithm. The previously presented implementation is straightforward but is inefficient because it involves too many iterations and wastes lots of valuable computing resources in an iteration. Algorithm 2 only implements joins on these instance triples whose PID is 0 or 1 in one iteration, while other instance triples are not processed concurrently. As a result, it needs $(n - 1)$ iterations to complete reasoning where n represents the length of the initial PCS. So we introduce a more efficient algorithm to greatly decrease the number of jobs and time required for reasoning computation.

In fact, we can perform more joins in an iteration if we set out a more flexible join requirement. Specifically, the join requirements contain two conditions.

(1) The PIDs of two triples' OPCs are adjacent strictly.

(2) The object of triple owning a smaller PID matches the other triple's subject.

For example, there are three instance triples as follows: $\mathbf{T_0}(\mathbf{Herb_0}$, treatment, $\mathbf{Disease_0})$, $\mathbf{T_1}(\mathbf{Disease_0}$, possibleDrug, $\mathbf{Drug_0})$, and $\mathbf{T_2}(\mathbf{Drug_0}$, hasTarget, $\mathbf{Target_0})$. As $\mathbf{T_1}$ meets join conditions both with $\mathbf{T_0}$ and $\mathbf{T_2}$, the Join Candidate Set should be $\{(\mathbf{T_0}, \mathbf{T_1}), (\mathbf{T_1}, \mathbf{T_2})\}$. So we derive two triples $\mathbf{T_3}$ ($\mathbf{Herb_0}$, $\mathbf{P_0} \otimes \mathbf{P_1}$, $\mathbf{Drug_0}$) and $\mathbf{T_4}(\mathbf{Disease_0}$, $\mathbf{P_1} \otimes \mathbf{P_2}$, $\mathbf{Target_0})$. It is obvious that $\mathbf{T_3}$ and $\mathbf{T_4}$ do not meet join conditions in next iteration. Therefore, we cannot derive the right result ($\mathbf{Herb_0}$, $\mathbf{P_0} \otimes \mathbf{P_1} \otimes \mathbf{P_2}$, $\mathbf{Target_0}$).

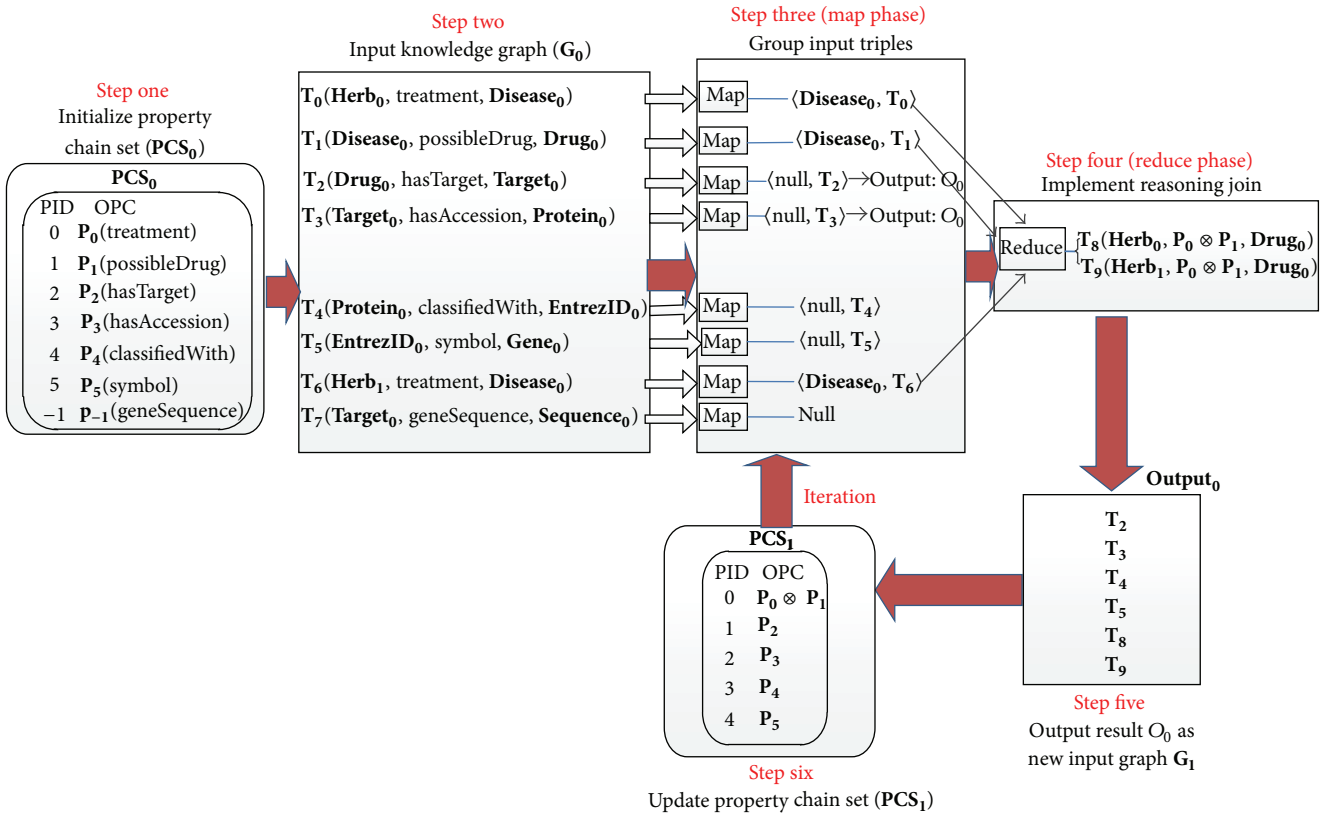


FIGURE 5: The workflow of naïve reasoning algorithm in the first iteration.

We are able to solve the problem if we add another restricted condition called Parity Judgment Rule to join requirement. Firstly, let us give the definition of Parity Judgment Rule.

Rule 1 (Parity Judgment Rule). We regard the Parity Judgment Rule as the third join condition. It is based on this principle that a triple (assuming k represents its PID and is an odd number) only performs joins with triples whose PID is $(\mathbf{k}-\mathbf{1})$. In particular, for an instance triple $\mathbf{T_k}(\mathbf{X_k}, \text{OPC}, \mathbf{Y_k})$, if PID of the OPC is an odd number $\mathbf{k}$, the join key is represented as $(\mathbf{k}-\mathbf{1})\_\mathbf{X_k}$. Otherwise, the join key is $\mathbf{k}\_\mathbf{Y_k}$. As for the above three triples, the join condition guarantees that $\mathbf{T_1}$ only connects with $\mathbf{T_0}$ in the first iteration. Then we can derive the right result in the second iteration.

As is shown in Figure 6, based on the above three join conditions, we can divide all biological entities except irrelevant contents (Sequence) into 3 ($\lceil \mathbf{N/2} \rceil$) groups where N represents the length of $\mathbf{PCS_0}$. Then we perform joins between the triples from the same group in an iteration. As a result, the derived triples will be the new input graph for the next iteration. Meanwhile, we halve the PCS by merging the two adjacent OPCs to one new OPC with the operation $\otimes$. Subsequently, we continue to apply similar method to reason until the length of PCS becomes 1. Obviously, this algorithm makes full use of the computing capacity of cluster nodes to limit the number of total iterations to 3 ($\log N$), which will greatly improve the efficiency of reasoning, compared to 5 $(N-1)$ iterations in the previous naïve algorithm.

Consider the same input quad Quad_0. In the first iteration, the Join Candidate Set is calculated as $\{(\mathbf{T_0}, \mathbf{T_1}), (\mathbf{T_1}, \mathbf{T_6}), (\mathbf{T_2}, \mathbf{T_3}), (\mathbf{T_4}, \mathbf{T_5})\}$ based on join conditions. Then new triples are derived as follows: $\{\mathbf{T_8}(\mathbf{Herb_0}, \mathbf{P_0 \otimes P_1}, \mathbf{Drug_0}), \mathbf{T_9}(\mathbf{Herb_1}, \mathbf{P_0 \otimes P_1}, \mathbf{Drug_0}), \mathbf{T_{10}}(\mathbf{Drug_0}, \mathbf{P_2 \otimes P_3}, \mathbf{Protein_0}), \mathbf{T_{11}}(\mathbf{Protein_0}, \mathbf{P_4 \otimes P_5}, \mathbf{Gene_0})\}$. Then we get a new graph $\mathbf{G_1} = \{\mathbf{T_8}, \mathbf{T_9}, \mathbf{T_{10}}, \mathbf{T_{11}}\}$. The PCS is also updated as $\mathbf{PCS_1} = \{\mathbf{P_0 \otimes P_1}, \mathbf{P_2 \otimes P_3}, \mathbf{P_4 \otimes P_5}\}$. So the first iteration ends up with a new smaller input quad Quad_1 =($\mathbf{G_1}$, $\mathbf{PCS_1}$, Herb, Gene). Similarly, in the second iteration, we work out the new Join Candidate Set which is expressed as $\{(\mathbf{T_8}, \mathbf{T_{10}}), (\mathbf{T_9}, \mathbf{T_{10}})\}$ and another new graph is recalculated as $G_2 = \{\mathbf{T_{11}}(\mathbf{Protein_0}, \mathbf{P_4 \otimes P_5}, \mathbf{Gene_0}), \mathbf{T_{12}}\ (\mathbf{Herb_0}, \mathbf{P_0 \otimes P_1 \otimes P_2 \otimes P_3 \otimes P_4}, \mathbf{Protein_0}), \mathbf{T_{13}}(\mathbf{Herb_1}, \mathbf{P_0 \otimes P_1 \otimes P_2 \otimes P_3 \otimes P_4}, \mathbf{Protein_0})\}$. The new PCS is also updated as $\{\mathbf{P_0 \otimes P_1 \otimes P_2 \otimes P_3}, \mathbf{P_4 \otimes P_5}\}$. Then we implement the last iteration. The Join Candidate Set is $(\mathbf{T_{12}}, \mathbf{T_{11}})$, $(\mathbf{T_{13}}, \mathbf{T_{11}})$. The desired triples are derived as follows: $\{\mathbf{T_{14}}(\mathbf{Herb_0}, \mathbf{P_0 \otimes P_1 \otimes P_2 \otimes P_3 \otimes P_4 \otimes P_5}, \mathbf{Gene_0}), \mathbf{T_{15}}(\mathbf{Herb_1}, \mathbf{P_0 \otimes P_1 \otimes P_2 \otimes P_3 \otimes P_4 \otimes P_5}, \mathbf{Gene_0})\}$. The length of $\mathbf{PCS_0}$ is 5. So the algorithm ends after 3 iterations. The first iteration scenario is shown in Figure 7.

When implemented in MapReduce, Mapper is used to group all triples that meet join conditions into a Reducer.

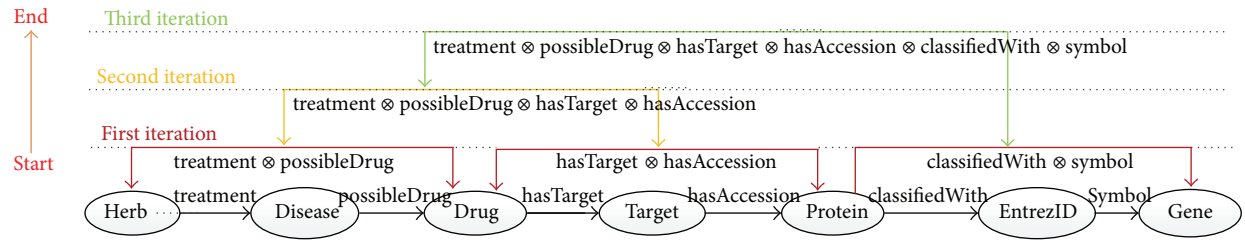


Figure 6: Parallel reasoning process based on property chain.

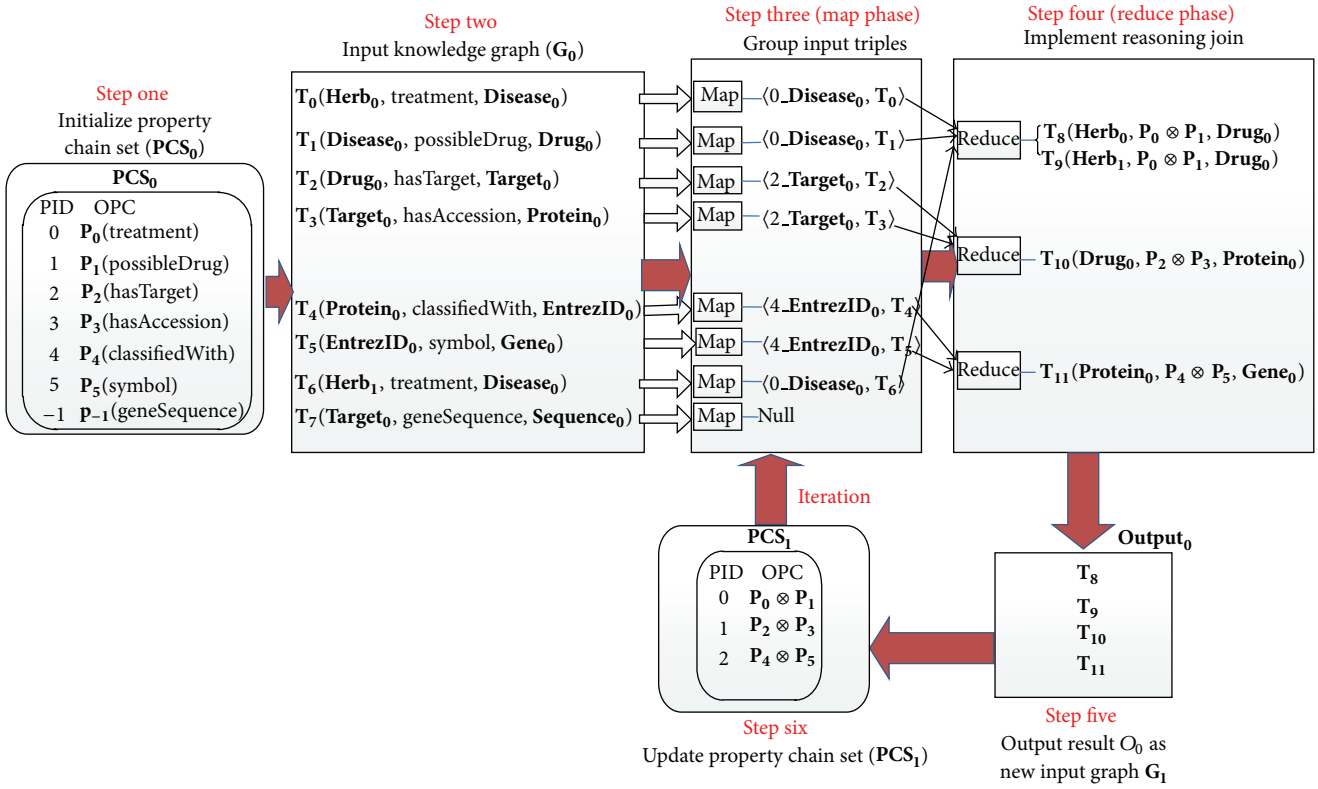


Figure 7: The workflow of parallel property chain reasoning algorithm in the first iteration.

Reducer is responsible for computing the Join Candidate Set and deriving new input triples for the next iteration. The algorithm is demonstrated in Algorithm 3.

In map function, we compute join key for every triple based on Parity Judgement Rule. The join key is used as intermediate key. Intermediate value is the triple itself. Each map process outputs several pairs of intermediate results $\langle ik, iv \rangle$.

In reduce function, we firstly divide input triples into two classes based on the parity of triple's PID. If PID is odd, we extract triple's object to a set called ObjectList. Otherwise, we add triple's subject to the set called SubjectList. We are able to get the Join Candidate Set based on ObjectList and SubjectList. Then we compute the shared OPC for all new derived triples. Subsequently, the output pairs $\langle ok, ov \rangle$ are written to HDFS (Hadoop Distributed File System) where ok is null and ov is the derived triples. The triples will form a new input triple graph for the next iteration.

At last, the PCS is updated by merging the two adjacent OPCs to one new OPC. Then another similar MapReduce job is launched until the length of PCS becomes 1.

3.6. Multichains Parallel Reasoning Algorithm. The previously described reasoning algorithms are intended to derive the association information among the entities from two specific biological classes in an iterative MapReduce job. A significant feature of the big biological network lies in the complex association relationships between biological data. Every property chain only represents the implicit associations between two specific biological classes. Meanwhile, there exist multiple property chains in the big biological network. If we want to get the associations between multiple pairs of biological classes, the reasoning process has to be repeated several times. This will result in low efficiency and waste I/O, network bandwidth, and CPU resources, where large-scale data must be reloaded and reprocessed at each iterated

```
Map(key, value)
    // key:linenumber(irrelevant)
    // value:instance triple
    PID = PCS·getPID(triple·predicate);
    //get the PID of the triple
    If PID == −1 then
        return;
    end if
    if PID == (len − 1)&len%2 == 1 then
        emit(null, triple);
        return;
    end if
    if PID%2 == 1 then
        key=(PID − 1)+"_"+triple·getSubject();
    else
        key=PID+"_"+triple·getObject();
    end if
    emit(key, value);
Reduce(key, value)
    // key:join key
    // value:triple
    subjectList = empty;
    objectList = empty;
    len=PCS·length;
    for each triple ∈ value do
        PID = PCS·getPID(triple·predicate);
        if PID%2 == 1 then
            subjectList·add(triple·subject);
        else
            objectList·add(triple·object);
        end if
    end for
    new_OPC=ComputeOPC();
    for each s ∈ subjectList do
        for each o ∈ objectList do
            emit(null, triple(s, new_OPC, o));
        end for
    end for
```

Algorithm 3: Efficient OWL reasoning algorithm based on MapReduce.

job. So to enhance the efficiency and parallel capability of the reasoning system, an improved multichains parallel reasoning algorithm is presented below.

First, we give two related definitions.

Definition 8 (OWL reasoning network (ORN)). An OWL reasoning network is a set of property chain sets (PCS). In previous example, the ORN only has a PCS element ($\mathbf{PCS_0}$). In this new reasoning scenario, the **i** in $\mathbf{PCS_{ij}}$ represents the *i*th element in ORN, while **j** denotes the new PCS after **j** iterations. Similarly, the **i** in $\mathbf{P_{ij}}$ represents which PCS the property belongs to, while **j** denotes its order in corresponding PCS.

Definition 9 (associated result set (ARS)). An associated result set is a collection of tuples like ($\mathbf{CLASS_1}$, $\mathbf{CLASS_2}$). Every tuple represents two biological classes that we want to discover implicit association information between them. Every reasoning rule chain or property chain set corresponds to an element in ARS. For example, $\mathbf{PCS_0}$ corresponds to the binary set (Herb, Gene).

So the multiple chains reasoning problem is defined formally as Problem 10.

Problem 10. Input a three tuple ($\mathbf{G}$, $\mathbf{ORN_0}$, $\mathbf{ARS_0}$), where $\mathbf{G}$ is the instance triple graph, $\mathbf{ORN_0}$ represents a concrete OWL reasoning network, and ARS_0 denotes the associated result set; the reasoner is required to solve the problem: find out the solution domain $S = \{(O_0, \mathrm{OPC}, O_k) \mid (O_0, O_k) \in \mathbf{ARS_0}$, OPC denotes the OWL property chain that links O_0 and O_k together.}.

For every reasoning rule chain in $\mathbf{ORN_0}$, the principle and process of reasoning are the same as Algorithm 3. The key task of multichains parallel reasoning is to ensure that every reasoning job can be executed simultaneously but does not affect the others.

Consider the $\mathbf{input_0}$ = ($\mathbf{G_0}$, $\mathbf{ORN_0}$, $\mathbf{ARS_0}$). $\mathbf{G_0}$ and ($\mathbf{ORN_0}$) are shown in Figure 8 (step one and step two). The $\mathbf{ARS_0}$ is {(herb, gene), (herb, ingredient)}. As the multiple reasoning rule chains in the $\mathbf{ORN_0}$ may intersect, the cross section (instance triples) should participate in the multiple reasoning jobs separately. Take $\mathbf{T_0}$ for example; since the $\mathbf{PCS_{00}}$ and $\mathbf{PCS_{10}}$ all contain property "treatment," the Mappers should emit two key/value pairs into different Reducers to isolate the two reasoning chains. For $\mathbf{T_3}$ and $\mathbf{T_6}$, as their properties "hasIngredient" and "classifiedWith" only exist in $\mathbf{PCS_{00}}$ or $\mathbf{PCS_{10}}$, Mappers only need to output a key/value pair. At the Mapper process, we add another optimization scheme for the triples whose PID is $n - 1$ (n is the length of corresponding PCS and n is an odd number). Because it is obvious that these triples do not meet join conditions, we only need to directly output the triples to HDFS without the processing of Reducer. Compared to Algorithm 3, Mappers only need to add a label reasoning chain identification to the intermediate key and Reducers remain almost unchanged. The number of iterations depends on the length of the longest PCS element in the $\mathbf{ORN_0}$. The first iteration process of the multichains parallel reasoning is shown in Figure 8.

4. Experiment Evolution

Our experiment aims at discovering the implicit associations between TCM and WM. In particular, it focuses on deriving the association information between Chinese herbs and western medical genes, drug ingredients. This information hidden in the big biological network is of quite value in promoting the development of novel drugs, TCM modernization, and understanding the complex biological system in whole. The distributed reasoner uses multichains parallel reasoning algorithm with two reasoning rule chains shown in Figure 4 (Chain 1 and Chain 2).

4.1. Data Preparation and Experimental Environment. As the data model, a big linked biological knowledge

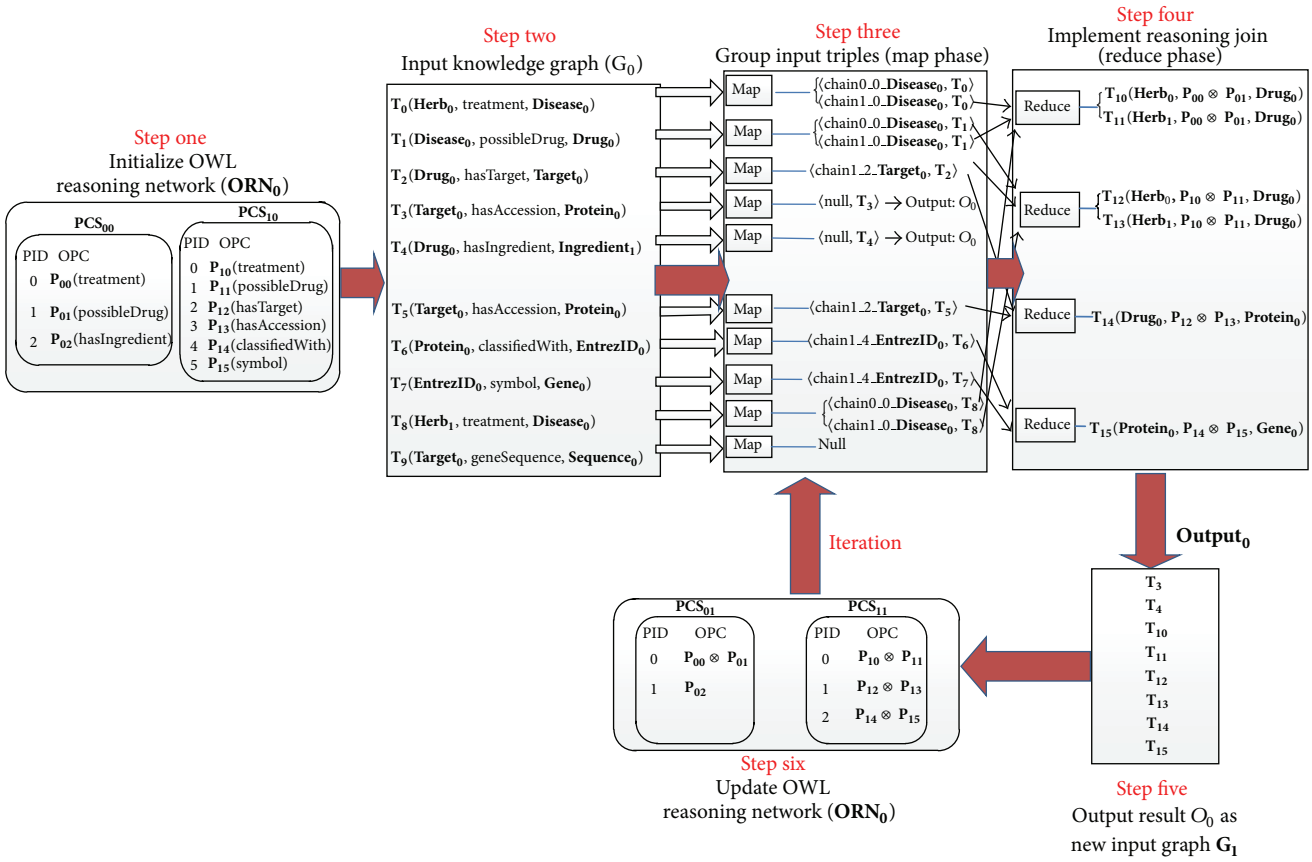


Figure 8: The workflow of multichains parallel reasoning algorithm in the first iteration.

network is constructed in Figure 9 (available at http://www.biotcm.org/mappingsearch/index.html; please click the buttons "List" and "Graph" to see the descriptions of all the ontology bases and overall knowledge graph, resp.). The linked biological knowledge network acts as the background database of the BioTCM (http://www.biotcm.org/), which is an integrated association discovery platform of modern biomedicine and Chinese medicine developed by us. It includes most of the typical biological ontologies across WM and TCM including Gene ontology [21], Disease ontology [22], Diseasome ontology [23], DrugBank [24], TCMGeneDit [25], TCMLS [26], UniProt [1], and NCBI Gene [27]. Every oval in the linked knowledge graph is marked up by a number which represents the triple number of the data set. The dashed ovals in Figure 9 indicate the experimental input data sets. The total triple number of the experimental input is more than 81 million triples, occupying 15 gigabytes. This is a so massive knowledge graph that all popular reasoners cannot process efficiently. At the same time, existing distributed reasoners such as WebPIE are also not able to fulfil the reasoning task over the big biological network, because they only can calculate the closure of large-scale triples based on fixed RDFS (resource description framework schema) or OWL rules [28].

We implemented the reasoning prototype system based on the Hadoop framework, which is an open-source Java implementation of MapReduce [29]. The experiment was conducted in a single node and several Hadoop clusters with the scale of 1 node (pseudodistributed model), 2 nodes, 3 nodes, 4 nodes, 5 nodes, and 6 nodes. One node in cluster acts as master (controlling node) and the left ones act as slaves (real computing nodes). The Hadoop version is 1.1.3. Each node has the same configuration, including Linux OS, 8 G RAM, 500 G disk capacity, and 8-core Intel(R) Xeon(R) CPU E5620 with 2.4 GHz. The nodes are connected by the network with the bandwidth of 1000 M/s. In experiment, as Reducer is responsible for the major computation, Reducer is dynamically set by the length of PCS in a MapReduce job. Each test is executed 5 times and the average computing time is recorded.

4.2. Evaluation Parameters. Because our ultimate goal is to develop an efficient reasoner to systemically explore the implicit relationships among biological entities from the big biological network for further analysis, the accuracy (high precision), efficiency (less processing time), scalability (larger input data), and effectivity (high practicality) will be critical. So we evaluate our reasoning system from the above several aspects. Accuracy evaluation is based on random sampling inspection. We selected a number of herb-gene pairs and herb-ingredient pairs from the results. Then three annotators with graduate degrees in biomedical and TCM

Table 2: Accuracy evaluation for selected genes.

Gene symbol	Sample size	TP	Precision	Total mappings
TNF	30	28	93.3%	34
PEP4	30	22	73.3%	101
HK1	30	24	80%	100
IL6	30	26	86.7%	178
NQO1	30	26	86.7%	77
Sum up	150	126	84%	490

Table 3: Accuracy evaluation for selected ingredients.

WM ingredient	Sample size	TP	Precision	Total mappings
Dasatinib	30	23	76.7%	57
Fluoxymesterone	30	26	86.7%	47
Paclitaxel	30	22	73.3%	114
Pindolol	30	24	80%	51
Trastuzumab	30	25	83.3%	78
Sum up	150	120	80%	347

Table 4: Accuracy evaluation for selected herbs.

Herb	Sample size	TP	Precision	Total mappings
Ganoderma lucidum	30	24	80%	310
Hypericum	30	23	76.7%	542
Salvia miltiorrhiza	30	26	86.7%	523
Artemisinin	30	23	76.7%	575
Ginkgo biloba	30	25	83.3%	788
Sum up	150	121	80.7%	2688

Table 5: Scalability over number of nodes.

Number of nodes	Time (minutes)	Speedup
1 node	Out of memory	
2 nodes	8.45	1
3 nodes	4.96	1.7
4 nodes	3.07	2.76
5 nodes	2.48	3.41
6 nodes	2.16	3.91

Table 6: Scalability over input data.

Input data (size)	Time (minutes)	Sizeup
1 time (15 G)	3.07	1
2 times (30 G)	5.75	1.87
3 times (45 G)	8.03	2.62
4 times (60 G)	11.26	3.67
6 times (90 G)	18.04	5.88
8 times (120 G)	24.09	7.85

domains independently examined whether each pair was correctly extracted by our system. Only the pairs agreed upon by all three curators were counted as true positives (**TP**). **Precision** is defined according to formula (1), where **TP** and **FP** are the numbers of true positives and false positives, respectively. Efficiency evaluation is conducted by comparing the running time of single node and the distributed reasoning system. According to formulas (2) and (3), **Speedup** and **Sizeup** are calculated for scalability evaluation. Effectivity evaluation is constructed by analysing the potential value of this association information. Consider

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \quad (1)$$

$$\textbf{Speedup} = \frac{\text{computing time on 1 computer}}{\text{computing time on cluster}} \quad (2)$$

$$\textbf{Sizeup} = \frac{\text{computing time for processing } m \times \text{data}}{\text{computing time for processing data}}. \quad (3)$$

4.3. Evaluation and Discussion

4.3.1. Accuracy. Our reasoning system derives 40,178 herb-gene pairs and 5,183 herb-ingredient pairs. As many mappings between herbs and western medical entities are still unproven by professional biochemistry experiments, there is no gold standard for determining the correct mapping space between herbs and western medical entities. If we choose those less studied genes, herbs, and ingredients, the calculated precision is underestimated significantly because we may mistake many **TP** for **FP** in the manual evaluation. For this reason, with the advice of related experts, we focused on some major reported genes, ingredients, and herbs in recent years. Then we randomly selected 30 pairs of associations (samples) for every selected entity from the reasoning results and used precision measurement to evaluate the performances. The accuracy evaluations of the association information are shown in Tables 2, 3, and 4. The results show that our system achieves high accuracy. The high accuracy provides strong evidence to support further results analysis for researchers. All the results, reasoners, and the unified ontologies are available online (https://github.com/hualichenxi/biological-knowledge-reasoner).

4.3.2. Efficiency. Table 5 shows that reasoning on a single node (not pseudodistributed model) leads to out-of-memory problem. When implemented in the distributed reasoning system, we are able to complete reasoning for 15 G data within several minutes. Especially when the scale of Hadoop cluster becomes bigger, the performance is improved significantly. Meanwhile, the multichains reasoning algorithm guarantees that reasoner can perform multiple reasoning tasks defined by users themselves in a MapReduce job. The high efficiency and flexibility make our reasoning system become an excellent reasoner for large-scale biological data.

4.3.3. Scalability. Table 5 shows how our approach scales with an increasing number of computing nodes, using the data from Figure 9 as a fixed input. We use the running time on the 2-node configuration as baseline because a single node cannot process all the data due to being out of memory. Table 6 shows how our approach scales with increasing

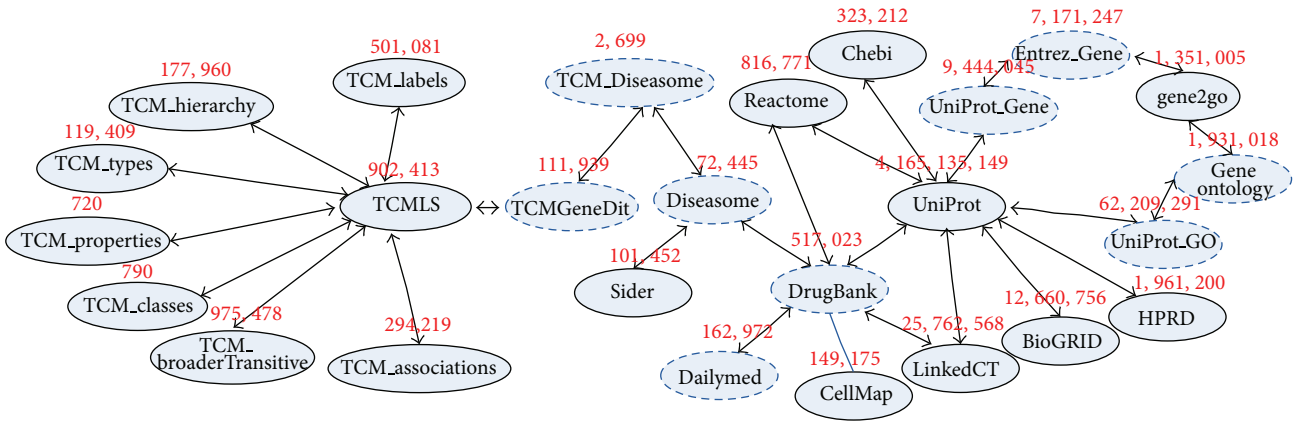


Figure 9: The big linked biological knowledge network.

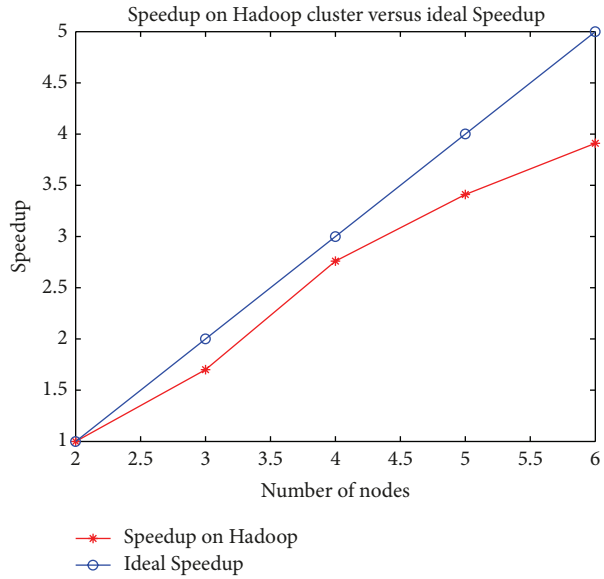


Figure 10: Speedup on Hadoop cluster.

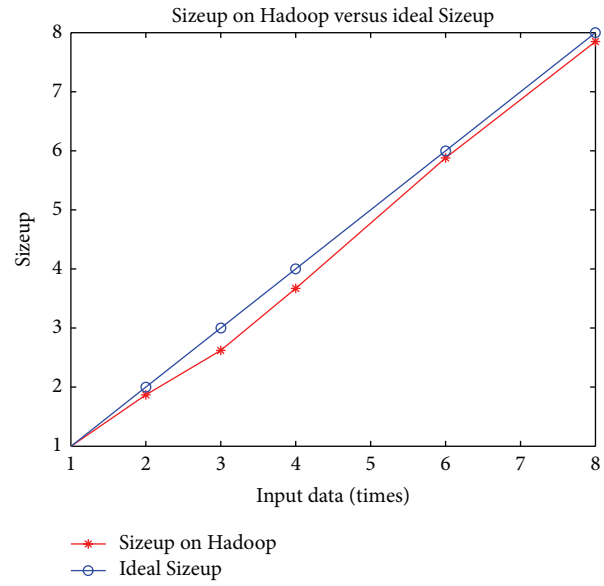


Figure 11: Sizeup on Hadoop cluster.

input size by doubling the original data (reasoning rules not changed), using a fixed configuration of 4 nodes. Speedup and Sizeup are shown in Figures 10 and 11, respectively.

From Figure 10 we can see that the Speedup increases strongly with increasing computing nodes, which means processing time is significantly reduced by adding more computing nodes. In theory, the processing time is supposed to grow linearly as the input data increases. That is to say, the processing time will increase **m** times when the input data increases **m** times. Figure 11 shows that the red line denoting the Sizeup on Hadoop is below the blue line representing the theoretical Sizeup, which shows that Sizeup of **m** times input is less than or equal to **m**. This means that execution time increases more slowly than input data size and our system works better in processing larger input set.

To sum up, considering the effects of the platform overhead, we conclude that the results show good scalability regarding the size of the input and number of nodes. Our reasoning system achieves excellent scalability. This advantage ensures that the reasoning system can be easily applied to the analysis of larger scale biological knowledge network.

4.3.4. Effectivity. The extracted association information consists of two parts: herb-gene pairs and herb-ingredient pairs. These associations are of great value to TCM and WM biologists in TCM modernization, new drug development, and so on.

Analysis for Herb-Gene Pairs. The derived herb-gene pairs could be used to provide some scientific evidences for TCM modernization from the perspective of modern biology by explaining the potential therapeutic mechanisms of herbs via gene regulations. Take gene *tumour necrosis factor* (TNF) for example. TNF as an important proinflammatory cytokine

plays a role in the regulation of cell differentiation, proliferation, and death which is closely correlated with tumour disease (http://www.ncbi.nlm.nih.gov/pubmed/21790707). Our experimental results reveal that TNF gene is associated with 34 herbs including *Ganoderma lucidum*, *Salvia miltiorrhiza*, and *Hypericum perforatum*. On the other hand, just as predicated by the results, according to chemical component analysis, most of these herbs (94%) contain anticancer compounds. The compounds can cause cancer cells to round up and die, inhibit tumor-induced blood supply development, and prevent tumor growth [30–32]. These derived associations suggest the possible therapeutic mechanisms involved by herbs, genes, and herb components. Besides, these herbs containing the anticancer components can inspire researchers into the development of new cancer drugs. The associations can also help biologists have a more comprehensive understanding of the functional mechanisms of the complex biological system as a whole from the two perspectives of TCM and modern biology.

Analysis for Herb-Ingredient Pairs. An increasing number of researchers are focusing their attention on developing drugs from traditional Chinese medicinal herbs and identifying the active ingredients of these herbs and their pharmacological mechanism of actions [33, 34]. The most successful herb example for TCM is the antimalarial drug *artemisinin*. Other famous TCM herbs (e.g., *Ginkgo biloba*, *Salvia miltiorrhiza*, *Hypericum perforatum*, and so on) are also widely used in WM for treating some complex diseases such as *Alzheimer* and *Asthma*. However, the active ingredients of many existing herbs have still remained unknown or uncertain for biologists. So besides regular chemical experiments, the extracted herb-ingredient pairs can also assist researchers to discover more information about some certain herb for revealing the mystery of herbs. Moreover, this TCM-inspired ingredient information can be further used to develop novel drugs. Take artemisinin for example; if biologists want to develop novel drugs for malaria, they can get some inspirations from these ingredients related to herb artemisinin. Our reasoning results show that the herb artemisinin is associated with 33 ingredients including adalimumab, docetaxel, and adenosine. Many of the ingredients have been proved to be effective for treating malaria [35–37]. So the mechanisms of action and chemical components of these ingredients can facilitate the development of new drug for malaria.

5. Related Work

5.1. Reasoning over Biological Data. Based on biological formal ontologies, we are able to make use of reasoning method from description logic to implement many biological applications, such as the discovery of new relationships, consistency checking, classification, and practical querying. Here are some examples which use OWL reasoning over biological data.

Holford et al. [38] used semantic web rules to reason with an ontology of pseudogenes to discover information about human pseudogene evolution. Volker et al. made use of existing reasoners Racer [11] to support reasoning with the foundational model of anatomy in OWL DL (description logic) [39]. Blondé et al. [40] applied relational closure rules to reason with bioontologies to enable practical querying.

So far, however, most of these applications only apply to relatively small data. When it comes to the data analysis of big integrated biological knowledge network, OWL reasoning faces the problems of low efficiency and out of memory [41].

5.2. Massive Biological Data Integration and Search Platforms. In recent years, several data integration and search platforms for the biological domain were presented, such as linked life data (LLD) (http://linkedlifedata.com), Bioportal [42], NCBI (http://www.ncbi.nlm.nih.gov/), and Bio2RDF [43]. LLD was a semantic data integration framework that enables access to multiple public biological databases. BioPortal was an open repository of biological ontologies that provided access via web services and web browsers to ontologies developed in OWL, RDF, OBO format, and Protege tool. The NCBI was a system of interlinked biological databases created by the US National Library of Medicine, which provided a series of search services for biological data. Bio2RDF was a mashup system to help the process of bioinformatics knowledge integration. But these systems lack a comprehensive ontology to model the entire biological network, including TCM and WM, making it hard to discover more implicit knowledge behind the big and complex biological network.

5.3. Large-Scale Semantic Data Reasoning Systems. Presently, some work of applying cloud computing to semantic data reasoning had been done to solve the problem of scalability. Urbani et al. [44] developed the MapReduce algorithms for materializing RDFS inference results. Liu et al. [45] extended the method to handle fuzzy pD reasoning. Oren et al. [46] presented a parallel and distributed platform for processing large amounts of RDF data on a network of loosely coupled peers. Heino and Pan [47] implemented RDFS reasoning on massively parallel hardware. The above systems mainly focus on computing closure for every domain based on RDFS or OWL rules by different cloud computing methods. None of them is dedicated to derive some implicit associations across multiple domains. However, in the data analysis of large-scale biological knowledge network, there are many problems across multiple biological domains. At this time, digging out meaningful knowledge from the big biological data network cannot be easily achieved using the above methods.

6. Conclusion

Confronted with the massive, disparate, and interlinked biological network, this paper presents a general biological data reasoning framework to model, integrate, and analyze the big biological network. We firstly summarize the basic requirements for a feasible framework. Then we give the overall OWL reasoning framework over big biological network and related modules. We construct a unified biological ontology to capture and model the complex biological network including modern biology and TCM. Based on the conceptual model,

a big biological linked knowledge network is formed to integrate and unify the heterogeneous data sources. Then for the data analysis of the big biological linked knowledge network, we propose three different kinds of reasoning algorithms and implement corresponding reasoning prototype systems which make full use of the advantages of MapReduce parallel programming model and OWL property chain reasoning method. Finally, we evaluate the reasoning prototype system on the big biological linked knowledge network, with its focus on discovering the implicit associations between TCM and WM. The results demonstrate that our prototype system achieves high efficiency, accuracy, scalability, and effectivity.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Thanks are due to Tong Yu, Bo Gao, Ling Zhu, and their colleagues from the China Academy of Chinese Medical Sciences for their help in constructing biological ontology and result evaluation. This work is funded by LY13F020005 of NSF of Zhejiang, NSFC61070156, YB2013120143, and Fundamental Research Funds for the Central Universities.

References

[1] R. Apweiler, A. Bairoch, C. H. Wu et al., "UniProt: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 32, supplement 1, pp. D115–D119, 2004.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.

[3] D. L. McGuinness and F. van Harmelen, "OWL web ontology language overview," W3C Recommendation, 2004.

[4] K.-H. Cheung, E. Prud'hommeaux, Y. Wang, and S. Stephens, "Semantic web for health care and life sciences: a review of the state of the art," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 111–113, 2009.

[5] R. Shearer, "OBO and OWL: leveraging semantic web technologies for the life sciences," in *Proceedings of the 6th International on the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference (ISWC '07/ASWC '07)*, vol. 4825, pp. 169–182, Springer, Busan, Republic of Korea, November.

[6] A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall, "Life sciences on the semantic web: the neurocommons and beyond," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 193–204, 2009.

[7] D. Brickley and R. V. Guha, fRDF Vocabulary Description Language 1.0: RDF Schema, 2004.

[8] S. Bechhofer, F. van Harmelen, J. Hendler et al., OWL web ontology language reference, 2004.

[9] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: a practical OWL-DL reasoner," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 2, pp. 51–53, 2007.

[10] D. Tsarkov and I. Horrocks, "FaCT++ description logic reasoner: system description," in *Automated Reasoning*, vol. 4130 of *Lecture Notes in Computer Science*, pp. 292–297, Springer, New York, NY, USA, 2006.

[11] V. Haarslev and R. Möller, "Racer: a core inference engine for the semantic web," in *Proceedings of the 2nd International Workshop on Evaluation of Ontology-Based Tools*, vol. 87, 2003.

[12] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[13] M. Mogi, M. Harada, H. Narabayashi, H. Inagaki, M. Minami, and T. Nagatsu, "Interleukin (IL)-1β, IL-2, IL-4, IL-6 and transforming growth factor-α levels are elevated in ventricular cerebrospinal fluid in juvenile parkinsonism and Parkinson's disease," *Neuroscience Letters*, vol. 211, no. 1, pp. 13–16, 1996.

[14] E. Levy, M. D. Carman, I. J. Fernandez-Madrid et al., "Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrhage, Dutch type," *Science*, vol. 248, no. 4959, pp. 1124–1126, 1990.

[15] F. E. Chen, C. Ooi, S. Y. Ha et al., "Genetic and clinical features of hemoglobin H disease in Chinese patients," *The New England Journal of Medicine*, vol. 343, no. 8, pp. 544–550, 2000.

[16] S. J. Leuenroth, D. Okuhara, J. D. Shotwell et al., "Triptolide is a traditional Chinese medicine-derived inhibitor of polycystic kidney disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4389–4394, 2007.

[17] K.-C. Cheng, H.-C. Huang, J.-H. Chen et al., "Ganoderma lucidum polysaccharides in human monocytic leukemia cells: from gene expression to network construction," *BMC Genomics*, vol. 8, no. 1, article 411, 2007.

[18] Y.-C. Hseu, F.-Y. Wu, J.-J. Wu et al., "Anti-inflammatory potential of *Antrodia camphorata* through inhibition of iNOS, COX-2 and cytokines via the NF-κB pathway," *International Immunopharmacology*, vol. 5, no. 13-14, pp. 1914–1925, 2005.

[19] C. Bizer and R. Cyganiak, "D2R server-publishing relational databases on the semantic web," in *Proceedings of the 5th International Semantic Web Conference*, p. 26, 2006.

[20] P. M. Barnes, E. Powell-Griner, K. McFann, and R. L. Nahin, "Complementary and alternative medicine use among adults: United States, 2002," in *Seminars in Integrative Medicine*, vol. 2, pp. 54–71, Elsevier, New York, NY, USA, 2004.

[21] M. A. Harris, J. Clark, A. Ireland et al., "The gene oncology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, pp. D258–D261, 2004.

[22] L. M. Schriml, C. Arze, S. Nadendla et al., "Disease ontology: a backbone for disease semantic integration," *Nucleic Acids Research*, vol. 40, pp. D940–D946, 2012.

[23] K.-I. Goh and I.-G. Choi, "Exploring the human diseasome: the human disease network," *Briefings in Functional Genomics*, vol. 11, no. 6, pp. 533–542, 2012.

[24] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, supplement 1, pp. D901–D906, 2008.

[25] Y.-C. Fang, H.-C. Huang, H.-H. Chen, and H.-F. Juan, "TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining," *BMC Complementary and Alternative Medicine*, vol. 8, no. 1, article 58, 2008.

[26] X. Zhou, Z. Wu, A. Yin, L. Wu, W. Fan, and R. Zhang, "Ontology development for unified traditional Chinese medical language system," *Artificial Intelligence in Medicine*, vol. 32, no. 1, pp. 15–27, 2004.

[27] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D54–D58, 2005.

[28] J. Urbani, S. Kotoulas, J. Maassen, F. van Harmelen, and H. Bal, "OWL reasoning with WebPIE: calculating the closure of 100 billion triples," *The Semantic Web: Research and Applications*, Springer, New York, NY, USA, vol. 6088, pp. 213–227, 2010.

[29] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST '10)*, pp. 1–10, Incline Village, Nev, USA, May 2010.

[30] H.-S. Zhang and S.-Q. Wang, "Salvianolic acid B from *Salvia miltiorrhiza* inhibits tumor necrosis factor-α (TNF-α)-induced MMP-2 upregulation in human aortic smooth muscle cells via suppression of NAD(P)H oxidase-derived reactive oxygen species," *Journal of Molecular and Cellular Cardiology*, vol. 41, no. 1, pp. 138–148, 2006.

[31] S. Y. Wang, M. L. Hsu, H. C. Hsu et al., "The antitumor effect of Ganoderma Lucidum is mediated by cytokines released from activated macrophages and T lymphocytes," *International Journal of Cancer*, vol. 70, no. 6, pp. 699–705, 1997.

[32] J. Barnes, L. A. Anderson, and J. D. Phillipson, "St John's wort (*Hypericum perforatum* L.): a review of its chemistry, pharmacology and clinical properties," *Journal of Pharmacy and Pharmacology*, vol. 53, no. 5, pp. 583–600, 2001.

[33] R. Graziose, M. A. Lila, and I. Raskin, "Merging traditional chinese medicine with modern drug discovery technologies to find novel drugs and functional foods," *Current Drug Discovery Technologies*, vol. 7, no. 1, pp. 2–12, 2010.

[34] W. L. W. Hsiao and L. Liu, "The role of traditional Chinese herbal medicines in cancer therapy from TCM theory to mechanistic insights," *Planta Medica*, vol. 76, no. 11, pp. 1118–1131, 2010.

[35] A. Golisade, J. Wiesner, C. Herforth, H. Jomaa, and A. Link, "Anti-malarial activity of N6-substituted adenosine derivatives. Part I," *Bioorganic and Medicinal Chemistry*, vol. 10, no. 3, pp. 769–777, 2002.

[36] V. Sinou, Y. Boulard, P. Grellier, and J. Schrevel, "Host cell and malarial targets for docetaxel (Taxotere) during the erythrocytic development of *Plasmodium falciparum*," *Journal of Eukaryotic Microbiology*, vol. 45, no. 2, pp. 171–183, 1998.

[37] S. Y. Shin, G. M. Seong, Y. R. Kim, J. W. Kang, and J. Kim, "An atypical Case of *Plasmodium vivax* malaria after initiating adalimumab therapy," *Journal of Rheumatic Diseases*, vol. 19, no. 3, pp. 160–162, 2012.

[38] M. E. Holford, E. Khurana, K. Cheung, and M. Gerstein, "Using semantic web rules to reason on an ontology of pseudogenes," *Bioinformatics*, vol. 26, no. 12, Article ID btq173, pp. i71–i78, 2010.

[39] C. Golbreich, S. Zhang, and O. Bodenreider, "The foundational model of anatomy in OWL: experience and perspectives," *Web Semantics*, vol. 4, no. 3, pp. 181–195, 2006.

[40] W. Blondé, V. Mironov, A. Venkatesan, E. Antezana, B. de Baets, and M. Kuiper, "Reasoning with bio-ontologies: using relational closure rules to enable practical querying," *Bioinformatics*, vol. 27, no. 11, pp. 1562–1568, 2011.

[41] J. Zhou, L. Ma, Q. Liu, L. Zhang, Y. Yu, and Y. Pan, "Minerva: a scalable OWL ontology storage and inference system," *The Semantic Web—ASWC 2006*, Springer, New York, NY, USA, vol. 4185, pp. 429–443, 2006.

[42] N. F. Noy, N. H. Shah, P. L. Whetzel et al., "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, vol. 37, supplement 2, pp. W170–W173, 2009.

[43] F. Belleau, M. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems," *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[44] J. Urbani, S. Kotoulas, E. Oren, and F. Van Harmelen, "Scalable distributed reasoning using MapReduce," *The Semantic Web—ISWC 2009*, Springer, New York, NY, USA, vol. 5823, pp. 634–649, 2009.

[45] C. Liu, G. Qi, H. Wang, and Y. Yu, "Large scale fuzzy pD* reasoning using MapReduce," in *The Semantic Web—ISWC 2011*, vol. 7031 of *Lecture Notes in Computer Science*, pp. 405–420, Springer, New York, NY, USA, 2011.

[46] E. Oren, S. Kotoulas, G. Anadiotis, R. Siebes, A. ten Teije, and F. van Harmelen, "Marvin: distributed reasoning over large-scale semantic web data," *Journal of Web Semantics*, vol. 7, no. 4, pp. 305–316, 2009.

[47] N. Heino and J. Z. Pan, "RDFS reasoning on massively parallel hardware," in *The Semantic Web—ISWC 2012*, vol. 7649 of *Lecture Notes in Computer Science*, pp. 133–148, Springer, New York, NY, USA, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 253128, 11 pages
http://dx.doi.org/10.1155/2014/253128

Research Article

A Knowledge-Driven Approach to Extract Disease-Related Biomarkers from the Literature

À. Bravo, M. Cases, N. Queralt-Rosinach, F. Sanz, and L. I. Furlong

Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Dr Aiguader 88, E-08003 Barcelona, Spain

Correspondence should be addressed to L. I. Furlong; lfurlong@imim.es

Received 3 November 2013; Revised 17 February 2014; Accepted 20 February 2014; Published 16 April 2014

Academic Editor: Farit Mochamad Afendi

The biomedical literature represents a rich source of biomarker information. However, both the size of literature databases and their lack of standardization hamper the automatic exploitation of the information contained in these resources. Text mining approaches have proven to be useful for the exploitation of information contained in the scientific publications. Here, we show that a knowledge-driven text mining approach can exploit a large literature database to extract a dataset of biomarkers related to diseases covering all therapeutic areas. Our methodology takes advantage of the annotation of MEDLINE publications pertaining to biomarkers with MeSH terms, narrowing the search to specific publications and, therefore, minimizing the false positive ratio. It is based on a dictionary-based named entity recognition system and a relation extraction module. The application of this methodology resulted in the identification of 131,012 disease-biomarker associations between 2,803 genes and 2,751 diseases, and represents a valuable knowledge base for those interested in disease-related biomarkers. Additionally, we present a bibliometric analysis of the journals reporting biomarker related information during the last 40 years.

1. Introduction

The Biomarkers Definition Working Group (formed by the US National Institutes of Health (NIH) and the US Food and Drug Administration (FDA), academia, and industry) defined biomarker as "*a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention*" [1]. With the advent of the genomics era, in April 2008, the FDA published in one of its "Guidance for Industry" documentations the specific definition of a genomic biomarker as "*a measurable DNA and/or RNA characteristic that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions*" [2]. More recently, Anderson and Kodukula [3] provided some definitions of different types of biomarkers (e.g., surrogate, clinical endpoint, diagnostic, prognostic, predictive, pharmacodynamic, efficacy, and toxicity/safety [4–6]) within their review of the role of biomarkers in pharmacology and drug discovery. All these definitions specify the requirements to be held by a biomarker, the different types that exist, their potential role in disease diagnosis and progression or in the therapeutic response control, and their utility for the assessment of new chemical entities as potential lead therapeutics [3].

Thousands of biomolecules are being investigated as potential biomarkers, but most of them do not advance effectively for diagnostic, prognostic, or therapeutic goals for different reasons (for a detailed discussion on this topic see [7–9]). The results of the research on potential biomarkers are widely reported on the biomedical literature. The MEDLINE database [10] has currently indexed more than 23 M articles, and since 1989 the MeSH term "Biological Markers" is applied to annotate those articles that provide data on "*measurable and quantifiable biological parameters (e.g., specific enzyme concentration, specific hormone concentration, specific gene phenotype distribution in a population, presence of biological substances) which serve as indices for health- and physiology-related assessments, such as disease risk, psychiatric disorders, environmental exposure and its*

effects, disease diagnosis, metabolic processes, substance abuse, pregnancy, cell line development, epidemiologic studies, etc." [11], and later on, in 2008, the MeSH term "Biomarkers, Pharmacological" was introduced to specifically annotate the "*measurable biological parameters that serve for drug development, safety and dosing (DRUG MONITORING)*" [12]. In particular, genomic biomarkers are frequently reported in the literature together with disease-related information. Thus, the MEDLINE database contains valuable knowledge for those interested in gathering information on biomarkers. In order to identify, extract, and analyse this information from the literature, automatic processing of the texts is required [13]. There are only few reports on text mining approaches in the biomarkers field [14–16]. Here, we present a knowledge-driven text mining approach for the extraction of disease-related biomarker information from the literature. Our approach, firstly, takes advantage of biomarker-specific MeSH terms annotations to retrieve a specific and comprehensive pool of publications from MEDLINE, secondly, applies our named entity recognition method (BioNER) to (1) identify genes and diseases as entities of interest, (2) filter ambiguous entities, (3) cluster equivalent terms to a certain concept, (4) characterize those genes as potential biomarkers based on terminology used, and, finally, (5) find associations between genes and diseases in single sentences, and ranks the associations based on their frequency in the literature. This approach, that allows the unique identification of genomic biomarkers and their associated diseases, was applied to the MEDLINE database resulting in a comprehensive knowledge base on disease-related biomarkers, which is publicly available at http://ibi.imim.es/biomarkers/. In addition, we provide an analysis of the results obtained and present an evaluation of the trend of biomarker research reporting as a topic in the scientific literature.

2. Material and Methods

We developed a text mining workflow aimed at extracting information on disease-related biomarkers from scientific publications. Briefly, after document selection, the text mining approach comprises as a first step the recognition and normalization of the *disease* and *biomarker* entities in biomedical publications by means of the biomedical named entity recognition (BioNER) system and, secondly, the identification of relationships between the aforementioned entities by their cooccurrence in sentences. For example, the following sentence (taken from PMID: 17397492), "**CK20** *is an important biomarker that can be used to identify* **TCC** *in urine cytology smears,*" contains the cooccurrence of the entities **CK20** (gene) and **TCC** (disease).

The different steps addressed in the text mining workflow are illustrated in Figure 1 and detailed below.

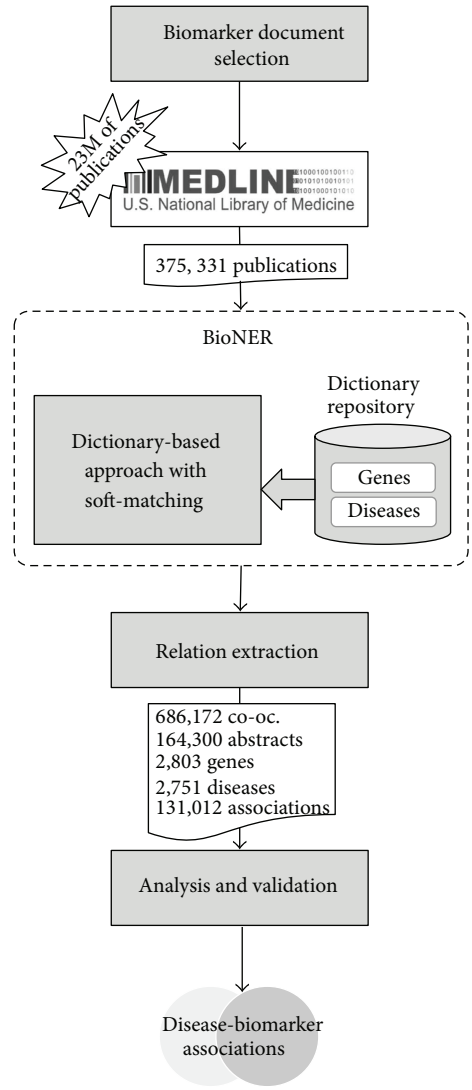


Figure 1: Text mining workflow.

2.1. Document Selection. To obtain a set of publications focused on biomarkers, we formulated the following PubMed query: ("Biological Markers" [MeSH Terms]) AND (has abstract [text]) AND (English [lang]) AND ("0001/01/01" [PDAT]: "2013/06/30" [PDAT]) AND "humans" [MeSH Terms], that resulted in 375,331 publications (September 30, 2013).

2.2. Development of Gene and Disease Dictionaries for Biomarker-Specific Information

Gene Dictionary. In order to collect the terms referring to human genes and proteins, we have integrated data from three biological databases: NCBI-Gene [17], HGNC [18], and UniProt [19, 20], followed by a semiautomatic curation process. These databases are cross-referenced between each other, providing a way to collect and integrate the terminology for a specific gene/protein entity from the different sources in a single entity. Figure 2 shows an example of terminology integration for the Lipocalin-2 gene. Note that we do not make a distinction between gene and protein mentions in the text, because in general both types of entities share their terminology. Thus, for the sake of simplicity, we refer to genes and proteins as genes.

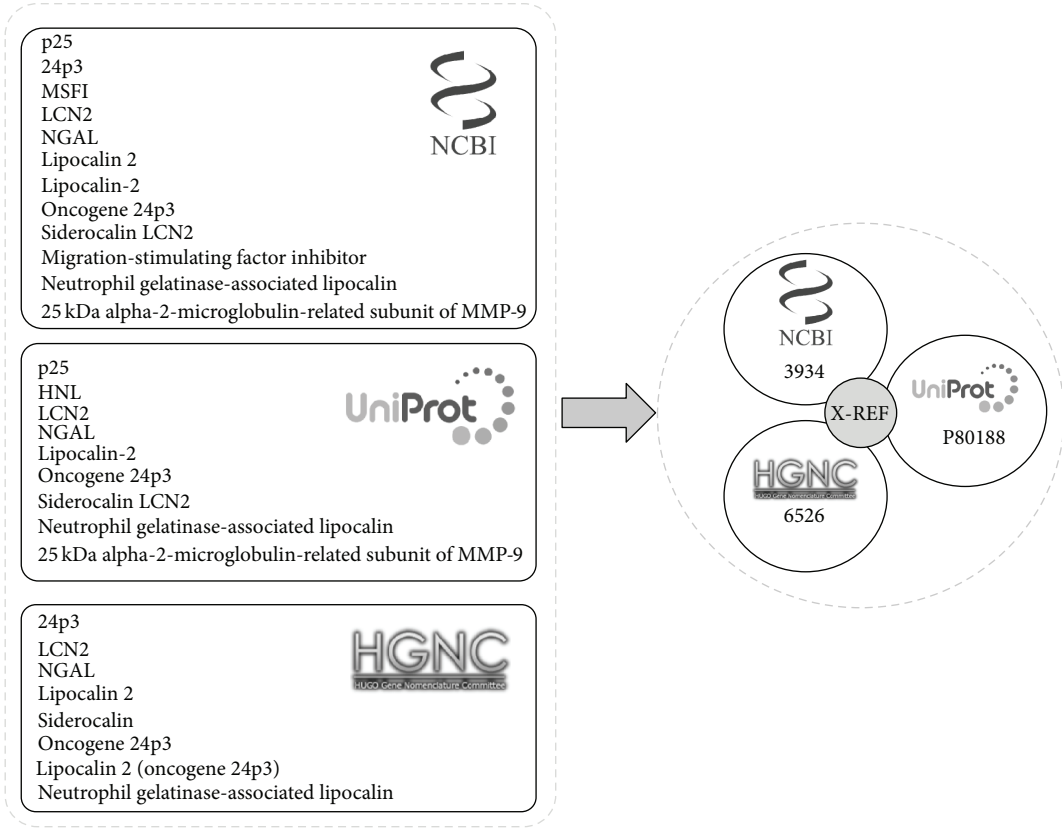


Figure 2: An example of the variability in terminology for genes depending on the primary sources.

Disease Dictionary. The Unified Medical Language System (UMLS) [21] database was used to create the disease dictionary. The UMLS Metathesaurus is a large, multipurpose, and multilingual thesaurus that contains millions of biomedical and health-related concepts, their synonymous names, and their known relationships. We selected all the concepts in English from the freely distributed vocabularies corresponding to the following semantic types: Congenital Abnormality (T019), Acquired Abnormality (T020), Disease or Syndrome (T047), Mental or Behavioral Dysfunction (T048), Experimental Model of Disease (T050), Sign or Symptom (T184), Anatomical Abnormality (T190), and Neoplastic Process (T191).

Both dictionaries were curated and extended semiautomatically using different rules to facilitate the matching task. Each dictionary has its own distinctive features; for example, the gene dictionary has a high prevalence of acronyms referring to genes (i.e., A2MP1, NOTCH1, and SF3B1), whereas long terms are prevalent in the disease dictionary (i.e., Alzheimer's disease, acute lymphoblastic leukemia, primary eosinophilic endomyocardial restrictive cardiomyopathy, and rheumatic tricuspid stenosis and insufficiency). In our curation process we defined the following rules with specific adjustments depending on the dictionary.

(1) To reduce ambiguity in the dictionary, the terms with a length smaller than three characters are removed.

(2) A specific number of characters are replaced by their general form; that is, the characters "à, ö, ç, û" are replaced by "a, o, c, u" (i.e., *Sjögren-Larsson syndrome* by *Sjogren-Larsson syndrome*).

(3) New variants are generated for gene symbols (i.e., *IL2*, *IL 2*, *IL (2)*, or *IL-2* is the same acronym referring to *interleukin 2*).

(4) Terms containing digits (Arabic numbers) can be written with roman numbers. New terms are generated by replacing Arabic with Roman numbers (*Adenylosuccinate lyase deficiency type 4* by *Adenylosuccinate lyase deficiency type IV*).

(5) Terms can contain Greek letters (such as HP1-alpha, HP1-beta, and HP1-gamma) or symbols (as HP1-α, HP1-β, and HP1-γ); both cases are considered.

(6) Prefix and suffix labels not used in natural language are removed from the terms (i.e., *[X]Gastric neurosis* or *Leber Aongenital Amaurosis [Disease/Finding]* by *Gastric neurosis* and *Leber Congenital Amaurosis*).

(7) All terms are transformed into lowercase characters (i.e., *FALDH deficiency* by *faldh deficiency*).

(8) All punctuation marks are removed to improve the fuzzy matching (i.e., *hnf-3-gamma* by *hnf 3 gamma*).

Particularly, the disease dictionary was also processed with Casper [22], a rule-based software that suppresses undesired terms from the UMLS Metathesaurus and generates additional synonyms and spelling variations, and, afterwards, manually curated in order to remove very general terms such DISEASE or SYNDROME.

As a final step, to select a set of putative biomarker genes from our *gene dictionary*, we conducted a text mining search for the genes that are mentioned together with biomarker terms in the same sentence (biomarker rule filtering) and then all entries of this set of genes were extracted from the dictionary as putative biomarkers-related terms. The rationale of this approach is that genes mentioned together with terms such as "marker" are very likely biomarkers themselves. The biomarker terms were collected from the concept "biological markers" present in the MeSH terminology [11]. This step retrieved a total of 3,533 genes which were mentioned together with at least one biomarker term, and they were collected from our *gene dictionary* to create the *biomarker-specific gene dictionary*. A similar procedure was applied to obtain a subset of the *disease dictionary* relevant to the biomarkers topic, the *biomarker-specific disease dictionary*. This filter allowed the selection of 3,122 diseases cooccurring with biomarker terms.

Table 1 shows the number of concepts, the ambiguity, and the variability for all the dictionaries to illustrate the effect of the curation and rules applied. The ambiguity quantifies how a term can refer to different concepts, while the variability reflects the average number of unique terms for each concept. The best curation process is the one that improves the variability minimizing the ambiguity of the dictionaries. In the case of the *gene dictionary*, the number of terms between the raw and curated dictionaries increases in 19% with a slight effect in the ambiguity. In the case of the *disease dictionary*, there are no major changes in ambiguity and variability after dictionary curation. Overall, the curation process keeps the ambiguity and improves the variability.

2.3. BioNER. The BioNER system applies the biomarker-gene and disease-biomarker dictionaries using fuzzy- and pattern-matching methods to find and uniquely identify entity mentions in the literature [23–25]. Firstly, our BioNER receives the dictionary type to extract mentions and a list of document identifiers (obtained in the document selection step). Each publication is recovered from a document repository and the abstracts are split into sentences, and a set of patterns is created from the selected dictionary (*biomarker-specific gene or disease* dictionaries), after removing a list of stop words. For each sentence, the BioNER extracts the longest term from the patterns without overlap. Then, each mention is normalized to its unique identifier using the dictionary.

2.4. Relation Extraction. In this study, we applied a relation extraction (RE) method based on cooccurrence findings, which assumes that a biomarker and a disease are associated if they are mentioned together in the same sentence. From 164,300 abstracts, 686,172 cooccurrences were found between 2,803 biomarkers and 2,751 diseases, resulting in 131,012 disease-biomarker different associations. Certainly, the title and the body of the abstract show different writing styles, in terms of both syntax and semantics. Generally, the title or the last part of the abstract tends to express more concisely the final message of the publication, whereas the rest of the abstract contains background information and more hypothetical discourses as contextual information of the study. In order to account for these differences and make a distinction depending on where the cooccurrence is detected in the text, the system separates each abstract into 3 parts: title, abstract body, and conclusions. Then, the associations are scored based on the frequency of each association in the literature represented by a variant of the Inverse Document Frequency model [26] as follows:

$$\text{Score}_{\text{DB}} = idf\,(\text{DB}, A) \cdot \sum_{i=1}^{|A|} af\left(\text{DB}, A_i\right), \tag{1}$$

$$idf\,(\text{DB}, A) = \log_{10} \frac{|A|}{|\{a \in A : \text{DB} \in a\}|}, \tag{2}$$

$$af\left(\text{DB}, A_i\right) = \frac{f\left(\text{DB}, A_i\right)}{\max\left\{f\left(XY, A_i\right) : XY \in A_i\right\}}, \tag{3}$$

where the score for the association between disease D and biomarker B (Score_{DB}, (1)) is obtained as the product between the inverse document frequency of the association between D and B ($idf(\text{DB}, A)$, (2)) and the normalized frequency of the association between D and B overall the documents ($af(\text{DB}, A_i)$, (3)).

The idf provides an indication of the popularity of the association across the corpus of documents under study, and it is obtained by dividing the total number of abstracts ($|A|$) by the number of abstracts containing the association between D and B ($|\{a \in A : \text{DB} \in a\}|$) and taking the logarithm of that quotient (2). The function $af(\text{DB}, A_i)$ in (3) is the frequency of the association between B and D in the ith abstract (A_i) and it is defined with a quotient between $f(\text{DB}, A_i)$, which is the number of times that the association between B and D occurs in A_i (multiplied by 2 if DB occurs in title or conclusion of A_i, or 1 if DB occurs in the body), and the maximum frequency of any association in A_i ($\max\{f(XY, A_i) : XY \in A_i\}$).

2.5. Analysis and Validation. In order to validate the disease-biomarker associations identified by text mining, we compared them to the biomarker information contained in the DisGeNET database, release 2.0 (July, 2012). DisGeNET is a database that integrates knowledge on the genes associated with human diseases from various expert curated databases and the literature [27, 28]. For this study we used the set of associations labelled as "biomarker" according to the DisGeNET gene-disease association ontology [29, 30]. We collected a list of 12,887 genes associated with 6,135 different disease terms stored in DisGeNET.

3. Results and Discussion

In this paper, we present a new methodology to extract disease-biomarker associations from the literature. One of

Table 1: Contents and statistics of gene, disease, and biomarker-gene and disease-biomarker dictionaries.

Dictionary	Number of concepts	Number of terms	Ambiguity	Variability
Gene	50,090	545,519	1.51	5.98
Gene curated/extended	50,090	649,414	1.46	12.96
Disease	79,781	378,616	1.01	4.14
Disease curated/extended	74,073	294,371	1.02	3.97
Biomarker-specific gene	3,533	89,236	1.27	25.26
Biomarker-specific disease	3,122	35,686	1.05	11.43

the major challenges that any text mining application faces is the variability of terms referring to the same concept; and then, consequently, the identification of entities in a nonambiguous manner (i.e., gene, protein, and disease). In this respect, biomedical terms gathered in domain-specific lexicons such as dictionaries, ontologies, and terms classifications (i.e., MeSH disease tree [31]) serve to organize synonymous terms into a central concept, facilitating both entity recognition and the hierarchical exploration of the results [32]. Another challenge in biomedical text mining is the identification of relationships between two entities [13]. Thus, our methodology faces both challenges by (1) the identification of *biomarker* and *disease* entities by means of the BioNER system and (2) the extraction of relationships between these entities by cooccurrence in sentences. An analysis of the associations between disease and biomarker is presented according to their mention frequency in MEDLINE, and they are evaluated by manual inspection and by comparison with the biomarker information integrated in the DisGeNET database.

The application of our text mining approach on a set of 375,331 publications pertaining to biomarkers (see Section 2.1) resulted in 686,172 disease-gene cooccurrences found in 164,300 abstracts. These cooccurrences represented associations between 2,803 genes and 2,751 diseases, giving rise to 131,012 unique disease-gene associations, which should be considered as potential disease-biomarker associations due to both the *document selection strategy* and the *biomarker rule filtering* addressed (see Section 2 for details and find examples of sentences including disease and biomarker concepts in Table 2). It is important to remark that the biomarker and disease mentions found in the text are linked to their corresponding identifiers in standard vocabularies (NCBI Gene for biomarkers and UMLS for diseases). This normalization of the entities extracted from the publications enables the unique identification of these entities and opens the possibility of integration of the extracted information with data from other standardized resources.

3.1. Distribution of Biomarker Information in the Biomedical Literature. From the approximately 23 M publications contained in the MEDLINE database, 375,331 are related to biomarkers and therefore have been annotated with the MeSH terms "biological markers" and "Biomarkers, Pharmacological" by PubMed curators. From these publications, 164,300 contain information on genes and proteins as biomarkers of a given disease in the abstract. The distribution

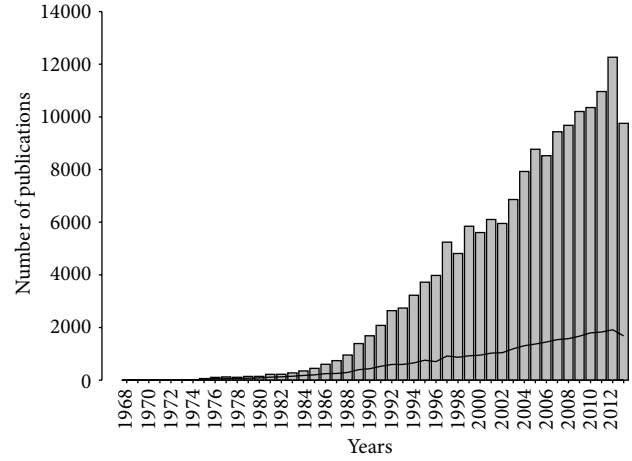


Figure 3: Number of publications (bars) and number of journals (line) by year.

of cooccurrences encountered in the title, the body of the abstract, and the conclusions section was 10, 85, and 5%, respectively. The evolution in reporting disease-biomarker related information throughout the years is presented in Figure 3. The document set under study represents publications in the field of biomarkers that contain information on genes and proteins from 3,983 different journals. Both the number of journals that publish disease-biomarkers-related data and the number of published articles show a progressive increase from the early 1980s. Only 5 of the journals include *marker* or *biomarker* in their journal name (*Int. J Biol. Markers* (336 abstracts), *Dis. Markers* (187), *Cancer Epidemiol. Biomarkers Prev.* (413), *Biomarkers* (11), and *Genet Test Mol. Biomarkers* (35)) and contribute to the disease-biomarker association list of this present study with a total of 2,253 disease-biomarker associations, which means only a 2% of the total list of associations identified in this present study.

A further analysis of the provenance of the cooccurrences, in terms of journals that report them, was carried out and results for the top 10 journals are represented in Figure 4. Concretely, these 10 journals report 13% (94,760) of the cooccurrences identified in the 12% (20,341) of the abstracts of the working set. Interestingly, the total number of cooccurrences is proportional to the number of disease-biomarker associations recorded from each of the top 10 journals. Note that the publication start year of each journal points out that not necessarily those journals reporting most

Table 2: Examples of sentences including disease-biomarker cooccurrences.

Disease (CUI)[a]	Biomarker (Gene ID)[b]	PMID (year)	Sentence
Hodgkin's lymphoma (C0019829)	Anti-Mullerian hormone (268)	17726078 (2007)	**Anti-Mullerian hormone** is a sensitive *serum marker* for gonadal function in women treated for **Hodgkin's lymphoma** during childhood. (TITLE)
TCC (C1861305)	CK20 (54474)	17397492 (2007)	**CK20** is an important *biomarker* that can be used to identify **TCC** in urine cytology smears. (CONCLUSIONS)
Autism (C0004352)	Brain-derived neurotrophic factor (BDNF) (627)	19119429 (2008)	To investigate levels of **brain-derived neurotrophic factor** (**BDNF**) in midpregnancy and neonatal blood specimens as early *biologic markers* for **autism**; we conducted a population-based case-control study nested within the cohort of infants born from July 2000 to September 2001 to women who participated in the prenatal screening program in Orange County, CA. (BODY)
Acute kidney injury (AKI) (C0022660)	Neutrophil gelatinase-associated lipocalin (NGAL) (3934), netrin-1 (9423)	21740336 (2011)	**Neutrophil gelatinase-associated lipocalin** (**NGAL**) and **netrin-1** have been proposed over the past years as emergent *biomarkers* for the early and accurate diagnosis and monitoring of **acute kidney injury** (**AKI**). (BODY)
Chronic heart failure (C0264716)	Cardiac troponin I (7137)	21751783 (2011)	Top-down quantitative proteomics identified phosphorylation of **cardiac troponin I** as a candidate *biomarker* for **chronic heart failure**. (TITLE)
Adenocarcinomas (C0001418)	MOC-31 (4072)	21732548 (2011)	**MOC-31** is an established *immunologic marker* to detect **adenocarcinomas**. (BODY)
Lung adenocarcinoma (C0152013)	ROM (6094)	21748260 (2012)	Hence, serum **ROM** level may be a useful *biomarker* for staging of **lung adenocarcinoma**. (BODY)

[a]Concept unique identifier at UMLS.
[b]NCBI gene identifier.

disease-biomarker associations in our working set started their publication earlier than others (i.e., Plos ONE). Most of the articles published in these top 10 journals describe basic laboratory, translational, and clinical investigations, and some of them have a special focus on specific therapeutic areas: hematology (*Blood*), immunology (*J Immunol.*), and oncology (*Cancer Res., Clin. Cancer Res., Cancer*, *Int. J Cancer*). In fact, over 300 journals of the list include the "clinical" word in their name, over 200 include the word "cancer" or the prefix "onco", and around 140 include the prefix "immun"; which are by far the main fields where biomarkers are being investigated.

Twenty-one percent of the disease-biomarker associations were identified in the top 10 journals (56% of diseases and 68% of biomarkers collected in this study, resp.; see Figure 4). Over 50% of the associations are retrieved from the first 100 journals (81% of diseases and 87% of biomarkers), and over 80% are from the first 500 journals (95% of diseases and 97% of biomarkers); and till we consider the first 1,000 journals we do not reach more than 90% of the total associations (98% of diseases and 99% of biomarkers).

This analysis shows that the number of journals and articles that report biomarkers information has increased during the last years, and this fact (i) expands the publication bias (few journals are specialized in biomarkers research and development, while most of journals include in their scope the biomarker topic or at least publish special issues devoted to biomarkers research), (ii) makes difficult the retrieval and exploitation of this information, and (iii) highlights the need of an improvement in the biomarker related data reporting [33] to ensure better quality of automatic extraction by means of mining techniques.

3.2. Analysis and Validation of the Disease-Biomarker Associations. The 131,012 disease-biomarker associations were scored based on their mention frequency in MEDLINE (see Table 3 for details of the associations distribution based on the score described in Section 2.4). The top 10 associations with higher score are shown in Table 3, where very well-studied disease biomarkers can be found (for instance, TP53 and ERBB2 for cancer and CD4 for immunodeficiencies).

Figure 5 shows the analysis of the associations based on the Score_{DB} (a) and the number of publications (b). The percentage of associations reported by different numbers of publications (from 1-2 publications to more than 2,000) in the corpus under study (131,012 associations, light grey bars) is represented. The caption shows the data for the associations reported by more than 100 publications, which represent a small percentage of all the associations. Note that most of

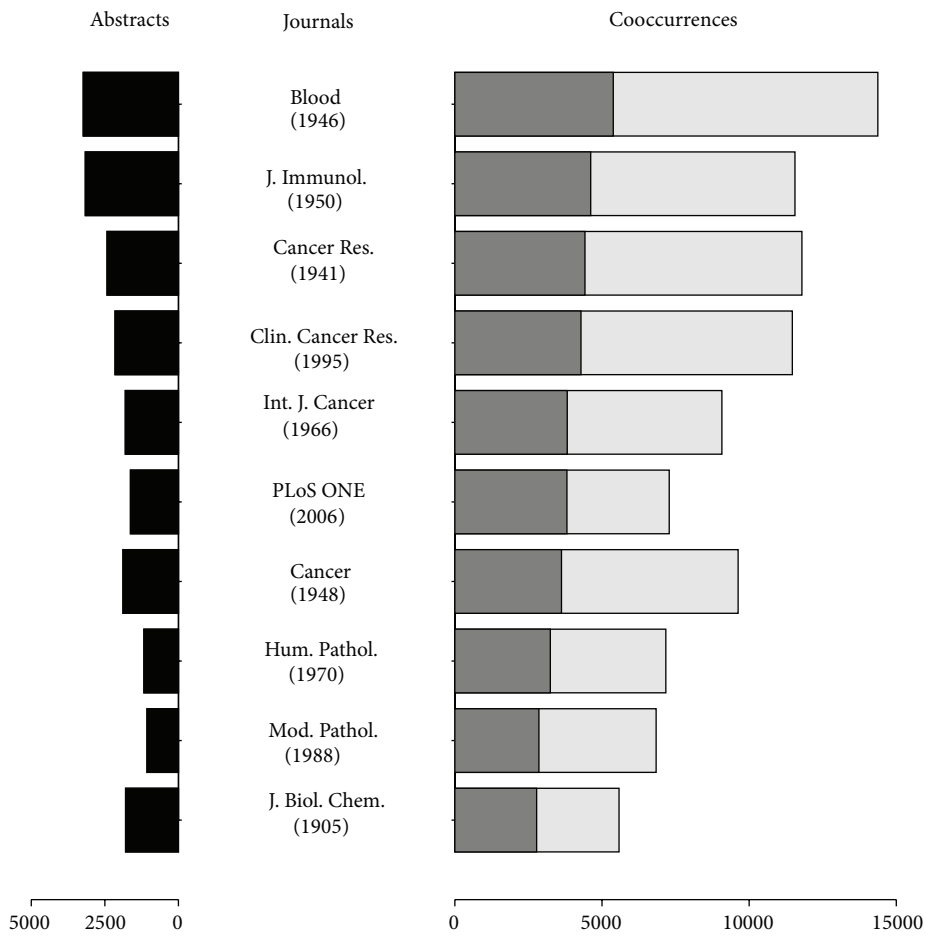


Figure 4: The top 10 journals sorted by unique disease-biomarker cooccurrences identified.

Table 3: The top 10 disease-biomarker associations. Disease-biomarker associations were ranked according to $Score_{DB}$ (see Section 2 for more details). The complete list of the associations is available at http://ibi.imim.es/biomarkers/.

Score	Disease name (CUI[a])	Gene symbol (Gene ID[b])	Number of abstracts
4076.42	NEOPLASM (C0027651)	TP53 (7157)	3,042
3930.25	NEOPLASM (C0027651)	ERBB2 (2064)	2,582
3441.32	NEOPLASM (C0027651)	CEACAM5 (1048)	2,234
2733.92	IMMUNODEFICIENCY DISORDER (C0021051)	CD4 (920)	1,548
2546.27	NEOPLASM (C0027651)	EGFR (1956)	1,710
2028.21	LEUKEMIA (C0023418)	CD34 (947)	1,071
1988.57	NEOPLASM (C0027651)	ESR1 (2099)	1,179
1943.15	NEOPLASM (C0027651)	AFP (174)	1,169
1915.15	NEOPLASM (C0027651)	CD34 (947)	1,108
1836.03	MALIGNANT NEOPLASTIC DISEASE (C0006826)	KLK3 (354)	936

[a]Concept unique identifier at UMLS.
[b]NCBI gene identifier.

these associations (more than 90%) have been reported in publications from the last three years.

In general, associations with high score are supported by a high number of publications (Figure 5(a)), and globally around 80% of the associations are supported by only 1 or 2 publications and have a low score. From this set, 35% corresponds to studies published in the last 3 years (Figure 5(b)). The novelty of these associations could explain the low number of supporting articles. Thus, it is likely that the remaining 65% of the associations supported by very few publications represent studies that could no longer be reproduced or are focused on very specific genes or diseases

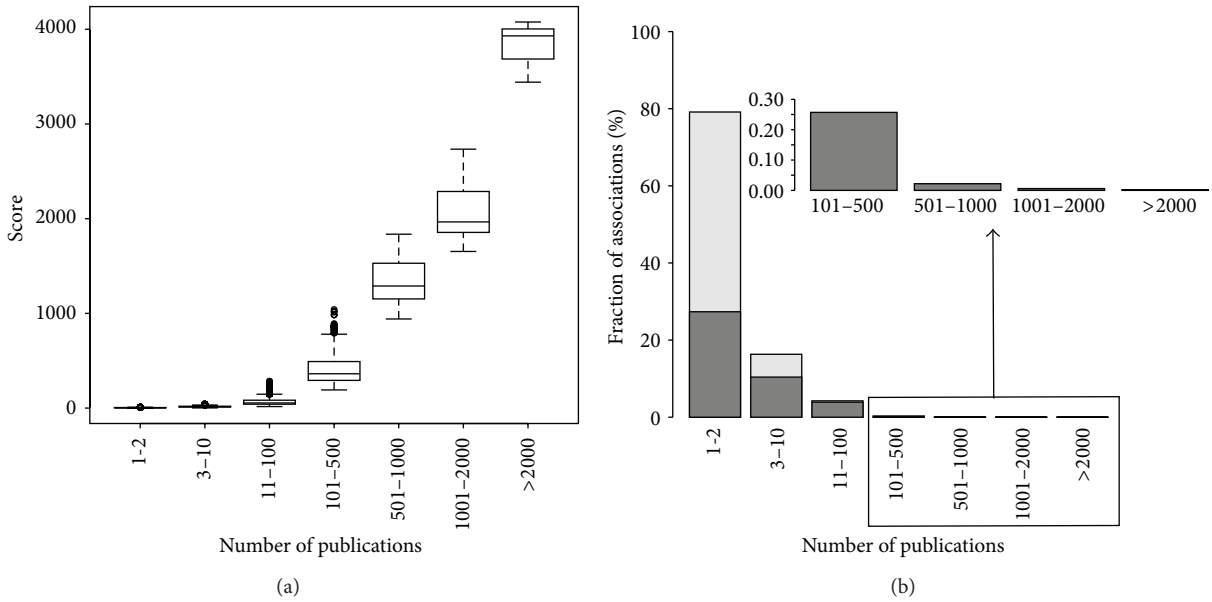


Figure 5: Associations analysis. (a) Boxplot showing the score *versus* number of publications supporting each disease-biomarker association. (b) Distribution of associations based on the number of publications that support each association. The fraction of the associations that were reported in the last three years is highlighted as dark grey bars.

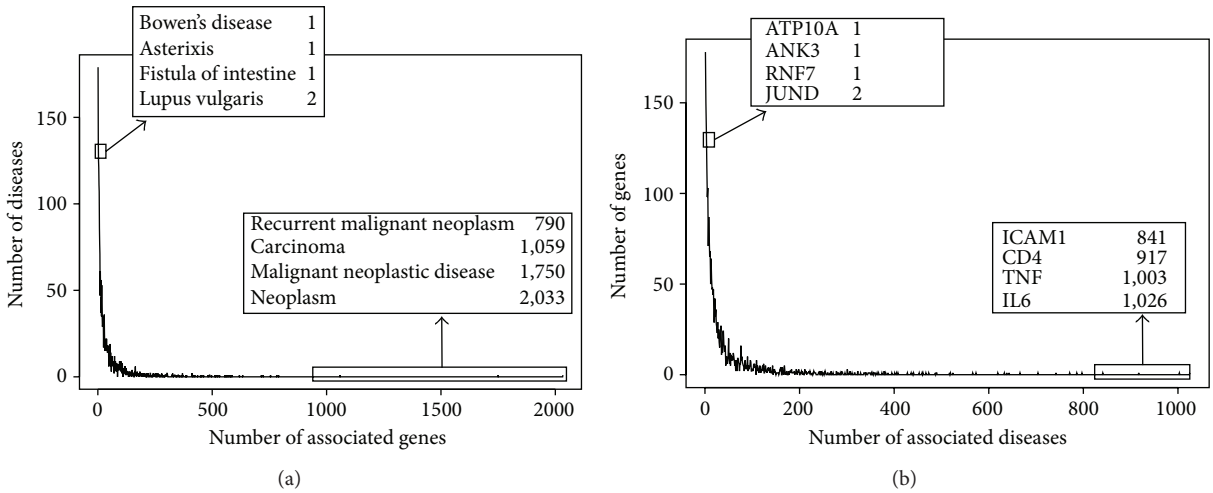


Figure 6: Distribution of the number of associated biomarkers (for diseases, (a)) and diseases (for biomarkers, (b)). Gene symbols from HGNC are used for the biomarkers.

that are not of widespread interest, as in the case of the most prevalent diseases such as some types of neoplasms. It is noteworthy that, for most of the associations supported by more than 10 articles, at least one of these articles has been published in the last 3 years (Figure 5(b)).

A further analysis of the results allows us to identify both the set of biomarkers associated with a large number of diseases (see Table 4) and the set of biomarkers associated with few (1 or 2) diseases. This information can be an indication of the "specificity" of a biomarker with respect to diseases. For example, a biomarker associated with many different diseases would be less specific than the other that has been studied in relation to a single disease, and, *vice versa*, the same consideration can be done for the diseases. The distribution of the number of associated diseases (for biomarkers) and biomarkers (for diseases) is depicted in Figure 6. For example, the genes PANK2, ANK3, and RNF7 appear as very specific biomarkers as they are associated with a single disease. On the other hand, several genes related to immune responses have been reported in associations with hundreds of diseases, such as IL6, TNF, and CD4 (Table 4 and Figure 6).

With respect to diseases, the results show that cancer is the therapeutic area that has more associated biomarkers (Table 5 and Figure 6). For instance, leukemia is associated with 782 biomarkers, and some of them (NOTCH1, SF3B1,

Table 4: The top 10 genes sorted by the number of associated diseases. The complete list of genes is available at http://ibi.imim.es/biomarkers/.

Gene symbol	Gene ID[a]	Gene name	Number of diseases
IL6	3569	Interleukin 6 (interferon, beta 2)	1,025
TNF	7124	Tumor necrosis factor	1,003
CD4	920	CD4 molecule	917
ICAM1	3383	Intercellular adhesion molecule 1	841
TP53	7157	Tumor protein p53	797
CRP	1401	C-reactive protein, pentraxin-related	786
CD8A	925	CD8a molecule	771
CD34	947	CD34 molecule	742
VEGFA	7422	Vascular endothelial growth factor A	704
ACE	1636	Angiotensin I converting enzyme	666

[a]NCBI gene identifier.

Table 5: The top 10 diseases sorted by the number of associated biomarkers. The complete list of diseases is available at http://ibi.imim.es/biomarkers/.

Disease name	CUI[a]	Number of genes
Neoplasm	C0027651	2,033
Malignant neoplastic disease	C0006826	1,750
Carcinoma	C0007097	1,059
Recurrent malignant neoplasm	C1458156	790
Leukemia	C0023418	782
Malignant melanoma	C0025202	755
Liver cell carcinoma	C2239176	723
Congenital deformity	C0000768	715
Tumor angiogenesis	C1519670	633
Tumor progression	C1519176	619

[a]Concept unique identifier at UMLS.

and BIRC3) have been found in recent literature reporting [34]. In contrast, few biomarkers have been identified for diseases like lupus vulgaris and Bowen's disease.

The disease-biomarker associations were also assessed according to the disease classes of the MeSH disease classification [31], indicating that neoplasm, nervous system diseases, and immune system diseases are by far the ones more investigated in the biomarkers research field (see Table 6).

In average, 11% of the disease-biomarker associations identified by our text mining approach were found in DisGeNET. Since DisGeNET contains information on the genetic determinants of human diseases and is not specially focused on biomarkers as defined in the present study, it is not surprising that only a small fraction of the information extracted from the literature is contained in DisGeNET database (July 2012 release). In addition, lag time in the population of the source databases by human curators may account for this difference. The dataset provided by the text mining approach here presented constitutes a large and valuable source of information on disease-related biomarkers, which can be used to populate specialized databases and to guide further research on biomarker validation. However, it is important to note that, based on the relation extraction approach used in this study, a proportion of the disease-biomarker associations found by this approach could be false positives. Future work will take in consideration the syntactic structure of the sentences in which a biomarker and a disease cooccur for the relation extraction process, with the aim of improving the precision of the approach. Search of semantic patterns reported in the abstracts' sentences will be checked in parallel to new data available from current and new disease-related biomarkers databases, with the aim of providing comprehensive and up-to-date knowledge to those biomedical researchers working in the disease-related biomarker field.

3.3. Related Work. Only few studies have proposed text mining approaches for extraction of biomarker related data [14–16]. For example, Younesi et al. presented a methodology for the retrieval of documents about biomarkers and showed as use cases the identification of markers for Alzheimer disease and multiple sclerosis [14]. Hui and Chunmei propose a finite state machine to identify pathways and diseases related to biomarkers [15]. We show in this study that a knowledge-driven approach is able to systematically exploit biomarker-specific information from large literature databases (e.g., MEDLINE) providing a comprehensive resource of biomarkers associated with diseases covering all the therapeutic areas.

4. Conclusions and Future Directions

The biomedical literature represents a rich resource for the identification of biomarker related information. However, both the size of the literature databases and the lack of standardization make difficult the automatic exploitation of the information contained in these resources. Text mining approaches have proven to be useful for the extraction of relations between entities, especially for the identification of interactions between proteins [13]. Here, we show that a knowledge-driven text mining approach can exploit a large literature database to extract a dataset of biomarkers related to diseases covering all therapeutic areas.

A bibliometric analysis of the journals reporting biomarker related information during the last 40 years highlighted the disparity among journals of different disciplines

Table 6: Comparison of disease-biomarkers pairs identified by the text mining (TM) approach with disease-biomarkers annotations in DisGeNET, based on MeSH disease classification [31].

MeSH disease class	MeSH disease class name	Number of disease-biomarker associations	The number validated with DisGeNET (%)
C01	Bacterial infections and mycoses	1,529	164 (10.73)
C02	Virus diseases	3,297	302 (9.16)
C03	Parasitic diseases	590	82 (13.90)
C04	**Neoplasms**	**31,627**	**5,264 (16.64)**
C05	Musculoskeletal diseases	5,771	388 (6.72)
C06	Digestive system diseases	8,154	1,156 (14.18)
C07	Stomatognathic diseases	2,531	195 (7.70)
C08	Respiratory tract diseases	5,460	735 (13.46)
C09	Otorhinolaryngologic diseases	770	40 (5.19)
C10	**Nervous system diseases**	**10,819**	**1,132 (10.46)**
C11	Eye diseases	2,513	226 (8.99)
C12	Male urogenital diseases	5,110	666 (13.03)
C13	Female urogenital diseases and pregnancy complications	6,432	863 (13.42)
C14	**Cardiovascular diseases**	**9,310**	**1,393 (14.96)**
C15	Hemic and lymphatic diseases	7,689	948 (12.33)
C16	**Congenital, hereditary, and neonatal diseases and abnormalities**	**10,382**	**397 (3.82)**
C17	Skin and connective tissue diseases	6,724	851 (12.66)
C18	Nutritional and metabolic diseases	6,314	711 (11.26)
C19	Endocrine system diseases	5,253	681 (12.96)
C20	**Immune system diseases**	**10,210**	**1,393 (13.64)**
C21	Disorders of environmental origin	2	0 (0.00)
C23	Pathological conditions, signs, and symptoms	8,212	606 (7.38)
C24	Occupational diseases	72	11 (15.28)
F01	Behavior and behavior mechanisms	594	24 (4.04)
F03	Mental disorders	2,810	613 (21.89)

which expands the publication bias, hampers the information retrieval and its exploitation, and, even, evidences the need of a standardization of the biomarker related data reporting to improve the quality of automatic extraction by means of mining techniques and gain confidence with their outcomes.

Our methodology focused on the extraction of disease-biomarker associations reported in the literature. This knowledge-driven approach takes advantage of the annotation of MEDLINE publications pertaining to biomarkers with MeSH terms, narrowing the search for specific publications and therefore minimizing the false positive ratio. The application of this methodology resulted in the identification of 131,012 disease-biomarker associations between 2,803 genes and 2,751 diseases and represents a valuable knowledge base for those interested in disease-related biomarkers. The results of this present study are available at http://ibi.imim.es/biomarkers/.

Future work in this area will focus on the identification of the type of association between the disease and the biomarker (for instance, distinguishing between the different levels of certainty that can be used to express an association or to specify the type of molecular change of the gene or protein associated with the disease). In addition, other types of molecules that can act as disease biomarkers could be identified as well.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

À. Bravo and M. Cases contributed equally to this paper.

Acknowledgments

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under Grants Agreements n°. [115002] (eTOX) and [115191]

(Open PHACTS), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. À. Bravo and L. I. Furlong received support from Instituto de Salud Carlos III Fondo Europeo de Desarollo Regional (CP10/00524). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

References

[1] A. J. Atkinson, W. A. Colburn, V. G. deGruttola et al., "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[2] Guidance for Industry-E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories, http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm129296.pdf.

[3] D. C. Anderson and K. Kodukula, "Biomarkers in pharmacology and drug discovery," *Biochemical Pharmacology*, vol. 87, no. 1, pp. 172–188, 2014.

[4] R. Frank and R. Hargreaves, "Clinical biomarkers in drug discovery and development," *Nature Reviews Drug Discovery*, vol. 2, no. 7, pp. 566–580, 2003.

[5] J. E. Dancey, K. K. Dobbin, S. Groshen et al., "Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents," *Clinical Cancer Research*, vol. 16, no. 6, pp. 1745–1755, 2010.

[6] C. N. A. M. Oldenhuis, S. F. Oosting, J. A. Gietema, and E. G. E. de Vries, "Prognostic versus predictive value of biomarkers in oncology," *European Journal of Cancer*, vol. 44, no. 7, pp. 946–953, 2008.

[7] G. Poste, "Bring on the biomarkers," *Nature*, vol. 469, no. 7329, pp. 156–157, 2011.

[8] J. R. Prensner, A. M. Chinnaiyan, and S. Srivastava, "Systematic, evidence-based discovery of biomarkers at the NCI," *Clinical & Experimental Metastasis*, vol. 29, no. 7, pp. 645–652, 2012.

[9] M. Cases, L. I. Furlong, and J. Albanell, "Improving data and knowledge management to better integrate health care and research," *Journal of International Medicine*, vol. 274, no. 4, pp. 321–328, 2013.

[10] MEDLINE, http://www.nlm.nih.gov/bsd/pmresources.html.

[11] Biological Markers MeSH term, http://www.ncbi.nlm.nih.gov/mesh/?term=biological+marker.

[12] Biomarkers, Pharmacological MeSH term, http://www.ncbi.nlm.nih.gov/mesh/?term=Biomarkers%2C+Pharmacological.

[13] U. Hahn, K. B. Cohen, Y. Garten, and N. H. Shah, "Mining the pharmacogenomics literature—a survey of the state of the art," *Briefings in Bioinformatics*, vol. 13, no. 4, pp. 460–494, 2012.

[14] E. Younesi, L. Toldo, B. Müller et al., "Mining biomarker information in biomedical literature," *BMC Medical Informatics & Decision Making*, vol. 12, article 148, 2012.

[15] L. Hui and L. Chunmei, "Biomarker identification using text mining," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 135780, 4 pages, 2012.

[16] W. J. Jessen, K. T. Landschulz, T. G. Turi, and R. Y. Reams, "Mining PubMed for biomarker-disease associations to guide discovery," *Nature Precedings*, 2012.

[17] NCBI-Gene, ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz.

[18] HGNC, http://www.genenames.org/cgi-bin/hgnc_downloads.

[19] UniProt human data, ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_human.dat.gz.

[20] UniProt IDmapping, ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping_selected.tab.gz.

[21] 2013AA UMLS Full Release Files, January 2013 version, http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.

[22] Casper, http://biosemantics.org/index.php?page=casper.

[23] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts," *Bioinformatics*, vol. 20, no. 18, pp. 3604–3612, 2004.

[24] W. W. Lau, C. A. Johnson, and K. G. Becker, "Rule-based human gene normalization in biomedical text with confidence estimation," in *Proceedings of the Life Sciences Society Computational Systems Bioinformatics Conference (CSB '07)*, vol. 6, pp. 371–379, San Diego, Calif, USA, August 2007.

[25] Q. C. Bui and P. M. Sloot, "A robust approach to extract biomedical events from literature," *Bioinformatics*, vol. 28, no. 20, pp. 2654–2661, 2012.

[26] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[27] A. Bauer-Mehren, M. Rautschka, F. Sanz, and L. I. Furlong, "DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks," *Bioinformatics*, vol. 26, no. 22, pp. 2924–2926, 2010.

[28] DisGeNET website, http://ibi.imim.es/DisGeNET/.

[29] DisGeNET gene-disease association ontology (dgdao), 2012, http://ibi.imim.es/DisGeNET-Dev/ontologies/GeneDiseaseAssociation_v4.owl.

[30] Semantics Science Integrated Ontology or SIO Ontology, http://code.google.com/p/semanticscience/wiki/SIO.

[31] MeSH Disease tree, http://www.nlm.nih.gov/mesh/trees.html.

[32] R. Xu and Q. Wang, "A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 827–834, 2012.

[33] S. M. Bentzen, F. M. Buffa, and G. D. Wilson, "Multiple biomarker tissue microarrays: bioinformatics and practical approaches," *Cancer and Metastasis Reviews*, vol. 27, no. 3, pp. 481–494, 2008.

[34] D. Rossi, "IX. Chronic lymphocytic leukaemia: new genetic markers as prognostic factors," *Hematological Oncology*, vol. 31, supplement 1, pp. 57–59, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 278748, 7 pages
http://dx.doi.org/10.1155/2014/278748

Research Article

Integrated Analysis of Gene Network in Childhood Leukemia from Microarray and Pathway Databases

Amphun Chaiboonchoe,[1,2] Sandhya Samarasinghe,[1,3] Don Kulasiri,[1,4] and Kourosh Salehi-Ashtiani[2]

[1] *Centre for Advanced Computational Solutions (CfACS), Lincoln University, Lincoln 7647, New Zealand*
[2] *Division of Science and Math, New York University Abu Dhabi and Center for Genomics and Systems Biology (CGSB), New York University Abu Dhabi Institute, P.O. Box 129188, Abu Dhabi, UAE*
[3] *Integrated Systems Modelling Group, Lincoln University, Lincoln 7647, New Zealand*
[4] *Department of Wine, Food & Molecular Biosciences, Lincoln University, Lincoln 7647, New Zealand*

Correspondence should be addressed to Sandhya Samarasinghe; sandhya.samarasinghe@lincoln.ac.nz

Received 22 November 2013; Revised 24 February 2014; Accepted 3 March 2014; Published 15 April 2014

Academic Editor: Shigehiko Kanaya

Glucocorticoids (GCs) have been used as therapeutic agents for children with acute lymphoblastic leukaemia (ALL) for over 50 years. However, much remains to be understood about the molecular mechanism of GCs actions in ALL subtypes. In this study, we delineate differential responses of ALL subtypes, B- and T-ALL, to GCs treatment at systems level by identifying the differences among biological processes, molecular pathways, and interaction networks that emerge from the action of GCs through the use of a selected number of available bioinformatics methods and tools. We provide biological insight into GC-regulated genes, their related functions, and their networks specific to the ALL subtypes. We show that differentially expressed GC-regulated genes participate in distinct underlying biological processes affected by GCs in B-ALL and T-ALL with little to no overlap. These findings provide the opportunity towards identifying new therapeutic targets.

1. Introduction

Childhood leukaemia, specifically, acute lymphoblastic leukaemia (ALL) can be divided into two subgroups: T-lineage and B-lineage. Glucocorticoids (GCs) are a type of steroid hormones. Synthetic glucocorticoids such as dexamethasone (Dex) and prednisolone (PRD) are the most important drugs that have been used extensively in the treatment of children with acute lymphoblastic leukaemia (ALL) because of their ability to induce apoptosis (cell death) in the lymphoid. There are more than 2000 studies on GC-induced apoptosis in lymphoid cells [1], of which some studies focus on identifying GC-regulated genes by using gene expression profiles from different types of glucocorticoid drugs in different clinical settings (*in vivo*, *in vitro*, and human samples) [2–4]; however, there are only a few overlapping genes identified from these studies. Therefore, there is a need to verify GC-regulated genes to better define the underlying network of genes involved in the actions of GCs.

In this study, we illustrated how a combination of existing bioinformatics tools can address the heterogeneity of ALL in three different levels: gene, gene set, and network and pathway. First, at the gene level, we identified GCs-regulated candidate genes for each subtype. The second level was to use a group of genes (gene sets) that may have similar functions. We aim to ascertain whether each gene set from each subtype is significantly enriched in a list of selected phenotypes. The third level was to construct gene networks of the proposed GCs regulated genes from three selected web-based tools: Ingenuity Pathway Analysis (IPA), Search Tool for The Retrieval of Interacting Genes (STRING), and Gene Multiple Association Network Integration Algorithm (GeneMANIA).

2. Materials and Methods

2.1. Dataset. Raw microarray data in the format of CEL files were obtained online from the Gene Expression Omnibus

(GEO) (http://www.ncbi.nlm.nih.gov/geo/). Raw data, comprising gene expression measurements for 13 patients (three T-ALL patients and ten B-ALL patients), contained gene expression measurements collected at three time points: 0 hour, 6/8 hours, and 24 hours. The experiments and analysis conducted on the data are described in Schmidt et al. [5]. Summary of dataset analysis methodology used in this study is shown in Figure 1.

2.2. Normalization. The normalization used in this study is Robust Multiarray Average (RMA) [6]. RMA uses a global correction, quantile normalization, and median polish summarization. There are many existing software and tools for RMA calculation available from both commercial and free sources. This study used R (http://www.r-project.org/) and BioConductor (http://www.bioconductor.org/). The differentially expressed genes were selected from the normalised data according to specified threshold activation (log-ratio of ±1 or higher or ±0.7 or higher).

2.3. Gene Set Enrichment Analysis and Enrichment Map. Gene set enrichment analyses (GSEA) [7, 8] are commonly used to determine the biological characterization, statistical significance, and concordant differences between an experimental gene set and a selected gene list from annotated gene sets knowledge base stored on Molecular Signatures Database (MSigDB). GSEA can be downloaded from http://www.broadinstitute.org/gsea/downloads.jsp. The Jaccard coefficient is used to compare the similarity between two sample gene sets A and B and defined as the intersection between group A and B divided by their union. The results from GSEA are then visualized through the enrichment map [9], a Cytoscape plugin for network visualization. The ranked experimental gene list along with the enriched gene sets from GSEA is used to build the network of gene sets (nodes) where edges represent their similarity. Size of a node varies by gene set size and the thickness of the edge represents the degree of correlation between two gene sets.

2.4. Networks and Pathway Analysis

2.4.1. Ingenuity Pathway Analysis Software (IPA). IPA (Ingenuity Systems, Redwood City, CA, USA, http://www.ingenuity.com/) is a web-based application that applies a systems biology approach to solve various biological problems. The knowledge base of IPA has been curated from journal articles, textbooks, and other data sources. This software has many applications; only functional analysis of genes and their networks were used in this study. The p value defines the significance of gene function in a network as well as gene to gene relation, and a p value less than 0.05 signifies a statistically significant and nonrandom association. The right-tailed Fisher Exact Test is used to calculate p value.

2.4.2. Analysis of Network Invoked by GC-Regulated Genes. The biological knowledge of gene and protein interactions is growing rapidly and there are many tools and curated databases available on a large scale. Insightful knowledge gained from studying gene sets rather than individual genes using network-based approaches can reveal network patterns and relevant molecular pathways from the experiment gene sets. In this study, we utilized two different freely accessible and user-friendly web tools as follows.

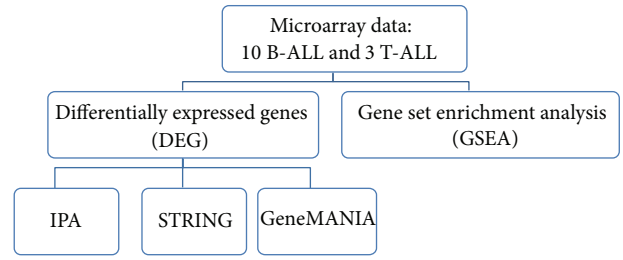


Figure 1: Summary of dataset analysis methodology used in this study.

Gene Multiple Association Network Integration Algorithm (GeneMANIA) [10, 11] (http://www.genemania.org/) is a web-based tool for prediction of gene function or implemented as a Cytoscape plugin tool. Based on single gene or gene set query from 7 organisms, it shows results for interactive functional associative network according to their coexpression data from Gene Expression Omnibus (GEO), physical and genetic interaction data derived from BioGRID, predicted protein interaction data based on orthology from I2D, colocalization, shared protein domain, and GO function.

Search Tool for the Retrieval of Interacting Genes (STRING) version 9.1 [12, 13] (http://string-db.org/) is an online protein-protein interaction database curated from literature and predicted associations from systemic genome comparisons. User can query using single or multiple name(s) and protein sequence(s). The protein interactions can be displayed according to their confidence, evidence, actions, or interactions.

3. Results and Discussion

3.1. Identification of GC-Regulated Genes through a Refined Analysis of Data. Our initial analysis of this time series gene expression data for differentially expressed gene identification followed the same method used by the original authors [5]. All 39 files were processed and normalised by Robust Multiarray Average (RMA) in R as in the original study. The selected normalization method may have an effect on downstream analysis, for example, reverse engineering analyses [14]; however, investigating this effect is beyond the scope of this study. We found that combining B-ALL and T-ALL data (as done by the original authors) compromises the accuracy of selection of differentially expressed. For example, the original authors found 22 differentially expressed genes from the combined dataset. However, a closer inspection revealed to us that only 8/22 candidate genes belonged to both B-ALL and T-ALL and 14/22 genes were found only in B-ALL or T-ALL.

Therefore, we separated the data from the two types of patients and a new set of differentially expressed genes was selected for T-ALL and B-ALL for each time point. Our new criteria used were log ratio of ±1 or higher for at least five

out of ten (50%) B-ALL patients and two out of three T-ALL patients. We also analysed data for early response (6 hours) and late response (24 hours), but we added an analysis of response between 6 and 24 hours because this can give more information about gene activity at different periods. The results are shown in Table 1.

Table 1 shows the number of differentially expressed genes 6 hours after treatment, between 6 hours and 24 hours, and 24 hours after treatment (before and after deleting cell cycle genes) with our new criteria. Before deleting the cell cycle genes, the final set had 237 probe sets (203 unique probe sets after removing repeats) for T-ALL for the combined time points and 257 probe sets (207 unique probe sets) for B-ALL and these were combined into one set. The final set contained 386 unique probe sets (24 unique probe sets were common to T-ALL and B-ALL, of which three probes were not found by the original authors. These were converted from probe set ID to gene symbol by using DAVID (the Database for Annotation, Visualization and Integrated Discovery). Then the cell cycle genes were deleted from this dataset (cell cycle gene list was retrieved from KEGG, Cell cycle database, and the original article).

After deleting cell cycle genes, T-ALL contained 222 probe sets (172 unique probe sets) and B-ALL contained 190 probe sets (155 unique probe sets) for the combined time points. The final set had 308 unique probe sets (304 genes) responsive to GCs (19 unique probe sets were common to T-ALL and B-ALL).

We then compared our new gene list with the genes reported in previous studies. None of the previous studies using the same drugs and the same times at which data were collected show similar results. Our main focus here is to use relevant experimental data from clinical samples, such as the study by Tissing et al. [15], where the investigators used the same drug on retrieved tissues from childhood leukemia patients, not cell cultures, but used a different normalization process, variance stabilization procedure (VSN), than of our study. They compared the primary childhood ALL cells treated with *in vivo* prednisone and leukemic cells of childhood ALL exposed to *in vitro* prednisolone, for finding early apoptosis responsive genes at 8 hours after treatment. Our 29 differentially expressed probe sets from T-ALL and 47 probe sets from B-ALL (Table 1, first two columns containing cell cycle genes) were compared with the 39 upregulated and 12 downregulated genes from Tissing et al. [15]. We found four common induced genes: BTG1 (T-ALL), FKBP5, ZBTB16, and SNF1LK (B-ALL). However, no common repressed genes were found between our results and Tissing et al. [15] for either B-ALL or T-ALL.

We selected another study, Thompson and Johnson [16], based on different chemotherapeutic drugs, different time points, and different tissues, but same gene selection criteria as ours. Thompson and Johnson [16] identified 39 upregulated genes and 21 downregulated genes in CEM (a cell line derived from human lymphoid cells). In addition, they proposed a time frame for apoptosis gene regulation after CEM-C7 were exposed to dexamethasone [16]. Comparing gene sets reported by Thompson and Johnson [16] with our differentially expressed genes from T-ALL and B-ALL patients, we found few overlapping genes, but more than what we found from comparison with Tissing et al. [15]. T-ALL had five overlapping genes (BCL2L1, SOCS1, BTG1, CD69, and NR3C1) and B-ALL had four overlapping genes (SOCS1, DFNA5, WFS1, and SLA). Of these two sets, BTG1 is the only common gene between our T-ALL, Tissing et al. [15] and Thompson and Johnson [16]. There were no common genes between our B-ALL, Tissing et al. [15] and Thompson and Johnson [16].

3.2. Extraction of Intrinsic Biological Patterns with Gene Enrichment Analysis Applied to Gene Expression Data. Identification of differentially expressed genes between the two classes has limited value in gaining biological insights unless it is integrated with other analyses. The gene set enrichment analysis (GSEA) method paves the way to interpret gene expression data by using prior knowledge databases to define functionally characterized gene sets and to reveal whether the identified gene sets have common biological functions or gene ontologies.

ALL gene expression data corresponding to all 54,675 probe sets on the chip were collapsed into 20,606 gene symbols. A gene set S, a subset derived from the gene symbols, was used to calculate the enrichment score, which placed the set S, according to statistical significance, at the top or bottom of the selected list L of gene sets from MSigDB (Molecular Signature Database, version 4.0 updated May 31, 2013). The list L that was used in this study consisted of two types: (1) functional set (C_2) 4722 gene sets collected from KEGG online pathway database and (2) GO gene set (C_5) 1454 gene sets according to the GO terms. The differentially expressed genes found in this study belonged to 176 gene sets and out of these 130 gene sets were upregulated in B-ALL and 46 gene sets were upregulated in T-ALL. For B- and T-ALL, 12 and 8 gene sets were significantly enriched at nominal p value <5%, respectively. Table 2 shows the different KEGG pathways where B- and T-ALL gene sets were enriched. The limitation of GSEA is the gene set redundancy and difficulty in interpreting the results from large gene lists. We used the enrichment map, a Cytoscape plugin, to build the network from GSEA Go term enrichment results as shown in Figure 2. Node size defines the number of genes in each gene set and edge thickness represents the proportion of overlap between gene sets, calculated using the Jaccard coefficient. Blue represents T-ALL gene sets and red represents B-ALL gene sets. From Table 2 and Figure 2, we differentiate the functional characteristics of metabolic pathways (KEGG) and biological processes (GO) in which B- and T-ALL are involved. These distinctions are as follows: B-ALL is likely to be involved in asthma, B-cell receptor signalling pathway, and phosphorylation, while T-ALL is involved in T-cell receptor signalling pathway, primary immunodeficiency process, and leucocyte.

3.3. Networks and Pathways Analysis

3.3.1. Inferring GR Gene Networks from GC-Induced Apoptosis Genes. Dataset containing expression values of GCs-regulated genes were uploaded and analyzed using Ingenuity Pathways Analysis software (Ingenuity Systems, http://www.ingenuity.com/). Ingenuity Pathway Analysis

Table 1: Differentially expressed probe sets 6 hours after treatment, between 6 hours and 24 hours, and 24 hours after treatment at ±1 log2 ratio fold change (before and after deleting cell cycle genes).

	0–6 hours				6–24 hours				0–24 hours			
	Before		After		Before		After		Before		After	
	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed
T-ALL	19	10	19	9	59	51	56	49	58	40	56	33
B-ALL	24	23	24	9	16	13	16	9	73	108	71	61

Table 2: Top 5 of KEGG enrichment term of B- and T-ALL.

Enriched in B-ALL	FDR	Enriched in T-ALL	FDR
KEGG_ASTHMA	0.001	KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	0.0003
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	0.002	KEGG_PRIMARY_IMMUNODEFICIENCY	0.07
KEGG_ANTIGEN_PROCESSING_AND_ PRESENTATIONKEGG_LEISHMANIA_INFECTION	0.003	KEGG_HEMATOPOIETIC_CELL_LINEAGE	0.209
KEGG_LEISHMANIA_INFECTION	0.009	KEGG_BIOSYNTHESIS_OF_UNSATURATED_FATTY_ACIDS	0.162
KEGG_TYPE_I_DIABETES_MELLITUS	0.058	KEGG_ALPHA_LINOLENIC_ACID_METABOLISM	0.210

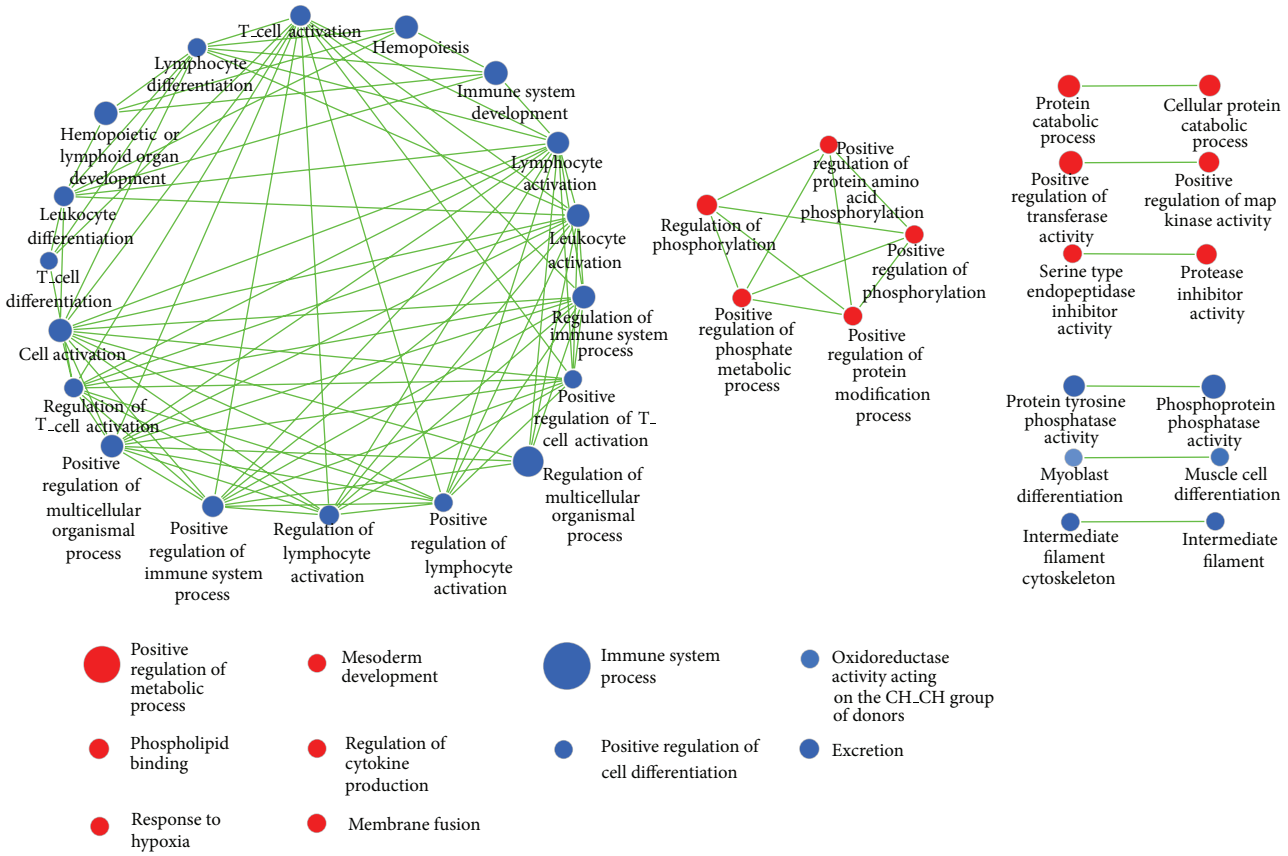


Figure 2: Gene set enrichment analysis delineates gene ontology (GO) that differentiates between B- and T-ALL with respect to biological processes. Gene set enrichment analysis (GSEA) comparing B-ALL (red) and T-ALL (blue) in ALL dataset, illustrating differentiation of gene ontology (biological processes) between two subgroups (5% FDR, $p = 0.05$). Cytoscape and enrichment map were used for visualization of the GSEA results; only gene sets from MSigDB C5 (gene ontology) were used. Nodes represent enriched GO gene sets, whose size reflects the total number of genes in that gene set. Edge thickness (green line) represents the number of overlapping genes between gene sets calculated using Jaccard coefficient. Single nodes and 2-node interactions for both B- and T-ALL, a 5 node-interaction for B-ALL, and interaction between a large number of nodes for T-ALL are shown.

software (IPA) maps the genes to pathways generating networks using an algorithm based on gene connectivity with cutoff of 35 molecules per network.

We used the gene list after deleting cell cycle genes to infer gene networks through IPA software. For T-ALL, there are 28 probe sets at 0–6 hours, 105 probe sets at 6–24 hours and 89 probe sets at 0–24 hours. For B-ALL there are 33 probe sets at 0–6 hours, 25 probe sets at 6–24 hours, and 132 probe sets at 0–24 hours. Results from IPA typically are a number of networks ranking from one to seven for each time point with a maximum of 35 molecules in each network consisting of those genes from our list plus those added by IPA. Processing of these networks were fairly time consuming, but we manually identified common genes active throughout the whole period (at least between two time points) which can be referred to as genes predominant in T- and B-ALL separately or common to both.

Overall, for T-ALL patients, 48 unique genes (23 genes were given as input and 25 were added by IPA) were found for the three different time points from IPA. For B-ALL patients, 47 unique genes (21 genes were given as input and 26 were added by IPA) were found.

In the next step, we only selected genes that were found at least in two of the three time intervals from our list to create gene networks. T- and B-ALL patients were quite different in the "molecular and cellular functions" and "canonical pathways and functions." For example, molecular and cellular functions of T-ALL were involved more in cell death, while those of B-ALL were more involved in cell cycle. This finding may imply that (i) apoptosis process in T-ALL may occur before B-ALL during the same period of treatment; (ii) there are many cells progressing through cell cycle (cycling cells) in B-ALL, while many noncycling cells are in T-ALL. Many pathways/steps are involved in cycling cells than in non-cycling cells in the apoptosis process after glucocorticoid treatment [17]. Genes found in both T-and B-ALL were involved in cancers functions.

3.3.2. Analysis of Network Invoked by GC-Regulated Genes. Network analysis can help understand the molecular and cellular interactions [18]. It can be visualized to represent entities (nodes) and their relationships (arcs). The advent of high throughput technology has led to a large increase in publicly available information. Each data type can capture different aspect of functional roles of interested genes. In this section, we investigated functional interaction among genes and proteins in the cell using available data and knowledge bases. We selected two different web-based network tools: GeneMANIA and STRING using the differentially expressed genes from early response of B- and T-ALL as a query gene sets. These toolsets integrates computational methods to predict the gene functions based on a collection of interaction networks.

GeneMANIA is a large collection of interaction networks from several data sources which identify genes and networks that are functionally associated (protein and genetic interactions, pathways, coexpression, colocalization, and protein domain similarity) with the query gene sets. Another advantage is that the user can run this as a plugin with the Cytoscape tool allowing the user to apply other tools to analyze the networks. STRING relies on the phylogeny to infer the functional interaction (protein networks) with direct interaction to score nodes, while GeneMANIA uses functional genomic data with label propagation to score nodes and generate gene networks. STRING uses precomputed networks, while those of GeneMANIA are not precomputed, and user can upload their own networks. STRING covers a large number of organisms, while GeneMANIA only covers 7 organisms but allows the user to upload or add more networks through the plugin. In addition, users can run enrichment analysis (GO, KEGG, PFAM, INTERPRO, and protein-protein interaction) on STRING.

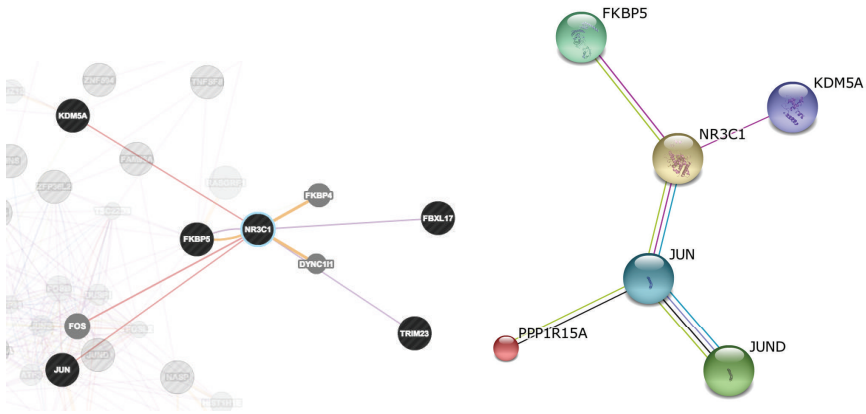
The results of genes and network that are functionally associated with the gene set from early response of T-ALL are shown in Figure 3. We compared the two networks from STRING and GeneMANIA based on the interactions they revealed; here we used NR3C1 as the centre gene in the comparison. All interactions found in STRING were found in GeneMANIA as described in more detail in Table 3. Comparison between the results from both tools can be used to confirm the functional associations of the interested gene sets. There is evidence of overlap and uniqueness in the interactions revealed by the two web-based tools.

4. Summary and Conclusions

We used a dataset from Schmidt et al. [5], the most prominent dataset at the time because it used gene expression data collected from childhood leukemia patients at two time points after treatment. We found that the selected dataset is reproducible and robust. The original differentially expressed genes were proposed by the authors for the combined dataset; however, only some of these genes were found in each subtype. To resolve the discrepancy, we proposed a new criteria for the two subtypes separately (log 2 ratio of ±1.0 for five out of ten B-ALL patients and two out of three T-ALL patients) and new gene sets were generated. Furthermore, we extended the analysis to find differentially expressed genes between 6 and 24 hours. In addition, we compared this gene set with differentially expressed gene sets from two previous studies and found only a few common genes possibly indicating that different chemotherapeutic agents and tissues may produce different results for the target gene set.

We identified common genes by manually extracting connections from inferred gene networks for each time interval. In addition, results from IPA showed different molecular and cellular functions, canonical pathways and functions between T- and B-ALL. T-ALL is more involved in cell death, while B-ALL is more involved in cell cycle.

Converting gene list to gene sets, we identified the known metabolic pathways that were enriched in each subtype using GSEA with enrichment map. We subsequently compared the top 5 gene ontology (GO) functions for B- and T-ALL. Then two network based tools: GeneMANIA and STRING were used to identify the gene network. STRING uses protein names to search for known and predicted protein interactions while GeneMANIA uses gene symbols to search for gene function prediction(s). Comparing gene interaction



(a) (b)

Figure 3: Gene network of T-ALL (early response) derived from GeneMANIA (a) and STRING (b) (NR3C1 interaction is highlighted). (a) A gene network from GeneMANIA shows the relationships for genes from the list (nodes) connected (with edges) according to the functional association networks from the databases. (b) The figure illustrates the protein interaction upon querying STRING protein network (evidence view) in *Homo sapiens* with 49 proteins. Additional information from other resources can be retrieved for each protein and interaction. Nodes represent proteins and different line colours denote the type of evidence for the interaction.

Table 3: Interactions for T-ALL (early response) network using GeneMANIA and STRING.

Interaction	GeneMANIA	STRING
NR3C1 → FKBP5	Coexpression Predicted	Coexpression Comentioned in PubMed abstracts
NR3C1 → JUN	Physical interaction	Comentioned in PubMed abstracts Association in curated databases, experiments data
NR3C1 → KDM5A	Physical interaction	Experiments data
NR3C1 → STS	Colocalization	—
NR3C1 → TRIM23	Coexpression	—
NR3C1 → EPM2AIP1	Coexpression	—
NR3C1 → CHST11	Coexpression	—
NR3C1 → FBXL17	Coexpression	—
FOS → NR3C1	Physical interaction	—
FKBP4 → NR3C1	Predicted	—
DYNC1I1 → NR3C1	Predicted	—
JUN → JUND	Physical interaction Coexpression Shared protein domains	Coexpression Comentioned in PubMed abstracts Association in curated databases Posttranslational modification
JUN → PPP1R15A	Coexpression Comentioned in PubMed abstracts	Coexpression Comentioned in PubMed abstracts

for T-ALL from GeneMANIA and STRING, we found some overlap.

In summary, we utilized the strengths of existing network/pathway tools and databases to gain insight into processes related to childhood leukemia subtypes; T-ALL and B-ALL have distinct molecular interaction patterns visible from various systems levels, including gene, gene set, molecular pathway, and gene networks. Discriminating between the two groups can help to improve the understanding of a drug's mechanism and further improve targeting in therapeutics drug research. In addition, the future research should consider combining RNA-Seq data [19–21] for identification of novel prognostic markers and therapeutic targets.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Amphun Chaiboonchoe designed the study, analyzed the data, and wrote the manuscript, and Sandhya Samarasinghe, Don Kulasiri, and Kourosh Salehi-Ashtiani contributed to the design of the study and co-wrote the paper. All authors have read and approved the final version of the paper.

Acknowledgments

Primary support for this work was provided by postgraduate research funding from Lincoln University, New Zealand; additional support was provided by New York University Abu Dhabi Institute Grant G1205 and Faculty Research Funds.

References

[1] I. Herr, N. Gassler, H. Friess, and M. W. Büchler, "Regulation of differential pro- and anti-apoptotic signaling by glucocorticoids," *Apoptosis*, vol. 12, no. 2, pp. 271–291, 2007.

[2] S. Schmidt, J. Rainer, C. Ploner, E. Presul, S. Riml, and R. Kofler, "Glucocorticoid-induced apoptosis and glucocorticoid resistance: molecular mechanisms and clinical relevance," *Cell Death and Differentiation*, vol. 11, supplement 1, pp. S45–S55, 2004.

[3] M. Tonko, M. J. Ausserlechner, D. Bernhard, A. Helmberg, and R. Kofler, "Gene expression profiles of proliferating versus G1/G0 arrested human leukemia cells suggest a mechanism for glucocorticoid-induced apoptosis," *The FASEB Journal*, vol. 15, no. 3, pp. 693–699, 2001.

[4] G. Cario, A. Fetz, C. Bretscher et al., "Initial leukemic gene expression profiles of patients with poor in vivo prednisone response are similar to those of blasts persisting under prednisone treatment in childhood acute lymphoblastic leukemia," *Annals of Hematology*, vol. 87, no. 9, pp. 709–716, 2008.

[5] S. Schmidt, J. Rainer, S. Riml et al., "Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia," *Blood*, vol. 107, no. 5, pp. 2061–2069, 2006.

[6] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[7] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[8] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, and J. P. Mesirov, "GSEA-P: a desktop application for gene set enrichment analysis," *Bioinformatics*, vol. 23, no. 23, pp. 3251–3253, 2007.

[9] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader, "Enrichment map: a network-based method for gene-set enrichment visualization and interpretation," *PLoS ONE*, vol. 5, no. 11, Article ID e13984, 2010.

[10] K. Zuberi, M. Franz, H. Rodriguez et al., "GeneMANIA prediction server 2013 update," *Nucleic Acids Research W*, vol. 41, no. 1, pp. W115–W122, 2013.

[11] D. Warde-Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research W*, vol. 38, supplement 2, pp. W214–W220, 2010.

[12] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D561–D568, 2011.

[13] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research D*, vol. 41, no. 1, pp. D808–D815, 2013.

[14] W. K. Lim, K. Wang, C. Lefebvre, and A. Califano, "Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks," *Bioinformatics*, vol. 23, no. 13, pp. i282–i288, 2007.

[15] W. J. E. Tissing, M. L. den Boer, J. P. P. Meijerink et al., "Genome-wide identification of prednisolone-responsive genes in acute lymphoblastic leukemia cells," *Blood*, vol. 109, no. 9, pp. 3929–3935, 2007.

[16] E. B. Thompson and B. H. Johnson, "Regulation of a distinctive set of genes in glucocorticoid-evoked apoptosis in CEM human lymphoid cells," *Recent Progress in Hormone Research*, vol. 58, pp. 175–197, 2003.

[17] K. King and J. Cidlowski, "Cell cycle regulation and apoptosis 1," *Annual Review of Physiology*, vol. 60, no. 1, pp. 601–617, 1998.

[18] A. Chaiboonchoe, W. Jurkowski, J. Pellet et al., "On different aspects of network analysis in systems biology," in *Systems Biology*, pp. 181–207, Springer, Dordrecht, The Netherlands, 2013.

[19] Z. K. Atak, V. Gianfelici, G. Hulselmans et al., "Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia," *PLoS Genetics*, vol. 9, no. 12, Article ID e1003997, 2013.

[20] I. Zwiener, B. Frisch, and H. Binder, "Transforming RNA-Seq data to improve the performance of prognostic gene signatures," *PLoS ONE*, vol. 9, no. 1, Article ID e85150, 2014.

[21] H. Lilljebjörn, H. Agerstam, C. Orsmark-Pietras et al., "RNA-seq identifies clinically relevant fusion genes in leukemia including a novel MEF2D/CSF1R fusion responsive to imatinib," *Leukemia*, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 194812, 11 pages
http://dx.doi.org/10.1155/2014/194812

Research Article

Tools and Databases of the KOMICS Web Portal for Preprocessing, Mining, and Dissemination of Metabolomics Data

Nozomu Sakurai,[1,2] Takeshi Ara,[1,2] Mitsuo Enomoto,[1,2] Takeshi Motegi,[1] Yoshihiko Morishita,[1] Atsushi Kurabayashi,[1] Yoko Iijima,[1,3] Yoshiyuki Ogata,[1,4] Daisuke Nakajima,[1] Hideyuki Suzuki,[1] and Daisuke Shibata[1]

[1] *Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan*
[2] *JST, National Bioscience Database Center (NBDC), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan*
[3] *Department of Nutrition and Life Science, Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Kanagawa 243-0292, Japan*
[4] *Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan*

Correspondence should be addressed to Nozomu Sakurai; sakurai@kazusa.or.jp

Received 5 December 2013; Revised 7 February 2014; Accepted 24 February 2014; Published 9 April 2014

Academic Editor: Shigehiko Kanaya

A metabolome—the collection of comprehensive quantitative data on metabolites in an organism—has been increasingly utilized for applications such as data-intensive systems biology, disease diagnostics, biomarker discovery, and assessment of food quality. A considerable number of tools and databases have been developed to date for the analysis of data generated by various combinations of chromatography and mass spectrometry. We report here a web portal named KOMICS (The Kazusa Metabolomics Portal), where the tools and databases that we developed are available for free to academic users. KOMICS includes the tools and databases for preprocessing, mining, visualization, and publication of metabolomics data. Improvements in the annotation of unknown metabolites and dissemination of comprehensive metabolomic data are the primary aims behind the development of this portal. For this purpose, PowerGet and FragmentAlign include a manual curation function for the results of metabolite feature alignments. A metadata-specific wiki-based database, Metabolonote, functions as a hub of web resources related to the submitters' work. This feature is expected to increase citation of the submitters' work, thereby promoting data publication. As an example of the practical use of KOMICS, a workflow for a study on *Jatropha curcas* is presented. The tools and databases available at KOMICS should contribute to enhanced production, interpretation, and utilization of metabolomic Big Data.

1. Introduction

A metabolome, which comprises comprehensive data on quantification of metabolites in an organism calculated using metabolomic technologies [9, 10], has been increasingly used for the analysis and practical applications of biological and environmental systems. Within the data-intensive systems biology discipline, metabolomics is particularly important compared to other "omics" (genome, transcriptome, and proteome) disciplines since metabolomes are more closely related to phenotype and regulate gene and protein expression networks [11–13]. Mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) are complementary techniques often used for the detection and identification of metabolites. MS technology has integrated separation techniques and is used in most cases because of its sensitivity, selectivity, speed, and broad applicability [14–16]. Owing to the wide range of chemical diversity, there is no ideal apparatus that is capable of analyzing all possible metabolites. Combinations of separation techniques with MS, such as liquid chromatography- (LC-) MS, gas chromatography- (GC-) MS, and capillary electrophoresis-

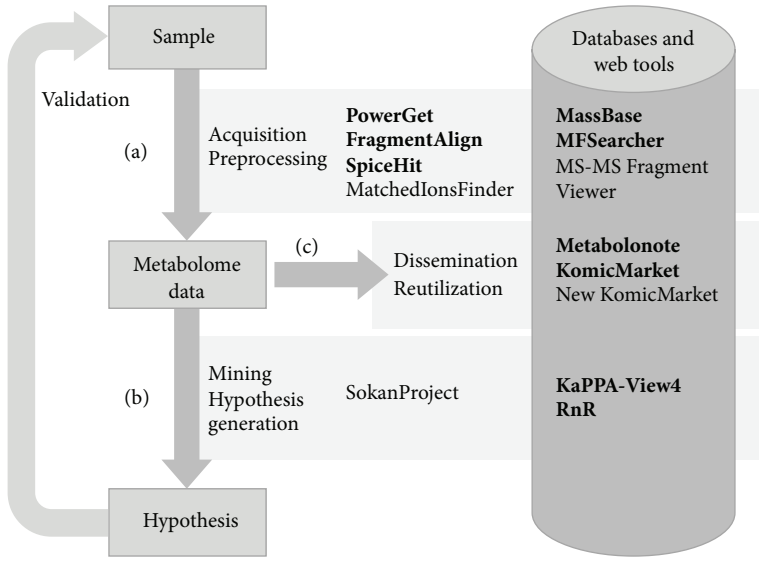


Figure 1: A typical workflow of a metabolomics study and KOMICS-relevant tools and databases. The process of data acquisition and preprocessing (a) is required for generating the metabolome data. A working hypothesis is generated by interpreting the metabolome data (b), and the cycle is completed after validating the hypotheses by further analyses (the arrow on the left side). The metabolome data are published in the databases (c) and utilized for preprocessing and data mining. The tools and databases introduced in the main text of this paper are shown in bold face.

(CE-) MS, are chosen according to a study's purpose [17–19]. Metabolomics technology, including instrumental analysis, detection and identification of metabolites, statistical interpretation, and generation of hypotheses with computational support, is used for a variety of studies, such as functional analysis of biological systems [20–22], biomarker discovery [23, 24], medical diagnostics [14, 25], quality assessment of foods [26, 27], evaluation of genetically modified crops [28, 29], and assessment of environmental pollution [30, 31].

A considerable number of software tools and databases have been developed for processing the complicated and multidimensional metabolome datasets generated by various types of MS-based instruments [32–35]. A typical workflow of metabolomic data analysis includes the following processes: (a) preprocessing of raw data for extraction of metabolite features, annotation of the metabolites, and finally generation of metabolome data; (b) mining and visualization of metabolome data for statistical interpretation of its nature and hypothesis generation; (c) storing and disseminating the data for further utilization and comparison (Figure 1). XCMS2 [36], MzMine2 [37], MathDAMP [38], MetAlign [39], and MET-IDEA [40] are typical tools for preprocessing including detection, alignment, and annotation of metabolite features. Some of these tools also provide statistical analysis functions for data interpretation. MassBank [41], METLIN [42], PRIMe [43], and HMDB [44] are available as references of mass spectra for metabolite annotation. The metabolite data are interpreted by means of the genome information from compound databases such as KNApSAcK [45], PubChem [46], and Chemspider (http://www.chemspider.com/) and by means of metabolic pathway databases including KEGG [47], BioCyc [48], and Reactome [49], which enable data visualization on pathway maps. The raw and processed data are stored publicly in databases such as PlantMetabolomics.org [50], GMD@CSB.DB [51], SetupX (currently not available), MetabolomeExpress [52], MetaboLights [53], and Metabolomics Workbench (http://www.metabolomicsworkbench.org/).

We report here a portal website named *KOMICS* (The Kazusa Metabolomics Portal, http://www.kazusa.or.jp/komics/), which hosts tools and databases that we developed for metabolomics. Although an increasing number of tools and databases have become available, two major issues remain to be resolved, that is, comprehensiveness of metabolites [54, 55] and data dissemination [53, 56, 57]. Our primary aim in developing data preprocessing tools is to help researchers with the manual annotation process that remains essential for nontarget metabolomics [54]. PowerGet for LC-high-resolution-MS and FragmentAlign for GC-MS are tools that enable curation of peak alignment results. SpiceHit is a high-throughput metabolite identification tool for CE-MS analysis using the selected ion monitoring (SIM) method. We have also developed data mining and visualization tools for the generation of working hypotheses (KaPPA-View and RnR). Real data is indispensable for comparative analysis and for the development and improvement of preprocessing tools [53, 58]. MassBase is one of the largest raw data repositories, and KomicMarket is a database of metabolic profiling data. We developed a metadata-specific database, Metabolonote, to promote data publication by researchers. These resources for a wide range of metabolome data processing are expected to contribute to improved production and utilization of metabolomic data.

2. Materials and Methods

The standalone tools for metabolome data production, PowerGet, FragmentAlign, and SpiceHit, were developed in Java (Oracle Corporation). The web-based tools and databases were developed and are run in Apache, PHP, Perl, MySQL, Java, and Tomcat on Linux servers. The KOMICS website was constructed using the content management system "Joomla!" running on a Linux server with Apache, PHP, and MySQL. The details of the development and license information are described in the individual introduction pages of KOMICS, in manuals, or in other relevant help resources. The tools and databases are freely available to academic users.

Details of the analytical methods for the evaluation of preprocessing tools are described in the Supplementary Material (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/194812).

3. Results and Discussion

The tools and databases we have developed and provided at the KOMICS web portal are classified into three categories according to the typical workflow of metabolomic data analysis, namely, (a) preprocessing tools, (b) data mining tools, and (c) databases for data dissemination (Figure 1). Here we describe several representative examples. All the currently available tools and databases are listed in Table 1. The number of records in each metabolomics-related database is shown in Table 2. The formats of input and output files and the availability of sample datasets are summarized in Table 3.

3.1. Data Preprocessing Tools

3.1.1. PowerGet. PowerGet is a standalone Java software package for detection, alignment, and annotation of metabolite features from data obtained using LC-high-resolution-MS (HRMS). Accurate mass values measured by HRMS, such as Fourier transform ion cyclotron resonance MS and Orbitrap MS (Thermo Fisher), allow users to predict the elemental composition of a metabolite. The intensity ratio of ^{13}C to ^{12}C isotopic ion peaks is useful for estimating the number of carbon atoms in a molecule. Estimation of ion adducts attached to the metabolites by coeluted ions is helpful for calculating elemental composition and for search of compound databases by mass values of nonionized molecules. The PowerFT module in PowerGet attaches these data automatically to all metabolite features in the LC-HRMS data. In the PowerMatch module, the metabolite features are aligned among the samples taking into account the similarity of MS/MS fragmentation patterns. A tool for refining the alignment results, MatchedIonsFinder [1], is also available via KOMICS.

To evaluate the accuracy of mass values of the peaks detected using PowerGet, the mass differences between a theoretical mass and a detected mass were compared to those of the peaks detected using the commercial software, Xcalibur (see Supplementary Method S1). PowerGet exhibited greater accuracy (0.579 ± 0.481 ppm (mean ± SD)) than Xcalibur (0.783 ± 0.563 ppm) in the evaluation of 143 standard compounds (Supplementary Table S1).

One of the unique functions of PowerGet is that the alignment results are manually editable: a user can promptly check metabolite's characteristics, such as mass chromatogram shape, existence of adjacent features, and MS/MS fragmentation patterns, by means of a graphical user interface (GUI), as shown in Figure 2. Alignment is essential for preparing matrices of samples to metabolite intensity data for further comparison and statistical analysis. Alignment is highly valuable when users need to annotate the metabolites, especially for unknown features. By comparing the features from several replicate samples, (1) the estimation error of the ion adducts is verified, (2) accuracy of mass measurement can be improved, and (3) reproducibly detected features are prioritized for further annotation. Therefore, alignment errors should be assessed and corrected during detailed annotation of unknown metabolites. PowerGet is utilized in preparing data for KomicMarket and Bio-MassBank (http://bio.massbank.jp/).

3.1.2. FragmentAlign. This is a standalone Java tool designed for GC-MS data analysis with functions for alignment and annotation of metabolite features. A GUI for editing the alignment results is also implemented in this software (Figure 3). The similarity of fragment ion patterns generated by electron ionization (EI) is taken into account in the alignment of metabolite features. The metabolite features can also be annotated based on EI fragment patterns, by comparing to patterns from standard compounds. The fragment pattern data of standard compounds can be imported and utilized when the data is written in the format defined by the National Institute of Standards and Technology (NIST), USA.

To evaluate the applicability of data matrices generated by FragmentAlign for further statistical analyses, a principal component analysis (PCA) was conducted using the GC-MS data obtained from 3 biological sources: *Arabidopsis* leaves, *Lotus japonicus* leaves, and *Arabidopsis* cultured cells. Five replicates of each source were mapped to similar positions, whereas the 3 sources were mapped separately from one another on the score plot (Supplementary Figure S1). High-correlation coefficients for peak intensity within the replicates were observed (Supplementary Figure S2). These results suggest that appropriate feature extraction and generation of data matrices can be performed successfully using FragmentAlign.

3.1.3. SpiceHit. The standalone Java tool SpiceHit is intended for high-throughput identification of metabolite features detected using the selected ion monitoring (SIM) method in CE-MS. The metabolite features are identified based on migration times relative to internal standard compounds and are compared to those of the standard compound library prepared in-house. The tool is designed for processing a large number of data files in a high-throughput manner; it requires checking and correcting the assignment errors manually.

To ascertain whether SpiceHit is applicable to practical data analysis, the accuracy of peak quantification was

Table 1: The tools and databases available at KOMICS (as of November 2013).

Name	Description	URL	Reference
	Standalone tools		
PowerGet	Metabolite detection, alignment, and annotation tool for LC-high-resolution-MS.	http://www.kazusa.or.jp/komics/software/PowerGet	
MatchedIonsFinder	Revising tool for metabolite alignment results from LC-MS analyses.	http://www.kazusa.or.jp/komics/software/MatchedIonsFinder	[1]
FragmentAlign	Metabolite alignment and annotation tool for GC-MS.	http://www.kazusa.or.jp/komics/software/FragmentAlign	
SpiceHit	High-throughput metabolite detection and annotation tool for SIM analysis in CE-MS.	http://www.kazusa.or.jp/komics/software/SpiceHit	
KAGIANA	Microsoft Excel-based tool for exploring the function of *Arabidopsis* genes.	http://webs2.kazusa.or.jp/kagiana/	[2]
SokanProject	A tool for calculating Pearson's correlation coefficients.	http://www.kazusa.or.jp/komics/software/SokanProject	
	Web tools		
MFSearcher	Web service for rapid prediction of elemental composition and database searching by accurate mass values.	http://webs2.kazusa.or.jp/mfsearcher/	[3]
DAGViz	Visualization tool for similarities of gene ontology annotations.	http://www.pgb.kazusa.or.jp/dagviz/	[4]
	Databases		
MassBase	Largest repository of metabolomics raw data.	http://webs2.kazusa.or.jp/massbase/	
KomicMarket	Sample-centric database for metabolomic profile data.	http://webs2.kazusa.or.jp/komicmarket/	
New KomicMarket temporary website	Developmental version of KomicMarket.	http://webs2.kazusa.or.jp/new_km_tmp/	
KaPPA-View4	Pathway database for visualizing metabolome and transcriptome data.	http://kpv.kazusa.or.jp/	[5]
Metabolonote	Metadata-specific Semantic MediaWiki-based database.	http://metabolonote.kazusa.or.jp/	
MS-MS Fragment Viewer	Database for MS/MS fragmentation data of 115 flavonoids.	http://webs2.kazusa.or.jp/msmsfragmentviewer/	
RnR	Database providing metabolite-to-gene relationships calculated from ~200 transgenic *Arabidopsis* cells.	http://webs2.kazusa.or.jp/kagiana/rnr	
CoP	Gene-to-gene coexpression database for 8 plant species calculated using the Confeito algorithm.	http://webs2.kazusa.or.jp/kagiana/cop0911	[6]
KATANA	Cross-search system for *Arabidopsis* genes.	http://www.kazusa.or.jp/katana/	[7]
ARTRA	Database of probe information of *Arabidopsis* DNA microarray data (developed by Takara Bio Inc.).	http://artra.kazusa.or.jp/artra/ARI3_101/	[8]
FuLoja	Database of *Lotus japonicus* full-length cDNA obtained in the NEDO project.	http://webs2.kazusa.or.jp/IntegrationDBRS/FuLoja/	
PMPj-Blast	Database of ESTs, cDNAs, and oligo DNA microarray probes for *Lotus japonicus* and some other plants.	http://webs2.kazusa.or.jp/IntegrationDBRS/pmpj-blast/	

Table 2: The number of records in metabolomics-related databases at KOMICS (as of November 2013).

Database name	Number	Description
MassBase	43959	Binary raw datasets
KomicMarket	85	Biological samples, including 251 instrumental analysis datasets
	215	Chemical samples, including 488 instrumental analysis datasets
New KomicMarket temporary website	16	Number of studies, including 166 analyzed datasets
Metabolonote	34	Number of studies, including metadata for 375 instrumental analysis datasets and 765 computational analysis datasets
MS-MS Fragment Viewer	115	Analyzed flavonoids
RnR	194	Metabolite features

Figure 2: The alignment-editing function of the PowerMatch module of PowerGet. (a) The Alignment Table shows the alignment results of the peaks detected in each sample. The intensity of peaks is summarized in another window (f). The details of the peak information (b), MS/MS fragments (c), appearance of peripheral peaks (d), and peak shape (e) are shown for the user-selected peaks. A misaligned peak (the blue colored cell in panel a) can be merged to an appropriate row using the Edit Alignment function (g), by immediately checking the detailed information for the peak (b–e).

compared to that acquired using the commercial software ChemStation (Agilent Technologies, Palo Alto, CA). In the detection of amino acids, the results from SpiceHit were strongly correlated with those from ChemStation, as well as with theoretical concentrations (Supplementary Table S4). Similar relative standard deviation (RSD) values for each amino acid in triplicate analyses were observed for SpiceHit and ChemStation (Supplementary Figure S3). Good linearity of peak areas common to SpiceHit and ChemStation was observed in the amino acid peaks detected in the biological samples (Supplementary Figure S4). These results suggest that the accuracy and the sensitivity of peak quantification by SpiceHit are similar to those of ChemStation and that SpiceHit is suitable for practical use.

3.1.4. MFSearcher. This is a web service that allows for rapid prediction of elemental composition from accurate mass values and for rapid searching of compound databases [3]. A GUI tool for MFSearcher queries is also provided as a module in PowerGet. PowerGet has a batch search function

Table 3: A summary of input and output file formats and availability of sample data for preprocessing tools. The precise formats are described in the instruction manuals for each tool.

Tool name	Input	Output	Availability of sample data
PowerGet	(1) PowerGet format (text file): MSGet tool is available for generating the text files from Xcalibur raw files (2) MassBase SMS format (text file) (3) mzXML file generated from the Xcalibur raw files using the ReAdW tool[*a]	(1) PowerGet format (text file): users can select the items and formats of the output file (2) TogoMD format (text file)	KOMICS website
FragmentAlign	*Deconvoluted peak data* (1) FragmentAlign format (text file, one of the NIST formats) (2) MassBase SMT format (text file) (3) GMD[*b] format (text file in NIST[*c] MSP format) (4).ELU file of AMDIS[*d] software *Chromatogram data for deconvolution* CSV file exported by Pegasus III (text file)	FragmentAlign format (text file)	KOMICS website A sample file of a standard compound library is included in the tutorial data
SpiceHit	*Electropherogram data* (1) CSV file (text file) (2) ChemStation .MS file (binary) *Standard compound library* Excel file (binary)	(1) Tab-delimited text file (2) Excel file	(1) KOMICS website (2) Included in the tool A sample of a standard compound library is included

[*a]ReAdW tool: available at http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW.
[*b]GMD: Golm Metabolome Database, http://gmd.mpimp-golm.mpg.de.
[*c]NIST: National Institute of Standards and Technology, http://www.nist.gov/.
[*d]AMDIS: available at http://chemdata.nist.gov/mass-spc/amdis/.

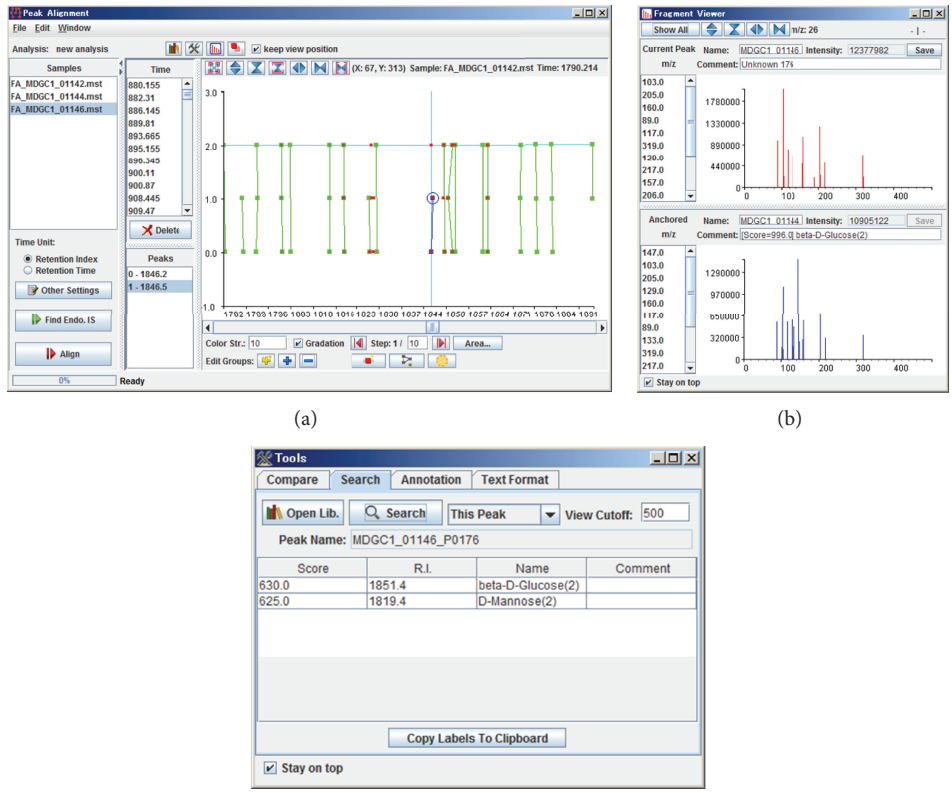


FIGURE 3: Screenshots of FragmentAlign. An alignment result from 3 samples is depicted (a). The electron ionization (EI) mode of a fragmentation pattern of the metabolite peak is presented in the Fragment Viewer panel (b). The metabolites are annotated by comparing the similarity of the fragment patterns to those obtained from standard compounds (c).

for querying thousands of detected metabolite features via MFSearcher.

Because MFSearcher is a RESTful web service, the query parameters for MFSearcher should be included in the description of a URL. Numerous sample queries are available as URL links on the MFSearcher website.

3.2. Data Mining Tools

3.2.1. KaPPA-View. This is a web-based tool for the visualization of metabolomic data on metabolic pathway maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [5]. The degree of change in metabolite abundance between several samples is expressed as hue of the compound symbols drawn on the KEGG pathway maps, based on the compound IDs assigned. Alterations in transcriptome data can be simultaneously depicted on the maps. This tool can be used for the integrated analysis of metabolomic, transcriptomic, and possibly proteomic data.

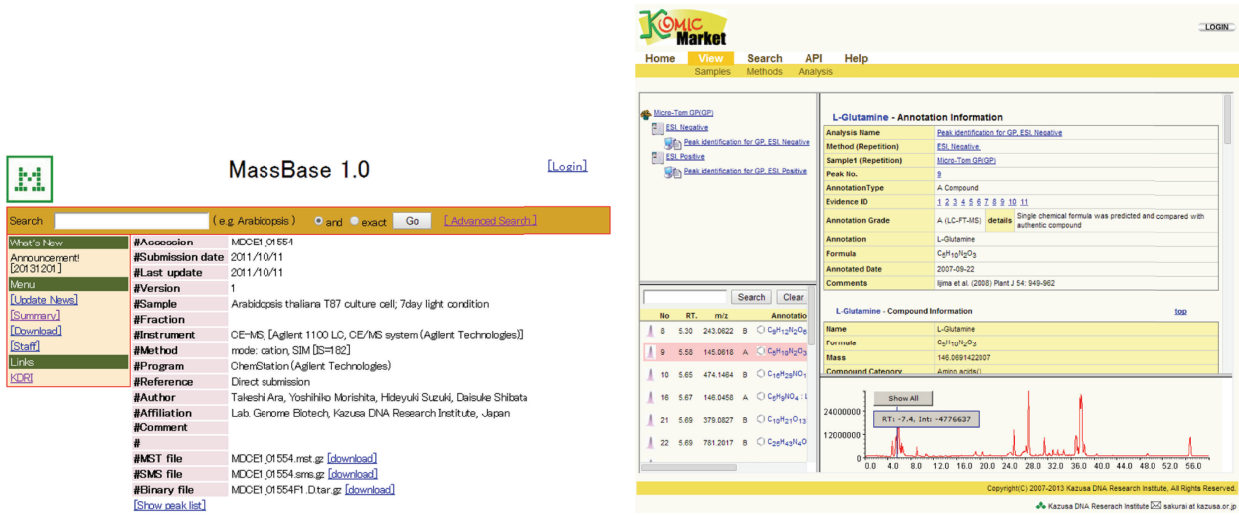
Sample data for testing the color representations on the pathway maps are available on the "Analysis" page of KaPPA-View. Users can select the items according to the directions presented on the page. Sample files for input data are available on the "Download" page.

3.2.2. RnR. This database contains data on the relationship between metabolites and genes; these data were generated via simultaneous measurement of the metabolome and transcriptome of approximately 200 transgenic cultured cell lines of the model plant *Arabidopsis*. The gene expression patterns and metabolite changes resulting from specific transgenes are compiled in the database. Users can search, for example, genes that can affect the abundance of the queried metabolites and vice versa. The database should contribute to knowledge discovery related to gene-to-metabolite regulatory networks in *Arabidopsis* cells.

To view an example dataset, a clickable pie chart of metabolites is presented on the main page of the RnR website. Clicking on a section of the pie chart will show a list of metabolites. After choosing a metabolite name, users will be able to view candidate genes that are related to the metabolite.

3.3. Databases for Data Dissemination

3.3.1. MassBase. The primary purpose of MassBase is the distribution of raw data generated by analytical instruments. Dissemination of raw data would enable the development and improvement of data analysis tools by bioinformatics researchers [53]. Binary raw data and near-raw text data exported from raw binary results are provided.



(a) (b)

FIGURE 4: Screenshots of the web interface of MassBase (a) and KomicMarket (b).

Users can search records by sample name, sample description, instrument type, and ionization mode on the "Advanced Search" page (Figure 4(a)). A summary of the records classified by species and instrument type is available on the "Summary" page.

3.3.2. KomicMarket. KomicMarket is a sample-centric database aimed at the distribution of metabolic profiles with and without metabolite annotations (Figure 4(b)). Previous results on the detection and annotation of metabolite features in certain samples can serve as good references for future metabolite annotations [56].

The records can be queried by keywords in the sample descriptions, including peak characteristics such as mass values, and in annotations of the peaks via the GUI on the KomicMarket website. The system provides application programming interfaces (APIs) for performing software-based querying of the data. Using the APIs, we employed the MFSearcher module of PowerGet to search metabolites in KomicMarket.

3.3.3. Metabolonote. This Semantic MediaWiki-based database is intended for managing metadata, which is the detailed information on experimental procedures accompanying the generation of data. Metabolonote is expected to accelerate publication of metabolomics data. The raw data obtained from the experiments or the processed data derived from them are not the target of Metabolonote, and the "actual data" are normally stored in other databases specifically built for a given purpose. Separation of the management of complicated metadata of metabolomics from the management of actual data makes it possible to share the same metadata among multiple actual databases such as raw data repositories, metabolic profile databases, reference libraries of MS/MS, and research papers. One-stop-shop management of complicated metadata of metabolomics eliminates the redundant management of metadata in several databases and reduces labor for data submitters. We defined a simple data format named *Togo metabolomics data format* (TogoMD) for easier description of metadata. Specifications of the TogoMD format are documented on the Metabolonote website (http://metabolonote.kazusa.or.jp/TogoMetabolomeDataFormat). Metabolonote provides application programming interfaces (APIs) for semantic searching of the records and retrieval of metadata. Because Metabolonote is a wiki system, it allows the submitters to attach free additional information about the metadata, such as images of the samples, video recordings of tricky analytical procedures, and links to a journal's website where the results are published. Therefore, metadata written by the submitter function as a hub of the web data resources related to the submitter's work. The increased presence of the submitters' published work on the web should increase the citations by others [58]. Therefore, the wiki system is expected to facilitate the dissemination of data to the public. Metabolonote is already linked to seven actual databases, including MassBase, KomicMarket, Bio-MassBank, and Riken PRIMe.

The metadata deposited in Metabolonote are listed on the "Public Pages." The registered metadata are semantically searchable by various items (and combinations thereof) on the "Metadata Search" page.

3.4. Practical Use. Here we present a workflow for a metabolomics study of *Jatropha curcas* L. [59, 60], a biofuel plant, to illustrate an example of the practical use of the KOMICS resources (Figure 5). LC-Orbitrap-MS analyses were conducted using 4 developmental stages of *J. curcas* fruit samples. The acquired data were primarily recorded as a binary raw file with commercial software (Xcalibur, Thermo Fisher). To analyze the data with PowerGet, the chromatogram data were exported to text files using the MSGet tool, which is also available on the KOMICS website. The raw files and extracted text files were published on MassBase. The text data were then processed using the PowerGet tool

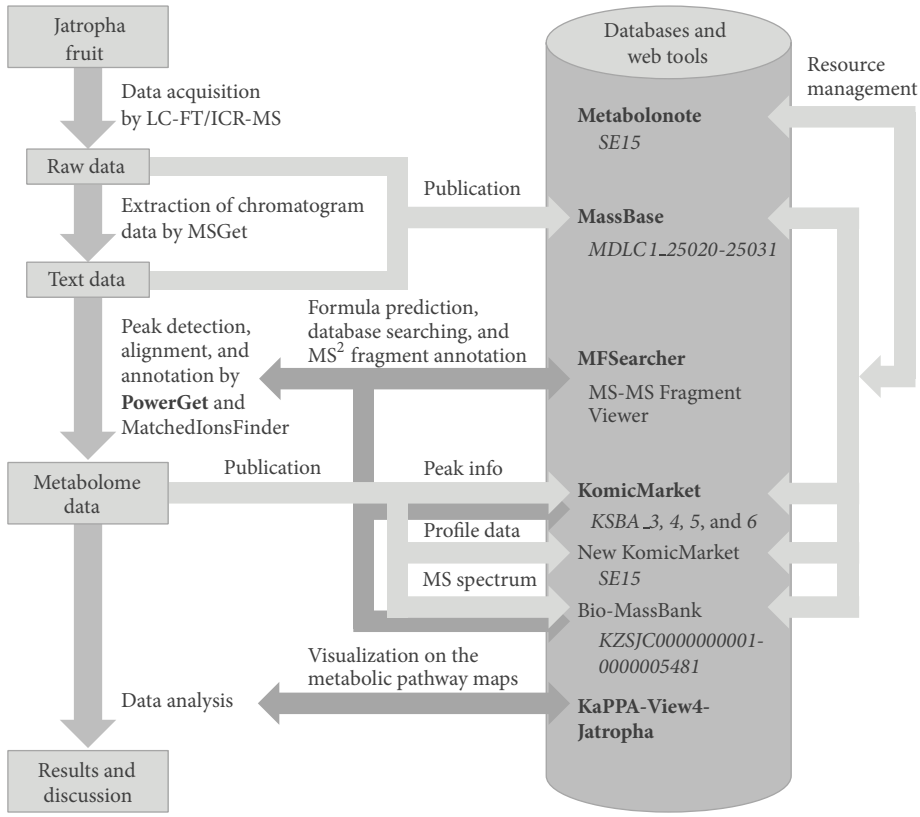


Figure 5: A schematic representation of the workflow for analysis of metabolomic changes in the developing fruit of *Jatropha curcas* L. The tools and databases introduced in the main text are shown in bold face. The accession IDs of the data in each database are shown in italics.

to generate the metabolomic data. MatchedIonsFinder was used to refine the alignment results. MFSearcher was used for high-throughput search of elemental composition and compound databases. MS-MS Fragment Viewer was used for interpreting MS/MS fragments in the metabolite annotations. The peak information, profile data (in the TogoMD format), and MS spectrum data were stored on KomicMarket (on the New KomicMarket temporary website) and on Bio-MassBank, respectively. These data were recursively used for metabolite annotations during the preprocessing step. Subsequently, the nature of the metabolomic data was interpreted by visualization on pathway maps using KaPPA-View4 and other statistical analyses. Consequently, a drastic change in metabolites during the maturation of *J. curcas* fruit was detected, and these data should contribute to further analysis of oil production by *J. curcas*. The record in Metabolonote (metadata ID: SE5) is a good guidepost for finding data resources deposited in various databases on the web.

4. Conclusions

We have developed various tools and databases for a wide range of processes related to metabolomic studies: preprocessing, data mining, and publication. To our knowledge, PowerGet and FragmentAlign are the first tools to allow users to curate alignment results via GUI. The unique concept of a metadata-specific database should accelerate data publication and dissemination. This infrastructure is expected to assist researchers to attain superior utilization of metabolomics' Big Data. Nonetheless, annotation of novel metabolites (the so-called unknown unknowns) remains a serious bottleneck in building comprehensive metabolomic datasets [16, 54]. Continuous efforts are needed to improve and automate annotation tasks. In addition, a systematic collection of annotation skills from experts will be necessary in the near future, as will the analysis and transfer of these skills to the public domain for education of fledgling annotators.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was partly supported by a grant from the New Energy and Industrial Technology Development Organization (NEDO, Japan) as part of the project "Development of Fundamental Technologies for Controlling the Material Production Process of Plants," a grant from the National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST) as part of the project titled "The Life-Science Database Integration Project," and a grant from the Kazusa DNA Research Institute.

References

[1] N. Yamamoto, T. Suzuki, T. Ara et al., "MatchedIonsFinder: a software tool for revising alignment matrices of spectrograms from liquid chromatography-mass spectrometry," *Plant Biotechnology*, vol. 29, no. 1, pp. 109–113, 2012.

[2] Y. Ogata, N. Sakurai, K. Aoki et al., "KAGIANA: an excel-based tool for retrieving summary information on *Arabidopsis* genes," *Plant and Cell Physiology*, vol. 50, no. 1, pp. 173–177, 2009.

[3] N. Sakurai, T. Ara, S. Kanaya et al., "An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values," *Bioinformatics*, vol. 29, no. 2, pp. 290–291, 2013.

[4] K. Yano, K. Aoki, H. Suzuki, and D. Shibata, "DAGViz: a directed acyclic graph browser that supports analysis of gene ontology annotation," *Plant Biotechnology*, vol. 26, no. 1, pp. 9–13, 2009.

[5] N. Sakurai, T. Ara, Y. Ogata et al., "KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D677–D684, 2011.

[6] Y. Ogata, H. Suzuki, N. Sakurai, and D. Shibata, "CoP: a database for characterizing co-expressed gene modules with biological information in plants," *Bioinformatics*, vol. 26, no. 9, pp. 1267–1268, 2010.

[7] K. Yano, T. Dansako, N. Sakurai, H. Suzuki, and D. Shibata, "KATANA: a web-based guide to public databases for *Arabidopsis* genomic information," *Plant Biotechnology*, vol. 22, no. 3, pp. 225–229, 2005.

[8] T. Ohba, K. Suzuki, T. Oura et al., "ARTRA: a new database of the *Arabidopsis* transcriptome and gene-specific sequences for microarray probes and RNAi triggers," *Plant Biotechnology*, vol. 26, no. 1, pp. 161–165, 2009.

[9] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, "Systematic functional analysis of the yeast genome," *Trends in Biotechnology*, vol. 16, no. 9, pp. 373–378, 1998.

[10] W. B. Dunn, "Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes," *Physical Biology*, vol. 5, no. 1, Article ID 011001, 2008.

[11] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, 2007.

[12] K. Hollywood, D. R. Brison, and R. Goodacre, "Metabolomics: current technologies and future trends," *Proteomics*, vol. 6, no. 17, pp. 4716–4723, 2006.

[13] J. Nielsen and S. Oliver, "The next wave in metabolome analysis," *Trends in Biotechnology*, vol. 23, no. 11, pp. 544–546, 2005.

[14] W. B. Dunn and T. Hankemeier, "Mass spectrometry and metabolomics: past, present and future," *Metabolomics*, vol. 9, no. 1, supplement, pp. 1–3, 2013.

[15] A. Zhang, H. Sun, P. Wang, Y. Han, and X. Wang, "Modern analytical techniques in metabolomics analysis," *Analyst*, vol. 137, no. 2, pp. 293–300, 2012.

[16] D. S. Wishart, "Computational strategies for metabolite identification in metabolomics," *Bioanalysis*, vol. 1, no. 9, pp. 1579–1596, 2009.

[17] Z. Lei, D. V. Huhman, and L. W. Sumner, "Mass spectrometry strategies in metabolomics," *The Journal of Biological Chemistry*, vol. 286, no. 29, pp. 25435–25442, 2011.

[18] R. D. Hall, "Plant metabolomics: from holistic hope, to hype, to hot topic," *New Phytologist*, vol. 169, no. 3, pp. 453–468, 2006.

[19] K. Saito and F. Matsuda, "Metabolomics for functional genomics, systems biology, and biotechnology," *Annual Review of Plant Biology*, vol. 61, pp. 463–489, 2010.

[20] S. H. Khoo and M. Al-Rubeai, "Metabolomics as a complementary tool in cell culture," *Biotechnology and Applied Biochemistry*, vol. 47, part 2, pp. 71–84, 2007.

[21] M. R. Mashego, K. Rumbold, M. de Mey, E. Vandamme, W. Soetaert, and J. J. Heijnen, "Microbial metabolomics: past, present and future methodologies," *Biotechnology Letters*, vol. 29, no. 1, pp. 1–16, 2007.

[22] J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, and M. Stitt, "Metabolite profiling in plant biology: platforms and destinations," *Genome Biology*, vol. 5, no. 6, article 109, 2004.

[23] A. Koulman, G. A. Lane, S. J. Harrison, and D. A. Volmer, "From differentiating metabolites to biomarkers," *Analytical and Bioanalytical Chemistry*, vol. 394, no. 3, pp. 663–670, 2009.

[24] J. Jansson, B. Willing, M. Lucio et al., "Metabolomics reveals metabolic biomarkers of Crohn's disease," *PLoS ONE*, vol. 4, no. 7, Article ID e6386, 2009.

[25] R. Madsen, T. Lundstedt, and J. Trygg, "Chemometrics in metabolomics—a review in human disease diagnosis," *Analytica Chimica Acta*, vol. 659, no. 1-2, pp. 23–33, 2010.

[26] M. A. Fitzgerald, S. R. McCouch, and R. D. Hall, "Not just a grain of rice: the quest for quality," *Trends in Plant Science*, vol. 14, no. 3, pp. 133–139, 2009.

[27] W. Pongsuwan, T. Bamba, K. Harada, T. Yonetani, A. Kobayashi, and E. Fukusaki, "High-throughput technique for comprehensive analysis of Japanese green tea quality assessment using ultra-performance liquid chromatography with time-of-flight mass spectrometry (UPLC/TOF MS)," *Journal of Agricultural and Food Chemistry*, vol. 56, no. 22, pp. 10705–10708, 2008.

[28] A. E. Ricroch, J. B. Bergé, and M. Kuntz, "Evaluation of genetically engineered crops using transcriptomic, proteomic, and metabolomic profiling techniques," *Plant Physiology*, vol. 155, no. 4, pp. 1752–1761, 2011.

[29] M. Kusano, H. Redestig, T. Hirai et al., "Covering chemical diversity of genetically-modified tomatoes using metabolomics for objective substantial equivalence assessment," *PLoS ONE*, vol. 6, no. 2, Article ID e16989, 2011.

[30] M. Krauss, H. Singer, and J. Hollender, "LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns," *Analytical and Bioanalytical Chemistry*, vol. 397, no. 3, pp. 943–951, 2010.

[31] C. Y. Lin, M. R. Viant, and R. S. Tjeerdema, "Metabolomics: methodologies and applications in the environmental sciences," *Journal of Pesticide Science*, vol. 31, no. 3, pp. 245–251, 2006.

[32] A. Fukushima and M. Kusano, "Recent progress in the development of metabolome databases for plant systems biology," *Frontiers in Plant Science*, vol. 4, article 73, 2013.

[33] T. Tohge and A. R. Fernie, "Web-based resources for mass-spectrometry-based metabolomics: a user's guide," *Phytochemistry*, vol. 70, no. 4, pp. 450–456, 2009.

[34] G. Blekherman, R. Laubenbacher, D. F. Cortes et al., "Bioinformatics tools for cancer metabolomics," *Metabolomics*, vol. 7, no. 3, pp. 329–343, 2011.

[35] M. Hur, A. A. Campbell, M. Almeida-de-Macedo et al., "A global approach to analysis and interpretation of metabolic data for plant natural product discovery," *Natural Product Reports*, vol. 30, no. 4, pp. 565–583, 2013.

[36] H. P. Benton, D. M. Wong, S. A. Trauger, and G. Siuzdak, "XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization," *Analytical Chemistry*, vol. 80, no. 16, pp. 6382–6389, 2008.

[37] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič, "MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data," *BMC Bioinformatics*, vol. 11, article 395, 2010.

[38] R. Baran, H. Kochi, N. Saito et al., "MathDAMP: a package for differential analysis of metabolite profiles," *BMC Bioinformatics*, vol. 7, article 530, 2006.

[39] A. Lommen, "Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing," *Analytical Chemistry*, vol. 81, no. 8, pp. 3079–3086, 2009.

[40] Z. Lei, H. Li, J. Chang, P. X. Zhao, and L. W. Sumner, "MET-IDEA version 2.06; improved efficiency and additional functions for mass spectrometry-based metabolomics data processing," *Metabolomics*, vol. 8, pp. 105–110, 2012.

[41] H. Horai, M. Arita, S. Kanaya et al., "MassBank: a public repository for sharing mass spectral data for life sciences," *Journal of Mass Spectrometry*, vol. 45, no. 7, pp. 703–714, 2010.

[42] C. A. Smith, G. O'Maille, E. J. Want et al., "METLIN: a metabolite mass spectral database," *Therapeutic Drug Monitoring*, vol. 27, no. 6, pp. 747–751, 2005.

[43] T. Sakurai, Y. Yamada, Y. Sawada et al., "PRIMe update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation," *Plant and Cell Physiology*, vol. 54, no. 2, article e5, 2013.

[44] D. S. Wishart, T. Jewison, A. C. Guo et al., "HMDB 3.0—the human metabolome database in 2013," *Nucleic Acids Research*, vol. 41, pp. D801–D807, 2013.

[45] F. M. Afendi, T. Okada, M. Yamazaki et al., "KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research," *Plant and Cell Physiology*, vol. 53, no. 2, article e1, 2012.

[46] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Research*, vol. 37, no. 2, pp. W623–W633, 2009.

[47] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, vol. 40, pp. D109–D114, 2012.

[48] R. Caspi, T. Altman, K. Dreher et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 38, no. 1, pp. D742–D753, 2009.

[49] D. Croft, G. O'Kelly, G. Wu et al., "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research*, vol. 39, no. 1, pp. D691–D697, 2011.

[50] P. Bais, S. M. Moon, K. He et al., "Plantmetabolomics.org: a web portal for plant metabolomics experiments," *Plant Physiology*, vol. 152, no. 4, pp. 1807–1816, 2010.

[51] J. Kopka, N. Schauer, S. Krueger et al., "GMD@CSB.DB: the Golm metabolome database," *Bioinformatics*, vol. 21, no. 8, pp. 1635–1638, 2005.

[52] A. J. Carroll, M. R. Badger, and A. H. Millar, "The MetaboloneExpress project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets," *BMC Bioinformatics*, vol. 11, article 376, 2010.

[53] K. Haug, R. M. Salek, P. Conesa et al., "MetaboLights-an open-access general-purpose repository for metabolomics studies and associated meta-data," *Nucleic Acids Research*, vol. 41, pp. D781–D786, 2013.

[54] B. P. Bowen and T. R. Northen, "Dealing with the unknown: metabolomics and metabolite atlases," *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 9, pp. 1471–1476, 2010.

[55] S. Neumann and S. Böcker, "Computational mass spectrometry for metabolomics: identification of metabolites and small molecules," *Analytical and Bioanalytical Chemistry*, vol. 398, no. 7-8, pp. 2779–2788, 2010.

[56] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: acquiring and understanding global metabolite data," *Trends in Biotechnology*, vol. 22, no. 5, pp. 245–252, 2004.

[57] T. Kind, M. Scholz, and O. Fiehn, "How large is the metabolome? A critical analysis of data exchange practices in chemistry," *PLoS ONE*, vol. 4, no. 5, Article ID e5440, 2009.

[58] H. A. Piwowar, R. S. Day, and D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate," *PLoS ONE*, vol. 2, no. 3, article e308, 2007.

[59] N. Sakurai, Y. Ogata, T. Ara et al., "Development of KaPPA-View4 for omics studies on Jatropha and a database system KaPPA-Loader for construction of local omics databases," *Plant Biotechnology*, vol. 29, no. 2, pp. 131–135, 2012.

[60] R. Sano, T. Ara, N. Akimoto et al., "Dynamic metabolic changes during fruit maturation in *Jatropha curcas* L," *Plant Biotechnology*, vol. 29, no. 2, pp. 175–178, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 831751, 15 pages
http://dx.doi.org/10.1155/2014/831751

Research Article

Supervised Clustering Based on DPClusO: Prediction of Plant-Disease Relations Using Jamu Formulas of KNApSAcK Database

Sony Hartono Wijaya,[1,2] Husnawati Husnawati,[3] Farit Mochamad Afendi,[4] Irmanida Batubara,[5] Latifah K. Darusman,[5] Md. Altaf-Ul-Amin,[1] Tetsuo Sato,[1] Naoaki Ono,[1] Tadao Sugiura,[1] and Shigehiko Kanaya[1]

[1] *Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan*
[2] *Department of Computer Science, Bogor Agricultural University, Kampus IPB Dramaga, Jl. Meranti, Bogor 16680, Indonesia*
[3] *Department of Biochemistry, Bogor Agricultural University, Kampus IPB Dramaga, Jl. Meranti, Bogor 16680, Indonesia*
[4] *Department of Statistics, Bogor Agricultural University, Kampus IPB Dramaga, Jl. Meranti, Bogor 16680, Indonesia*
[5] *Biopharmaca Research Center, Bogor Agricultural University, Kampus IPB Taman Kencana, Jl. Taman Kencana No. 3, Bogor 16151, Indonesia*

Correspondence should be addressed to Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 30 November 2013; Accepted 18 February 2014; Published 7 April 2014

Academic Editor: Samuel Kuria Kiboi

Indonesia has the largest medicinal plant species in the world and these plants are used as Jamu medicines. Jamu medicines are popular traditional medicines from Indonesia and we need to systemize the formulation of Jamu and develop basic scientific principles of Jamu to meet the requirement of Indonesian Healthcare System. We propose a new approach to predict the relation between plant and disease using network analysis and supervised clustering. At the preliminary step, we assigned 3138 Jamu formulas to 116 diseases of International Classification of Diseases (ver. 10) which belong to 18 classes of disease from National Center for Biotechnology Information. The correlation measures between Jamu pairs were determined based on their ingredient similarity. Networks are constructed and analyzed by selecting highly correlated Jamu pairs. Clusters were then generated by using the network clustering algorithm DPClusO. By using matching score of a cluster, the dominant disease and high frequency plant associated to the cluster are determined. The plant to disease relations predicted by our method were evaluated in the context of previously published results and were found to produce around 90% successful predictions.

1. Introduction

Big data biology, which is a discipline of data-intensive science, has emerged because of the rapid increasing of data in omics fields such as genomics, transcriptomics, proteomics, and metabolomics as well as in several other fields such as ethnomedicinal survey. The number of medicinal plants is estimated to be 40,000 to 70,000 around the world [1] and many countries utilize these plants as blended herbal medicines, for example, China (traditional Chinese medicine), Japan (Kampo medicine), India (Ayurveda, Siddha, and Unani), and Indonesia (Jamu). Nowadays, the use of traditional medicines is rapidly increasing [2, 3]. These medicines consist of ingredients made from plants, animals, minerals, or combination of them. The traditional medicines have been used for generations for treatments of diseases or maintaining health of people and the most popular form of traditional medicine is herbal medicine. Blended herbal medicines as well as single herb medicines include a large number of constituent substances which exert effects on human physiology through a variety of biological pathways. The KNApSAcK Family database systems can be used to comprehensively understand the medicinal usage of plants based upon traditional and modern knowledge [4, 5]. This

Table 1: List of diseases using International Classification of Diseases ver. 10 (class of disease IDs correspond to Table 2).

ID	Disease	Class of disease
1	Abdominal pain	3
2	Abdominal pain, diarrhea	3
3	Acne	16
4	Acne, skin problems (cosmetics)	16
5	Amenorrhoea, dysmenorrhea	6
6	Amenorrhoea, irregular menstruation	6
7	Anaemia	1
8	Appendicitis, urinary tract infection, tonsillitis	3
9	Arthralgia	11
10	Arthralgia, arthritis	11
11	Asthma	15
12	Benign prostatic hyperplasia (Bph)	10
13	Breast disorder	6
14	Bromhidrosis	16
15	Bronchitis	15
16	Cancer	2
17	Cancer pain	2
18	Cancer, inflammation	2
19	Colic abdomen, bloating (in infant)	3
20	Common cold	15
21	Common cold, dyspepsia, insect bites	15, 3, 16
22	Common cold, influenza	15
23	Cough	15
24	Degenerative disease	14
25	Dermatitis, urticaria, erythema	16
26	Diabetes	14
27	Diabetic gangrene	16
28	Diarrhea	3
29	Diarrhea, abdominal pain	3
30	Diseases of the eye	5
31	Disorders in pregnancy	6
32	Dysmenorrhea	6
33	Dysmenorrhea, irregular menstruation	6
34	Dysmenorrhea, menstrual syndrome	6
35	Dyspepsia	3
36	Dyspnoea	15
37	Dyspnoea, cough, orthopnoea	15
38	Fatigue	11
39	Fatigue, anaemia, loss appetite	1
40	Fatigue, lack of sexual function	6
41	Fatigue, low back pain	11
42	Fatigue, myalgia, arthralgia	11
43	Fatigue, osteoarthritis	11
44	Fertility problem	6, 10
45	Fever	0
46	Gastritis, gastric ulcer	3
47	Haemorrhoids	1
48	Headache	13
49	Heart diseases	8
50	Heartburn	3, 8
51	Hepatitis, other diseases of liver	3
52	Hypercholesterolaemia	14
53	Hypertension	8
54	Hypertension, diabetes	14
55	Hypertension, hypercholesterolaemia	14
56	Hyperuricemia	1
57	Immunodefficiency	9
58	Indigestion (K.30)	3
59	Indigestion, lose appetite	3
60	Infertility	6, 10
61	Irregular menstruation, menstruation syndrome	6
62	Kidney diseases	17
63	Lactation problems	6
64	Leukorrhoea (Vaginalis)	6
65	Leukorrhoea (Vaginalis), dysmenorrhoea	6
66	Lose appetite	3
67	Lose appetite, underweight	14
68	Low back pain, myalgia, arthralgia	11
69	Low back pain, myalgia, constipation	11
70	Low back pain, urinary tract infection	17
71	Lung diseases	15
72	Malaise and Fatigue	11
73	Malaise and Fatigue, Constipation	11
74	Malaise and Fatigue, Fertility Problems	10, 11
75	Malaise and Fatigue, Low Back Pain	11
76	Malaise and Fatigue, Sexual Dysfunction	11, 6, 10
77	Malaise and Fatigue, Skin Problems (Cosmetics)	16
78	Malaria, anaemia	1
79	Meno-metrorrhagia	6
80	Menopausal syndrome	6
81	Menopause/menstrual syndrome, leukorrhoea (vaginalis)	6
82	Menstrual syndrome	6
83	Menstrual syndrome, fatigue	6
84	Migraine	13
85	Mood disorder	18
86	Myalgia, arthralgia	11
87	Nausea/vomiting of pregnancy	6
88	Osteoarthritis	11
89	Osteoarthritis, fatigue	11

Table 1: Continued.

ID	Disease	Class of disease
90	Overweight, obesity	14
91	Paralysis	13
92	Post partum syndrome	6
93	Prevent from overweight	14
94	Respiratory infection due to smoking	15
95	Respiratory tract infection	15
96	Rheumatoid arthritis, gout	11
97	Secondary amenorrhea	6
98	Secondary amenorrhea, irregular menstruation	6
99	Sexual dysfunction, fatigue	6, 10
100	Skin diseases	16
101	Skin problems (cosmetics)	16
102	Sleeping and Mood Disorders	18
103	Sleeping disorders	18
104	Stomatitis	3
105	Stomatitis, gingivitis, tonsilitis	3
106	Stone in kidney (N20.0)	17
107	Stone in kidney (N20.0), urinary bladder stone (N21.0)	17
108	Tonsilitis	4
109	Tonsilofaringitis	4
110	Toothache	13
111	Typhoid, dyspepsia	3
112	Ulcer of anus and rectum	3
113	Underweight, lose appetite	3
114	Urinary tract infection (urethritis)	17
115	Vaginal discharges	6
116	Vaginal diseases	6

database has information about the selected herbal ingredients, that is, the formulas of Kampo and Jamu, omics information of plants and humans, and physiological activities in humans. Jamu is generally composed based on the experience of the users for decades or even hundreds of years. However, versatile scientific analyses are needed to support their efficacy and their safety. Attaining this objective is in accordance with the 2010 policy of the Ministry of Health of Indonesian Government about scientification of Jamu. Thus, it is required to systemize the formulations and develop basic scientific principles of Jamu to meet the requirement of Indonesian Healthcare System. Afendi et al. initiated and conducted scientific analysis of Jamu for finding the correlation between plants, Jamu, and their efficacy using statistical methods [6–8]. They used Biplot, partial least squares (PLS), and bootstrapping methods to summarize the data and also focused on prediction of Jamu formulations. These methods give a good understanding about relationship between plants, Jamu, and their efficacy. Among 465 plants used in 3138 Jamu, 190 plants were shown to be effective for at least one efficacy and these plants were considered to be the main ingredients of Jamu. The other 275 plants are considered to be supporting ingredients in Jamu because their efficacy has not been established yet.

Network biology can be defined as the study of the network representations of molecular interactions, both to analyze such networks and to use them as a tool to make biological predictions [9]. This study includes modelling, analysis, and visualizations, which holds important task in life science today [10]. Network analysis has been increasingly utilized in interpreting high throughput data on omics information, including transcriptional regulatory networks [11], coexpression networks [12], and protein-protein interactions [13]. We can easily describe relationship between entities in the network and also concentrate on part of the network consisting of important nodes or edges. These advantages can be adopted for analyzing medicinal usage of plants in Jamu and diseases. Network analysis provides information about groups of Jamu that are closely related to each other in terms of ingredient similarity and thus allows precise investigation to relate plants to diseases. On the other hand, multivariate statistical methods such as PLS can assign plants to efficacy by global linear modeling of the Jamu ingredients and efficacy. However, there is still lack of appropriate network based methods to learn how and why many plants are grouped in certain Jamu formula and the combination rule embedding numerous Jamu formulas.

It is needed to explore the relationship between Indonesian herbal plants used in Jamu medicines and the diseases which are treated using Jamu medicines. When effectiveness of a plant against a disease is firmly established, then further analysis about that plant can be proceeded to molecular level to pinpoint the drug targets. The present study developed a network based approach for prediction of plant-disease relations. We utilized the Jamu data from the KNApSAcK database. A Jamu network was constructed based on the similarity of their ingredients and then Jamu clusters were generated using the network clustering algorithm DPClusO [14, 15]. Plant-disease relations were then predicted by determining the dominant diseases and plants associated with selected Jamu clusters.

2. Methods

2.1. Concept of the Methodology. Jamu medicines consist of combination of medicinal plants and are used to treat versatile diseases. In this work we exploit the ingredient similarity between Jamu medicines to predict plant-disease relations. The concept of the proposed method is depicted in Figure 1. In step 1 a network is constructed where a node is a Jamu medicine and an edge represents high ingredient similarity between the corresponding Jamu pair. In Figure 1, the nodes of the same color indicate the Jamu medicines used for the same disease. The similarity is represented by Pearson correlation coefficient [16, 17]; that is,

$$\mathrm{corr}(X, Y) = \frac{\sum_{i=1}^{l} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{l} (x_i - \overline{x})^2 \sum_{i=1}^{l} (y_i - \overline{y})^2}}, \quad (1)$$

Table 2: Distribution of Jamu formulas according to 18 classes of disease (classes of diseases are determined by NCBI in ID1 to ID16 and by the present study in ID17 and ID18 represented by asterisks in Ref. columns).

ID	Class of disease (NCBI)	Ref.	Number of Jamu	Percentage
1	Blood and lymph diseases	NCBI	201	6.41
2	Cancers	NCBI	32	1.02
3	The digestive system	NCBI	457	14.56
4	Ear, nose, and throat	NCBI	2	0.06
5	Diseases of the eye	NCBI	1	0.03
6	Female-specific diseases	NCBI	382	12.17
7	Glands and hormones	NCBI	0	—
8	The heart and blood vessels	NCBI	57	1.82
9	Diseases of the immune system	NCBI	22	0.70
10	Male-specific diseases	NCBI	17	0.54
11	Muscle and bone	NCBI	649	20.68
12	Neonatal diseases	NCBI	0	—
13	The nervous system	NCBI	32	1.02
14	Nutritional and metabolic diseases	NCBI	576	18.36
15	Respiratory diseases	NCBI	313	9.97
16	Skin and connective tissue	NCBI	163	5.19
17	The urinary system	*	90	2.87
18	Mental and behavioral disorders	*	21	0.67
	The number of Jamu classified into multiple disease classes		119	3.79
	The number of Jamu unclassified		4	0.13
	Total Jamu formulas		3138	100.00

where x_i is the weight of plant-i in Jamu X, y_i is the weight of plant-i in Jamu Y, $\overline{x}$ is mean of Jamu X, and $\overline{y}$ is mean of Jamu Y. The higher similarity between Jamu pairs the higher the correlation value. In the present study, x_i and y_i are assigned as 1 or 0 in cases the ith plant is, respectively, included or not included in the formula. Under such condition, Pearson correlation corresponds to fourfold point correlation coefficient; that is,

$$\text{corr}(X, Y) = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}, \quad (2)$$

where a, b, c, and d represent the numbers of plants included in both X and Y, in only X, in only Y, and in neither X nor Y, respectively.

In step 2 the Jamu clusters are generated using network clustering algorithm DPClusO. DPClusO can generate clusters characterized by high density and identified by periphery; that is, the Jamu medicines belonging to a cluster are highly cohesive and separated by a natural boundary. Such clusters contain potential information about plant-disease relations.

In step 3 we assess disease-dominant clusters based on matching score represented by the following equation:

$$\text{matching score} = \frac{\text{number of Jamu belonging to the same disease}}{\text{total number of Jamu in the cluster}}. \quad (3)$$

Matching score of a cluster is the ratio of the highest number of Jamu associated with a single disease to the total number of Jamu in the cluster. We assign a disease to a cluster for which the matching score is greater than a threshold value. In step 4, we determine the frequency of plants associated with a cluster if and only if a disease is assigned to it in the previous step. The highest frequency plant associated to a cluster is considered to be related to the disease assigned to that cluster. True positive rates (TPR) or sensitivity was used to evaluate resulting plants. TPR is the proportion of the true positive predictions out of all the true predictions, defined by the following formula [18]:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (4)$$

where true positive (TP) is the number of correctly classified and false negative (FN) is the number of incorrectly rejected entities. We refer to the proposed method as supervised clustering because after generation of the clusters we narrow down the candidate clusters for further analysis based on supervised learning and thus improve the accuracy of prediction of the proposed method.

3. Result and Discussion

3.1. Construction and Comparison of Jamu and Random Networks. We used the same number of Jamu formulas from previous research [6], 3138 Jamu formulas, and the set union

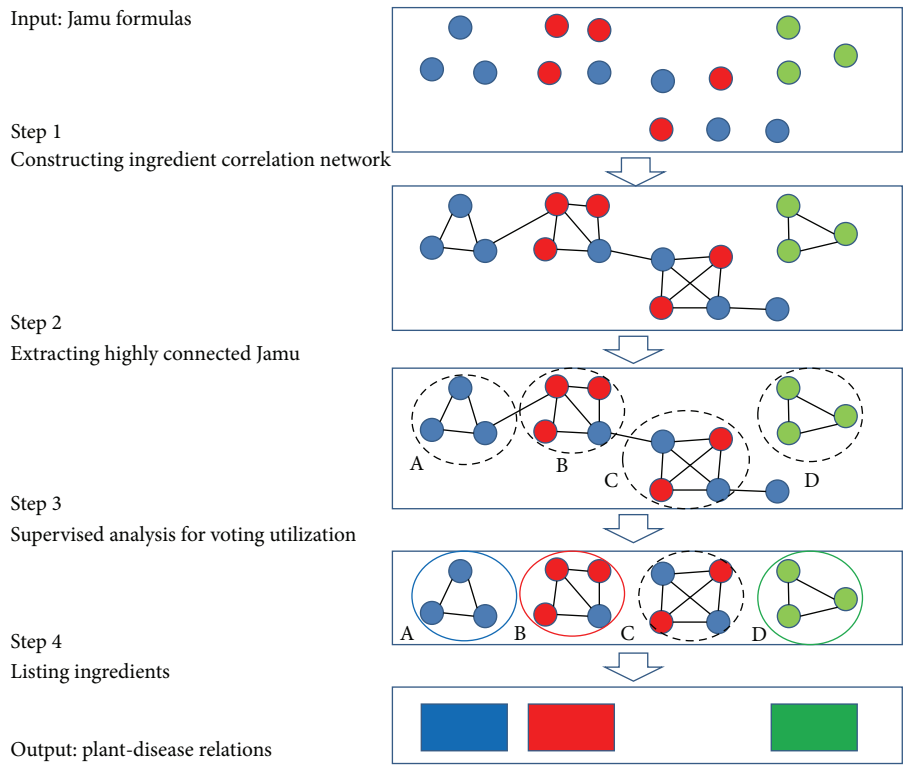


Figure 1: Concept of the methodology: network construction based on ingredient similarity between individual Jamu medicines, network clustering, and classification of medicinal plants to dominant disease.

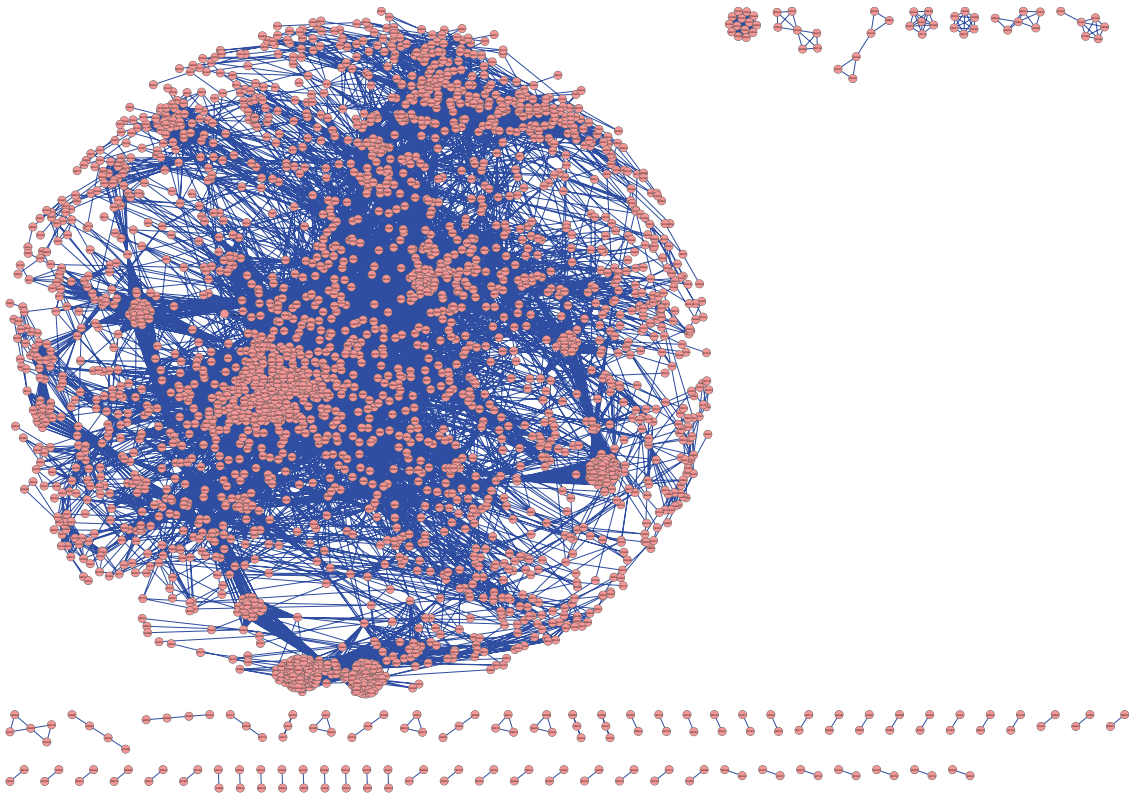


Figure 2: The network consisting of 0.7% Jamu pairs (correlation value above or equal to 0.596).

Table 3: Statistics of three datasets.

	Parameters	0.7%	0.5%	0.3%
Network statistics	Total pairs	34,454	24,610	14,766
	Minimum correlation	0.596	0.665	0.718
	Number of Jamu formulas	2,779	2,496	2,085
	Average degree	24.8	19.7	14.2
	(Random network: ER)	(24.8 ± 0.0)	(19.7 ± 0.0)	(14.2 ± 0.0)
	(Random network: BA)	(24.7 ± 0.1)	(19.7 ± 0.1)	(14.1 ± 0.1)
	(Random network: CNN)	(24.7 ± 0.4)	(19.7 ± 0.4)	(14.0 ± 0.4)
	Clustering coefficient	0.521	0.520	0.540
	(Random network: ER)	(0.009 ± 0.000)	(0.008 ± 0.000)	(0.007 ± 0.000)
	(Random network: BA)	(0.030 ± 0.001)	(0.028 ± 0.001)	(0.026 ± 0.001)
	(Random network: CNN)	(0.246 ± 0.008)	(0.239 ± 0.008)	(0.233 ± 0.010)
	Number of connected components	69	119	254
	(Random networks: ER, BA, CNN)	(1)	(1)	(1)
	Network diameter	15	17	20
	(Random network: ER)	(4.0 ± 0.0)	(4.0 ± 0.0)	(5.0 ± 0.0)
	(Random network: BA)	(10.8 ± 0.8)	(11.2 ± 1.5)	(10.8 ± 0.9)
	(Random network: CNN)	(14.6 ± 1.9)	(14.1 ± 1.4)	(14.7 ± 1.3)
	Network density	0.008	0.008	0.007
	(Random network: ER)	(0.009 ± 0.000)	(0.008 ± 0.000)	(0.007 ± 0.000)
	(Random network: BA)	(0.009 ± 0.000)	(0.008 ± 0.000)	(0.007 ± 0.000)
	(Random network: CNN)	(0.009 ± 0.000)	(0.008 ± 0.000)	(0.007 ± 0.000)
DPClusO	Total number of clusters	1,746	1,411	938
	Number of clusters with more than 2 Jamu	1,296	873	453
	(%)	(74.2)	(61.9)	(48.3)
	Number of Jamu formulas in the biggest cluster	118	104	89

of all formulas consists of 465 plants. We assigned 3138 Jamu formulas to 116 diseases of International Classification of Diseases (ICD) version 10 from World Health Organization (WHO, Table 1) [19]. Those 116 diseases are mapped to 18 classes of disease, which contains 16 classes of disease from National Center for Biotechnology Information (NCBI) [20] and 2 additional classes. Table 2 shows distribution of 3138 Jamu into 18 classes of disease. According to this classification, most Jamu formulas are useful for relieving muscle and bone, nutritional and metabolic diseases, and the digestive system. Furthermore, there is no Jamu formula classified into glands and hormones and neonatal disease classes. We excluded 4 Jamu formulas which are used to treat fever in the evaluation process because this symptom is very general and almost appeared in all disease classes. Jamu-plant-disease relations can be represented using 2 matrices: first matrix is Jamu-plant relation with dimension 3138×465 and the second matrix is Jamu-disease relation with dimension 3138×18.

After completion of data acquisition process, we calculated the similarity between Jamu pairs using correlation measure. The similarity measures between Jamu pairs were determined based on their ingredients. Corresponding to K (3138 in present case) Jamu formulas, there can be maximum $(K \times (K - 1)/2) = (3138 \times (3137/2)) = 4{,}921{,}953$ Jamu pairs. We sorted the Jamu pairs based on correlation value using descending order and selected top-n (0.7%, 0.5%, and 0.3%) pairs of Jamu formula to create 3 sets of Jamu pairs. The number of Jamu pairs for 0.7%, 0.5%, and 0.3% datasets is 34,454 pairs, 24,610 pairs, and 14,766 pairs and the corresponding minimum correlation values are 0.596, 0.665, and 0.718, respectively. The three datasets of Jamu pairs can be regarded as three undirected networks (step 1 in Figure 1) consisting of 2779, 2496, and 2085 Jamu formulas, respectively (Table 3). Figure 2 shows visualization of 0.7% Jamu networks using Cytoscape Spring Embedded layout. We verified that the degree distributions of the Jamu networks are somehow close to those of scale-free networks, that is, roughly are of power law type. However, in the high-degree region the power law structure is broken (Figure 3). Nearly accurate relation of power laws between medicinal herbs and the number of formulas utilizing them was observed in Jamu system but not in Kampo (Japanese crude drug system) [4]. The difference of formulas between Jamu and Kampo can be explained by herb selection by medicinal researchers based on the optimization process of selection [4]. Thus, the broken structure of power law corresponding to Jamu networks is associated with the fact that selection of Jamu pairs based on ingredient correlation leads to nonrandom selection. We also constructed random networks according

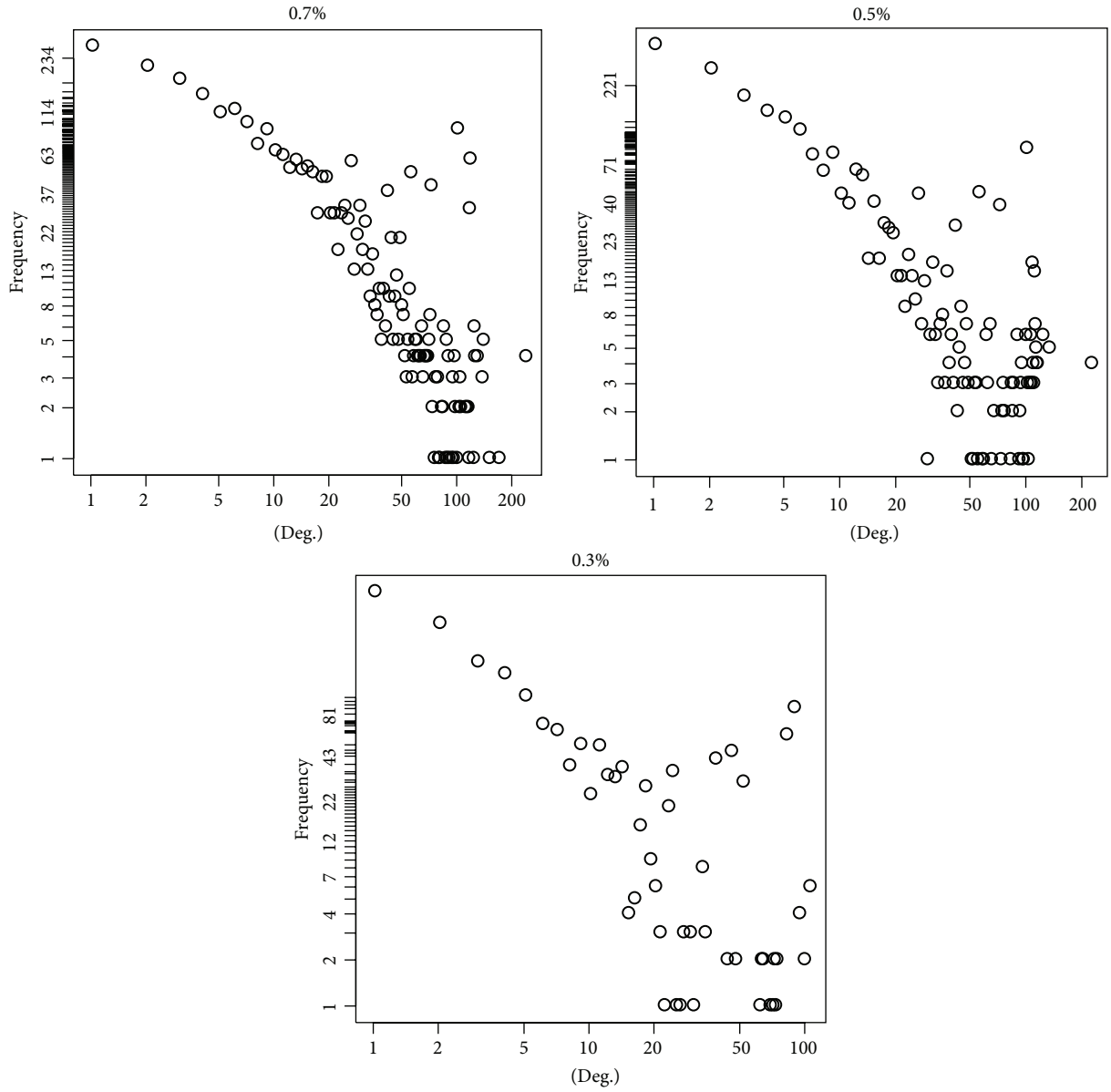


Figure 3: Degree distributions of three Jamu networks roughly follow power law. The x-axis corresponds to the log of degree of a node in the Jamu network and the y-axis corresponds to the log of the number of Jamu.

to Erdős-Rényi (ER) model [21], Barabási-Albert (BA) model [22], and Vazquez's Connecting Nearest Neighbor (CNN) model [23] of the same size corresponding to each of the real Jamu network. We used Cytoscape Network Analyzer plugin [24] and R software for analyzing the characteristics of both the Jamu and the random networks.

We determined five statistical indexes, that is, average degree, clustering coefficient, number of connected component, network diameter, and network density of each Jamu network and also of each random network. The clustering coefficient C_n of a node n is defined as $C_n = 2e_n/(k_n(k_n - 1))$, where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n. The network diameter is the largest distance between any two nodes. If a network is disconnected, its diameter is the maximum of all diameters of its connected components. A network's density is the ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes (which is $n(n-1)/2$, where n is the number of vertices, for an undirected graph). The average number of neighbors and the network density are the same for the real and random networks of the same size as it is shown in Table 3. In case of 0.7% and 0.5% real networks, the clustering coefficient is roughly the same and in case of 0.3% the clustering coefficient is somewhat larger. The number of connected components and the diameter of the Jamu networks gradually decrease as the network grows bigger by addition of more nodes and edges.

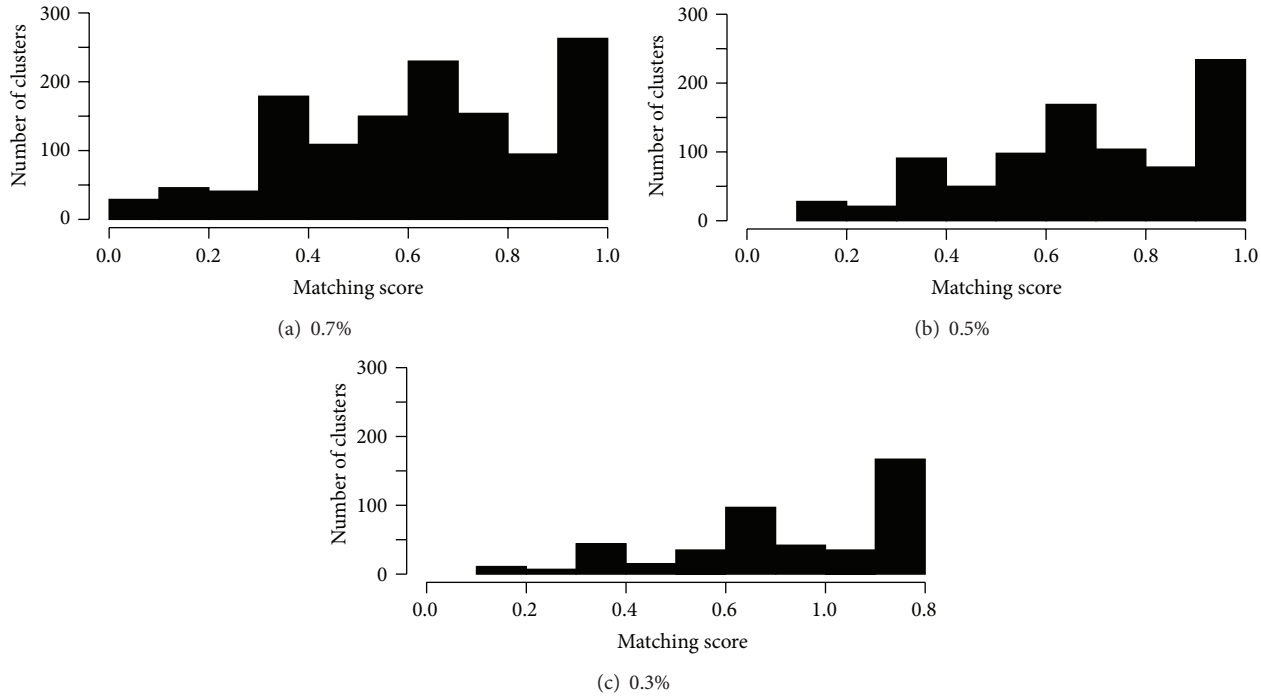


Figure 4: Distribution of clusters based on matching score.

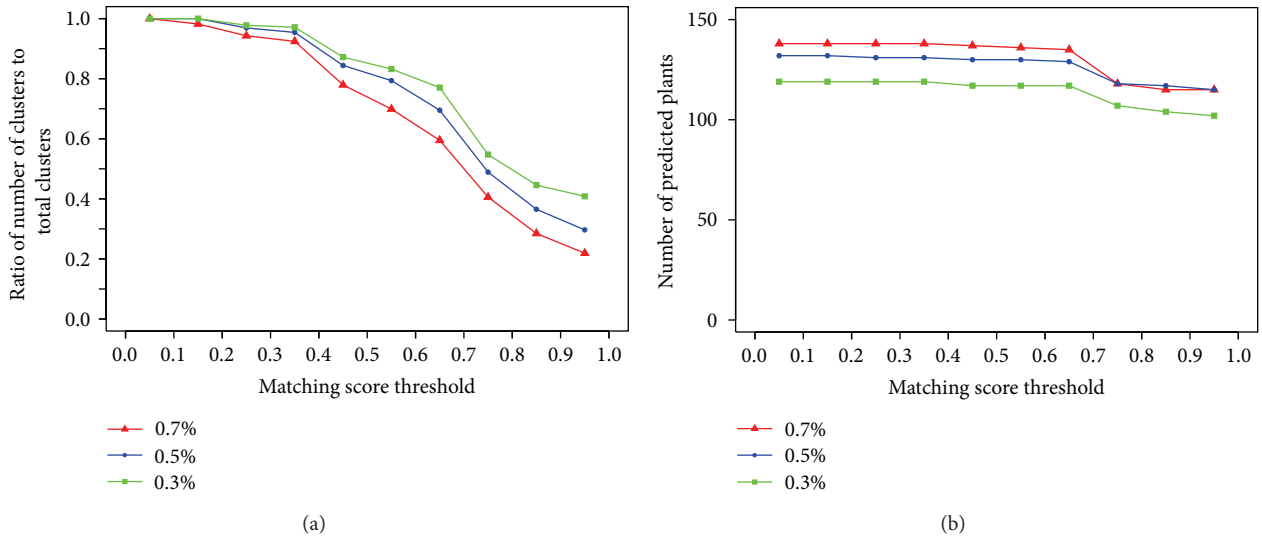


Figure 5: (a) Success rate and (b) number of predicted plants with respect to matching score thresholds.

Very different values corresponding to clustering coefficient, connected component, and network diameter imply that the Jamu networks are quite different from all 3 types of random networks. The differences between Jamu networks and ER random networks are the largest. Random networks constructed based on other two models are also substantially different from Jamu networks. Based on the fact that the random networks constructed based on all three types of models are different from the Jamu networks, it can be concluded that structure of Jamu networks is reasonably biased and thus might contain certain information about plant-disease relations. Specially, much higher value corresponding to clustering coefficient indicates that there are clusters in the networks worthy to be investigated. To extract clusters from the Jamu networks (step 2 in Figure 1) we applied DPClusO network clustering algorithm [14] to generate overlapping clusters based on density and periphery tracking.

3.2. Supervised Clustering Based on DPClusO. DPClusO is a general-purpose clustering algorithm and useful for finding overlapping cohesive groups in an undirected simple graph

Table 4: List of plants assigned to each disease.

Number	Plants name	Hit-miss status	
	A. Disease: blood and lymph diseases		
1	*Tamarindus indica*	*Hit*	*
2	*Allium sativum*	*Hit*	*
3	*Tinospora tuberculata*	*Hit*	*
4	*Piper retrofractum*	*Hit*	
5	*Syzygium aromaticum*	*Hit*	*
6	*Bupleurum falcatum*	*Hit*	
7	*Graptophyllum pictum*	*Hit*	
8	*Plantago major*	*Hit*	
9	*Zingiber officinale*	*Hit*	*
10	*Cinnamomum burmannii*	*Hit*	*
11	*Soya max*	*Miss*	*
12	*Kaempferia galanga*	*Hit*	
13	*Curcuma longa*	*Hit*	*
14	*Piper nigrum*	*Hit*	
15	*Zingiber aromaticum*	*Hit*	*
16	*Phyllanthus urinaria*	*Hit*	*
17	*Oryza sativa*	*Hit*	
18	*Myristica fragrans*	*Hit*	*
19	*Alstonia scholaris*	*Hit*	*
20	*Syzygium polyanthum*	*Miss*	
21	*Andrographis paniculata*	*Hit*	*
22	*Sida rhombifolia*	*Miss*	
23	*Cyperus rotundus*	*Hit*	
24	*Sonchus arvensis*	*Miss*	
25	*Curcuma aeruginosa*	*Hit*	*
26	*Curcuma xanthorrhiza*	*Hit*	
	B. Disease: cancers		
1	*Catharanthus roseus*	*Hit*	
	C. Disease: the digestive system		
1	*Foeniculum vulgare*	*Hit*	
2	*Glycyrrhiza uralensis*	*Hit*	*
3	*Imperata cylindrica*	*Hit*	
4	*Zingiber purpureum*	*Hit*	*
5	*Physalis peruviana*	*Hit*	
6	*Punica granatum*	*Hit*	*
7	*Echinacea purpurea*	*Hit*	
8	*Zingiber officinale*	*Hit*	*
9	*Psidium guajava*	*Hit*	
10	*Baeckea frutescens*	*Hit*	*
11	*Amomum compactum*	*Hit*	
12	*Cinnamomum burmannii*	*Hit*	*
13	*Melaleuca leucadendra*	*Hit*	
14	*Caesalpinia sappan*	*Hit*	*
15	*Parkia roxburghii*	*Hit*	
16	*Rheum tanguticum*	*Hit*	
17	*Kaempferia galanga*	*Hit*	
18	*Coriandrum sativum*	*Hit*	
19	*Curcuma longa*	*Hit*	
20	*Zingiber aromaticum*	*Hit*	
21	*Phyllanthus urinaria*	*Hit*	
22	*Myristica fragrans*	*Hit*	
23	*Hydrocotyle asiatica*	*Hit*	*
24	*Carica papaya*	*Hit*	
25	*Mentha arvensis*	*Hit*	
26	*Lepiniopsis ternatensis*	*Hit*	
27	*Helicteres isora*	*Hit*	
28	*Andrographis paniculata*	*Hit*	
29	*Symplocos odoratissima*	*Hit*	
30	*Schisandra chinensis*	*Hit*	
31	*Blumea balsamifera*	*Hit*	
32	*Silybum marianum*	*Hit*	*
33	*Cinnamomum sintoc*	*Hit*	
34	*Elephantopus scaber*	*Hit*	
35	*Curcuma aeruginosa*	*Hit*	
36	*Kaempferia pandurata*	*Hit*	
37	*Curcuma xanthorrhiza*	*Hit*	
38	*Curcuma mangga*	*Hit*	*
39	*Curcuma zedoaria*	*Hit*	
40	*Daucus carota*	*Hit*	*
41	*Matricaria chamomilla*	*Hit*	*
42	*Cymbopogon nardus*	*Hit*	*
	D. Disease: female-specific diseases		
1	*Foeniculum vulgare*	*Hit*	
2	*Imperata cylindrica*	*Hit*	
3	*Tamarindus indica*	*Hit*	
4	*Pluchea indica*	*Hit*	*
5	*Piper retrofractum*	*Hit*	
6	*Punica granatum*	*Hit*	
7	*Uncaria rhynchophylla*	*Hit*	
8	*Zingiber officinale*	*Hit*	
9	*Guazuma ulmifolia*	*Hit*	*
10	*Nigella sativa*	*Hit*	
11	*Terminalia bellirica*	*Hit*	
12	*Baeckea frutescens*	*Hit*	
13	*Phaseolus radiatus*	*Hit*	
14	*Amomum compactum*	*Hit*	*
15	*Sauropus androgynus*	*Hit*	
16	*Usnea misaminensis*	*Hit*	
17	*Cinnamomum burmannii*	*Hit*	
18	*Melaleuca leucadendra*	*Hit*	
19	*Parameria laevigata*	*Hit*	
20	*Parkia roxburghii*	*Hit*	
21	*Piper cubeba*	*Hit*	
22	*Kaempferia galanga*	*Hit*	

Table 4: Continued.

Number	Plants name	Hit-miss status	
23	*Coriandrum sativum*	*Hit*	
24	*Kaempferia angustifolia*	*Hit*	
25	*Curcuma longa*	*Hit*	
26	*Zingiber aromaticum*	*Hit*	
27	*Languas galanga*	*Hit*	
28	*Galla lusitania*	*Hit*	
29	*Quercus lusitanica*	*Hit*	
30	*Hydrocotyle asiatica*	*Hit*	
31	*Areca catechu*	*Hit*	
32	*Lepiniopsis ternatensis*	*Hit*	
33	*Helicteres isora*	*Hit*	*
34	*Piper betle*	*Hit*	
35	*Elephantopus scaber*	*Hit*	*
36	*Kaempferia pandurata*	*Hit*	
37	*Curcuma xanthorrhiza*	*Hit*	
38	*Sesbania grandiflora*	*Hit*	
	E. Disease: the heart and blood vessels		
1	*Allium sativum*	*Hit*	
2	*Curcuma longa*	*Hit*	*
3	*Morinda citrifolia*	*Hit*	*
4	*Homalomena occulta*	*Hit*	*
5	*Hydrocotyle asiatica*	*Hit*	
6	*Alstonia scholaris*	*Hit*	*
7	*Syzygium polyanthum*	*Miss*	*
8	*Andrographis paniculata*	*Hit*	*
9	*Apium graveolens*	*Miss*	
10	*Imperata cylindrica*	*Hit*	
	F. Disease: male-specific diseases		
1	*Cucurbita pepo*	*Miss*	
2	*Serenoa repens*	*Miss*	
3	*Baeckea frutescens*	*Hit*	
4	*Phaseolus radiatus*	*Hit*	
5	*Curcuma longa*	*Hit*	
6	*Elephantopus scaber*	*Hit*	
	G. Disease: muscle and bone		
1	*Foeniculum vulgare*	*Hit*	
2	*Clausena anisum-olens*	*Hit*	*
3	*Zingiber purpureum*	*Hit*	
4	*Allium sativum*	*Hit*	
5	*Strychnos ligustrina*	*Hit*	
6	*Tinospora tuberculata*	*Hit*	*
7	*Piper retrofractum*	*Hit*	
8	*Syzygium aromaticum*	*Hit*	
9	*Cola nitida*	*Hit*	*
10	*Ginkgo biloba*	*Hit*	*
11	*Panax ginseng*	*Hit*	
12	*Equisetum debile*	*Hit*	*
13	*Zingiber officinale*	*Hit*	

Table 4: Continued.

Number	Plants name	Hit-miss status	
14	*Ganoderma lucidum*	*Hit*	
15	*Nigella sativa*	*Hit*	
16	*Terminalia bellirica*	*Hit*	*
17	*Baeckea frutescens*	*Hit*	*
18	*Amomum compactum*	*Hit*	
19	*Cinnamomum burmannii*	*Hit*	
20	*Melaleuca leucadendra*	*Hit*	
21	*Parameria laevigata*	*Hit*	*
22	*Psophocarpus tetragonolobus*	*Hit*	*
23	*Parkia roxburghii*	*Hit*	
24	*Piper cubeba*	*Hit*	*
25	*Kaempferia galanga*	*Hit*	
26	*Coriandrum sativum*	*Hit*	
27	*Cola acuminata*	*Hit*	
28	*Coffea arabica*	*Hit*	
29	*Orthosiphon stamineus*	*Hit*	
30	*Curcuma longa*	*Hit*	
31	*Piper nigrum*	*Hit*	
32	*Alpinia galanga*	*Hit*	
33	*Vitex trifolia*	*Hit*	
34	*Zingiber amaricans*	*Hit*	*
35	*Zingiber zerumbet*	*Hit*	
36	*Zingiber aromaticum*	*Hit*	
37	*Languas galanga*	*Hit*	
38	*Massoia aromatica*	*Hit*	
39	*Morinda citrifolia*	*Hit*	
40	*Carum copticum*	*Hit*	*
41	*Panax pseudoginseng*	*Hit*	*
42	*Oryza sativa*	*Hit*	
43	*Myristica fragrans*	*Hit*	
44	*Pandanus amaryllifolius*	*Hit*	
45	*Eurycoma longifolia*	*Hit*	
46	*Hydrocotyle asiatica*	*Hit*	
47	*Areca catechu*	*Hit*	*
48	*Mentha arvensis*	*Hit*	*
49	*Lepiniopsis ternatensis*	*Hit*	
50	*Pimpinella pruatjan*	*Hit*	
51	*Andrographis paniculata*	*Hit*	
52	*Blumea balsamifera*	*Hit*	
53	*Cymbopogon nardus*	*Hit*	
54	*Sida rhombifolia*	*Hit*	
55	*Cinnamomum sintoc*	*Hit*	
56	*Piper betle*	*Hit*	*
57	*Talinum paniculatum*	*Hit*	
58	*Elephantopus scaber*	*Hit*	
59	*Cyperus rotundus*	*Hit*	
60	*Curcuma aeruginosa*	*Hit*	
61	*Kaempferia pandurata*	*Hit*	*

Table 4: Continued.

Number	Plants name	Hit-miss status	
62	*Curcuma xanthorrhiza*	*Hit*	
63	*Tribulus terrestris*	*Hit*	
64	*Corydalis yanhusuo*	*Hit*	
65	*Pausinystalia yohimbe*	*Hit*	
H. Disease: nutritional and metabolic diseases			
1	*Foeniculum vulgare*	*Hit*	
2	*Glycyrrhiza uralensis*	*Hit*	
3	*Zingiber purpureum*	*Hit*	
4	*Allium sativum*	*Hit*	
5	*Tinospora tuberculata*	*Hit*	
6	*Pandanus conoideus*	*Hit*	
7	*Syzygium aromaticum*	*Hit*	
8	*Punica granatum*	*Hit*	
9	*Zingiber officinale*	*Hit*	
10	*Guazuma ulmifolia*	*Hit*	
11	*Nigella sativa*	*Hit*	
12	*Amomum compactum*	*Hit*	*
13	*Cinnamomum burmannii*	*Hit*	
14	*Parameria laevigata*	*Hit*	
15	*Caesalpinia sappan*	*Hit*	
16	*Soya max*	*Hit*	*
17	*Cocos nucifera*	*Hit*	
18	*Rheum tanguticum*	*Hit*	
19	*Piper cubeba*	*Hit*	*
20	*Murraya paniculata*	*Hit*	
21	*Kaempferia galanga*	*Hit*	*
22	*Coffea arabica*	*Hit*	*
23	*Orthosiphon stamineus*	*Hit*	
24	*Curcuma longa*	*Hit*	
25	*Piper nigrum*	*Hit*	*
26	*Zingiber aromaticum*	*Hit*	
27	*Aloe vera*	*Hit*	
28	*Phaleria papuana*	*Hit*	
29	*Galla lusitania*	*Hit*	
30	*Quercus lusitanica*	*Hit*	
31	*Morinda citrifolia*	*Hit*	
32	*Myristica fragrans*	*Hit*	*
33	*Momordica charantia*	*Hit*	
34	*Areca catechu*	*Hit*	
35	*Lepiniopsis ternatensis*	*Hit*	
36	*Alstonia scholaris*	*Hit*	
37	*Hibiscus sabdariffa*	*Hit*	
38	*Laminaria japonica*	*Hit*	
39	*Syzygium polyanthum*	*Hit*	
40	*Andrographis paniculata*	*Hit*	
41	*Sindora sumatrana*	*Hit*	*
42	*Cassia angustifolia*	*Hit*	
43	*Woodfordia floribunda*	*Hit*	

Table 4: Continued.

Number	Plants name	Hit-miss status	
44	*Piper betle*	*Hit*	
45	*Spirulina*	*Hit*	
46	*Stevia rebaudiana*	*Hit*	
47	*Theae sinensis*	*Hit*	
48	*Sonchus arvensis*	*Hit*	
49	*Curcuma heyneana*	*Hit*	
50	*Curcuma aeruginosa*	*Hit*	
51	*Kaempferia pandurata*	*Hit*	*
52	*Curcuma xanthorrhiza*	*Hit*	
53	*Curcuma zedoaria*	*Hit*	*
54	*Olea europaea*	*Hit*	
I. Disease respiratory diseases			
1	*Foeniculum vulgare*	*Hit*	
2	*Clausena anisum-olens*	*Hit*	
3	*Glycyrrhiza uralensis*	*Hit*	
4	*Zingiber purpureum*	*Hit*	
5	*Piper retrofractum*	*Hit*	*
6	*Syzygium aromaticum*	*Hit*	
7	*Gaultheria punctata*	*Hit*	
8	*Panax ginseng*	*Hit*	
9	*Equisetum debile*	*Hit*	*
10	*Zingiber officinale*	*Hit*	
11	*Citrus aurantium*	*Hit*	*
12	*Nigella sativa*	*Hit*	*
13	*Amomum compactum*	*Hit*	
14	*Cinnamomum burmannii*	*Hit*	
15	*Melaleuca leucadendra*	*Hit*	
16	*Parkia roxburghii*	*Hit*	
17	*Cocos nucifera*	*Hit*	
18	*Piper cubeba*	*Hit*	
19	*Kaempferia galanga*	*Hit*	
20	*Coriandrum sativum*	*Hit*	
21	*Curcuma longa*	*Hit*	
22	*Piper nigrum*	*Hit*	
23	*Zingiber aromaticum*	*Hit*	
24	*Languas galanga*	*Hit*	
25	*Mentha piperita*	*Hit*	
26	*Oryza sativa*	*Hit*	*
27	*Myristica fragrans*	*Hit*	
28	*Pandanus amaryllifolius*	*Hit*	*
29	*Hydrocotyle asiatica*	*Hit*	*
30	*Mentha arvensis*	*Hit*	
31	*Lepiniopsis ternatensis*	*Hit*	
32	*Helicteres isora*	*Hit*	
33	*Blumea balsamifera*	*Hit*	
34	*Cymbopogon nardus*	*Hit*	
35	*Piper betle*	*Hit*	
36	*Curcuma xanthorrhiza*	*Hit*	

Table 4: Continued.

Number	Plants name	Hit-miss status	
37	*Salix alba*	*Hit*	*
38	*Matricaria chamomilla*	*Miss*	*
	J. Disease: skin and connective tissue		
1	*Strychnos ligustrina*	*Hit*	
2	*Merremia mammosa*	*Hit*	*
3	*Piper retrofractum*	*Hit*	*
4	*Santalum album*	*Hit*	
5	*Zingiber officinale*	*Hit*	*
6	*Citrus aurantium*	*Hit*	
7	*Citrus hystrix*	*Hit*	
8	*Cassia siamea*	*Hit*	
9	*Cocos nucifera*	*Hit*	
10	*Trigonella foenum-graecum*	*Hit*	
11	*Orthosiphon stamineus*	*Hit*	
12	*Curcuma longa*	*Hit*	
13	*Vetiveria zizanioides*	*Hit*	
14	*Aloe vera*	*Hit*	
15	*Rosa chinensis*	*Hit*	
16	*Jasminum sambac*	*Hit*	
17	*Phyllanthus urinaria*	*Hit*	
18	*Mentha piperita*	*Hit*	
19	*Oryza sativa*	*Hit*	
20	*Myristica fragrans*	*Hit*	*
21	*Hydrocotyle asiatica*	*Hit*	
22	*Lepiniopsis ternatensis*	*Hit*	
23	*Alstonia scholaris*	*Hit*	
24	*Andrographis paniculata*	*Hit*	
25	*Cymbopogon nardus*	*Hit*	
26	*Piper betle*	*Hit*	
27	*Theae sinensis*	*Hit*	
28	*Curcuma heyneana*	*Hit*	
29	*Kaempferia pandurata*	*Hit*	*
30	*Curcuma xanthorrhiza*	*Hit*	
31	*Melaleuca leucadendra*	*Hit*	
32	*Matricaria chamomilla*	*Miss*	*
	K. Disease: the urinary system		
1	*Foeniculum vulgare*	*Hit*	*
2	*Imperata cylindrica*	*Hit*	*
3	*Strychnos ligustrina*	*Hit*	*
4	*Plantago major*	*Hit*	
5	*Zingiber officinale*	*Hit*	*
6	*Cinnamomum burmannii*	*Hit*	*
7	*Strobilanthes crispus*	*Hit*	
8	*Kaempferia galanga*	*Hit*	*
9	*Orthosiphon stamineus*	*Hit*	
10	*Phyllanthus urinaria*	*Hit*	
11	*Blumea balsamifera*	*Hit*	*
12	*Sonchus arvensis*	*Hit*	
13	*Curcuma xanthorrhiza*	*Hit*	

*indicates that plant will not assigned if we use matching score >0.7.

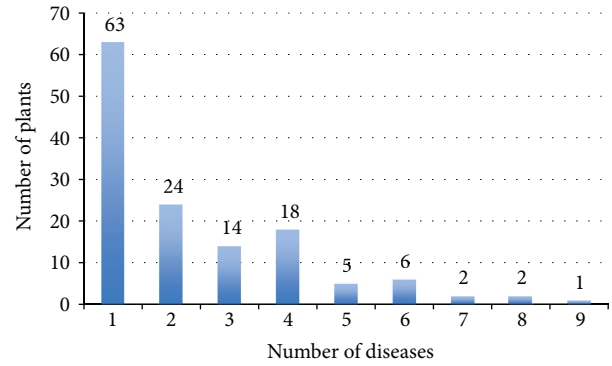


Figure 6: Distribution of 135 plants assigned based on 0.7% dataset with respect to the number of diseases they are assigned to.

for any type of application. It ensures coverage and performs robustly in case of random addition, removal, and rearrangement of edges in protein-protein interaction (PPI) networks [14]. While applying DPClusO, the parameter values of density and cluster property that we used in this experiment are 0.9 and 0.5, respectively [15]. Table 3 shows the summary of clustering result by DPClusO. Because clusters consisting of two Jamu formulas are trivial clusters, for the next steps we only use clusters each of which consists of 3 or more Jamu formulas. The number of total clusters increases along with the larger dataset, although the threshold correlation between Jamu pairs decreases. We evaluated the clustering result using matching score to determine dominant disease for every cluster (step 3 in Figure 1). Matching score of a cluster is the ratio of the highest number of Jamu associated with the same disease to the total number of Jamu in the cluster. Thus matching score is a measure to indicate how strongly a disease is associated to a cluster. Figure 4 shows the distribution of the clusters with respect to matching score from three datasets. All datasets have the highest frequency of clusters at matching score >0.9 and overall most of the clusters have higher matching score, which means most of the DPClusO generated clusters can be confidently related to a dominant disease. Furthermore the number of clusters with matching score >0.9 is remarkably larger compared to the same in other ranges of matching score in case of the 0.3% dataset (Figure 4(c)). If we compare the ratio of frequency of clusters at matching score >0.9 for every dataset, the 0.3% dataset has the highest ratio with 40.84% (of 453), compared to 29.67% (of 873) and 21.91% (of 1296), in case of 0.5% and 0.7% datasets, respectively. Thus, the most reliable species to disease relations can be predicted at matching score >0.9 corresponding to the clusters generated from 0.3% dataset.

Figure 5(a) shows the success rate for all 3 datasets with respect to threshold matching scores. Success rate is defined as the ratio of the number of clusters with matching score larger than the threshold to the total number of clusters. As expected it tends to produce lower success rate if we decrease correlation value to create the datasets. However more clusters are generated and more information can be extracted when we lower the threshold correlation value. The success rate increases rapidly as the matching score decreases

Table 5: Relation between disease classes in NCBI and efficacy classes reported by Afendi et al. [6].

Class of disease	Ref.	Efficacy class
D1 Blood and lymph diseases	NCBI	**E7** Pain/inflammation (PIN)
D2 Cancers	NCBI	**E7** Pain/inflammation (PIN)
D3 The digestive system	NCBI	**E4** Gastrointestinal disorders (GST) **E7** Pain/inflammation (PIN)
D4 Ear, nose, and throat	NCBI	**E7** Pain/inflammation (PIN)
D5 Diseases of the eye	NCBI	**E7** Pain/inflammation (PIN)
D6 Female-specific diseases	NCBI	**E5** Female reproductive organ problems (FML)
D7 Glands and hormones	NCBI	**E7** Pain/inflammation (PIN)
D8 The heart and blood vessels	NCBI	**E7** Pain/inflammation (PIN)
D9 Diseases of the immune system	NCBI	**E7** Pain/inflammation (PIN)
D10 Male-specific diseases	NCBI	**E6** Musculoskeletal and connective tissue disorders (MSC)
D11 Muscle and bone	NCBI	**E6** Musculoskeletal and connective tissue disorders (MSC)
D12 Neonatal diseases	NCBI	**E7** Pain/inflammation (PIN)
D13 The nervous system	NCBI	**E7** Pain/inflammation (PIN)
D14 Nutritional and metabolic diseases	NCBI	**E2** Disorders of appetite (DOA) **E4** Gastrointestinal disorders (GST)
D15 Respiratory diseases	NCBI	**E8** Respiratory disease (RSP) **E7** Pain/inflammation (PIN)
D16 Skin and connective tissue	NCBI	**E9** Wounds and skin infections (WND)
D17 The urinary system	*	**E1** Urinary related problems (URI)
D18 Mental and behavioural disorders	*	**E3** Disorders of mood and behavior (DMB)

from 0.9 to 0.6 and after that the slope of increase of success rate decreases. Therefore in this study we empirically decide 0.6 as the threshold matching score to predict plant-disease relations.

3.3. Assignment of Plants to Disease. By using DPClusO resulting clusters, we assigned plants to classes of disease. Based on a threshold matching score we assigned dominant disease to a cluster. Then we assign a plant to a cluster by way of analyzing the ingredients of the Jamu formulas belonging to that cluster and determining the highest frequency plant, that is, the plant that is used for maximum number Jamu belonging to that cluster (step 4 in Figure 1). Thus we assign a disease and a plant to each cluster having matching score greater than a threshold. Our hypothesis is that the disease and the plant assigned to the same cluster are related.

The total number of assigned plants depends on matching score value. Figure 5(b) shows the number of predicted plants that can be assigned to diseases in the context of matching score. With higher matching score value, the number of predicted plants assigned to classes of disease is supposed to remain similar or decrease but the reliability of prediction increases. In Figure 5(b) a sudden change in the number of predicted plants is seen at matching score 0.6 which we consider as empirical threshold in this work. Based on the 0.7% dataset, the largest number of plants (135 plants, Table 4) was assigned to diseases. There are 63 plants assigned to only one class of disease, whereas the other 72 plants are assigned to at least two or more classes of disease (Figure 6).

3.4. Evaluation of the Supervised Clustering Based on DPClusO. We used previously published results [6] as gold standard to evaluate our results. The previous study assigned plants to 9 kinds of efficacy whereas we assigned the plants to 18 disease classes (16 from NCBI and 2 additional classes). For the sake of evaluation we got done a mapping of the 18 disease classes to 9 efficacy classes by a professional doctor, which is shown in Table 5. Table 6 shows the prediction result of plant-disease relations for all 3 datasets, corresponding to clusters with matching score greater than 0.6. Table 6 also shows corresponding efficacy, the number of assigned plants, number of correctly predicted plants, and true positive rates (TPR), respectively.

We determined TPR corresponding to a disease/efficacy class by calculating the ratio of the number of correct prediction to the number of all predictions. When a disease corresponds to more than one kind of efficacy, the highest TPR can be considered the TPR for the corresponding disease. For all 3 datasets the TPR corresponding to each disease is roughly 90% or more. The 0.3% dataset consists of Jamu pairs with higher correlation values and based on this dataset 117 plants are assigned to 14 disease classes. The 0.7% dataset contains more Jamu pairs and assigned plants to 11 disease classes, one less disease class compared to 0.5% dataset. The two disease classes covered by 0.3% dataset but not covered by 0.5% and 0.7% datasets are the nervous system (D13) and disease of the immune system (D9). The only disease class covered by 0.3% and 0.5% datasets but not covered by 0.7% dataset is mental and behavioural disorders (D18). The larger dataset network tends to have

Table 6: The prediction result of plant-disease relations using matching score >0.6.

Class of disease	Corresponding efficacy	0.7% dataset			0.5% dataset			0.3% dataset		
		Number of assigned plants	Correct prediction	True positive rate	Number of assigned plants	Correct prediction	True positive rate	Number of assigned plants	Correct prediction	True positive rate
D1	E7	26	22	0.85	24	20	0.83	24	20	0.83
D2	E7	1	1	1.00	5	5	1.00	1	1	1.00
D3	E4	42	42	1.00	33	33	1.00	28	28	1.00
	E7		38	0.90		30	0.91		25	0.89
D4	E7	0	0	—	0	0	—	0	0	—
D5	E7	0	0	—	0	0	—	0	0	—
D6	E5	38	38	1.00	37	37	1.00	32	32	1.00
D7	E7	0	0	—	0	0	—	0	0	—
D8	E7	10	8	0.80	8	7	0.88	6	5	0.83
D9	E7	0	0	—	0	0	—	1	1	1.00
D10	E6	6	4	0.67	2	0	—	3	1	0.33
D11	E6	65	65	1.00	71	71	1.00	60	60	1.00
D12	E7	0	0	—	0	0	—	0	0	—
D13	E7	0	0	—	0	0	—	5	5	1.00
D14	E2	54	44	0.81	45	36	0.80	35	26	0.74
	E4		54	1.00		45	1.00		35	1.00
D15	E7	38	37	0.97	34	34	1.00	33	33	1.00
	E8		31	0.82		30	0.88		29	0.88
D16	E9	32	31	0.97	32	32	1.00	27	27	1.00
D17	E1	13	13	1.00	9	9	1.00	8	8	1.00
D18	E3	0	0	—	5	5	1.00	4	4	1.00
Total assigned plants			135			129			117	

lower coverage of disease classes. The number of Jamu pairs, that is, the number of edges in the network, affect the number of DPClusO resulting clusters and number of Jamu formulas per cluster. As a consequence, for the larger dataset networks, the success rate becomes lower and the coverage of disease classes is lower but prediction of more plant-disease relations can be achieved.

4. Conclusions

This paper introduces a novel method called supervised clustering for analyzing big biological data by integrating network clustering and selection of clusters based on supervised learning. In the present work we applied the method for data mining of Jamu formulas accumulated in KNApSAcK database. Jamu networks were constructed based on correlation similarities between Jamu formulas and then network clustering algorithm DPClusO was applied to generate high density Jamu modules. For the analysis of the next steps potential clusters were selected by supervised learning. The successful clusters containing several Jamu related to the same disease might be useful for finding main ingredient plant for that disease and the lower matching score value clusters will be associated with varying plants which might be supporting ingredients. By applying the proposed method important plants from Jamu formulas for every classes of disease were determined. The plant to disease relations predicted by proposed network based method were evaluated in the context of previously published results and were found to produce a TPR of 90%. For the larger dataset networks, success rate and the coverage of disease classes become lower but prediction of more plant-disease relations can be achieved.

Conflict of Interests

The authors declare that there is no financial interest or conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Bioscience Database Center in Japan and the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Grant-in-Aid for Scientific Research on Innovation Areas "Biosynthetic Machinery. Deciphering and Regulating the System for Creating Structural Diversity of Bioactivity Metabolites (2007)").

References

[1] R. Verporte, H. K. Kim, and Y. H. Choi, "Plants as source of medicines," in *Medicinal and Aromatic Plants*, R. J. Boger, L. E. Craker, and D. Lange, Eds., chapter 19, pp. 261–273, 2006.

[2] A. Furnharm, "Why do people choose and use complementary therapies?" in *Complementary Medicine: An Objective Appraisal*, E. Ernst, Ed., pp. 71–88, Butterworth-Heinemann, Oxford, UK, 1996.

[3] E. Ernst, "Herbal medicines put into context," *British Medical Journal*, vol. 327, no. 7420, pp. 881–882, 2003.

[4] F. M. Afendi, T. Okada, M. Yamazaki et al., "KNApSAcK family databases: integrated metabolite—plant species databases for multifaceted plant research," *Plant and Cell Physiology*, vol. 53, no. 2, p. e1, 2012.

[5] F. M. Afendi, N. Ono, Y. Nakamura et al., "Data mining methods for omics and knowledge of crude medicinal plants toward big data biology," *Computational and Structural Biotechnology Journal*, vol. 4, no. 5, Article ID e201301010, 2013.

[6] F. M. Afendi, L. K. Darusman, A. Hirai et al., "System biology approach for elucidating the relationship between Indonesian herbal plants and the efficacy of Jamu," in *Proceedings of the 10th IEEE International Conference on Data Mining Workshops (ICDMW '10)*, pp. 661–668, Sydney, Australia, December 2010.

[7] F. M. Afendi, L. K. Darusman, A. H. Morita et al., "Efficacy of Jamu formulations by PLS modeling," *Current Computer-Aided Drug Design*, vol. 9, pp. 46–59, 2013.

[8] F. M. Afendi, L. K. Darusman, M. Fukuyama, M. Altaf-Ul-Amin, and S. Kanaya, "A bootstrapping approach for investigating the consistency of assignment of plants to Jamu efficacy by PLS-DA model," *Malaysian Journal of Mathematical Sciences*, vol. 6, no. 2, pp. 147–164, 2012.

[9] W. Winterbach, P. V. Mieghem, M. Reinders, H. Wang, and D. de Ridder, "Topology of molecular interaction networks," *BMC Systems Biology*, vol. 7, article 90, 2013.

[10] C. Bachmaier, U. Brandes, and F. Schreiber, "Biological network," in *Handbook of Graph Drawing and Visualization*, pp. 621–651, CRC Press, 2013.

[11] X. Chen, M. Chen, and K. Ning, "BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network," *Bioinformatics*, vol. 22, no. 23, pp. 2952–2954, 2006.

[12] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, article 559, 2008.

[13] A. Martin, M. E. Ochagavia, L. C. Rabasa, J. Miranda, J. Fernandez-de-Cossio, and R. Bringas, "BisoGenet: a new tool for gene network building, visualization and analysis," *BMC Bioinformatics*, vol. 11, article 91, 2010.

[14] M. Altaf-Ul-Amin, M. Wada, and S. Kanaya, "Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking," *ISRN Biomathematics*, vol. 2012, Article ID 726429, 11 pages, 2012.

[15] M. Altaf-Ul-Amin, H. Tsuji, K. Kurokawa, H. Asahi, Y. Shinbo, and S. Kanaya, "DPClus: a density-periphery based graph clustering software mainly focused on detection of protein complexes in interaction networks," *Journal of Computer Aided Chemistry*, vol. 7, pp. 150–156, 2006.

[16] S. K. Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, Radius Press, New York, NY, USA, 1991.

[17] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlations coefficient," *The American Statiscian*, vol. 42, pp. 59–66, 1995.

[18] M. Li, J.-E. Chen, J.-X. Wang, B. Hu, and G. Chen, "Modifying the DPClus algorithm for identifying protein complexes based on new topological structures," *BMC Bioinformatics*, vol. 9, article 398, 2008.

[19] World Health Organization, "International Classification of Diseases (ICD) 10," 2010, http://www.who.int/classifications/icd/en/.

[20] National Center for Biotechnology Information, *Genes and Disease*, NCBI, Bethesda, Md, USA, 1998.

[21] P. Erdos and A. Renyi, "On the evolution of random graph," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[23] A. Vázquez, "Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 67, no. 5, Article ID 056104, 15 pages, 2003.

[24] Max Planck Institut Informatik, "NetworkAnalyzer," 2013, http://med.bioinf.mpi-inf.mpg.de/netanalyzer/index.php.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 205239, 12 pages
http://dx.doi.org/10.1155/2014/205239

Research Article

A Novel Feature Selection Strategy for Enhanced Biomedical Event Extraction Using the Turku System

Jingbo Xia,[1,2] Alex Chengyu Fang,[2,3] and Xing Zhang[2,3]

[1] *College of Science, Huazhong Agricultural University, Wuhan, Hubei 430070, China*
[2] *Department of Chinese, Translation and Linguistics, City University of Hong Kong, Kowloon, Hong Kong*
[3] *The Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong, Kowloon, Hong Kong*

Correspondence should be addressed to Alex Chengyu Fang; acfang@cityu.edu.hk

Received 14 January 2014; Revised 22 February 2014; Accepted 3 March 2014; Published 6 April 2014

Academic Editor: Shigehiko Kanaya

Feature selection is of paramount importance for text-mining classifiers with high-dimensional features. The Turku Event Extraction System (TEES) is the best performing tool in the GENIA BioNLP 2009/2011 shared tasks, which relies heavily on high-dimensional features. This paper describes research which, based on an implementation of an accumulated effect evaluation (AEE) algorithm applying the greedy search strategy, analyses the contribution of every single feature class in TEES with a view to identify important features and modify the feature set accordingly. With an updated feature set, a new system is acquired with enhanced performance which achieves an increased F-score of 53.27% up from 51.21% for Task 1 under strict evaluation criteria and 57.24% according to the approximate span and recursive criterion.

1. Introduction

Knowledge discovery based on text mining technology has long been a challenging issue for both linguists and knowledge engineering scientists. The application of text mining technologies based on large collections of known texts, such as the MEDLINE data base, has become especially popular in the area of biological and medical information processing [1, 2]. However, the rate of data accumulation is ever increasing at an astonishing speed [3]. The published literature grows exponentially and huge amounts of scientific information such as protein property and gene function are widely hidden in prohibitively large collections of text. For example, the literature in PubMed grows at a speed of two printed pages per second. As a result, it is practically impossible to manually curate experimental results from texts. As a result of this information explosion, text mining and linguistic methods have been used to perform automatic named entity recognition (NER) including biomedical NER [4], the extraction of clinical narratives [5] and clinical trials [6], analysis of the similarity in gene ontology [7], and the prioritization of vital function genes [8].

While entity recognition has been exploited as a powerful approach towards automatic NER retrieval, there has recently been an increased interest to find more complex structural information and more abundant knowledge in documents [9]. Hence, as a more recent development, moving beyond the purpose of entity recognition, the GENIA task in the BioNLP 2009/2011 shared tasks [10, 11] was set to identify and extract nine biomedical events from GENIA-based corpus texts [12], including gene expression, transcription, protein catabolism, localization, binding, phosphorylation, regulation, positive regulation, and negative regulation. The GENIA task is a publicly accepted task within the BioNLP communities and it is a leading pioneer in the area of structured event extraction.

Designed by Bjorne et al. [13], the Turku Event Extraction System (TEES) was the leading participant tool in both the BioNLP 2009 and 2011 shared tasks. Among the twenty-four participants of BioNLP shared tasks, TEES ranked first place in the GENIA task in 2009. While Riedel et al. performed best (with an F-score of 56.0%) for the GENIA task in 2011 ST [14], TEES was the third for the GENIA task and ranked number one for the other four of the eight subtasks. In 2011 and 2012, Björne et al. [15,

16] enhanced TEES and the F-score was increased further from 52.86% to 53.15%. The system was kept updated, and its 1.01 version achieved 55.65% in F-score. Meanwhile, TEES was also utilized for practical database searching. The system was applied to all PubMed citations and a number of analyses of the extracted information were illustrated [17–19]. TEES is now available at http://bionlp.utu.fi/. In 2011, the BioNLP shared task expanded from the GENIA task to eight generalized tasks including GENIA task, epigenetics and posttranslational modification task, infectious disease task, bacteria biotope task, and bacteria interaction task [11]. As the only system that participated in all the tasks in 2011, TEES was again ranked first place for four out of the eight tasks. Afterwards, Bjorne and Salakoski updated TEES 1.01 to TEES 2.0 in August 2012. The updated system is now capable of handling the DDI'11 (drug-drug interaction extraction) challenge (http://jbjorne.github.io/TEES/) in addition to all the BioNLP 2011 shared tasks. At the moment, the most up-to-date version is TEES 2.1 released in 2013 [20] and a major change is the wider coverage for serving more XML format corpus, and the core algorithm remains unchanged.

As a sophisticated text mining classifier, TEES uses enormous feature classes. For the BioNLP 2009 shared tasks, it produced 405,165 features for trigger detection and 477,840 features for edge detection out of the training data, which not only consume large amounts of processing time but also create undesirable noises to affect system performance. To address this issue and achieve even better system performance in terms of processing time and recognition accuracy, a natural starting point is to perform an in-depth examination of the system processes for feature selection (FS) and feature reduction (FR).

The integration of FS and FR has been demonstrated to hold good potentials to enhance an existing system [21–24]. Feature reduction can delete redundant features, avoid overfitting, provide more efficient modelling, and help gain a deeper insight into the importance of features. It aims to obtain a subset through an optimization between the feature size and the performance [25]. Amongst all of the feature reduction algorithms, the greedy search algorithm, also called hill climbing algorithm, is such an optimal technique for the identification and selection of important features [26]. van Landeghem et al. [27] introduced a feature selection strategy in the Turku system for a different testing data set and illustrated how feature selection could be applied to better understand the extraction framework. In their scheme, candidate events were trained to create a feature selector for each event type. The classifier was then built with the filtered training samples. By doing this, different trigger words for different event types were discriminated and those individual features with a higher occurrence frequency were assigned higher weights. Tag clouds of lexical vertex walks showed certain properties that are interesting from a linguistic point of view.

The purpose of this paper is to focus on the feature selection rules of TEES and aim to improve its performance for the GENIA task. By designing an objective function in the greedy search algorithm, we propose an accumulated effect evaluation (AEE) algorithm, which is simple and effective and can be used to numerically evaluate the contribution of each feature separately concerning its performance in the combined test. Moreover, we make further changes to the core feature class by incorporating new features and merging redundant features in the system. The updated system was evaluated and found to produce a better performance in both Tasks 1 and 2. In Task 1, our system achieves a higher F-score of 53.27% according to the "strict evaluation" criterion and 57.24% according to the "approximate span and recursive" criterion. In Task 2, the new strategy achieved an F-score of 51.77% under the "strict evaluation" criterion and 55.79% under the "approximate span and recursive" criterion. These represent the best performance scores till now for event recognition and extraction.

2. Materials and Method

2.1. GENIA Task and Scheme of Turku TEES System

2.1.1. GENIA Task and the Evaluation Criteria. Different from the previous NER task, the GENIA task aims to recognize both the entities and the event relationship between such entities. Extended from the idea of semantic networks, the recognition task includes the classification of entities (nodes) and their associated events (edges). The participants of the shared task were expected to identify nine events concerning given proteins, that is, gene expression, transcription, protein catabolism, localization, binding, phosphorylation, regulation, positive regulation, and negative regulation. The mandatory core task, Task 1, involves event trigger detection, event typing, and primary argument recognition [10]. An additional optional task, Task 2, involves the recognition of entities and the assignment of these entities. Finally, Task 3 targets the recognition of negation and speculation. Because of their fundamental importance and vast potential for future developments, researchers mainly focus on Tasks 1 and 2.

Like the other text mining systems, the performance of TEES is evaluated by precision, recall, and F-score. The first measure equals the fraction of obtained relevant documents and the retrieved documents represent the correctness of the extraction system; that is,

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}. \quad (1)$$

The second measure, recall, is defined as

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}. \quad (2)$$

Recall is used to assess the fraction of the documents relevant to the query that are successfully retrieved. Precision and recall indicators are well-known performance measures in text mining, while F-score is a third measure that combines

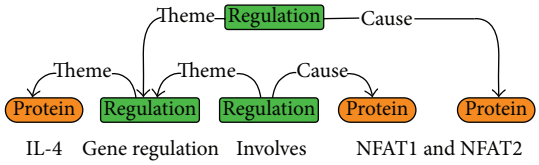


FIGURE 1: The graphical representation of a complex biological event (refer to [5]).

precision and recall and is the harmonic mean of precision and recall:

$$F\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (3)$$

The F-score evenly weighs the precision and recall and forms a reliable measure to evaluate the performance of a text mining system. For more information, refer to [28].

2.1.2. Scheme of TEES. The core of TEES consists of two components, that is, classification-style trigger/event detection and rich features in graph structure. By mapping the tokenized word and entity to the node and mapping event relation to edge between entities, TEES regards the event extraction as a task of recognizing graph nodes and edges as shown in Figure 1.

Generally, TEES first convert the node recognition to a problem of 10-class classification, which corresponds to 9 events defined in the shared task and another class for negative case. This procedure is defined as trigger detection. Thereafter, the edge detection is defined by the recognition of concrete relationships between entities, including semantic direction and theme/cause relation.

As in Figure 1, the word "IL-4" is assigned to class "protein," while "involves" is assigned to class "regulation." The edge between "IL-4" and "regulation" is labelled as "theme." Hence we obtain a simple regulation event, namely, "IL-4 regulation," which means "IL-4" is the theme of "regulation." Similarly, the representation in Figure 1 indicates that the simple event "IL-4 regulation" is the theme of "involves" and the protein "NFAT1" is the cause of "involves," from which a complex regulation event can be extracted; that is, "IL-4 regulation" regulates protein "NFAT1." For more detailed information, refer to Bjorne's work [5] and the website of TEES, http://bionlp.utu.fi/eventextractionsoftware.html.

The data set used in TEES consists of four files in GENIA corpus [3]. Each file is given in a stand-off XML format with split sentences, annotation of known proteins, parts of speech (POS), and syntactic dependency tree information. One is train123.xml, which contains 800 abstracts with 7,482 sentences; another is devel123.xml, which contains 150 abstracts with 1,450 sentences. The file everything123.xml is the sum of the two files above, and test.xml is the test data set, which contains 950 abstracts with 8,932 sentences.

The scheme of TEES system consists of three phases. First, linguistic features are generated in the feature generation phase. Second, in the training phase, train123.xml is used as the training data set, devel123.xml is used as development set, and optimum of parameters is obtained. Then in the third phase, using everything123.xml (sum of train123.xml and devel123.xml) as the training set, test.xml is used as unknown data set for event prediction. Events are extracted from this unknown set and accuracy is subsequently evaluated. See Figure 2.

Mainly, there are two parameters in grid searching at the training stage. The first parameter is C in polynomial kernel function of support vector machine. Second, in order to set a proper precision-recall trade-off, a parameter β ($\beta > 0$) is introduced in trigger detection. For "no trigger" class, the given classification score is multiplied by β so as to increase the possibility of tokens falling into "trigger" class. Optionally, β is set as 0.6, and it will be put into grid searching for obtaining optimum value.

2.2. Definition of Feature Generation Rule of Turku System. Features used for trigger detection are designed in a rational way, and abundant features are generated from training data set.

In terms of the training data set with GENIA format, the dependency relation of each sentence is output by Stanford parser [29], which addresses the syntactic relations between word tokens and thus converts the sentence to a graph, where the node denotes a word token and the edge corresponds to the grammatical relation between two tokens. By doing this, a directed acyclic graph (DAG) is constructed based on the dependency parse tree, and the shortest dependency path (SDP) is located.

For a targeted word in sentence which represents a node in graph, the purpose of trigger detection is to recognize the event type belonging to the word. Meanwhile, edge detection is to recognize the theme/cause type between two entities. Therefore, both detections can be considered as multiclass classification.

The features produced in trigger detection are categorized into six classes, as listed in Figure 3, and the whole feature generation rule is listed in Supplementary Appendix A in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/205239. First, for a sentence in which the target word occurs, token information is counted for the whole sentence as a bag of words (BOW). This feature class is defined as "sentence feature." The second feature class is "main feature" of the target word, including part of speech (POS) and stem information output by Porter Stemmer [29]. The third class, "linear order feature," focuses on the information about the neighboring word tokens in natural order of the sentence. TEES also maintains a fourth class about the microlexical information of the word token, for example, upper or lower case of the word, existence of digital number, double letter in the word, and three letters in the word. These features constitute a "content feature" class, while the "attached edge feature" class focuses on the information about the neighboring word tokens of the target word in SDP. Finally, there is a sixth "chain feature" class, which focuses on the information of the whole SDP instead.

Here, the features are well structured from the macro- and microperspectives about the sentence. Basically, the feature

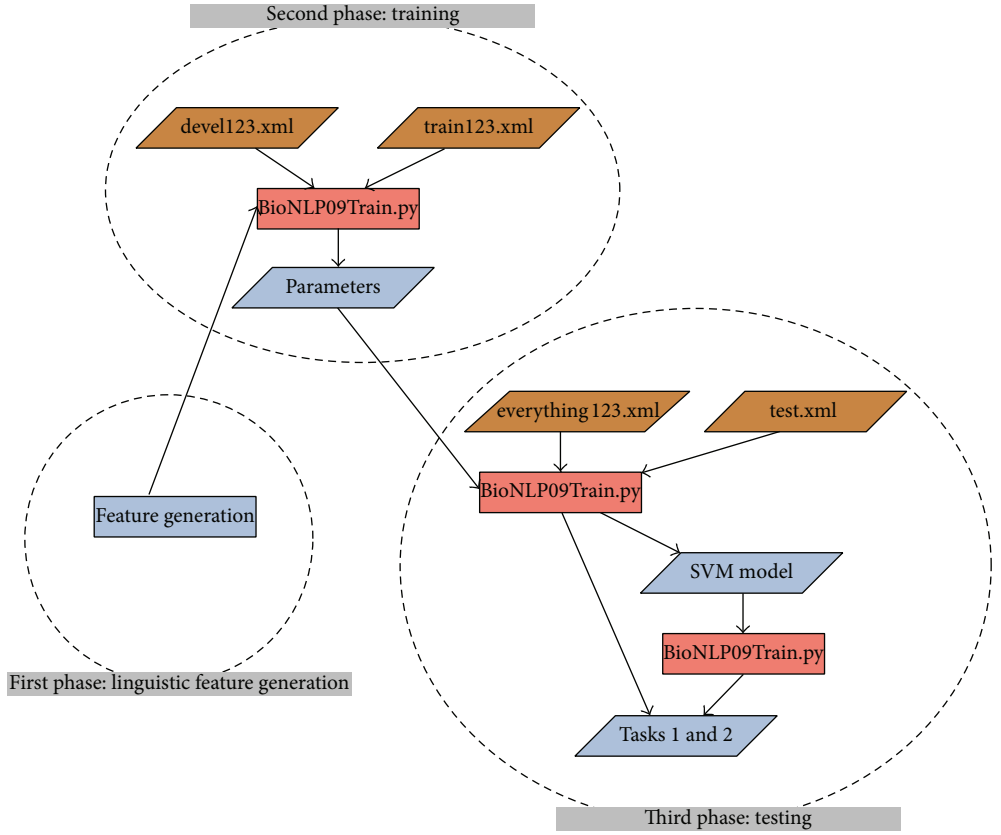


Figure 2: Data set and pipeline of TEES, cited from TEES Toolkit. Location: TurkuEvent ExtractionSystem readme.pdf, p. 7 in the zip file, http://bionlp.utu.fi/static/event-extractor/TurkuEventExtractionSystem-1.0.zip.

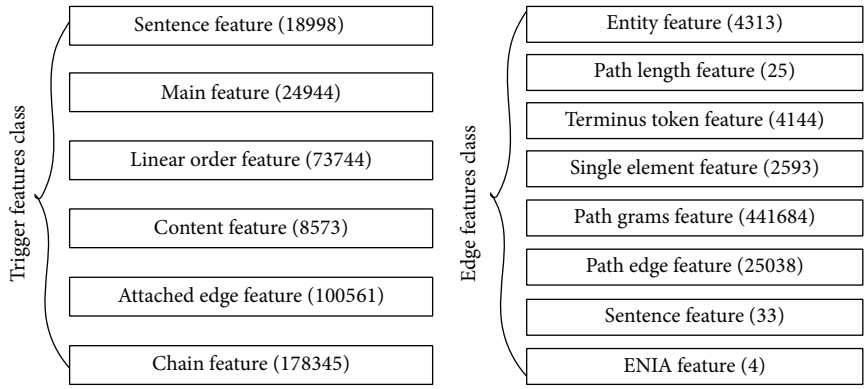


Figure 3: Feature class in trigger detection and edge detection.

generation rule of "main feature" and "content feature" mostly relies on microlevel information about the word token itself, while "sentence feature" and "linear order feature" rely on the macrolevel information about the sentence. Differently, "attached edge" and "chain feature" rely on the dependency tree and graph information, especially the SDP.

Similarly, features used for edge detection can be classified into 8 classes, namely, entity feature, path length feature, terminus token feature, single element feature, path grams feature, path edge feature, sentence feature, and GENIA feature. Among the features above, the third feature is omitted, since it is considered part of the first feature.

2.3. Feature Evaluation Method: Accumulated Effect Evaluation (AEE) Algorithm. A quantitative method is designed to evaluate the importance of feature classes. Here, the feature combination methods are ranked by F-score, and the

Step 1. Categorize features into n feature classes and there are $2^n - 1$ possibilities for all kinds of feature combination. Denote $N = 2^n - 1$, and run all of classifying test. F-score is recorded for each experiment.
Step 2. Sort N feature combinations according to F-score with the descending order.
Step 3. FOR $i = 1$ to n, $j = 1$ to N,
$O_{ij} = 0$, $c_{ij} = 0$;
END FOR
// O_{ij} is used to calculate the occurrence of the ith feature among j experiments.
// c_{ij} evaluates the partial importance of the ith feature after j times experiments.
Step 4. FOR $j = 1$ to N, // For each feature combination in the jth experiment.
FOR $i = 1$ to n, // For each feature.
IF the ith feature occur in the jth feature combination,
O_{ij}++;
// For each experiment, the occurrence of the ith feature is accumulated.
END IF
$c_{ij} = O_{ij}/j$;
// By dividing O_{ij} by j, we get a numerical value to evaluate the importance
// of the ith feature. Since the items are ranked by F-score with a descending
// order, the first ranked features combination corresponds to a smaller j and
// hence a bigger c_{ij}.
END FOR
END FOR
Step 5. FOR $i = 1$ to n,
$$\text{AEE1}(i) = \sum_{j=1}^{n} c_{ij};$$
END FOR
// By summing up c_{ij} for a fixed i, we get the accumulated effect of the ith feature.
Step 6. AEE1(i) is the objective function of greedy search algorithm, and we sort AEE1(i) to get the most important feature.

Algorithm 1: AEE1 algorithm.

occurrence of ith feature class in the top jth combinations (j runs from the top one to the last one) is counted, the rate of occurrence is computed, and finally the sum of the occurrence is calculated. The total calculation is shown in Supplementary Material Appendix C. The algorithm is denoted as AEE1 algorithm as shown in Algorithm 1.

Here, AEE1(i) is the contribution value of the ith feature and also the objective function of the greedy search algorithm, which reflects the accumulated effect of the ith feature among the top classifiers. Since AEE1(i) makes sense in a comparative way, the theoretical maximum and minimum of AEE1(i) are calculated by

$$\begin{aligned}\text{Max\_AEE1} &= 1 \times 2^{n-1} + \frac{2^{n-1}}{2^{n-1}+1} + \frac{2^{n-1}}{2^{n-1}+2} \cdots + \frac{2^{n-1}}{2^n - 1},\\ \text{Min\_AEE1} &= 0 \times \left(2^{n-1} - 1\right) + \frac{1}{2^{n-1}} + \frac{2}{2^{n-1}+1} \cdots + \frac{2^{n-1}}{2^n - 1}.\end{aligned} \tag{4}$$

The idea of AEE1 comes from the understanding that the top classifiers with higher F-scores include more efficient and important features. However, this consideration disregards the feature size in terms of those top classifiers.

For better understanding, a simple case is considered. Assume there are two feature classes that could be used in the classifier and so there are three feature combinations for the classifier, namely, 1, 2, and 1&2. Without loss of generality, we assume that the best classifier uses the feature class 1&2, the second one uses the feature class 1, and the worst classifier uses the feature class 2. Here, we denote the rank list 1&2, 1, and 2 as Rank Result A. According to the third column in Table 1(a), AEE value will be calculated as AEE1(1) = 1/1 + 2/2 + 2/3 = 2.667.

As another example, we assume a rank list 1, 1&2, and 2, which is denoted as Rank Result B and shown in Table 1(b). As can be seen from Table 1(b), AEE1(1) equals 2.667, the same as in Table 1(a). However, this is not a reasonable result, since feature 1 is more significant in Rank Result A than in Rank Result B if the feature class size is considered.

Therefore, an alternative algorithm, AEE2, is proposed by updating $c_{ij} = O_{ij}/j$ in Step 4 of AEE1 as $c_{ij} = O_{ij}/\sum_{i=1}^{n} O_{ij}$. Here, the size of feature class is considered for the computation of c_{ij} and a classifier with higher performance and smaller feature size will ensure a high score for the feature it owns. As an example, column 4 in Tables 1(a) and 1(b) shows that AEE2(1) = 1.667 in Rank Result A and AEE2(1) = 2.167 in Rank Result B. The result ensures the comparative advantage for feature class 1 in Rank Result A. Similarly, AEE2(i) mainly makes sense in a comparative way with the theoretical maximum and minimum values, which are also similarly computed.

Table 1: Examples for AEEi algorithm.

(a) AEEi result for Ranking Result A

Rank	Result A	O_{ij}			AEE1 c_{ij}			AEE2 c_{ij}	
		$i = 1$	$i = 2$		$i = 1$	$i = 2$		$i = 1$	$i = 2$
$j = 1$	1&2	1	1		1/1	1/1		1/2	1/2
$j = 2$	1	2	1		2/2	1/2		2/3	1/3
$j = 3$	2	2	2		2/3	2/3		2/4	2/4
				AEE1(i)	2.667	2.167	AEE2(i)	1.667	1.333

(b) AEEi result for Ranking Result B

Rank	Result B	O_{ij}			AEE1 c_{ij}			AEE2 c_{ij}	
		$i = 1$	$i = 2$		$i = 1$	$i = 2$		$i = 1$	$i = 2$
$j = 1$	1	1	0		1/1	0/1		1/1	0/1
$j = 2$	1&2	2	1		2/2	1/2		2/3	1/3
$j = 3$	2	2	2		2/3	2/3		2/4	2/4
				AEE1(i)	2.667	1.167	AEE2(i)	2.167	0.833

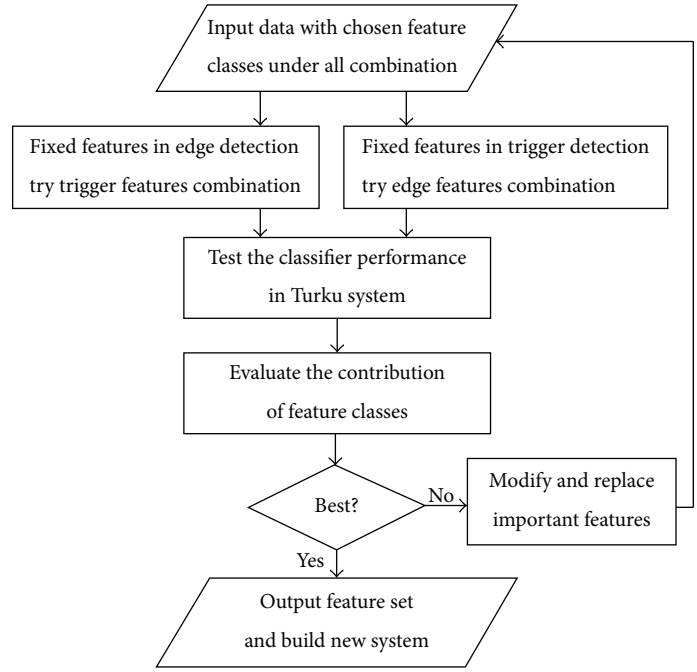


Figure 4: Flowchart of the research.

Considering AEE1(1) = 2.667 > AEE1(2) = 2.167 and AEE2(1) = 1.667 > AEE2(2) = 1.333 in Table 1(a), the importance of feature class 1 prevails over that of feature class 2 in both cases. This ensures a reasonable consistency. Actually, both AEE1 and AEE2 algorithms ensure a correct weighting rank among different feature classes and a hybrid of the two methods is used in the experiments.

2.4. Flowchart of the Research. Unlike a trial-and-error procedure, the scheme of this research is oriented towards better feature selection so that important feature classes are identified by evaluating the contribution of the individual features.

Accordingly, codes are written to enhance vital features. Thereafter, the classifiers with new updated features are tested and better performance is presumed. Compared with the previous TEES system, our main contribution is to import feature selection strategy in Phase 1, that is, "linguistic feature selection," as shown in Figure 2. The flowchart of our research is shown in Figure 4.

3. Results

3.1. Evaluation of Individual Feature Classes by Quantitative Method. Based on the AEE algorithm, all of the combinations of the feature classes are used in classifiers and their

Table 2: Top 10 best classifiers with corresponding feature classes in combination experiment.

Rank	F-score (%)	Trigger feature combination with fixed trigger feature	F-score (%)	Edge feature combination with fixed edge feature
1st	51.34	1&2&3&4&5	52.16	1&2&4&5&6&7&8
2nd	51.21	1&2&3&4&5&6	51.81	1&2&4&5&6&8
3rd	50.71	1&2&4&5	51.29	1&2&5&6&7&8
4th	50.19	1&2&3&4&6	51.18	1&2&5&6&8
5th	49.9	1&2&4&5&6	50.33	1&2&4&6&7&8
6th	49.74	1&3&4&5	50.23	1&2&4&6&8
7th	49.44	1&4&5	49.26	2&4&5&6&8
8th	49.16	1&2&3&4	49.02	1&2&4&5&6
9th	48.82	1&2&4&6	48.32	2&4&5&6&7&8
10th	47.82	1&2&4	47.42	2&5&6&8

Table 3: Score of features in trigger detection.

Feature ID	1	2	3	4	5	6		
Feature name	Sentence feature	Main feature	Linear order feature	Content feature	Attached edge feature	Chain feature	Theoretical maximum	Theoretical minimum
AEE1 score	40.83	39.94	32.99	52.23	36.12	30.09	53.43	10.27
AEE2 score	10.80	10.82	8.99	14.16	9.80	8.43	20.52	3.28

Table 4: Score of features in edge detection.

Feature ID	1	2	4	5	6	7	8		
Feature Name	Entity Feature	Path length Feature	Single element Feature	Path grams Feature	Path edge Feature	Sentence Feature	GENIA Feature	Theoretical Maximum	Theoretical Minimum
AEE1 score	77.60	83.84	68.45	74.88	77.76	56.09	66.35	107.61	20.08
AEE2 score	18.63	19.57	16.57	17.51	17.88	13.40	15.45	35.80	5.54

corresponding F-scores are ranked so as to evaluate the contribution of the individual feature classes.

Using the quantitative algorithm, the importance of feature classes is addressed in the two-phase TEES procedure that involves trigger detection and edge detection. For better understanding, the top ten classifiers with respective feature combinations are shown in Table 2, and the total calculation is shown in Supplementary Material Appendix B. The final results are collected in Tables 3 and 4.

During the trigger detection, the results show that the 4th feature performs best and 6th feature performs worst. By calculating AEEi value of features, Figures 5 and 6 show plots of the best and the worst feature classes in trigger detection.

The comparison between the two features shows how the 4th feature performs better than 6th feature. And 52.23 and 30.09 also correspond to the value of area below the curve. The AEE1 and AEE2 plot of the best and worst features in trigger and edge detections are shown in Figures 5 and 6.

3.2. Modification of Features and New Experiment Results. Combinations of features for trigger detection show that the "content feature" in trigger detection is the most important one, and the "chain feature" is the worst. This result shows that, in terms of identifying a target word token, the target itself provides more information than the neighboring tokens.

Taking the feature generation rule of 4th feature into consideration, the "content feature" class contains four features, "upper," "has," "dt," and "tt." Specifically, "upper" is to identify the upper case or lower case of letters, "has" is to address the existence of a digit or hyphen, "dt" is to record the continuous double letters, and "tt" is to record three continuous letters. Since the content feature is vital for trigger detection, the feature generation rule could be strengthened similarly. Accordingly, a new "ft" feature is inserted into the "content feature" class by which the consecutive four letters in the word are considered. Moreover, modification is performed on the 6th feature class by merging similar features related to dependency trees in both the 5th and the 6th feature classes. For simplicity, the updated features are denoted as $4'$ and $6'$.

Furthermore, if we compare the best performance between classifiers with trigger feature comprising the original features, $4'$ features, $6'$ features, or $4'\&6'$ features

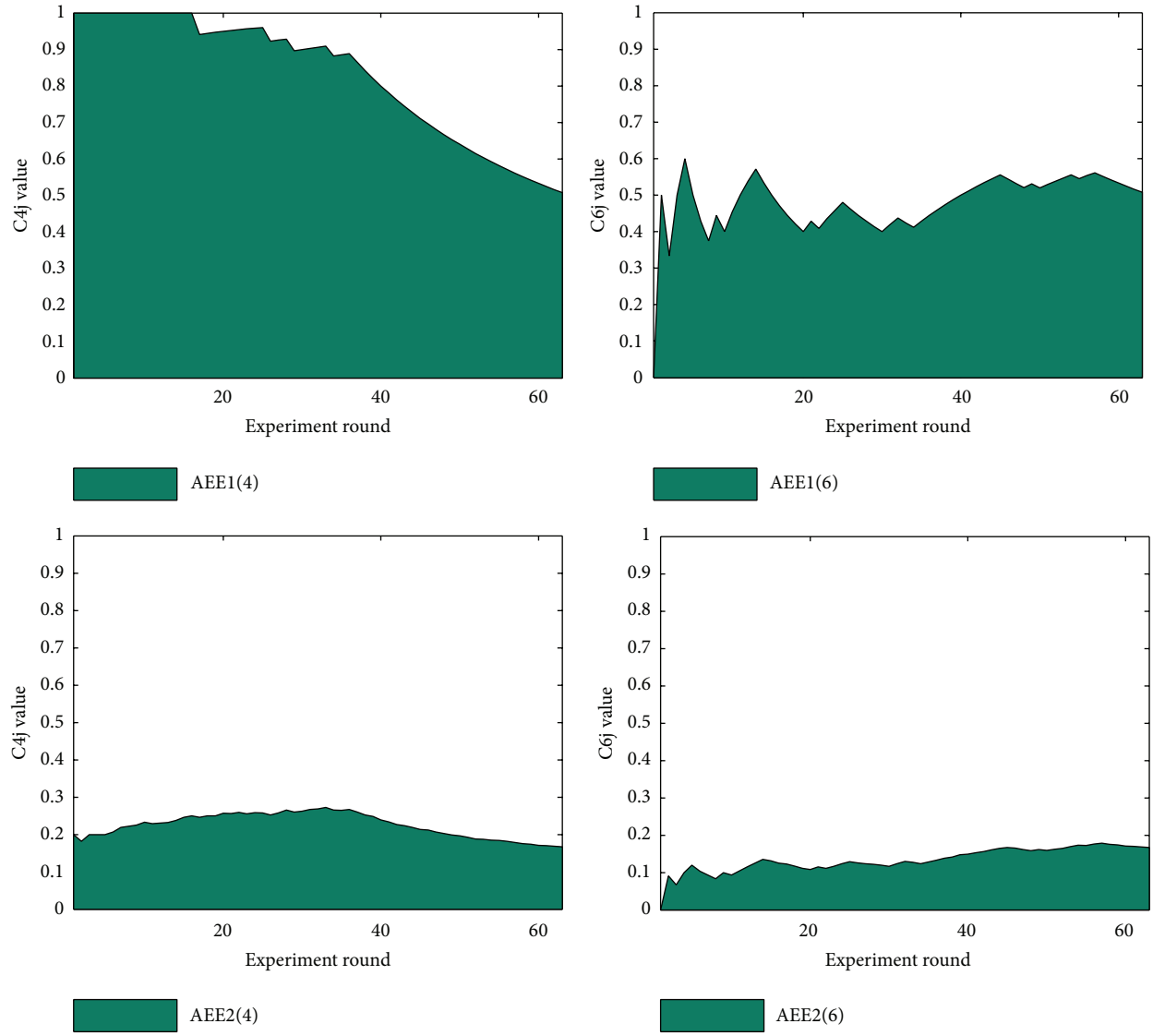


Figure 5: AEE1 and AEE2 plots of the best/worst feature in trigger detection.

(all include original edge features), we get that the best ones for trigger features are 1&2&3&4&5, 1&2&3&4′&5, 1&2&3&4&5&6′, and 1&2&3&4′&5&6′, respectively, while the F-score reaches 51.34, 52.21, 51.93, and 51.99. The new combination result is listed in Supplementary Material Appendix D.

The complete combination experiment shows that the best combination of trigger feature is 1&2&3&4′&5, with an F-score of 52.21. Generally, the best feature combination covers the major part of feature sets. An interesting phenomenon in the best combination is the absence of the 6th or 6′th feature class, which indicates that this feature class is redundant and can be done without it.

Similarly, for feature selection in edge detection, various experiments are carried out based on the best combination of trigger features. Here, the best feature and the worst feature (2nd and 7th feature) are chosen to be modified in edge detection, and we denote the new feature classes as 2′ and 7′. With the fixed trigger feature 1&2&3&4′&5, we test the classifier with the original edge features, 2′ feature, 7′ feature, and 2′&7′ feature. We obtain the best classifier in each combination experiment. The best classifier in each combination owns a feature set 1&2&4&5&6&7&8, 1&2′&4&5&6&7&8, 1&2&4&5&6&7′&8, or 1&2′&4&5&6&7′&8, separately, and the achieved F-score is 52.16, 52.37, 52.47, or 52.68, respectively.

In the above experiments, we test the performance of trigger feature class by fixing edge features. Likewise, we test edge features by fixing trigger features. We observe that feature modifications in this phase are indeed capable of achieving improvement, where all of the best combinations perform better than the result of the best trigger. Finally, we use the best trigger feature (1&2&3&4′&5) and best edge feature (1&2′&4&5&6&7′&8), and eventually the best combination of feature set achieved the highest score of 53.27, which is better than the best performance of 51.21 previously

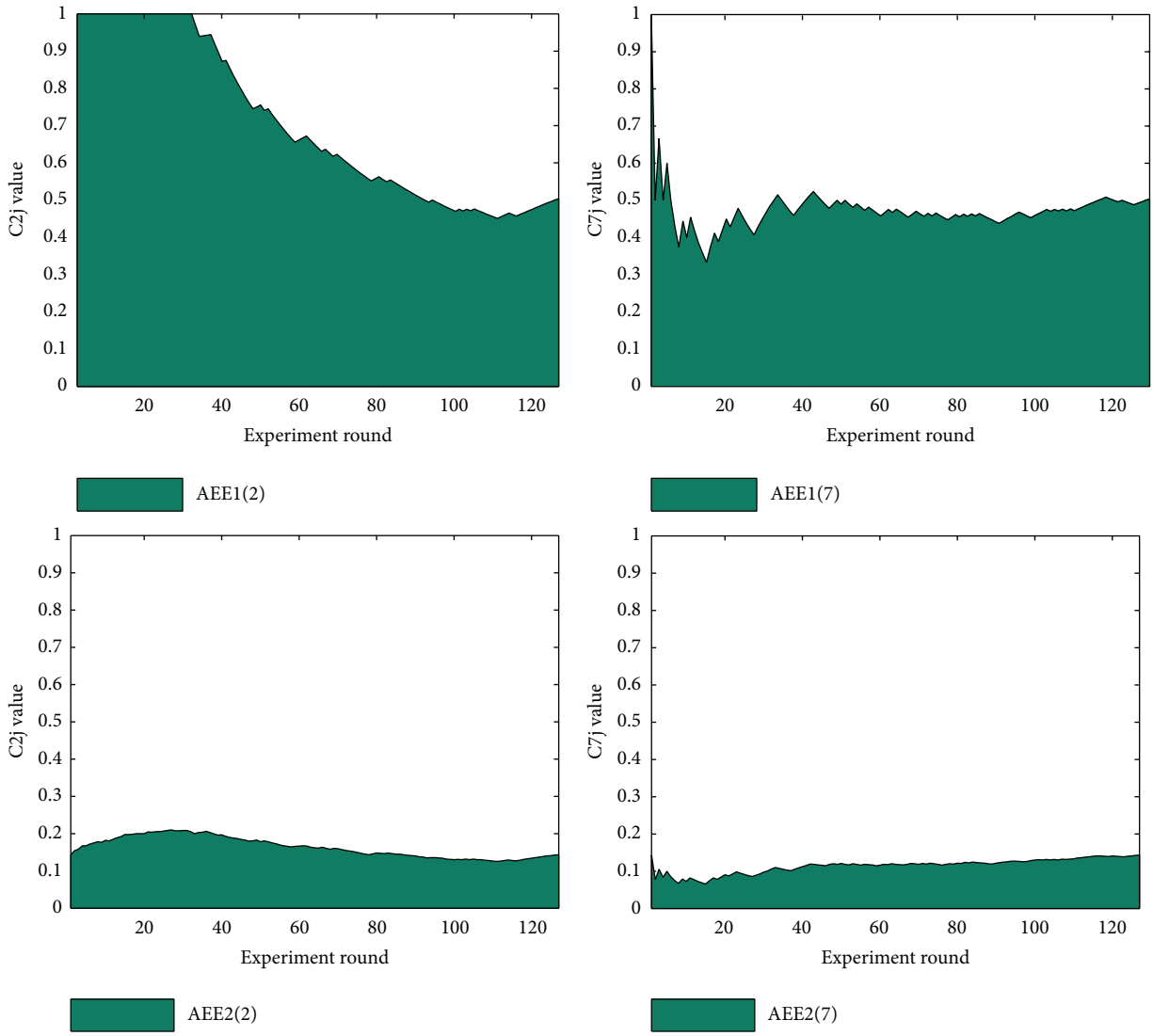


Figure 6: AEE1 and AEE2 plots of the best/worst feature in edge detection.

reported for TEES 2.0. Therefore, it is concluded that the best classifier has a feature set with trigger feature 1&2&3&4′&5 and edge feature 1&2′&4&5&6&7′&8, where trigger-c = 250000, edge-c = 28000, and β = 0.7. The final result is listed in Table 5.

Comparing with the 24 participants in GENIA task of BioNLP 2009 and historical progress of Bjorne's work, a contour performance is given in Figure 7.

As Figure 6 indicates, our system ranks the first among the 24 systems. Comparisons of F-scores are listed in Table 6.

4. Discussions and Conclusions

4.1. An Analysis of the Contribution of Features. In this research, we designed a feature selection strategy, AEE, to evaluate the performance of individual feature classes to identify the best performing feature sets. An important finding is that the greatest contribution comes from the content feature class in trigger detection. In this section, a routine analysis of the contribution is shown, which yields the same finding and supports the same conclusion that the content feature class contributes the most towards event recognition and extraction.

First, retaining one feature class in the classifier, we can get separate F-scores based on these features. Dividing the F-score by feature size, we calculate the average contribution roughly. This value partly reflects the contribution by feature classes in terms of class size. The result in Table 6 shows that the average contribution of the 4th feature is 0.003, which is the greatest score achieved by individual feature class.

Second, we observe all of the double combinations involving the ith feature and observe that, in most cases, when the ith feature is combined with the 4th feature, it reaches the best performance score. Even the worst performance of the double feature combinations involving the 4th feature performs much better than the other settings.

Table 5: The best feature combination after choosing the best trigger feature (1&2&3&4′&5) and best edge feature (1&2′&4&5&6&7′&8).

Task 1	Recall	Precision	F-score
Strict evaluation mode	46.23	62.84	53.27
Approximate span and recursive mode	49.69	67.48	57.24
Event decomposition in the approximate span mode	51.19	73.21	60.25
Task 2	Recall	Precision	F-score
Strict evaluation mode	44.90	61.11	51.77
Approximate span and recursive mode	48.41	65.81	55.79
Event decomposition in the approximate span mode	50.52	73.15	59.76

Recall = TP/(TP + FN), Precision = TP/(TP + FP), and F-score = 2((Recall × Precision)/(Recall + Precision)).

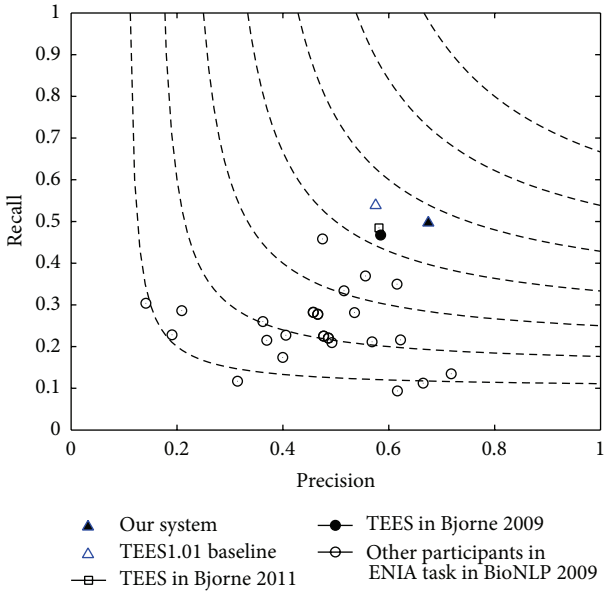


Figure 7: F-score contour performance of participants in GENIA task. This F-score is evaluated under approximate span and recursive mode. Our current system is marked with a full triangular label.

Third, a similar phenomenon occurs in the case of three-feature-combination experiment and four-feature-combination experiment. In all cases, when ith feature is combined with the 4th feature, it reaches the best performance. The same as before, even the worst performance of double feature combination involving the 4th feature is much better than the other combinations. See Table 7.

Finally, in yet another analysis, we observe the cases where the ith feature is cancelled out. The results show that the combination without the 4th class performs worst, which in turn confirms the importance of the 4th feature.

Through the routine analysis above, there is ample evidence arguing in support of the importance of the 4th feature in trigger detection. Compared with the results in numerical scores, the contribution value of the 4th feature is greater than the others, which confirms the judgment. Furthermore, we can sort these features according to their contribution values. These results can further prove our decision to modify the 4th feature and thereafter enhance system performance.

It is interesting that the routine analysis shows the substantial positive evaluation for the 4th trigger feature, which is proved by the results of quantitative analysis of AEE algorithm. This shows a consistent tendency of feature importance, which in turn proves the reliability of the AEE algorithm. Since it is clumsy to use routine analysis to analyze all of the features, we expect that the AEEi algorithm makes sense in generalized circumstances.

4.2. Linguistic Analysis of Chosen Trigger Features and Importance of Microlexical Feature. The effectiveness of the 4th trigger feature motivated the feature class modification by inserting "ft" features. One should also note that all these features in the 4th feature class are related to the spelling of word tokens, which is similar to stems but contains more abundant information than stems. Besides, we can also analyze and ascertain the importance of other features, like POS information, through feature combinations. Here, the edge features are fixed, and only a smaller section of the trigger features is tested through combination experiments.

The full combinations are listed in Supplementary Material Appendix E and Table 8 lists the top 5 important features. Here, "stem" and "nonstem" are the stem part or left part of stem after using Porter Stemmer [28]. It is also generated in microlexical level, similar to "dt" and "tt." The results in this table show that the lexical feature generation rule affects the token information extraction in a decisive way.

4.3. Strategy Discussion in Feature Selection under Machine Learning Strategy. As a matter of fact, trigger features could be analyzed according to their generation rule, namely, sentence feature, main feature, linear order feature, content feature, attached edge feature, and chain feature. This is a state-of-the-art strategy in feature selection. TEES is a nice system based on machine learning, which, however, does not perform intensive feature selection. The absence of a feature selection strategy in previous research mainly stems from two reasons. The first is that the natural core idea of machine learning is just to put enough features into the classifier as a black box; the second is that the performance of a classifier with huge sizes of features is always better in accuracy and F-score. Previously, the features used in TEES have been mostly chosen by trial and error. By adding some codes to produce additional features and seeing how they impact the performance of the system, TEES always achieves a higher F-score but with unsure directions. This strategy introduces useful but uncertain features and produces a large number of features. By introducing a feature selection and evaluation method, the importance of different feature classes could be ranked in a proper way, which helps to identify the important features to be modified effectively. Therefore, in terms of the current research regarding the development of the Turku Event Extraction System, we believe that our research

Table 6: Comparison of F-score performance in Task 1 with other systems (under primary criterion: approximate span and recursive).

	Bjorne et al. 2009 [13]	Bjorne et al. 2011 [15]	Bjorne et al. 2012 [16]	TEES1.01	Riedel et al. 2011 [14]	Ours
F-score (%)	51.95	52.86	53.15	55.65	56.00	57.24

Table 7: The best feature combination after choosing features dynamically in trigger and edge detection.

	1	2	3	4	5	6
Feature in trigger	#Sentence feature	#Main feature	#Linear order Feature	#Content feature	#Attached edge Feature	#Chain feature
Feature size	18998	24944	73744	8573	100561	178345
(Merely one feature analysis)						
F-score under one class	0	42.05	3.50	27.11	7.33	5.48
Average contribution	0	0.001	$4e-5$	0.003	$7e-5$	$3e-5$
(Double feature combination analysis)						
Best performance	(+4)	(+4)	(+4)	(+1)	(+2)	(+4)
Involving ith feature	45.39	36.65	22.29	45.39	28.27	23.48
Worst performance	(+2)	(+1)	(+1)	(+3)	(+1)	(+1)
Involving ith feature	0	0	0	22.29	0.91	3.10
(Three-feature combination analysis)						
Best performance	(+4, 5)	(+1, 4)	(+1, 4)	(+1, 5)	(+1, 4)	(+1, 4)
Involving ith feature	49.44	47.82	45.13	49.44	49.44	45.51
Worst performance	(+2, 3)	(+1, 3)	(+1, 2)	(+3, 6)	(+1, 3)	(+1, 3)
Involving ith feature	0.11	0.11	0.11	20.76	1.98	2.40
(Four-feature combination analysis)						
Best performance	(+2, 4, 5)	(+1, 4, 5)	(+1, 4, 5)	(+1, 2, 5)	(+1, 2, 4)	(+1, 2, 4)
Involving ith feature	50.71	50.71	49.74	50.71	50.71	48.82
Worst performance	(+2, 3, 6)	(+1, 3, 6)	(+1, 2, 6)	(+3, 5, 6)	(+1, 2, 3)	(+1, 2, 3)
Involving ith feature	5.77	5.77	5.77	21.13	7.02	5.77
(Five-feature combination analysis)						
Performance						
Without ith feature	34.22	47.1	49.90	16.37	50.19	51.34

Table 8: Analysis of average contribution of lexical features.

	Feature	Feature class	F-score	Feature size	Average contribution
1	Nonstem	Main feature	6.43	154	0.04175
2	POS	Main feature	1.52	47	0.03234
3	dt	Content feature	20.13	1172	0.01718
4	tt	Content feature	27.63	7395	0.00374
5	Stem	Main feature	35.45	11016	0.00322

reported in this paper is helpful for further improving this system. We also believe that our strategy will also serve as an example for feature selection in order to achieve enhanced performance for machine learning systems in general.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Research described in this paper was supported in part by grant received from the General Research Fund of the University Grant Council of the Hong Kong Special Administrative Region, China (Project no. CityU 142711), City University of Hong Kong (Project nos. 6354005, 7004091, 9610283, 7002793, 9610226, 9041694, and 9610188), the Fundamental Research Funds for the Central Universities of China (Project no. 2013PY120), and the National Natural Science Foundation

of China (Grant no. 61202305). The authors would also like to acknowledge supports received from the Dialogue Systems Group, Department of Chinese, Translation and Linguistics, and the Halliday Center for Intelligent Applications of Language Studies, City University of Hong Kong. The authors wish to express their gratitude to the anonymous reviewers for their constructive and insightful suggestions.

References

[1] R. Zhang, M. J. Cairelli, M. Fiszman et al., "Using semantic predications to uncover drug—drug interactions in clinical data," *Journal of Biomedical Informatics*, 2014.

[2] W. Liu, K. Miao, G. Li, K. Chang, J. Zheng, and J. C. Rajapakse, "Extracting rate changes in transcriptional regulation from MEDLINE abstracts," *BMC Bioinformatics*, vol. 15, 2, p. s4, 2014.

[3] B. Shen, "Translational biomedical informatics in the cloud: present and future," *BioMed Research International*, vol. 2013, Article ID 658925, 8 pages, 2013.

[4] R. Patra and S. K. Saha, "A kernel-based approach for biomedical named entity recognition," *The Scientific World Journal*, vol. 2013, Article ID 950796, 7 pages, 2013.

[5] P. Jindal, *Information Extraction for Clinical Narratives*, University of Illinois at Urbana-Champaign, Champaign, Ill, USA, 2014.

[6] T. Hao, A. Rusanov, M. R. Boland, and C. Weng, "Clustering clinical trials with similar eligibility criteria features," *Journal of Biomedical Informatics*, 2014.

[7] G. K. Mazandu and N. J. Mulder, "Information content-based gene ontology semantic similarity approaches: toward a unified framework theory," *BioMed Research International*, vol. 2013, Article ID 292063, 11 pages, 2013.

[8] J. Xia, X. Zhang, D. Yuan, L. Chen, J. Webster, and A. C. Fang, "Gene prioritization of resistant rice gene against Xanthomas oryzae pv. oryzae by using text mining technologies," *BioMed Research International*, vol. 2013, Article ID 853043, 9 pages, 2013.

[9] R. Faiz, M. Amami, and A. Elkhlifi, "Semantic event extraction from biological texts using a kernel-based method," in *Advances in Knowledge Discovery and Management*, pp. 77–94, Springer, Berlin, Germany, 2014.

[10] J. D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP '09 shared task on event extraction," in *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL, 2009 Workshop*, pp. 1–9, 2009.

[11] J. D. Kim, S. Pyysalo, T. Dhta, R. Bossy, N. Nguyen, and J. Tsujii, "Overview of BioNLP shared task 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 1–6, Portlan, Ore, USA, 2011.

[12] J. D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC Bioinformatics*, vol. 9, article 10, 2008.

[13] J. Bjorne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski, "Extracting complex biological events with rich graph-based feature sets," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 10–18, 2009.

[14] S. Riedel, D. Mcclosky, M. Surdeanu, A. Mccallum, and C. D. Manning, "Model combination for event extraction in BioNLP," in *Proceedings of the BioNLP Shared Task Workshop*, pp. 51–55, 2011.

[15] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski, "Extracting contextualized complex biological events with rich graph-based feature sets," *Computational Intelligence*, vol. 27, no. 4, pp. 541–557, 2011.

[16] J. Bjorne, G. Filip, and T. Salakoski, "University of Turku in the BioNLP '11 Shared Task," *BMC Bioinformatics*, vol. 13, 11, p. s4, 2012.

[17] J. Bjorne, S. V. Landeghem, S. Pyysalo et al., "Scale event extraction for post translational modifications, epigenetics and protein structural relations," *Proceedings of BioNLP*, pp. 82–90, 2012.

[18] J. Bjorne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Scaling up biomedical event extraction to the entire PubMed," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (ACL '10)*, pp. 28–36, 2010.

[19] J. Bjorne and T. Salakoski, "Generalizing biomedical event extraction," in *Proceedings of the BioNLP shared Task Workshop*, pp. 183–191, 2011.

[20] J. Bjorne and T. Salakoski, "TEES 2.1: automated annotation scheme learning in the BioNLP Shared Task," in *Association For Computational Linguistics (ACL '13)*, pp. 16–25, 2013.

[21] J. Zou, C. Lippert, D. Heckerman, M. Aryee, and J. Listgarten, "Epigenome-wide association studies without the need for cell-type composition," in *Nature Methods*, vol. 11, pp. 309–311, 2014.

[22] S. Fong, K. Lan, and R. Wong, "Classifying human voices by using hybrid SFX time-series preprocessing and ensemble feature selection," *BioMed Research International*, vol. 2013, Article ID 720834, 27 pages, 2013.

[23] D. Wang, L. Yang, Z. Fu, and J. Xia, "Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction," *Protein and Peptide Letters*, vol. 18, no. 7, pp. 684–689, 2011.

[24] S. O. Cao and J. H. Manton, "Unsupervised optimal discriminant vector based feature selection method," *Mathematical Problems in Engineering*, vol. 2013, Article ID 396780, 7 pages, 2013.

[25] L. Carlos Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: a survey and experimental evaluation," in *Proceedings of the IEEE International Conference on Data Mining (ICDM '02)*, pp. 306–313, December 2002.

[26] C. Yang, W. Zhang, J. Zou, S. Hu, and J. Qiu, "Feature selection in decision systems: a mean-variance approach," *Mathematical Problems in Engineering*, vol. 2013, Article ID 268063, 8 pages, 2013.

[27] S. van Landeghem, T. Abeel, Y. Saeys, and Y. van de Peer, "Discriminative and informative features for biomolecular text mining with ensemble feature selection," *Bioinformatics*, vol. 26, no. 18, Article ID btq381, pp. i554–i560, 2010.

[28] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[29] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 765648, 11 pages
http://dx.doi.org/10.1155/2014/765648

Research Article

A Novel Bioinformatics Method for Efficient Knowledge Discovery by BLSOM from Big Genomic Sequence Data

Yu Bai,[1] Yuki Iwasaki,[2] Shigehiko Kanaya,[1] Yue Zhao,[3] and Toshimichi Ikemura[2]

[1] *Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan*
[2] *Department of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama-shi, Shiga-ken 526-0829, Japan*
[3] *Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang, Liaoning 110004, China*

Correspondence should be addressed to Yuki Iwasaki; b105023@nagahama-i-bio.ac.jp

Received 24 October 2013; Accepted 14 February 2014; Published 3 April 2014

Academic Editor: Md. Altaf-Ul-Amin

With remarkable increase of genomic sequence data of a wide range of species, novel tools are needed for comprehensive analyses of the big sequence data. Self-Organizing Map (SOM) is an effective tool for clustering and visualizing high-dimensional data such as oligonucleotide composition on one map. By modifying the conventional SOM, we have previously developed Batch-Learning SOM (BLSOM), which allows classification of sequence fragments according to species, solely depending on the oligonucleotide composition. In the present study, we introduce the oligonucleotide BLSOM used for characterization of vertebrate genome sequences. We first analyzed pentanucleotide compositions in 100 kb sequences derived from a wide range of vertebrate genomes and then the compositions in the human and mouse genomes in order to investigate an efficient method for detecting differences between the closely related genomes. BLSOM can recognize the species-specific key combination of oligonucleotide frequencies in each genome, which is called a "genome signature," and the specific regions specifically enriched in transcription-factor-binding sequences. Because the classification and visualization power is very high, BLSOM is an efficient powerful tool for extracting a wide range of information from massive amounts of genomic sequences (i.e., big sequence data).

1. Introduction

Genome sequences, both protein coding and non-coding parts of the sequences, contain a wealth of information. The G + C content (G + C%) is a fundamental characteristic of individual genomes and used for a long period as a basic phylogenetic parameter to characterize inter- and intragenomic differences. The G + C%, however, is too simple to differentiate wide varieties of genomes. Many groups have reported that the oligonucleotide composition, which is an example of high-dimensional data, varies significantly among genomes and can be used to study genome diversity [1–9], and the oligonucleotide compositions, including dinucleotide composition, are called the "genome signature" of each species. Various linguistic tools for analyzing DNA sequence have been developed [8, 9]. Unsupervised neural network algorithm, Kohonen's Self-Organizing Map (SOM), is a powerful tool for clustering and visualizing high-dimensional complex data on a two-dimensional map [10–12]. On the basis of batch learning SOM, we have previously developed a modification of the conventional SOM for genome and gene sequence analyses, which makes the learning process and resulting map independent of the order of data input: BLSOM [13–15]. Importantly, BLSOM is suitable for actualizing high-performance parallel-computing and, therefore, can analyze big sequence data such as millions of genomic sequences simultaneously [16].

When we constructed BLSOMs for di-, tri-, and tetranucleotide composition in 10 kb genomic sequences derived from a wide range of prokaryotic and eukaryotic genomes, the sequences were clustered (i.e., self-organized) according to species without any information regarding the species during the BLSOM calculation, and increasing the length of the oligonucleotides from di- to tetranucleotides increased the clustering power [15]. An apparent causative factor for the genome signature is the context-dependent DNA mutation

and repair mechanisms. It should also be noted that oligonucleotides especially longer than trinucleotides often represent motif sequences responsible for sequence-specific protein binding (e.g., transcription factor binding). The occurrence of such motif oligonucleotides in the genome should differ from the level expected from the mononucleotide composition in the respective genome and may differ among genomic portions of one genome. We have recently found that DegPenta and DegHexa for the human genome can effectively detect characteristic occurrence patterns of many transcription-factor-binding motifs in pericentromeric heterochromatin regions [17].

In the present study, in order to clarify vertebrates' genome signatures, we first analyzed pentanucleotide compositions in 100 kb genomic sequences derived from a wide range of vertebrates and then those from human and mouse genomes in order to investigate the power to detect differences between the closely related genomes.

2. Materials and Methods

2.1. BLSOM. BLSOM is an unsupervised neural network algorithm that implements a characteristic nonlinear projection from the high-dimensional space of input data onto a two-dimensional array of weight vectors [10, 12]. We have previously modified the conventional SOM for genome informatics to make the learning process and resulting map independent of the order of data input and established BLSOM [13–15]. Here, we explain the BLSOM method developed by Kanaya et al. [13].

In the original Kohonen's SOM, the initial vectorial data were set by random values, but in the BLSOM the initial vectors are set based on the widest scale of the sequence distribution in the oligonucleotide frequency space with the principal component analysis (PCA) [13]. Weights in the first dimension (I) were arranged into lattices corresponding to a width of five times the standard deviation ($5\sigma_1$) of the first principal component: the second dimension (J) was defined by the nearest integer greater than $\sigma_2/\sigma_1 \times I$; and I was set in the present study as the average number of sequence data per neuron which becomes approximately four. σ_1 and σ_2 were the standard deviations of the first and second principal components, respectively. The weight vector on the ijth lattice ($\mathbf{w}_{ij}$) was represented as follows:

$$\mathbf{w}_{ij} = \mathbf{x}_{\text{av}} + \frac{5\sigma_1}{I}\left[\mathbf{b}_1\left(i - \frac{I}{2}\right) + \mathbf{b}_2\left(j - \frac{J}{2}\right)\right], \quad (1)$$

where $\mathbf{x}_{\text{av}}$ is the average vector for oligonucleotide frequencies of all input vectors, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are eigenvectors for the first and second principal components. In Step 2, the Euclidean distances between the input vector $\mathbf{x}_k$ and all weight vectors $\mathbf{w}_{ij}$ were calculated; then $\mathbf{x}_k$ was associated with the weight vector (called $\mathbf{w}_{i'j'}$) with minimal distance. After associating all input vectors with weight vectors, updating was done according to Kanaya et al. [13].

BLSOM learning for oligonucleotide composition was conducted as described previously [15]. BLSOM program was obtained from Niigata Univ. (takaabe@ie.niigata-u.ac.jp) or UNTROD, Inc. (y_wada@nagahama-i-bio.ac.jp).

2.2. U-Matrix. Distances of weight vectors between neighboring lattice points on BLSOM can be visualized as black levels with a U-matrix method [19], and this provides information regarding similarity of oligonucleotide composition in local areas on BLSOM; the areas composed of lattice points with similar or distinct oligonucleotide composition can be recognized as low or high black level, respectively.

2.3. Genome Sequences. Genome DNA sequences were obtained from UCSC ftp site (http://www.ncbi.nlm.nih.gov/genomes/). When the number of undetermined nucleotides (Ns) in a fragment sequence (e.g., 100 kb) exceeded 20% of the sequence, the sequence was omitted from the analysis. In the case where the number of Ns was less than 20%, the oligonucleotide frequencies were normalized to the length without Ns and included in the analysis.

3. Results

3.1. Characteristics of BLSOM Clustering. In the era of extensive genome sequencing, it is important to develop novel bioinformatics tools to support an efficient knowledge discovery from massive amounts of genomic sequences. Analyses on the species-specific oligonucleotide composition "genome signature" (e.g., penta- and hexanucleotide compositions) may provide *in silico* information concerning important signal sequences such as transcription factor binding sequences [17]. To show the clustering ability of BLSOM for vertebrate genome sequences and to explain the basal features of BLSOM clustering patterns, we first analyze pentanucleotide compositions in 100 kb sequence fragments derived from 10 vertebrate genomes.

In DNA databases, only one strand of each pair of complementary sequences is registered. Previous analysis of prokaryotic species that was done by Abe et al. [15] revealed that sequences (e.g., 10 kb sequences) from a single prokaryotic genome were often split vertically into two territories according to the transcriptional direction of the genes present in the fragment. However, to study general characteristics of genomic sequences such as the genome signature, differences in the oligonucleotide composition between two complementary strands are not necessarily important. Therefore, we construct a BLSOM in which the frequencies of a pair of complementary pentanucleotides (e.g., AAAAC and GTTTT) in each fragment are summed [17]. The BLSOM for this degenerate set of a pair of complementary pentanucleotides is designated as DegPenta.

On the BLSOM, lattice points containing sequences from a single species are indicated in a color specifying the species; those containing sequences from multiple species are indicated in black. Because most lattice points are colored, a high separation power is apparent for DegPenta (Figure 1(a)), with no information concerning species during the BLSOM calculation. We next explain the basal characteristics of BLSOM separation observed for the vertebrate sequences. G + C% has long been used as a fundamental value that characterizes both inter- and intragenomic differences. For example, on a warm-blooded vertebrate genome, there exists a long-range

segmental G + C% distribution "isochores," which have been connected with chromosomal bands [7, 20–23]. Figure 1(b) presents the G + C% that is calculated from pentanucleotide composition at each lattice point in the DegPenta. Sequences with high and low G + C% (wine red or green in Figure 1(b)) are located on the left and right side of the map, respectively, showing that the G + C% level is reflected primarily in the horizontal direction. The territory of each species is often split into several subterritories, which should relate at least in part to isochore structures because the G + C% level differs between subterritories of a single species, for example, chicken and human territories.

BLSOMs can visualize diagnostic oligonucleotides responsible for species-specific clustering (self-organization). We first calculate the pentanucleotide frequencies expected from the mononucleotide composition that is obtained from the vectorial data (i.e., pentanucleotide composition) at each lattice point and indicate the observed/expected ratio as follows: red (overrepresented), blue (underrepresented), and white (moderately represented) (Figure 1(c)). This observed/expected ratio is useful in unveiling genome signatures, since it allows us to examine the oligonucleotide composition at each lattice point, independently of a simple effect derived from its mononucleotide composition [17]. For various pentanucleotides, transitions between red and blue often coincide exactly with species-specific territory borders. AACAT + ATGTT, ACAAC + GTTGT, ATTTA + TAAAT, and CAGCG + CGCTG are overrepresented in fishes (Fugu, Medaka, Stickleback, Tetraodon, and Zebrafish) but not in almost all tetrapods (Human, Lizard, Mouse, Chicken, and Xenopus). ACCCT + AGGGT and CCAAG + CTTGG are overrepresented in tetrapods but not in fishes. AACCC + GGGTT is underrepresented in chicken and a part of fish (Fugu, Stickleback, and Tetraodon). GAAGA + TCTTC is underrepresented in Xenopus and Zebrafish. These findings show that BLSOM can recognize the species-specific oligonucleotide composition and identify the combinatorial diagnostic oligonucleotides responsible for species-specific clustering; that is, a combination of not a few but many pentanucleotides contributes to the accurate clustering (self-organization) of genomic sequences according to species.

3.2. BLSOMs for Human and Mouse Genomes. We have next constructed DegPenta with 100 kb sequences derived from the human and mouse genomes (Figure 2). This enables us to examine a BLSOM power for separating the species with a relatively close phylogenetic relationship and to clarify difference in the genome signatures of the closely related species. Lattice points that contain sequences derived from human and mouse are indicated in red and blue, respectively, and those that include sequences from both human and mouse are indicated in black. With no information regarding species during the BLSOM calculation, the species-specific clustering (self-organization) of the 100 kb sequences is clear.

In Figure 2(b), the observed/expected ratios of individual pentanucleotides calculated as explained in Figure 1(c) are illustrated in red (overrepresented), blue (underrepresented), and white (moderately represented). Transitions between red (overrepresentation) and blue (underrepresentation) for various pentanucleotides often coincide exactly with species territory borders, showing that BLSOM recognizes the species-specific combination of oligonucleotide frequencies that is the representative signature of one genome and enables us to identify the frequency patterns that are characteristics of individual genomes.

Seven examples of the pentanucleotides diagnostic for the species territory formation are presented (Figure 2(b)). AAATT + AATTT, ATCAC + GTGAT, and TTCAA + TTGAA are preferred in the human genome but not in the mouse genome. On the other hand, AACAC + GTGTT, ACAAC + GTTGT, ACAAG + CTTGT, and ACACT + AGTGT are preferred in the mouse genome but not in the human genome. It should be stressed that a complex combination of many pentanucleotides contributes to the species-specific clustering (i.e., self-organization) of sequence fragments. Some of these diagnostic pentanucleotides, if not all, may have important biological significances, which should be related to functions.

3.3. Characteristics of Sequences Belonging to Specific Zones. While most 100 kb sequences are classified primarily into species-specific territories, there are a few minor human zones (red) that are located within the mouse territory (blue) and are surrounded with white lattice points. In addition, there is a nub-type human zone that is located in the border region between human and mouse territories and also is surrounded by white lattice points. In Figure 2(a), lattice points with no genomic sequence assigned after the BLSOM calculation are left white. It should be mentioned that Abe et al. [15] and Iwasaki et al. [17] have previously shown that lattice points containing genomic sequences whose oligonucleotide composition is very distinct from other genomic sequences tend to be surrounded by lattice points containing no genomic sequence.

Similarity in oligonucleotide composition between neighboring lattice points in BLSOM (and thus between sequences belonging to neighboring lattice points) can be visualized using a U-matrix [19] with a level of blackness (Figure 3(a)), as described in Section 2. On the U-matrix, borders between human and mouse territories are visualized as black lines, which represent distinct pentanucleotide compositions between human and mouse sequences. Furthermore, there are small dark black zones and gray zones surrounded by a black circle, which should contain sequences with peculiar oligonucleotide composition distinct from the compositions from other genomic sequences; the respective zones composed of human sequences are numbered as Sz-H1 and Sz-H2 and that of mouse sequences is specified as Sz-M (Figure 3(b)). Importantly, these numbered zones primarily correspond to zones surrounded by white lattice points in Figure 2(a), confirming that the sequences in these specific zones have peculiar oligonucleotide compositions very distinct from a major portion of the respective genome. Actually, occurrence levels of individual pentanucleotides in the specific zones are clearly different from those in the major portion of the respective genome (Figure 2(b)). AATCT + AGATT and AGATA + TATCT are preferred in Sz-H2 but not

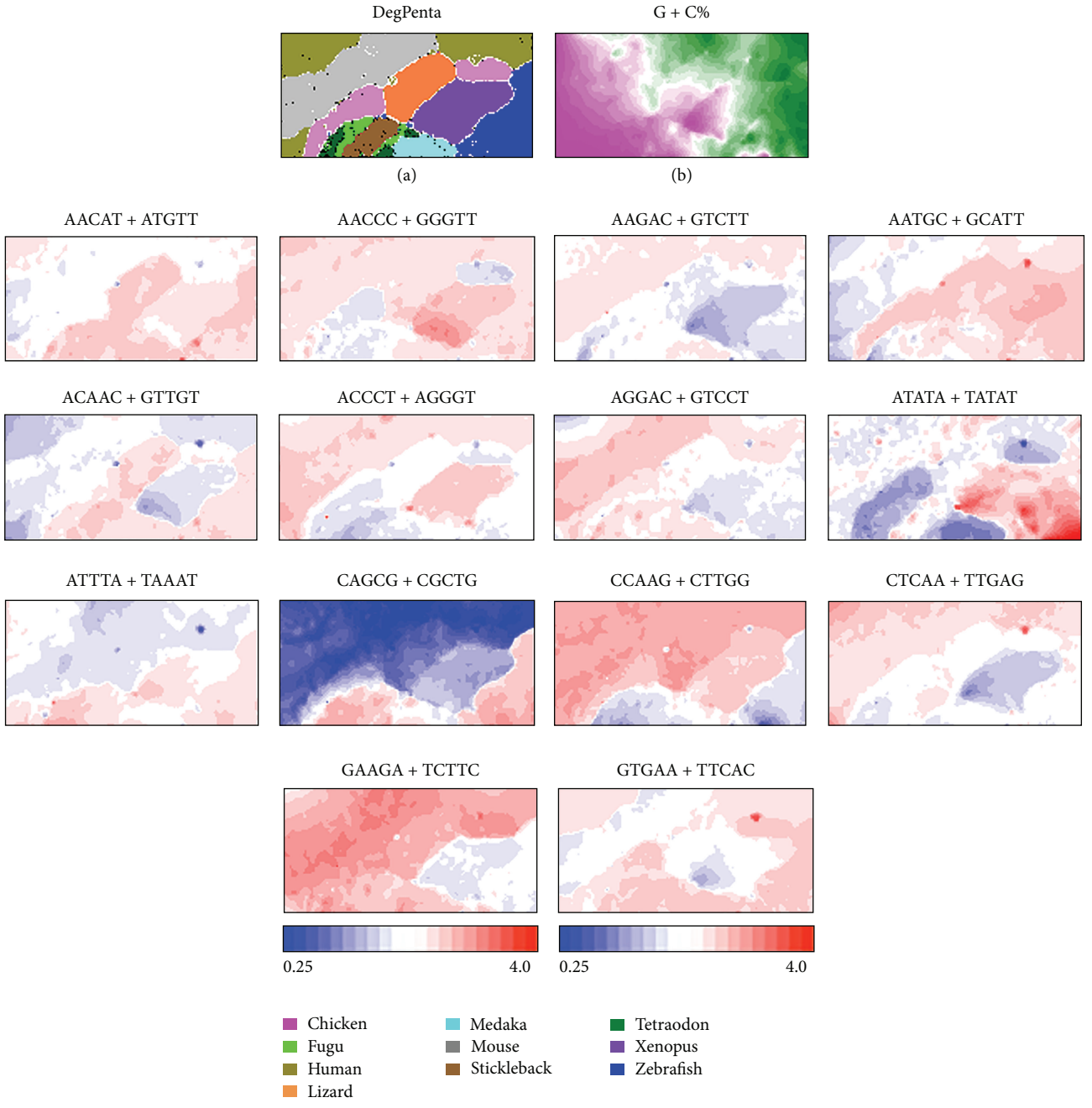


Figure 1: BLSOMs for 100 kb sequences derived from 10 vertebrata genomes. (a) DegPenta. Lattice points containing sequences from multiple species are indicated in black and those containing sequences from a single species are indicated in color as shown in the keys. (b) G + C%. For each lattice point in the DegPenta, G + C% was calculated and divided into 21 categories with an equal number of lattices. The lattice points belonging to the categories of the highest, middle, and lowest G + C% are shown in wine red, white, and green, respectively. (c) Diagnostic pentanucleotides responsible for species-specific clustering. Occurrence of each pentanucleotide for each lattice point was calculated and normalized with occurrence expected from the mononucleotide composition for the respective lattice point [16, 17]. This observed/expected ratio is indicated in color presented under the panel. This ratio has been shown to be useful in unveiling genome signatures because the oligonucleotide composition can be analyzed independently of a simplex effect reflecting the mononucleotide composition of genomic sequences [16–18].

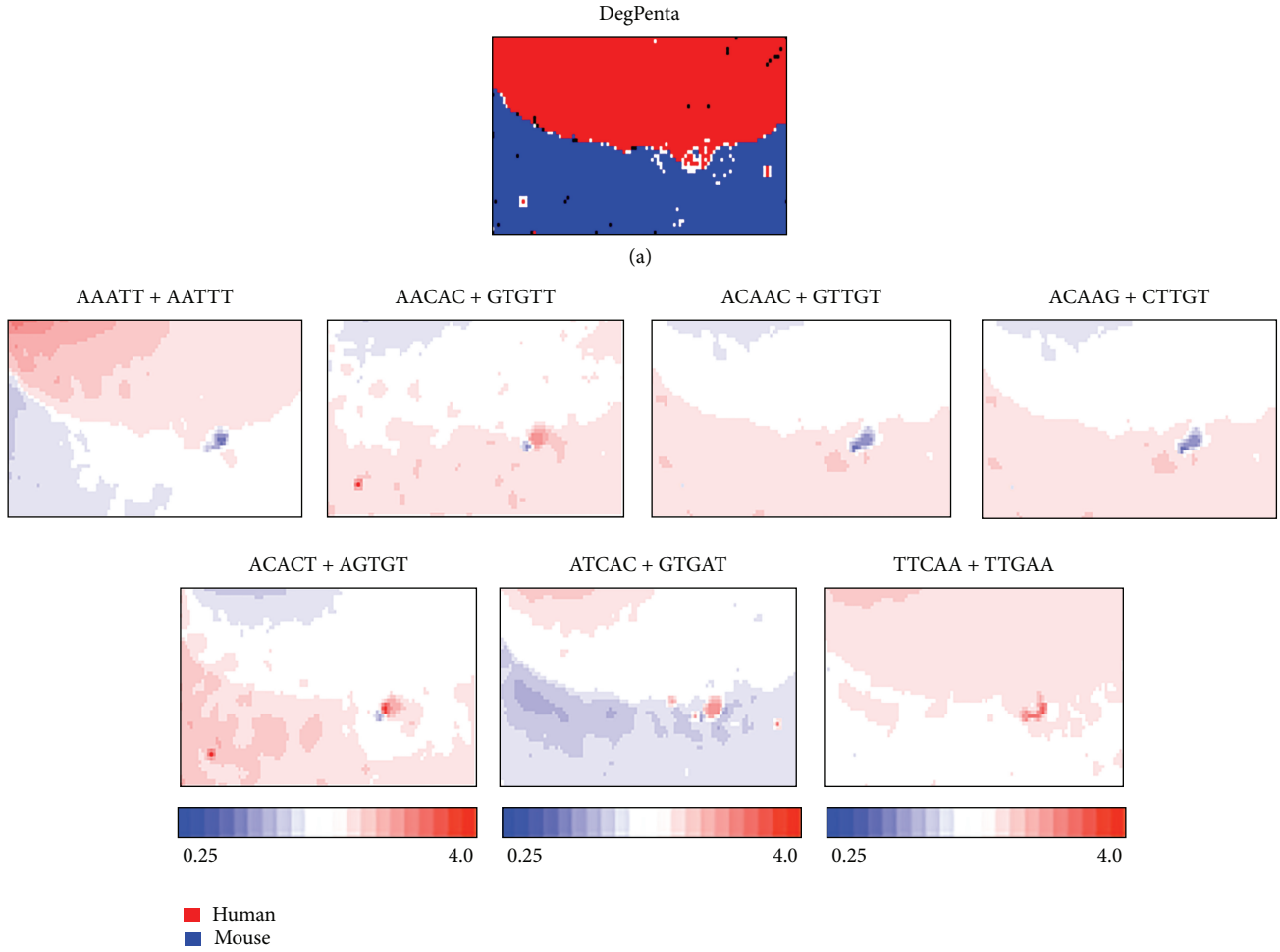


Figure 2: BLSOMs for 100 kb sequences derived from the human and mouse genomes. (a) DegPenta. Lattice points containing sequences from human and mouse are indicated in black and those containing sequences from a single species are indicated in color as shown in the keys. (b) Occurrence level of each pair of complimentary pentanucleotides in the DegPenta. Level of a complimentary pentanucleotide pair for each lattice point is calculated and normalized with the level expected from the mononucleotide composition for the lattice point. The observed/expected ratio is indicated in colors presented at the bottom of the figure. Seven examples of the pentanucleotides diagnostic for species-specific separations are presented.

in Sz-H1 and Sz-M. ATTGA + TCAAT is preferred in Sz-H1 and Sz-H2 but not in Sz-M. ATTGG + CCAAT is preferred in Sz-H1 but not in Sz-H2 and Sz-M. The pentanucleotides listed in Figure 2(b) correspond to human transcription-factor-binding (TFB) motifs and the reason why these motif pentanucleotides are chosen is explained below.

The oligonucleotides such as penta- and hexanucleotides often provide the binding sites of proteins such as transcription factors. When we consider the oligonucleotides that can function as important signal sequences such as TFB motifs, their occurrence levels in genomic sequences should be biased significantly from the levels expected from random sequences. Therefore, the overrepresentation of a certain oligonucleotide only in a restricted portion of the BLSOM (and thus of the genomic sequences) is thought to provide useful information for understating the biological significance of the respective sequence, especially when a biological function of the oligonucleotide of interest is known.

In our previous study [17], we have shown that oligonucleotide BLSOM such as DegPenta can be used for studying sequences derived even from one genome. In that study, addition of computer-generated random sequences to real human sequences can successfully separate the specific sequences with distinct oligonucleotide composition from a major portion of the human genome; that is, these specific sequences are displaced well from the major portion of human sequences and surrounded by the random sequences. Interestingly, the specific human sequences thus found are derived mainly from pericentromeric regions and enriched by TFB motif sequences [17]. Instead of the human plus random sequences used in the previous study, human plus mouse sequences are analyzed in the present study, and the addition of the closely related species appears to effectively assign the 100 kb sequences with peculiar oligonucleotide compositions very distinct from those in the major portion of the respective genome (Figure 2(a)). In order to clarify

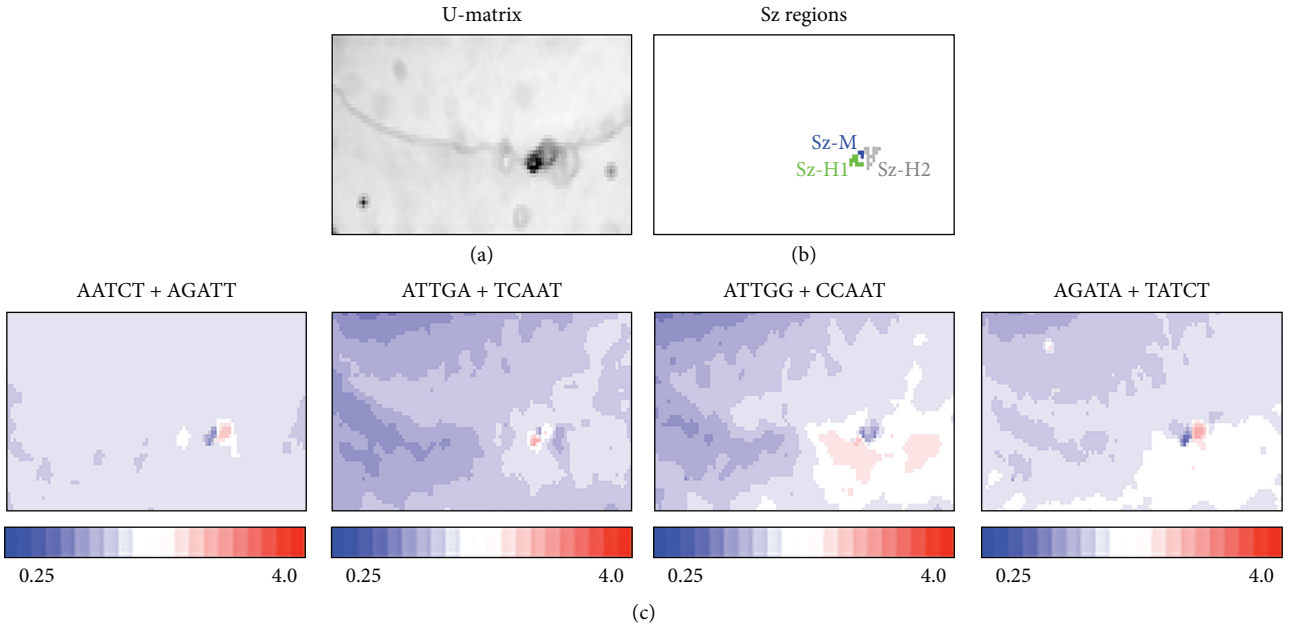


Figure 3: Pentanucleotides specifically enriched in specific region. (a) *U*-matrix for the BLSOM listed in Figure 2(a). (b) Sz regions. Sz-H1 and Sz-H2 regions of human sequences are indicated in green and gray letters, respectively. A very tiny Sz-M region of mouse sequences is indicated in blue. (c) Pentanucleotides specifically enriched in Sz. The observed/expected ratio for each pair of complimentary pentanucleotides is calculated as described in Figure 1(c) and indicated in colors presented under the panel.

the characteristics of the specific sequences found in this study and to compare with those found previously, we have analyzed the occurrence of the pentanucleotides corresponding to human TFB motifs analyzed in the previous paper. All of the TFB pentanucleotides are overrepresented (red) in a certain specific zone but underrepresented (blue) in almost all other human 100 kb sequences, confirming the previous result. When we examine their occurrences in the mouse territory, AATCT + AGATT, ATTGA + TCAAT, and TATCA + TGATA are underrepresented in a major portion of the mouse genome. However, ATTGG + CCAAT and AGATA + TATCT are underrepresented only in a half portion of the mouse genome, indicating that the biological function of these two pentanucleotides may differ from that for human. Comparative analyses of the closely related species can provide this type of information concerning a possible evolutionary change in functional signal sequences such as TFB motif sequences, but the addition of the computer-generated ransom sequences cannot provide the information concerning the molecular evolutionary change. The reason why the specific zones of mouse on DegPenta are less evident than human will be discussed below.

4. Discussion

4.1. Repeat and Unique Sequences. Vertebrate genomes are composed of repeat and nonrepeat, unique sequences, which have distinct biological functions. Since repeat sequences usually have peculiar oligonucleotide composition, there exists a possibility that the specific zones' sequences with peculiar oligonucleotide compositions distinct from a major portion of the genome are repeat sequences, and this possibility is examined as follows. In the UCSC database, repeat sequences identified by RepeatMasker and Tandem Repeats Finder are specified in lower-case letters for distinguishing from unique sequences specified in upper-case letters. We first concatenated unique or repeat sequences separately, divided these concatenated sequences into 100 kb sequences, and counted pentanucleotide composition in each 100 kb sequence.

Clear separation between species and between repeat and unique sequences is observed on DegPenta (Figure 4(a)). Interestingly, human repeat sequences (pink) forms one satellite-type minor territory located at the lowest part of the map and the mapping of the specific zones' sequences marked in Figure 3(b) shows that these specific sequences are mainly located in the minor territory of human repeat sequences (Figure 4(b)). Therefore, the specific sequences actually belong to the repeat category. However, it has been separately shown that these specific sequences are different from the ubiquitously distributed human repeat sequences such as Alu and LI (our unpublished data). As another separate analysis, we have found that these sequences are also different from alphoid sequences, which are a major component of human centromeric regions. Core parts of human centromeric regions mainly composed of alphoid sequences have not been included in the standard human genome sequences currently available, because of the difficulty to get contiguous sequences. The minor human territory of interest is colored in black on *U*-matrix (Figure 4(c)) and appears to be split into two parts: a very dark small part and its adjacent gray part.

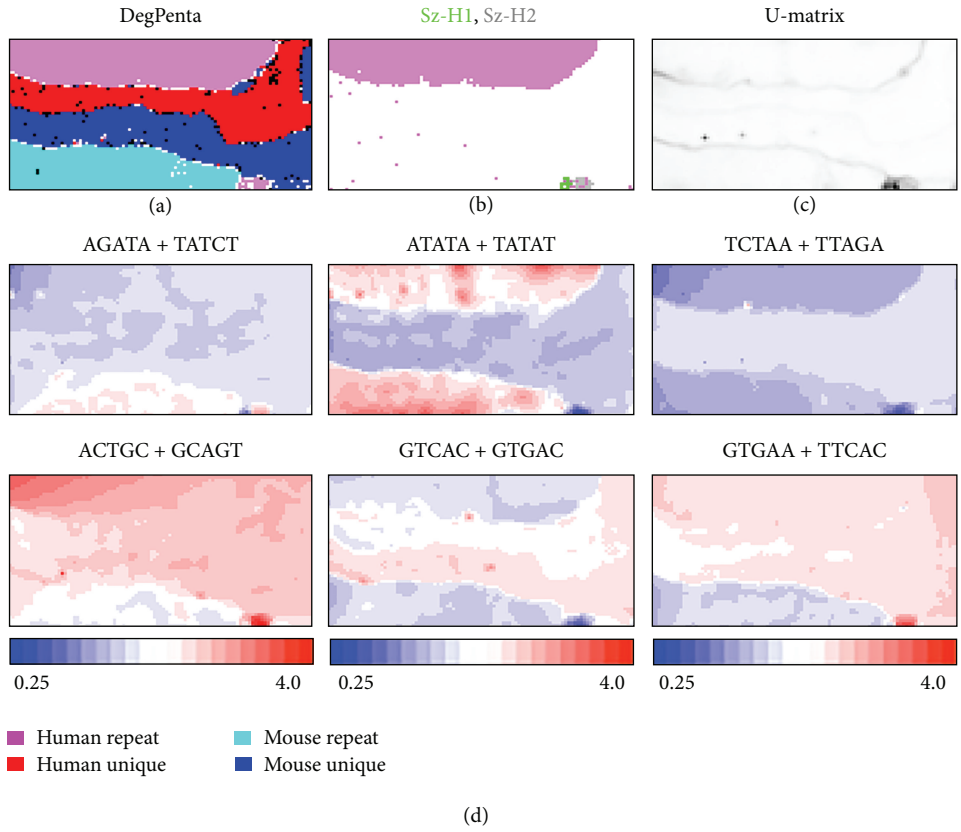


Figure 4: BLSOMs for 100 kb repeat and unique sequences derived from the human and mouse genomes. (a) DegPenta. Lattice points containing sequences from more than one category are indicated in black and those containing sequences from a single category are indicated in color as shown in the keys. (b) Human repeat sequences. Human Sz-H1 and Sz-H2 sequences defined in Figure 3(b) are mapped and indicated in green and gray, respectively. (c) *U*-matrix for the DegPenta listed in (a). (d) Diagnostic pentanucleotides responsible for species-specific clustering. The observed/expected ratio for each pair of complimentary pentanucleotides is calculated as described in Figure 1(c) and indicated in color presented under the panel.

In Figure 4(d), we list six examples of pentanucleotides (including a TFB motif) diagnostic for separation between species and/or between repeat and unique sequences. Interestingly, all pentanucleotides show a very high (dark red) or very low (dark blue) occurrence in the minor human repeat territory, again showing its very peculiar oligonucleotide composition. The specific characteristics in this minor repeat territory are further confirmed when we examine the dinucleotide CG-containing pentanucleotides (Figure 5). All these CG-containing pentanucleotides listed are specifically overrepresented in the very dark, small region visualized in *U*-matrix (Figure 4(c)) but evidently underrepresented in all other regions. When we examine the occurrence of all CG-containing pentanucleotides in detail, the CGA-containing pentanucleotides are particularly enriched in the very dark, small region in *U*-matrix, and almost all examples listed in Figure 5 correspond to the CGA-containing pentanucleotides.

The evident underrepresentation of the CG dinucleotide (i.e., CG suppression) is well known in vertebrate genomes and the CG suppression is believed to relate to methylation at CG dinucleotide, which is a well-characterized epigenetic marker. Concerning the CG occurrence level, CpG islands, in which the CG occurrence is clearly higher than in other genomic regions, are well known to have important roles in transcriptional regulation. The sizes of the CpG islands are known to be a few or several hundred bp and, therefore, are clearly different from the size of specific sequences found in the present study (a 100 kb level). Furthermore, CpG islands belong primarily to the unique sequence regions. Therefore, the 100 kb level sequences enriched with the CG-containing pentanucleotides are not the CpG island sequences. As noted above, the CG dinucleotide is a target of methylation and this C methylation is known to have important roles in epigenetic systems. The 100 kb level specific sequences may have important roles that are different from but possibly related to the function of CpG islands. The finding that the CGA-containing pentanucleotides are more preferred in the specific sequences than other CG-containing pentanucleotides may give information for clarifying biological functions of the 100 kb level sequences of interest.

4.2. Possible Biological Functions of Sequences with Peculiar Oligonucleotide Composition. As a separate analysis, we have examined the chromosomal locations of sequences belonging to the human specific zones and found a major portion of

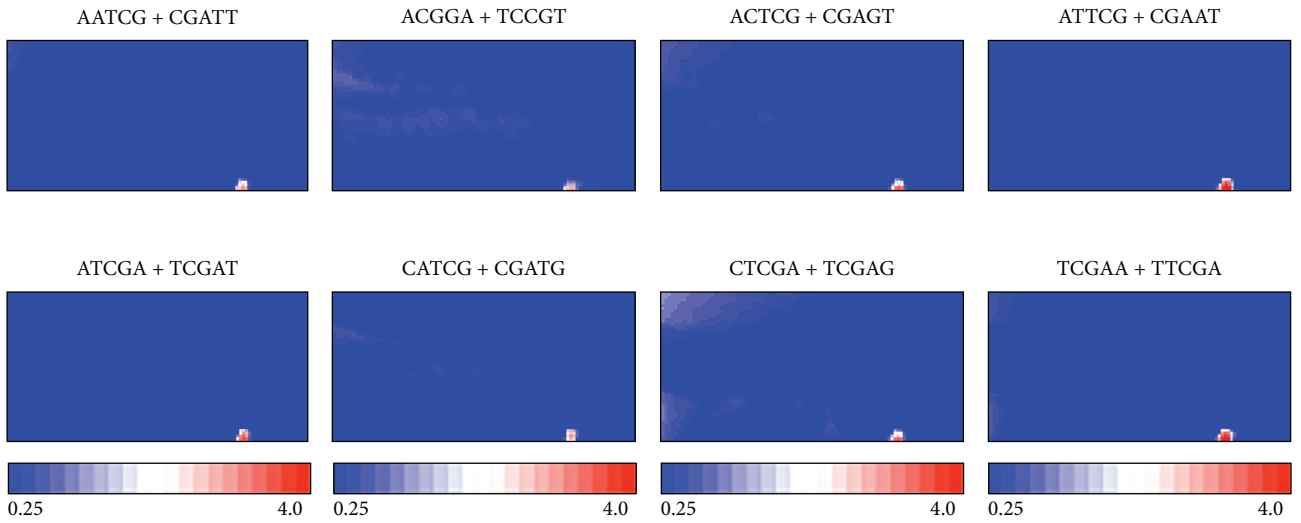


Figure 5: CG-containing pentanucleotides enriched in Sz region. The observed/expected ratio for each pair of complimentary pentanucleotides is calculated as described in Figure 1(c) and indicated in color presented under the panel. The small red region for each pair of complimentary pentanucleotides corresponds to the dark black region in the U-matrix listed in Figure 4(c).

these 100 kb specific sequences to be derived from pericentromeric heterochromatin regions (data not shown), as supporting the previous finding [17]. Pericentromeric regions form the heterochromatin structure "chromocenter" in interphase nuclei. Chromocenter was once thought to be stable in composition and transcriptionally inert but has recently been shown to be surprisingly dynamic [24–28]. Mouse centromere-derived double-stranded transcripts appear to be involved in establishing the heterochromatin structure [24], and Dicer-related RNA interference machinery is involved in the formation of the centromeric heterochromatin structure in higher vertebrate cells [29]. A strand-specific burst in transcription of mouse pericentromeric satellites is required for chromocenter formation during early mouse development [27], and long nuclear noncoding RNA transcribed from the periphery of pericentromeric heterochromatin has recently been reported [30]. Because the centromere RNA has been shown to be a key component for the assembly of nucleoproteins at the nucleolus and centromere [31, 32], the notable clustering of TFB motifs in the pericentromeric regions should provide novel knowledge about the higher order of nuclear organization.

In Figure 2, specific regions are mainly observed for the human genomic sequences. This appears to be related to the finding that the human specific sequences are mainly derived from the pericentromeric heterochromatin regions. In the case of mice, their chromosomes are acrocentric and the highly repetitive sequences in their pericentromeric regions are less represented in the reported genome sequence than for the human genome. When more sequences of the mouse pericentromeric regions will become available, comparative analyses of their sequences should provide novel information concerning biological significance of 100 kb level sequences with the very peculiar oligonucleotide compositions. In the present study, we have analyzed 100 kb sequences, but the analyses of 50 kb sequences give similar results (data not shown).

4.3. Other Applications of BLSOM and Future Prospects. BLSOM can classify genomic sequences according to species with no information other than oligonucleotide frequencies. Because the classification and visualization power is very high, BLSOM is a powerful bioinformatics tool for extracting a wide range of information from a large amount of genomic sequences. A wide variety of oligonucleotide sequences function as genetic signals (e.g., regulatory signals for gene expression). We have found that occurrence levels of oligonucleotide sequences corresponding to important functional signals (e.g., TFB motif sequences) are often biased significantly from the occurrence levels found in a major portion of the human genome and are diagnostic for the specific zones visualized in Figures 2 and 3. When we systematically characterize in advance the known signal sequences of various species with enough experimental data with BLSOMs, we may develop an *in silico* method of signal prediction, which is most useful for genomes that are sequenced but for which little additional experimental data are available. Because the number of such genomes has increased rapidly, development of the *in silico* method has become increasingly important. Functional signals, such as transcription-regulatory signals, are typically longer than pentanucleotides, and therefore analyses of longer oligonucleotides become important. To conduct BLSOM with longer oligonucleotides such hexa- and heptanucleotides (4,096- and 16,384-dimensional data) for a massive amount of genome sequences currently available, a large-scale computation using a high-performance supercomputer will become essential, and the BLSOM algorithm is suitable for a high level of parallel computing.

One important application of BLSOM to genome informatics is the use for metagenome analyses. Most environmental microorganisms cannot be cultured easily under laboratory conditions. Genomes of uncultured organisms have remained mostly uncharacterized and are thought to contain a wide range of novel genes of scientific and industrial interest [33–38]. Metagenomic approaches, which are

analyses of mixed populations of uncultured microbes, have been developed to identify novel and industrially useful genes and to study microbial diversity in a wide variety of environments. With the metagenomic approach, genomic DNAs are extracted directly from an environmental sample containing multiple organisms, and the DNA fragments are cloned and sequenced. This is a powerful strategy for comprehensive analysis of biodiversity in an ecosystem. However, for a simple collection of many sequence fragments, the conventional phylogenetic method cannot predict from what phylotypes individual sequences are derived or the phylogenetic novelty of the individual sequences. Traditional methods of phylogenetic assignment have been based on sequence homology searches and therefore inevitably focused on well-characterized genes, for which orthologous sequences required for constructing a reliable phylogenetic tree are available. However, most of the well-characterized genes are not industrially attractive. BLSOM is an alignment-free clustering method, and thus is the most suitable method for this metagenomics analysis.

For phylogenetic classification of species-unknown sequences obtained from environmental and clinical samples, we have constructed BLSOMs in advance with all available sequences from species-known prokaryotes and eukaryotes, as well as from viruses and organelles, and found that the sequences are clustered (self-organized) according to phylotypes with high accuracy [16]. By mapping a large number of environmental metagenomic sequences on the large-scale BLSOM, we can predict phylotypes of these environmental sequences [39]. Because BLSOM does not require orthologous sequence sets, this alignment-free method can provide a systematic strategy for revealing microbial diversity and relative abundance of different phylotype members of uncultured microorganisms including viruses in an environmental sample [39]. Actually, as collaborative studies with experimental research groups, we have used the BLSOM for phylogenetic classification of genomic sequence fragments obtained from mixed genomes of uncultured microbes in environmental samples [18, 40, 41]. We have recently found that the addition of a large number of computer-generating random sequences can classify the metagenomic sequences according to phylotypes [42]. In addition, BLSOM with oligopeptide composition can classify protein sequences mainly according to function [18].

5. Conclusions

Because of the remarkable progress of various high-throughput measuring instruments, a massive amount of various data other than sequence data has been accumulated. Complex data can be represented by a high-dimensional multivariate data. BLSOM can analyze a massive amount of high-dimensional multivariate data because the algorithm is suitable for high-level parallel computing. BLSOM can support efficient knowledge discoveries from such big data, showing that the BLSOM is a timely bioinformatics method in the era of big data studies in bioscience. In the present study, we characterized vertebrate genomes using BLSOM. We first analyzed pentanucleotide compositions in 100 kb sequences derived from a wide range of vertebrate genomes and then the compositions in the human and mouse genomes in order to investigate a method for detecting differences between the closely related genomes. BLSOM can recognize the species-specific key combination of oligonucleotide frequencies in each genome, which is called a "genome signature," and the specific regions specifically enriched by transcription-factor-binding sequences.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Bioscience Database Center in Japan; the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Grant-in-Aid for Scientific Research on Innovation Areas "Biosynthetic Machinery. Deciphering and Regulating the System for Creating Structural Diversity of Bioactivity Metabolites (2007)"), the Grant-in-Aid for Scientific Research (C) and for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan; and the Grant-in-Aid for JSPS Fellows (JSPS KAKENHI no. 24•9979). The computation was done in part with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

[1] R. Nussinov, "Doublet frequencies in evolutionary distinct groups," *Nucleic Acids Research*, vol. 12, no. 3, pp. 1749–1763, 1984.

[2] G. J. Phillips, J. Arnold, and R. Ivarie, "Mono-through hexanucleotide composition of the *Escherichia coli* genome: a markov chain analysis," *Nucleic Acids Research*, vol. 15, no. 6, pp. 2611–2626, 1987.

[3] S. Karlin, "Global dinucleotide signatures and analysis of genomic heterogeneity," *Current Opinion in Microbiology*, vol. 1, no. 5, pp. 598–610, 1998.

[4] S. Karlin, A. M. Campbell, and J. Mrázek, "Comparative DNA analysis across diverse genomes," *Annual Review of Genetics*, vol. 32, pp. 185–225, 1998.

[5] E. P. C. Rocha, A. Viari, and A. Danchin, "Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons," *Nucleic Acids Research*, vol. 26, no. 12, pp. 2971–2980, 1998.

[6] A. J. Gentles and S. Karlin, "Genome-scale compositional comparisons in eukaryotes," *Genome Research*, vol. 11, no. 4, pp. 540–546, 2001.

[7] G. Bernardi, *Structural and Evolutionary Genomics: Natural Selection in Genome Evolution*, Elsevier, New York, NY, USA, 2004.

[8] S. Vinga and J. Almeida, "Alignment-free sequence comparison: a review," *Bioinformatics*, vol. 19, no. 4, pp. 513–523, 2003.

[9] A. Bolshoy, "DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity," *Appl Bioinformatics*, vol. 2, no. 2, pp. 103–112, 2003.

[10] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[11] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[12] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceedings of the IEEE*, vol. 84, no. 10, pp. 1358–1384, 1996.

[13] S. Kanaya, Y. Kudo, T. Abe, T. Okazaki, D. C. Carlos, and T. Ikemura, "Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome," *Genome Informatics Series. Workshop on Genome Informatics*, vol. 13, pp. 369–371, 1998.

[14] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency," *Genome Informatics Series. Workshop on Genome Informatics*, vol. 13, pp. 12–20, 2002.

[15] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for unveiling hidden genome signatures," *Genome Research*, vol. 13, no. 4, pp. 693–702, 2003.

[16] T. Abe, H. Sugawara, S. Kanaya, and T. Ikemura, "Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator," *Journal of the Earth Simulator*, vol. 6, pp. 17–23, 2006.

[17] Y. Iwasaki, K. Wada, Y. Wada, T. Abe, and T. Ikemura, "Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance," *Chromosome Research*, vol. 21, pp. 461–474, 2013.

[18] T. Abe, S. Kanaya, H. Uehara, and T. Ikemura, "A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses," *DNA Research*, vol. 16, no. 5, pp. 287–298, 2009.

[19] A. Ultsch, "Self organized feature maps for monitoring and knowledge acquisition of a chemical process," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN '93)*, S. Gielen and B. Kappen, Eds., pp. 864–867, Springer, London, UK, 1993.

[20] G. Bernardi, B. Olofsson, and J. Filipski, "The mosaic genome of warm-blooded vertebrates," *Science*, vol. 228, no. 4702, pp. 953–958, 1985.

[21] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms," *Molecular Biology and Evolution*, vol. 2, no. 1, pp. 13–34, 1985.

[22] T. Ikemura and S. Aota, "Global variation in G + C content along vertebrate genome DNA. Possible correlation with chromosome band structures," *Journal of Molecular Biology*, vol. 203, no. 1, pp. 1–13, 1988.

[23] T. Ikemura and K.-N. Wada, "Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data," *Nucleic Acids Research*, vol. 19, no. 16, pp. 4333–4339, 1991.

[24] C. Maison, D. Bailly, A. H. F. M. Peters et al., "Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component," *Nature Genetics*, vol. 30, no. 3, pp. 329–334, 2002.

[25] C. Maison and G. Almouzni, "HP1 and the dynamics of heterochromatin maintenance," *Nature Reviews Molecular Cell Biology*, vol. 5, no. 4, pp. 296–304, 2004.

[26] A. V. Probst, E. Dunleavy, and G. Almouzni, "Epigenetic inheritance during the cell cycle," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 3, pp. 192–206, 2009.

[27] A. V. Probst, I. Okamoto, M. Casanova, F. El Marjou, P. le Baccon, and G. Almouzni, "A Strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development," *Developmental Cell*, vol. 19, no. 4, pp. 625–638, 2010.

[28] A. V. Probst and G. Almouzni, "Heterochromatin establishment in the context of genome-wide epigenetic reprogramming," *Trends in Genetics*, vol. 27, no. 5, pp. 177–185, 2011.

[29] T. Fukagawa, M. Nogami, M. Yoshikawa et al., "Dicer is essential for formation of the heterochromatin structure in vertebrate cells," *Nature Cell Biology*, vol. 6, no. 8, pp. 784–791, 2004.

[30] C. Maison, D. Bailly, D. Roche et al., "SUMOylation promotes de novo targeting of HP1 ± to pericentric heterochromatin," *Nature Genetics*, vol. 43, no. 3, pp. 220–227, 2011.

[31] Y. Du, C. N. Topp, and R. K. Dawe, "DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA," *PLoS Genetics*, vol. 6, no. 2, Article ID e1000835, 2010.

[32] L. H. Wong, K. H. Brettingham-Moore, L. Chan et al., "Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere," *Genome Research*, vol. 17, no. 8, pp. 1146–1160, 2007.

[33] R. I. Amann, W. Ludwig, and K.-H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiological Reviews*, vol. 59, no. 1, pp. 143–169, 1995.

[34] P. Hugenholtz and N. R. Pace, "Identifying microbial diversity in the natural environment: a molecular phylogenetic approach," *Trends in Biotechnology*, vol. 14, no. 6, pp. 190–197, 1996.

[35] M. R. Rondon, P. R. August, A. D. Bettermann et al., "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms," *Applied and Environmental Microbiology*, vol. 66, no. 6, pp. 2541–2547, 2000.

[36] P. Lorenz, K. Liebeton, F. Niehaus, and J. Eck, "Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space," *Current Opinion in Biotechnology*, vol. 13, no. 6, pp. 572–577, 2002.

[37] E. F. DeLong, "Microbial population genomics and ecology," *Current Opinion in Microbiology*, vol. 5, no. 5, pp. 520–524, 2002.

[38] P. D. Schloss and J. Handelsman, "Biotechnological prospects from metagenomics," *Current Opinion in Biotechnology*, vol. 14, no. 3, pp. 303–310, 2003.

[39] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples," *DNA Research*, vol. 12, no. 5, pp. 281–290, 2005.

[40] H. Hayashi, T. Abe, M. Sakamoto et al., "Direct cloning of genes encoding novel xylanases from the human gut," *Canadian Journal of Microbiology*, vol. 51, no. 3, pp. 251–259, 2005.

[41] R. Nakao, T. Abe, A. M. Nijhof et al., "A novel approach, based on BLSOMs (Batch Learning Self-Organizing Maps), to the microbiome analysis of ticks," *The ISME Journal*, vol. 7, pp. 1003–1015, 2013.

[42] H. Uehara, Y. Iwasaki, C. Wada, T. Ikemura, and T. Abe, "A novel bioinformatics strategy for searchingindustrially useful genome resources frommetagenomic sequence libraries," *Genes and Genetic Systems*, vol. 86, no. 1, pp. 53–66, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 439476, 11 pages
http://dx.doi.org/10.1155/2014/439476

Review Article

Applied Graph-Mining Algorithms to Study Biomolecular Interaction Networks

Ru Shen[1] and Chittibabu Guda[2,3]

[1] *Department of Computer Science, University at Albany, 1400 Washington Avenue, Albany, NY 12222, USA*
[2] *Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, College of Medicine, Omaha, NE 68198-5145, USA*
[3] *Bioinformatics and Systems Biology Core, University of Nebraska Medical Center, Omaha, NE 68198, USA*

Correspondence should be addressed to Chittibabu Guda; babu.guda@unmc.edu

Received 14 January 2014; Accepted 19 February 2014; Published 2 April 2014

Academic Editor: Altaf-Ul-Amin

Protein-protein interaction (PPI) networks carry vital information on the organization of molecular interactions in cellular systems. The identification of functionally relevant modules in PPI networks is one of the most important applications of biological network analysis. Computational analysis is becoming an indispensable tool to understand large-scale biomolecular interaction networks. Several types of computational methods have been developed and employed for the analysis of PPI networks. Of these computational methods, graph comparison and module detection are the two most commonly used strategies. This review summarizes current literature on graph kernel and graph alignment methods for graph comparison strategies, as well as module detection approaches including seed-and-extend, hierarchical clustering, optimization-based, probabilistic, and frequent subgraph methods. Herein, we provide a comprehensive review of the major algorithms employed under each theme, including our recently published frequent subgraph method, for detecting functional modules commonly shared across multiple cancer PPI networks.

1. Introduction

Recent advances in systems biology research have generated a wealth of data on physical and genetic interactions capable of revealing relationships between biomolecules. For example, high-throughput screening methods, such as two-hybrid analysis [1] and mass spectrometry [2], have produced volumes of data on protein-protein interactions (PPI). PPI networks provide the basis for understanding the modular organization of molecular interactions. Still, computational algorithms are required to process this data for large-scale PPI networks. Such networks provide the basis for understanding the modular organization of molecular interactions. Analyzing PPI networks using graph-theory-based algorithms and graph-mining methods has become commonplace in systems biology research. Similarly, for comparative analysis of PPI networks in cancer, both graph comparison and module detection have been used to determine the structure of networks [3]. Here, we review various graph comparison and module detection algorithms that have been widely used for analyzing PPI networks in different systems biology applications.

The first group of algorithms of interest, graph comparison, is the process of comparing and contrasting graph-based networks in order to determine the PPI network similarities or detect common or distinct substructures (i.e., subnetworks or subgraphs). Thus, graph-theory-based methods are widely used in the comparative study of the molecular interaction networks (MINs). These methods have been applied in various studies and analyzed in previous review articles. For example, in 2006, Sharan and colleagues published a review of the applications of graph comparison methods to analyze MINs [4]. In our paper, we focus on more current algorithmic details of graph comparison methods. Our primary focus is on the following two most widely published graph comparison algorithms: graph kernels and graph alignments.

The second group of algorithms, module detection, involves the identification of functionally important substructures within a larger PPI network, which is one of the most widely studied topics in PPI network analyses. Considering that biological interactions do not operate based on sequence homology between partners, sequence homology-based methods such as Basic Local Alignment Search Tool

(BLAST) [5] are generally not useful for detecting interacting modules. Instead, network analysis has become a key approach to understanding the functional relationships between interacting proteins. Because subunits of a molecular complex generally function towards a common biological goal, predicting an uncharacterized protein as part of a known complex increases the confidence in the annotation of that protein [6]. Thus, detecting important modules in PPI networks is a well-studied problem in graph analysis. In this paper, we review the following five categories of module detection methods: (1) seed-and-extend, (2) hierarchical clustering, (3) optimization-based, (4) probabilistic, and (5) frequent subgraph.

Graph comparison can be used to search conserved regions representing functional, orthologous modules across different species or biological systems. In contrast, module detection algorithms can be applied to graph alignment to find the optimal local alignment between protein networks [7]. Sometimes, both types of algorithms are also used in combination to perform PPI network analysis. For example, finding the common PPI network modules in multiple cancer networks requires the combination of graph comparison and module detection algorithms [3]. Hence, graph comparison and module detection are interrelated rather than isolated topics. This paper is organized as follows. First, we review graph comparison strategies that include graph kernel and graph alignment methods. Next, we discuss module detection strategies that include five subsections covering seed-and-extend, hierarchical clustering, optimization-based, probabilistic, and frequent subgraph methods. The final section includes a summary of this review along with our conclusions.

2. Graph Comparison Strategies

Graph comparison is an important tool for understanding PPI networks. For instance, by measuring the discrepancy between PPI networks of healthy and diseased individuals it is possible to predict disease outbreak and progression [8]. Also, through the alignment of networks we can identify evolutionarily conserved patterns in biological pathways [9]. The key to accurate graph comparison is to find a suitable scoring function that can correctly measure differences between networks. Some of the early distance-based methods, including the Maximal Common Subgraph (MCS) [10, 11] and edit distance-based [12] methods, focus on the global comparison of networks. For the MCS method, similarity is measured based on the percentage of the overall network that the maximal common subgraph occupies. For the graph edit distance method, similarity is measured by what it takes to transform one network into another by substituting, deleting, and inserting nodes and edges. While distance-based methods provide an intuitive means of comparing graphs, these comparisons are based on the exact matching of substructures and therefore cannot be generalized to identify approximate similarities required for the analyses of many biological networks.

Graphlets are smaller units of subgraphs of distinct sizes. Graphlet degree distribution has been established as a more-comprehensive model compared to early graph comparison methods. Graphlet degree distribution is a generalization of degree distribution of larger networks. In 2006, Przulj reported that agreement in graphlet degree distribution could be effectively used to compare biological networks [13]. This distribution is calculated as the number of nodes attached to each of a predefined number of graphlets. Due to the topological differences of nodes on graphlets, connections at unequal positions of graphlets are considered different attachments. For example, for the 30 predefined graphlets in Przulj's study, there are 73 types of attachments (also known as orbits), which give rise to 73 distributions, one for each orbit. Given two networks for a particular orbit, the distribution agreement of the orbit is the inverse of the Euclidean distance between the two distributions. The distribution agreement between two networks is the arithmetic or geometric mean of the distribution agreement of all orbits. The graphlet degree distribution method was the first of a group of graphlet-based methods. The concepts of graphlet degree and the measurement of similarity by graphlet degree vectors provide the foundation for the subsequent graphlet-based comparison methods, including GRAph ALigner (GRAAL) [14] and its variations, H-GRAAL [15] and MI-GRAAL [16], discussed later in this review.

Like the graphlet degree distribution method that uses graphlets to compare PPI networks, graph kernels decompose networks into subunits and use the subunit information to calculate PPI network similarities.

2.1. Graph Kernels. As proposed by Haussler, graph kernels can be viewed as special cases of R-convolution kernels [17]. Graph kernels can offer an analysis of networks by comparing nontrivial substructures. A fundamental issue in graph comparison is the problem of subgraph isomorphism. In this case, given two graphs, G and H, for example, we need to determine whether G contains a subgraph that is isomorphic to H. Subgraph isomorphism is proven to be NP-complete, which presents a large challenge with respect to computational complexity and run time. In the context of graph kernels, in order to accurately compare two PPI networks, an exhaustive comparison of all of their subgraphs is needed. However, performing such a comparison using graph kernels requires the same computational complexity as subgraph isomorphism. Therefore, a key reason for using graph kernels is to closely approximate an exhaustive comparison while maintaining computational tractability. Among various graph kernels, *walk-based kernels* [18, 19] compare graphs by counting the number of matching walks within two input graphs. A walk is a sequence of edges from the graph that connect to a sequence of vertices. Vertices are allowed to repeat in a walk but not in a path. *Path-based kernels* [20] use paths instead of walks to avoid "tottering" problems related to walk-based kernels (i.e., high similarity values resulting from cycles or loops in the graphs). On the *other hand*, subtree kernels compare all pairs of matching substructures in subtree patterns [21, 22]. While subtrees are more expressive structures compared to paths or walks, constructing subtrees is computationally expensive. Here we will highlight a few methods that use fast algorithms for efficient computation of kernels.

In 2005 and 2007, Borgwardt and colleagues proposed fast algorithms for computing random walk kernels [8, 23]. This type of computation performs random walks on graphs in order to decompose these graphs into multiple paths and compute the number of matching paths. Various incarnations of these kernels use different metrics for computing similarities between paths [18, 23], but performance issues when computing large networks plague almost all of these kernels. As a noteworthy advancement, Borgwardt and colleagues introduced the fast random walk kernel that increases the speed of performance to up to three orders of magnitude. In Borgwardt's method, the kernel is defined by the product graph, which is a composite graph whose nodes are tuples of vertices from each network. Edges exist only if the corresponding vertices are adjacent in both networks. Three efficient schemes are utilized to decrease the computation time. These schemes include reducing the kernel to the problem of solving a generalized Sylvester equation [24], using the conjugate gradient methods to solve the kernel and implementing a fixed-point iteration method to speed up the computation time [25].

Given that decomposing networks to small substructures is an expensive process, many state-of-the-art graph kernels do not scale to large graphs. To address this issue, in 2009 Shervashidze and colleagues proposed a statistical approach to compare graphs based on the distribution of graphlets [26]. Here, the graphlet kernel is calculated as the product of the graphlet distribution vectors. Graphlet distribution could be used to represent the distribution of the graph, especially when the graph is large. Given that the exhaustive enumeration of graphlets is prohibitively expensive, two theoretically grounded alternatives have been proposed. First, sampling a fixed number of graphlets suffices to bind the deviation of the empirical estimates of the graphlet distribution from the true distribution. Second, for graphs of a bounded degree, the exact number of all graphlets of size k can be determined at a computational complexity that is close to polynomial time. Because the sampling technique of graphlet kernels is independent of the graph size, graphlet kernels are scalable to larger graphs.

In 2007, Shervashidze and colleagues proposed another algorithm for fast computation of subtree kernels [27]. The fast subtree algorithm is built upon the Weisfeiler-Lehman test of isomorphism [28]. The key idea of this algorithm is to first assign the nodes with a sorted set of labels from neighboring nodes and then compress this sorted set of labels to short labels. Given the limited range from elements of the set, this method applies a counting sort on the set of labels to achieve linear complexity. Not only is the runtime reduced, but also the accuracy of fast subtree kernel is competitive with state-of-the-art kernels on several graph classification benchmark data sets. Following the fast subtree kernel, a family of Weisfeiler-Lehman kernels was later designed [29], including the Weisfeiler-Lehman edge kernel and the Weisfeiler-Lehman shortest path kernel. In terms of runtime on large graphs, these kernels outperform other kernels, including the recently developed random walk kernels [8] and graphlet kernels [26].

Graph kernels bridge the gap between graph-structured data and a large spectrum of machine-learning algorithms [29]. Graph kernels have quickly developed into an independent branch of graph-mining methods and are widely used in various areas such as computational biology and social network analysis. Nevertheless, with a single value being produced as a result of the comparison, graph kernels cannot provide detailed information on the node or edge mapping. On the other hand, graph alignment methods are designed to address the above limitations in graph kernel methods. Graph alignment methods can provide a more in-depth knowledge of the comparison, with the tradeoff being a longer processing time.

2.2. Graph Alignment Methods. Graph alignment is the process of mapping nodes and edges between graphs such that conserved subgraphs can be identified. Graph alignment adopts a similar concept as that used for sequence alignment. However, in contrast to sequence alignment, which aligns linear sequences to identify regions of similarity, graph alignment must be able to handle data from multiple dimensions of the graph. Graph alignment involves a subgraph isomorphism test that is proven to be NP-complete. Similar to graph kernels, getting an exact solution for graph alignment is not feasible for even moderate sized graphs. Thus, most graph alignment methods resort to heuristic solutions to reduce the cost of computation.

Similar to sequence alignment strategies, graph alignment can be local or global. Local graph alignment matches nodes and edges to maximize the local alignment score. In 2006, Koyuturk and colleagues proposed a local alignment framework for PPI networks based on the duplication/divergence evolutionary model [7]. In this framework, in a given pair of PPI networks, the alignment score between two PPI networks is calculated based on matches, mismatches, and duplications. This method is heuristic-based and aims to locate all maximal protein-subset pairs such that the alignment score is locally maximized. Another example of local alignment is the modular subgraph alignment algorithm [30], where each larger network is decomposed into a collection of smaller subnetworks in order to compute the alignment of the two networks as the optimal alignment of the subnetworks.

A series of NetworkBLAST algorithms have been reported for the global alignment of networks. For example, PathBLAST [9] is the first one of such algorithms to search for high-probability pathway alignments between two PPI networks [31]. A later version, NetworkBLAST [32, 33], constructs a general framework for comparing more than two protein networks in order to search for conserved patterns such as short linear pathways and dense clusters of complexes. This search algorithm exhaustively identifies high scoring subnetwork seeds and uses them to expand the search. NetworkBLAST is an exhaustive approach and is limited in its application to the alignment of only up to three networks, while an extended version, NetworkBLAST-M, can handle multiple networks [34]. NetworkBLAST-M progressively constructs a layered alignment graph, with each layer corresponding to a network. Connections between layers indicate

similar proteins across different protein networks. The set of potentially orthologous proteins is represented by a subnet, which includes a vertex from each of the layers. NetworkBLAST-M computes a local alignment by readily finding subnets of high, local conservation based on inferred phylogeny. With the novel representation of a layered alignment graph, NetworkBLAST-M can achieve dramatic reductions in run time and memory requirements for multiple network alignments.

Other methods use biological features of interacting proteins for graph alignment. For example, using integer quadratic programming, Li and colleagues proposed a PPI alignment algorithm in 2006 based on similarities in both the protein sequence and network architecture [35]. In this method, the alignment of PPI networks is formulated as a combination of the sequence similarity score of proteins and the matching score of protein interactions. A coefficient is introduced to balance the weight between the node and edge similarity of the networks. Integer quadratic programming is used to maximize the alignment score among all feasible combinations of matching scores. Nodes without alignment gaps are selected to construct a minimally connected subgraph within each network; these subgraphs are regarded as conserved patterns.

The GHOST alignment method [36] developed by Patro and Kingsford uses the spectrum of the graph adjacency matrix to measure topological similarities between networks. GHOST performs a global network alignment using a two-phased approach. The first phase employs a seed-and-extend strategy to align high scoring node pairs with their neighbors, which is similar to the way that BLAST functions [5]. The second phase uses a local search method to realign nodes in order to achieve better topological or biological quality. NETAL is another global graph alignment method that was developed by Neyshabur and colleagues [37]. NETAL constructs an alignment score matrix and uses a broad search to find the best alignment between networks. The alignment score matrix is constructed from the similarity and interaction score matrices. The similarity score matrix indicates both topological and biological similarities between nodes; the interaction score matrix represents the approximated number of conserved interactions incident to the nodes. During the process of a broad search, node pairs with maximum alignment scores are selected and aligned. As a result, the interaction score matrix is updated, which in turn impacts the alignment score matrix. The process continues until all of the nodes of one network are aligned to at least some nodes of the other network.

GRAAL (GRAph ALigner) is a global alignment method based solely on the network topology [14]. For each node in a network, a vector of "graphlet degrees" is used to record the number of each kind of graphlet that the node touches. Signature similarity is computed as the distance between two vectors. Alignment of two networks is completed by matching pairs of nodes originating in different networks based on the similarity of their signatures. Although this algorithm operates based on local alignment, it produces global alignment results. H-GRAAL, which is a variation of GRAAL [15], uses the Hungarian algorithm to solve assignment problems and determine the optimal alignments between networks. Given the cost of aligning two nodes, the Hungarian algorithm locates the assignment of all pairs of nodes that yield a minimized total cost of alignment. Additionally, there are other graphlet-based alignment algorithms including MI-GRAAL [16] and the latest C-GRAAL [38]. Like GRAAL, MI-GRAAL is a seed-and-extend approach. However, unlike GRAAL and H-GRAAL, which are purely based on topological information, MI-GRAAL can integrate any type of similarity measures into the model. From these similarity measures, MI-GRAAL computes the alignment confidence score between pairs of nodes. High scoring pairs are inserted into a priority queue and used as seeds during the alignment. Similar to GRAAL and MI-GRAAL, C-GRAAL also uses a seed-and-extend approach. In contrast to GRAAL, which is based on graphlet degree information, C-GRAAL aligns networks based on common information of neighboring graphlets.

Regardless of if the calculation is for local or global alignments, performing efficient and accurate alignments on multiple networks continues to be challenging. Nevertheless, the Graemlin algorithm was the first algorithm capable of performing scalable, multiple network alignments [39]. Graemlin starts from pairwise network alignments. It uses a seed-and-extend algorithm to first identify clusters of proteins as "seeds," and then it broadly extends the alignment to yield a maximal increase in alignment score. In aligning multiple networks, Graemlin successively aligns the closest pair of networks obtained from the pairwise alignment phase, resulting in the construction of new networks from the alignments. In practice, Graemlin avoids an exponential run time because the constructed networks have small overlaps. The newer version, Graemlin 2.0 [40], is a global alignment algorithm that can adjust scoring function parameters and perform multiple network alignments.

IsoRank [41] is another multiple network alignment method that aims to correspond nodes and edges of input networks to maximize the global "match" between PPI networks. The maximum match is a combination of the following two factors: (1) the size of the common graph determined by mapping and (2) the aggregate sequence similarity between nodes mapped to one another. The IsoRank algorithm first associates a functional similarity score with each possible match between nodes of two networks. Functional similarity shapes the tradeoff between the twin objectives of topological overlapping and high sequence similarity between mapped nodes. This similarity is resolved through eigenvalue computation. In the second stage, mappings between networks are extracted from the functional similarity scores. To align multiple networks, the above processes are repeated for each pair of networks. An improved version of IsoRank, IsoRankN [42], was developed to perform more efficient and highly accurate global alignments over multiple networks. Through the use of spectral partitioning algorithm, IsoRankN can find dense, clique-like clusters that are considered conserved regions among the networks.

As a summary of graph comparisons, we have reviewed an array of methods from early single-feature, distance-based algorithms to the most current multiple network alignment graphs. Early distance-based algorithms are founded on strict

Table 1: A summary of graph comparison methods by the strategy employed.

Methods	Comparison strategy	Specification				References
		Local	Global	Pairwise	Multiple	
MCS	Distance-based		x	x		[10, 11]
Editing distance	Distance-based		x	x		[12]
Graphlet	Graphlet		x	x		[13]
Fast random walk kernel	Graph kernel			x		[8]
Graphlet kernel	Graph kernel			x		[26]
Fast subtree kernel	Graph kernel			x		[27]
Weighted alignment	Graph alignment	X		x		[7]
Substructure-based alignment	Graph alignment	X		x		[30]
Class of NetworkBLAST	Graph alignment	X		x	x	[9, 31–34]
Quadratic programming	Graph alignment	X		x		[35]
Class of GRAAL	Graph alignment		x	x		[14–16, 38]
Class of Graemlin	Graph alignment		x		x	[39, 40]
Class of IsoRank	Graph alignment		x		x	[41, 42]
GHOST	Graph alignment		x	x		[36]
NETAL	Graph alignment		x	x		[37]

matching of network structures, and thus these algorithms are only suitable for simple network comparisons. Graph kernels compare graphs by decomposing and comparing graph substructures, which are based on well-supported statistical analyses and mathematical derivations. This results in more accurate and meaningful comparisons, particularly for approximate structural similarities. However, with a single value being produced as the comparison result, graph kernels cannot provide substantial internal details for the comparison, such as the node and edge mapping. Therefore, graph kernels are most suitable for solving classification problems for small to medium sized graphs. For large sized graph comparison, graph kernels are at disadvantage, because on one hand the kernel calculations are very time consuming for large networks and on the other hand the calculated values are less informative than those of smaller sized networks. To perform detailed comparisons of networks, graph alignments are preferred. Local alignments align subnetworks to maximize the local alignment score. Global alignments, on the other hand, focus on maximizing the overall alignment score. While local alignments may be ambiguous, global alignments typically produce unique mapping between nodes. In recent years, several multiple network alignment methods have been developed. Compared to pairwise alignments, multiple network alignment methods provide greater proof of conservation of the identified subnetworks. Graph alignments are very effective for network comparison and identification of conserved regions in networks. However, due to the multidimensional nature and complexity of graph data, graph alignment algorithms rely on heuristics to derive the optimal solution. The drawbacks in graph alignments are that different heuristics usually result in very different solutions and there are no standards or benchmarks like those available for sequence alignment.

A summary of different strategies used for graph comparisons is provided in Table 1. Here, the first column lists the name of the method, and the second column specifies the comparison strategy used. The third column annotates the methods with keywords such as local, global, pairwise, or multiple. Local versus global indicates if the method is for local graph alignment or global graph alignment. Pairwise versus multiple indicates if the method is a pairwise graph comparison or a multiple graph comparison. The last column of the table refers to the reference number listed in the reference section of this paper.

3. Module Detection Strategies

One of the most important applications of biological network analysis is the identification of functionally relevant modules in PPI networks. Similar to social networks and internet-based networks, PPI networks are conjectured to exhibit a power law degree distribution [43]. Proving that PPI networks follow a power law degree distribution requires more rigorous statistical data analysis than that available today [44]; however, it is clear that the connectivity of the PPI networks is centralized around a small number of hub nodes. Such a connectivity pattern indicates that the subgraphs (modules) centered on these hub nodes are important for accomplishing specific biological functions. These modules may be mapped to biological pathways or physically interacting complexes. A variety of module detection algorithms exist in the literature. Here, we review the following five methods on module detection: (1) seed-and-extend, (2) hierarchical clustering, (3) optimization-based, (4) probabilistic, and (5) frequent subgraph methods. Many of these methods are also discussed in a recent review on community detection in graphs by Fortunato [45]. In our paper, we also discuss our recently published work on frequent subgraph method for detecting common functional modules among multiple PPI networks involved in nine different cancers [3].

3.1. Seed-and-Extend Approaches. Seed-and-extend approaches predict functional protein modules based on

the density of PPI networks. The functional modules are generally initiated from single nodes deemed as central nodes or "seeds," and new nodes are added to "extend" the subnetworks. Different algorithms have specific metrics for determining when the subnetworks will reach convergence.

The first seed-and-extend approach we will review is Molecular Complex Detection (MCODE). This method detects densely connected regions in large PPI networks that may represent molecular complexes [6]. MCODE generates weights for all vertices based on their local network density and identifies high weight seed proteins. Seeds are expanded outwards by including vertices in complexes whose weights are above a given threshold. This process continues until no more vertices can be added to the complex. The Speed and Performance in Clustering (SPICi) method [46] is another seed-and-extend algorithm used for clustering large biological networks. SPICi uses a heuristic approach to build clusters from an initial seed-connected pair of vertices (S) with the highest weight degrees. During the expansion stage, SPICi searches for a vertex with the maximum value of support amongst all the nonclustered vertices adjacent to S. The procedure is repeated until all vertices in the graph are clustered.

Both the MCODE and SPICi methods are purely based on network topology. In contrast, Maraziotis and colleagues presented a new method that discovers functional modules from weighted graphs [47]. These modules are obtained by clustering proteins according to their gene expression profiles and then measuring the distances between clusters. First, the seed proteins are selected from complexes that have at least six members and more than 80% data coverage. The neighbors of the seed protein are sorted in a descending degree of significance, and this subset of nodes is named the kernel. Adjacent nodes are iteratively added to the selected kernel. Still, of all the seed-and-extend methods, SPICi is the most useful to cluster larger networks due to its efficient memory utilization algorithm [46].

3.2. Hierarchical Clustering. Hierarchical clustering is another group of clustering algorithms widely used for biological data analysis. Hierarchical clustering methods are often applied to gene expression data to determine coexpressed genes, clusters, and outliers [48]. Hierarchical clustering methods can also be applied to PPI networks to identify potential modules from within the networks. Similar to seed-and-extend methods, hierarchical clustering algorithms assume the unbalanced distribution of nodes and edges in networks. These methods hierarchically group objects based on the distance among the objects.

The Protein Distance Based on Interactions (PRODISTIN) method was developed based on the principle that the greater the two proteins in a network share common interactions, the more likely it is that they are functionally related [49]. Using the number of common and distinctive interacting members of the two proteins in an interaction network, PRODISTIN computes the functional distance using the Czekanovski-Dice distance formula [50], which reflects the symmetrical difference between the two proteins in an interaction network. The distance values are clustered using a variation of the neighbor-joining algorithm [51] to generate a hierarchical tree. In 2004, Lu and colleagues presented ADJW and Hall clustering algorithms [52]. ADJW employs the adjacency matrix of the network as the similarity matrix for the clustering. Densities of the edges between node groups are computed from the similarity matrix. Using single linkage clustering, each step clusters two groups having maximum edge densities into one group. Hall clustering projects proteins into Euclidian space according to their connectivity. The Euclidian distances are used as a metric to measure the topological distance of vertices in the network. Because the two groups with the smallest sum form a new group at each step, the groups closer in distance are selected earlier in the clustering process.

The hierarchical clustering methods discussed above are agglomerative methods in which the additions of edges are used to construct hierarchical trees. In hierarchical clustering, there is another class of methods called divisive methods that construct hierarchical trees by removing edges. Divisive methods attempt to find the least similar connected pairs of vertices from the network of interest and then remove the edges between the pairs. An example of a divisive method is Newman and colleagues' hierarchical clustering for finding community structures in networks [53]. This method looks for edges with the highest "betweenness," where betweenness is a measure that favors edges that lie between communities and disfavors edges that lie inside communities. By removing the edges with highest betweenness, this method divides the network into smaller components.

Hierarchical clustering methods have primarily focused on grouping nodes. An unconventional method proposed by Ahn and colleagues in 2010 focuses on edge clustering [54]. Rather than assuming that a module is a set of nodes connected to one another, edge clustering defines modules as sets of closely interrelated links. Similarity scores are calculated for each pair of links that share a node; this calculation is used to build a link dendrogram. An objective function is defined to compute the link density at each level of the dendrogram to help determine the best way to cut the tree, so to speak. Compared to node-based clustering, edge-based clustering can simultaneously reveal hierarchical and overlapping relationships. In another work related to edge-based clustering, Solava and colleagues proposed a measure of edge graphlet-degree-vector (GDV) similarity [55]. Edge GDV counts the number of different graphlets that touch an edge. Because their extended network neighbors are compared, the edge GDV of two edges gives a sensitive measure of their topological similarity. When edge GDV similarity is used to substitute the original edge similarity measure, as is the case in Ahn and colleagues' hierarchical clustering method [54], the predication accuracy outperforms the original algorithm.

3.3. Optimization Methods. In addition to seed-and-extend approaches and hierarchical clustering, module detection can also be formulated as an optimization problem. In 2004, King and colleagues completed work on predicting protein complexes via cost-based clustering [56]. Here, the module

detection problem is transformed to finding the optimum partitioning of the network in order to minimize the value of the clustering cost function. One method, namely, Restricted Neighborhood Search Clustering (RNSC), is a local search algorithm based loosely on the Tabu search metaheuristic [57]. To search for low cost graph partitioning, RNSC uses a simple integer-valued cost function, referred to as the naive cost function, as a preprocessor. Next, a more expressive real-valued cost function, referred to as the scaled cost function, is used to evaluate clustering. RNSC iteratively moves a node from one cluster to another in a randomized fashion to reduce the clustering cost.

Genes with significant changes in expression have immediate and wide interest as markers of disease and the stage of disease development, as well as markers for a variety of other cellular phenotypes [58]. Genes with correlated expression changes in many conditions are likely involved in similar functions or cellular processes. Such correlated expression patterns in networks can be uncovered by subnetwork searching. For example, the Cytoscape plug-in jActiveModules searches for such active subnetworks (i.e., connected regions of the network that show significant changes in expression over particular subsets of conditions) [59]. The jActiveModules method first adopts a statistical scoring system to capture a change in gene expression for a given subnetwork. Then, jActiveModules identifies the highest scoring subnetworks, which are the active modules in the network studied. Because the problem of finding the maximal-scoring-connected subgraph is NP-hard, the heuristic simulated annealing algorithm is used.

In 2010, Zhang and colleagues introduced a new method that uses graph modularity density to detect functional modules in PPI networks [60]. Graph modularity measures the fraction of edges in the network that connect vertices of the same type (community) against the expected value of connections in a random network. The larger the value of modularity density is, the more accurate the partition would be. Therefore, the community detection problem can be viewed as a problem of finding a network partition that has maximal modularity density. The optimization problem of finding maximized modularity density is first attempted using the simulated annealing (SA) technique. SA allows one to complete an exhaustive search of networks and minimize the problem of finding suboptimal partitions. The result of running SA over modularity and modularity density shows that maximizing the density can provide more detailed and valid results.

HOTNET, published by Raphael's lab in 2011, is another framework for de novo identification of significantly mutated subnetworks [61]. HOTNET first formulates an influence measure between pairs of genes based on their topological relationships in the network. Next, HOTNET builds an influence graph that incorporates information from only the mutated genes in the neighboring network. One clear way to detect significant subnetworks from the influence graph is to identify sets of nodes that are connected through a high influence measure and correspond to mutated genes. However, finding such node sets is a difficult problem that cannot be solved in polynomial time by any known algorithms. Alternatively, a computationally efficient approach is based on the concept of enhancing the influence measure by the number of mutations observed in each of these genes. Thus, the strength of connections in the enhanced influence graph is a function of both the interaction between the nodes and the number of mutations observed in their corresponding genes. Next, a threshold is set for the weight of connections, and the influence graph is decomposed into connected components by removing edges with weights smaller than the threshold. The significance of the subnetworks discovered depends on the choice of threshold. The computational complexity of the algorithm is linear to the size of the graph.

3.4. Probabilistic Methods. In recent years, probabilistic-based machine learning methods have been developed and successfully used in many areas in bioinformatics. Here, we review a few machine learning methods developed for network module detection. In 2011, Shi and colleagues used a "semisupervised" method for detecting protein complexes in PPI networks [62]. This method uses topological features such as degree statistics, edge weight statistics, clustering coefficients, and biological features like protein length and polarity of amino acids to construct a two-layer, feed-forward neural network. Shi first obtained a weighted PPI network, a set of known protein complexes, and nonprotein complexes to set up the training model. Using the initial model, this method builds new complexes and uses them to train the model iteratively until no more proteins can be added. PPI networks typically contain large amounts of false negative (missing data) and false positive connections. Because neural networks are regarded for their high tolerance for noisy data, this method offers a suitable model for PPI module detection.

In 2008, Qi and colleagues presented a Bayesian network (BN) algorithm for detecting protein complexes from PPI networks [63]. In this supervised learning approach, a probabilistic Bayesian network mimics each complex subgraph. Specific topological and biological features are selected for representative properties of the protein complex. These features include node size, degree statistics, edge weight statistics, and protein weight or size statistics. Another method, the Markov Clustering Method (MCL), is based on the probability of landing on different vertices through random walks in the network [64]. The premises underlying MCL are that (1) the number of paths between two vertices is larger when the two vertices belong to the same cluster and (2) random walks have a higher probability of traversing within the same cluster than traversing across different clusters. The algorithm starts by creating a Markov matrix, which is an adjacency matrix normalized to 1. Two operators, expansion and inflation, are then used iteratively to recompute the transition probabilities. Expansion corresponds to matrix multiplication, which is responsible for creating new edges, while inflation increases the contrasts between existing differences of probability. The iterative process converges quickly, and the resulting matrix represents a nonoverlapping cluster of the network.

3.5. Frequent Subgraph Methods. Most module detection algorithms are based on either network connectivity or

Table 2: A summary of module detection methods by the strategy employed.

Methods	Module detection strategy	Specification		References
		Topological	Both	
MCODE	Seed-and-extend	x		[6]
SPICi	Seed-and-extend	x		[46]
Kernel set	Seed-and-extend		x	[47]
PRODISTIN	Hierarchical clustering		x	[49]
ADJW and Hall	Hierarchical clustering	x		[52]
Divisive	Hierarchical clustering	x		[53]
Edge clustering	Hierarchical clustering	x		[54, 55]
RNSC	Optimization	x		[56]
jActiveModules	Optimization		x	[59]
Modularity density	Optimization	x		[60]
HOTNET	Optimization		x	[61]
Semi-supervised	Probabilistic		x	[62]
Bayesian network	Probabilistic		x	[63]
MCL	Probabilistic	x		[64]
Frequent subgraph	Frequency-based method	x		[3, 67, 68]

the density of subgraphs. In contrast, we recently developed a novel method that predicts functional modules based on the frequency of subgraphs [3]. We compared nine cancer PPI networks to identify common and frequent substructures among the networks. Given their unusual frequency in multiple cancer-related PPI networks, these substructures strongly appear to be functionally relevant to cancer. Our method begins with assigning canonical labels to subgraphs, where subgraphs with the same canonical labels are isomorphic to one another. Starting from small sized subgraphs, the canonical labels are compared, and infrequent edges are pruned from the networks. Frequent and common substructures are recorded and included in the search for larger sized modules. The process iterates as the subgraph size increases until no more substructures can be discovered. From the nine cancer PPI networks, frequent and common substructures have been discovered from two to ten edges. Gene Ontology (GO) semantic similarity scores of the substructures discovered have been compared with those of randomly generated patterns [65]. Common substructures exhibit significantly higher scores compared to random substructures at all edge levels, indicating that the discovered subgraphs are functionally significant. A survey of frequency-based subgraph mining algorithms was previously reviewed by Jiang and colleagues [66]. In general, Apriori-based approach and pattern growth approach are the two major groups of algorithms for identifying frequent subgraphs. Apriori-based algorithms such as FSG [67] generate candidate subgraphs of larger size by joining two smaller subgraphs. On the other hand, pattern growth approach [68] extends patterns directly from a single pattern, instead of joining two smaller subgraphs. Nevertheless, both groups of algorithms are restricted by the graph size due to the subgraph isomorphism problem.

In summary, we reviewed multiple methods for module detection. The performance of each of these methods varies. In Brohee and colleagues' review paper, the performance of MCL, RNSC, and MCODE is compared in terms of robustness, sensitivity, and the results of clustering [69]. In general, RNSC and MCL outperform the MCODE under most conditions. Similarly, Dhara and colleagues performed a comparative behavioral analysis of RNSC and MCL on power law distribution graphs [70]. According to their analysis, RNSC is preferred to MCL in terms of both cost and quality.

Among the numerous methods for module detection, different metrics are used to evaluate the weights of modules. Some metrics are based on the connectivity density, such as MCODE and edge clustering, while some are based on vertex scoring, such as RNSC and jActiveModules. Seed-and-extend methods such as MCODE and SPICi assume that hub nodes always exist as the centers of modules. Such assumptions may limit the types of modules that can be discovered by seed-and-extend algorithms. On the other hand, hierarchical clustering algorithms, including both agglomerative and divisive, do not effectively use the topological information of the networks. Because the distances between nodes or edges determine how the clusters are drawn, using different distance metrics in the algorithm will lead to different clustering results. Optimization methods have their limitations too. Each run of optimization methods may generate different results, depending on the initial settings. Therefore multiple runs of optimization methods are required to achieve a relatively consistent result. Finally, frequency-based methods look for recurring patterns in PPI networks. Frequency-based methods are plagued by the performance issue because frequent pattern matching involves subgraph isomorphism tests, and it is proven that subgraph isomorphism problem is NP-complete.

Table 2 provides a summary of these different methods. Similar to Table 1, the first and second columns list the name of the method and module detection strategy, respectively. The third column annotates each method with specifications; topological versus both indicates if the method is purely based on topological information or both topological and biological information. Depending on the theoretical

assumptions and metrics included, these methods are capable of uncovering substructures that represent specific biological functions. Deciding which method to use depends on the specific biological context of the problem.

4. Conclusions

Graph comparison and module detection are two commonly used strategies for analyzing PPI networks. Among the algorithms for graph comparisons, graph kernels compare graphs by decomposing the graphs to nontrivial subunits. Similarity scores between the graphs can be derived through comparing these subunits. In contrast to other graph comparison methods, graph kernels have the advantage of speed. In the development of graph kernels, efficiency is the key issue addressed. In contrast to graph kernels that can only produce limited information from the comparison, graph alignment provides in-depth analysis of the mappings between graphs. Graph alignment adopts concepts from sequence alignment; the alignment scores are adjusted to reflect topological or relational information for the graphs. For the purpose of graph comparison, graph kernels are suitable for classification tasks that require high-speed computation and intuitive measurement of distance. Graph alignment methods are suitable for determining conserved regions between PPI networks. Note that graph alignments can be local or global, and graph comparisons can be between two networks (pairwise) or greater than two networks (multiple).

Among the module detection algorithms, seed-and-extend methods identify modules by first selecting their core nodes and then expanding the core nodes with new nodes that increase the subgraph density. Hierarchical clustering creates clusters hierarchically based on distances between the clusters. Optimization-based and probabilistic approaches use mathematical derivations to determine best scoring modules. The frequent subgraph approach searches for common and frequent substructures among PPI networks. Different methods tackle the problem from different perspectives. For example, seed-and-extend methods use connection density and neighboring network information to find heavily connected modules in PPI networks. Hierarchical clustering methods use distances between nodes or edges as the key factor for clustering. Optimization-based methods represent network divisions using mathematical models. Modules are detected through the optimal division of the network. Probabilistic methods use statistics of graph data to construct training models and to determine the state transitions of the algorithm. Finally, for frequency-based algorithms, the frequency of subgraphs becomes the key criterion for detecting modules. The method selection for graph analysis depends on the interpretation of the problem and the perspective of the investigator tackling the problem. As the methods are developed from different perspectives, they produce complementary views of graph data. The future development of graph analysis will be likely focused on integrated analysis using an ensemble of methods because no single method can perform well for all types of comparisons. Because most of the interaction network studies require both graph comparison and module detection, such analyses will benefit from the integration of methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported in part by Grants from National Institutes of Health [1R01GM086533-01A1 to CG] and startup funds to CG from University of Nebraska Medical Center. The authors also thank Ms. Melody Montgomery at the UNMC Research Editorial Office for help in the professional editing of this paper.

References

[1] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[2] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[3] R. Shen, N. C. W. Goonesekere, and C. Guda, "Mining functional subgraphs from cancer protein-protein interaction networks," *BMC Systems Biology*, vol. 6, supplement 2, 2012.

[4] R. Sharan and T. Ideker, "Modeling cellular machinery through biological network comparison," *Nature Biotechnology*, vol. 24, no. 4, pp. 427–433, 2006.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[6] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[7] M. Koyutürk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama, "Pairwise alignment of protein interaction networks," *Journal of Computational Biology*, vol. 13, no. 2, pp. 182–199, 2006.

[8] K. M. Borgwardt, H. P. Kriegel, S. V. N. Vishwanathan et al., "Graph kernels for disease outcome prediction from protein-protein interaction networks," in *Proceedings of the Pacific Symposium on Biocomputing.*, pp. 4–15, 2007.

[9] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, "PathBLAST: a tool for alignment of protein interaction networks," *Nucleic Acids Research*, vol. 32, pp. W83–W88, 2004.

[10] H. Bunke and K. Shearer, "A graph distance metric based on the maximal common subgraph," *Pattern Recognition Letters*, vol. 19, no. 3-4, pp. 255–259, 1998.

[11] M.-L. Fernández and G. Valiente, "A graph distance metric combining maximum common subgraph and minimum common supergraph," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 753–758, 2001.

[12] A. Sanfeliu and K.-S. Fu, "A distance measure between attributed relational graphs for pattern recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, no. 3, pp. 353–362, 1983.

[13] N. Pržulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, no. 2, pp. e177–e183, 2007.

[14] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, "Topological network alignment uncovers biological function and phylogeny," *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1341–1354, 2010.

[15] T. Milenković, W. L. Ng, W. Hayes, and N. Pržulj, "Optimal network alignment with graphlet degree vectors," *Cancer Informatics*, vol. 9, pp. 121–137, 2010.

[16] O. Kuchaiev and N. Pržulj, "Integrative network alignment reveals large regions of global network similarity in yeast and human," *Bioinformatics*, vol. 27, no. 10, Article ID btr127, pp. 1390–1396, 2011.

[17] D. Haussler, "Convolutional kernels on discrete structures," Technical Report UCSC-CRL-99-10, Computer Science Department, Santa Cruz, Calif, USA, 1999.

[18] T. Gärtner, P. Flach, and S. Wrobel, "On graph kernels: hardness results and efficient alternatives," in *Proceedings of the 16th Annual Conference on Learning Theory and 7th Kernel Workshop (COLT/Kernel '03)*, pp. 129–143, August 2003.

[19] H. Kashima, K. Tsuda, and A. Inokuchi, "Marginalized Kernels Between Labeled Graphs," in *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, pp. 321–328, Washington, DC, USA, August 2003.

[20] K. M. Borgwardt and H.-P. Kriegel, "Shortest-path kernels on graphs," in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM '05)*, pp. 74–81, November 2005.

[21] P. Mahé and J.-P. Vert, "Graph kernels based on tree patterns for molecules," *Machine Learning*, vol. 75, no. 1, pp. 3–35, 2009.

[22] J. Ramon and T. Gartner, "Expressivity versus efficiency of graph kernels," in *Proceedings of the 1st International Workshop on Mining Graphs, Trees and Sequences*, pp. 65–74, 2003.

[23] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. 1, pp. i47–i56, 2005.

[24] J. D. Gardiner, A. J. Laub, J. J. Amato, and C. B. Moler, "Solution of the Sylvester matrix equation $AXB^T+CXD^T = E$," *ACM Transactions on Mathematical Software*, vol. 18, no. 2, pp. 223–231, 1992.

[25] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, 1999.

[26] N. Shervashidze, S. V. N. Vishwanathan, T. H. Petri et al., "Efficient graphlet kernels for large graph comparison," in *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, vol. 5, 2009.

[27] N. Shervashidze and K. M. Borgwardt, "Fast subtree kernels on graphs," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '09)*, pp. 1660–1668, December 2009.

[28] B. Weisfeiler and A. A. Lehman, "A reduction of a graph to a canonical form and an algebra arising during this reduction," *Nauchno-Technicheskaya Informatsia*, vol. 2, pp. 12–16, 1968.

[29] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-Lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, pp. 2539–2561, 2011.

[30] F. Towfic, M. Heather, W. Greenlee et al., "Aligning biomolecular networks using modular graph kernels," in *Proceedings of the 9th International Conference on Algorithms in Bioinformatics*, pp. 345–361, 2009.

[31] B. P. Kelley, R. Sharan, R. M. Karp et al., "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 20, pp. 11394–11399, 2003.

[32] R. Sharan, S. Suthram, R. M. Kelley et al., "Conserved patterns of protein interaction in multiple species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1974–1979, 2005.

[33] M. Kalaev, M. Smoot, T. Ideker, and R. Sharan, "NetworkBLAST: comparative analysis of protein networks," *Bioinformatics*, vol. 24, no. 4, pp. 594–596, 2008.

[34] M. Kalaev, V. Bafna, and R. Sharan, "Fast and accurate alignment of multiple protein networks," *Journal of Computational Biology*, vol. 16, no. 8, pp. 989–999, 2009.

[35] Z. Li, Y. Wang, S. Zhang et al., "Alignment of protein interaction network by integer quadratic programming," in *Proceedings of the 28th EMBS Annual International Conference*, New York, NY, USA, 2006.

[36] R. Patro and C. Kingsford, "Global network alignment using multiscale spectral signatures," *Bioinformatics*, vol. 28, pp. 3105–3114, 2012.

[37] B. Neyshabur, A. Khadem, and S. Hashemifar, "NETAL: a new graph-based method for global alignment of protein-protein interaction networks," *Bioinformatics*, vol. 29, pp. 1654–1662, 2013.

[38] V. Memisevic and N. Przulj, "C-GRAAL: common-neighbors-based global GRAph ALignment of biological networks," *Integrative Biology*, vol. 4, pp. 734–743, 2012.

[39] J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou, "Græmlin: general and robust alignment of multiple large interaction networks," *Genome Research*, vol. 16, no. 9, pp. 1169–1181, 2006.

[40] J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou, "Automatic parameter learning for multiple local network alignment," *Journal of Computational Biology*, vol. 16, no. 8, pp. 1001–1022, 2009.

[41] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12763–12768, 2008.

[42] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, "IsoRankN: spectral methods for global alignment of multiple protein networks," *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.

[43] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[44] M. P. H. Stumpf and M. A. Porter, "Critical truths about power laws," *Science*, vol. 335, no. 6069, pp. 665–666, 2012.

[45] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[46] P. Jiang and M. Singh, "SPICi: a fast clustering algorithm for large biological networks," *Bioinformatics*, vol. 26, no. 8, Article ID btq078, pp. 1105–1111, 2010.

[47] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos, "Growing functional modules from a seed protein via integration of protein interaction and gene expression data," *BMC Bioinformatics*, vol. 8, article 408, 2007.

[48] J. Seo, M. Bakay, P. Zhao et al., "Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 3, pp. 461–464, 2003.

[49] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guénoche, and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network," *Genome Biology*, vol. 5, no. 1, article R6, 2003.

[50] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[51] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[52] H. Lu, X. Zhu, H. Liu et al., "The interactome as a tree: an attempt to visualize the protein-protein interaction network in yeast," *Nucleic Acids Research*, vol. 32, no. 16, pp. 4804–4811, 2004.

[53] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, Article ID 026113, 2004.

[54] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[55] R. W. Solava, R. P. Michaels, and T. Milenkovic, "Graphlet-based edge clustering reveals pathogen-interacting proteins," *Bioinformatics*, vol. 28, pp. i480–i486, 2012.

[56] A. D. King, N. Pržulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[57] F. Glover, "Tabu search. Part I," *ORSA Journal on Computing*, vol. 1, pp. 190–206, 1989.

[58] R. B. Altman and S. Raychaudhuri, "Whole-genome expression analysis: challenges beyond clustering," *Current Opinion in Structural Biology*, vol. 11, no. 3, pp. 340–347, 2001.

[59] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. 1, pp. S233–S240, 2002.

[60] S. Zhang, X.-M. Ning, C. Ding, and X.-S. Zhang, "Determining modular organization of protein interaction networks by maximizing modularity density," *BMC Systems Biology*, vol. 4, no. 2, article 10, 2010.

[61] F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for detecting significantly mutated pathways in cancer," *Journal of Computational Biology*, vol. 18, no. 3, pp. 507–522, 2011.

[62] L. Shi, X. Lei, and A. Zhang, "Protein complex detection with semi-supervised learning in protein interaction networks," *Proteome Science*, vol. 9, no. 1, article S5, 2011.

[63] Y. Qi, F. Balem, C. Faloutsos, J. Klein-Seetharaman, and Z. Bar-Joseph, "Protein complex identification by supervised graph local clustering," *Bioinformatics*, vol. 24, no. 13, pp. i250–i268, 2008.

[64] S. M. V. Dongen, *Graph clustering by flow simulation [Ph.D. thesis]*, University of Utrecht, 2002.

[65] The Gene Ontology Consortium, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[66] C. Jiang, F. Coenen, and M. Zito, "A survey of frequent subgraph mining algorithms," *The Knowledge Engineering Review*, vol. 28, no. 01, pp. 75–105, 2013.

[67] M. Kuramochi and G. Karypis, "An efficient algorithm for discovering frequent subgraphs," Technical Report, University of Minnesota, Department of Computer Science, 2002.

[68] X. Yan and J. Han, "gSpan: graph-based substructure pattern mining," in *Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM '02)*, pp. 721–724, December 2002.

[69] S. Brohée and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, article 488, 2006.

[70] M. Dhara and K. K. Shukla, "Comparative performance analysis of RNSC and MCL algorithms on power-law distribution," *Advanced Computing*, vol. 3, pp. 19–34, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 154594, 8 pages
http://dx.doi.org/10.1155/2014/154594

Research Article

An Unsupervised Approach to Predict Functional Relations between Genes Based on Expression Data

Md. Altaf-Ul-Amin, Tetsuo Katsuragi, Tetsuo Sato, Naoaki Ono, and Shigehiko Kanaya

Computational Systems Biology Lab, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan

Correspondence should be addressed to Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 1 November 2013; Revised 31 January 2014; Accepted 3 February 2014; Published 31 March 2014

Academic Editor: Farit Mochamad Afendi

This work presents a novel approach to predict functional relations between genes using gene expression data. Genes may have various types of relations between them, for example, regulatory relations, or they may be concerned with the same protein complex or metabolic/signaling pathways and obviously gene expression data should contain some clues to such relations. The present approach first digitizes the log-ratio type gene expression data of *S. cerevisiae* to a matrix consisting of 1, 0, and –1 indicating highly expressed, no major change, and highly suppressed conditions for genes, respectively. For each gene pair, a probability density mass function table is constructed indicating nine joint probabilities. Then gene pairs were selected based on linear and probabilistic relation between their profiles indicated by the sum of probability density masses in selected points. The selected gene pairs share many Gene Ontology terms. Furthermore a network is constructed by selecting a large number of gene pairs based on FDR analysis and the clustering of the network generates many modules rich with similar function genes. Also, the promoters of the gene sets in many modules are rich with binding sites of known transcription factors indicating the effectiveness of the proposed approach in predicting regulatory relations.

1. Introduction

The cell works as a system governed by integrated action of the genes indicating that genes are functionally related; for example, they may have regulatory relations between each other or they may be concerned with the same protein complex or metabolic/signaling pathways and so on. Determining functional relations between genes enables development of a genetic network which leads to the prediction of the complex rolls of the genes in different systems in the cell. Nucleotide and/or amino acid sequence similarities have been extensively used to predict functional relation between genes [1, 2]. Affinity purification [3, 4] and yeast two-hybrid assays [5, 6] are employed to determine physical association between proteins which are gene products. Synthetic lethal screens [7] measure the tendency for genes to compensate the loss of other genes. Scientists have performed numerous studies in an attempt to better understand and classify digenic epistatic relationships [8]. In [9] a probabilistic functional network of yeast genes was constructed by integrating diverse genomic data. In [10] an algorithm was proposed for regulatory networks of gene modules that combines information from genome wide location and expression data sets. Constraint-based Bayesian Structure Learning (BSL) techniques, namely, (a) PC Algorithm, (b) Grow-shrink (GS) algorithm, and (c) Incremental Association Markov Blanket (IAMB), were used to model the functional relationships between genes associated with differentiation potential of aged myogenic progenitors in the form of acyclic networks from the clonal expression profiles [11]. Attempts have been made not only to determine functional relationship between individual genes but also to measure functional relationship between gene sets [12]. Many more similar studies can be cited. Microarray gene expression data incorporating with other information have been extensively used for predicting regulatory relation between genes [13–15]. However it is logical to assume that expression data contains information about various types of functional relations between genes. In the present work we

propose an approach for estimating integrated linear and probabilistic relations between expression profiles of genes and applied the concept to determine functional relations between yeast genes solely based on gene expression data. The proposed method successfully detected functionally related gene pairs that share many GO terms. The method also shows promise to be utilized in the process of detecting regulatory relations between genes.

2. Materials and Methods

2.1. Data Used in This Work. The data used in this work was previously used in other works [16–19]. The data is a 2467 × 79 matrix containing some missing values. Each data point produced by a DNA microarray hybridization experiment represents the log of the ratio of expression levels of a particular gene under two different experimental conditions. The result, from an experiment with n genes on a single chip, is a series of n log-transformed expression-level ratios. Typically, the numerator of each ratio is the expression level of the gene in the varying condition of interest, whereas the denominator is the expression level of the gene in some reference condition. The expression measurement is positive if the gene is induced (turned up) with respect to the reference state and negative if it is repressed (turned down). The data were collected at various time points during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks.

2.2. Missing Value Imputation. In microarray gene expression data missing values often occur due to various reasons, such as insufficient resolution, image corruption, dust, or scratches on the slide. Usually, microarray datasets are estimated to have more than 5% missing values and up to 90% of genes are affected [20, 21]. The gene expression data considered in this work contains 3760 missing values. The missing values were filled based on principal component analysis (PCA) by using the R package pcaMethods [22]. Using PCA we can model a matrix M by defining two parameter matrices, the scores, T, and the loadings, P, such that when multiplied with each other they well reconstruct the original matrix as follows:

$$M = 1 \times \overline{m} + TP^t + E, \tag{1}$$

where E is the error matrix and $1 \times \overline{m}$ denotes the original variable averages. Now if M contains missing values but P and T can be completely estimated, then we can use

$$\widehat{M} = 1 \times \overline{m} + TP^t \tag{2}$$

as an estimate for M_{ij} if M_{ij} is missing.

2.3. Digitization of Gene Expression Matrix. After missing value imputation, let us denote the gene expression data matrix as M. For each row of M we calculate the average and standard deviation. Let for the ith row the average and standard deviations be denoted as avg_i and sd_i. Now, the digitized matrix D is created as follows:

$$\begin{aligned} D_{ij} &= 1 \quad \text{if } M_{ij} \geq \text{avg}_i + \text{th} \times \text{sd}_i \\ D_{ij} &= -1 \quad \text{if } M_{ij} \leq \text{avg}_i - \text{th} \times \text{sd}_i \\ D_{ij} &= 0 \quad \text{otherwise.} \end{aligned} \tag{3}$$

In the above equations "th" is a threshold which should be a real number and in most practical cases it is within 0 to 2. We digitized the data using the values of threshold "th" as 0.5, 1, and 1.5. For each case the distribution of the genes with respect to the count of 1 s in their profiles is shown in Figure 1. In case of th = 0.5, the distribution approaches roughly normal and we observed similar trend in case of −1. Hence in this work we considered th = 0.5 for the digitization of the gene expression data.

Table 1: Nine joint probabilities calculated for each gene pair.

a/b	1	0	−1
1	$P(1, 1)$	$P(1, 0)$	$P(1, -1)$
0	$P(0, 1)$	$P(0, 0)$	$P(0, -1)$
−1	$P(-1, 1)$	$P(-1, 0)$	$P(-1, -1)$

2.4. Probability Density Mass Function Table. Based on a digitized matrix containing only 1, 0, and −1 a probability density mass function table can be constructed corresponding to each gene pair indicating nine joint probabilities as shown in Table 1.

Any element of the above table $P(k, k')$ (corresponding to two genes say, gene a and gene b) where $k, k' \in \{1, 0, -1\}$ can be calculated by assuming TRUE = 1 and FALSE = 0 in (4) as follows:

$$P\left(k, k'\right) = \frac{\sum_{i=1}^{N} D_{ai} == k \text{ AND } D_{bi} == k'}{N}. \tag{4}$$

Here N is the width of matrix D.

We assume that the joint probabilities of Table 1 and corresponding conditional probabilities contain important clues to estimate functional relations between genes.

2.5. Hypothesis. In this work we hypothesize that when gene a is positively functionally related to gene b, then $P(b = 1 \mid a = 1)$ should be statistically high. Using Bayes rule we can write $P(b = 1 \mid a = 1) = P(a = 1, b = 1)/P(a = 1)$. Now if $P(a = 1)$ is very small, then $P(b = 1 \mid a = 1)$ can be very high and that can sometimes happen because of noisy data. To avoid this problem we can consider $P(b = 1, a = 1)$ as an indicator that gene a is positively functionally related to gene b. To further strengthen the case we consider that when both $P(b = 1, a = 1)$ and $P(b = 1, a = 1) + P(b = 0, a = 0) + P(b = -1, a = -1)$ are statistically significant then gene a and gene b are positively functionally related. Considering other joint probability masses might be useful for finding functional relations between some multi function genes. By intuition we can realize that the sum of

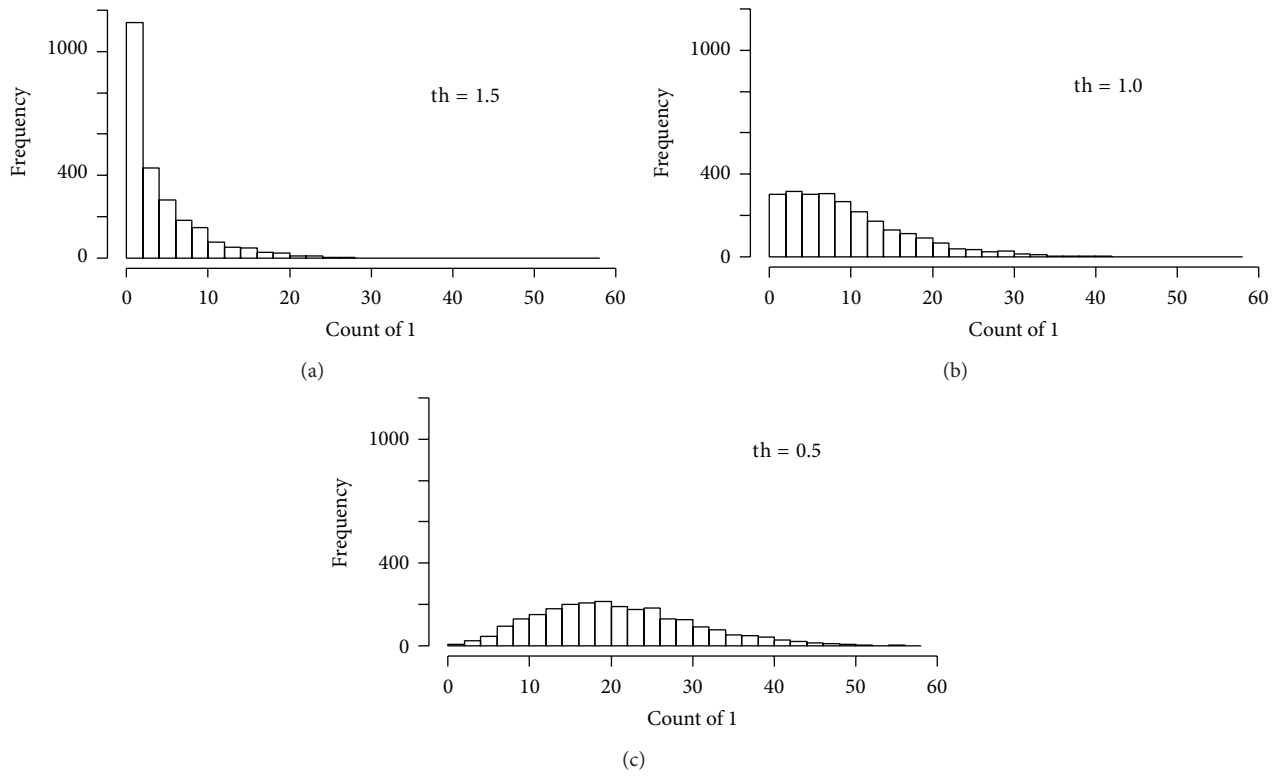


Figure 1: Distribution of the genes with respect to the count of 1 in their profiles in the context of the digitized matrix.

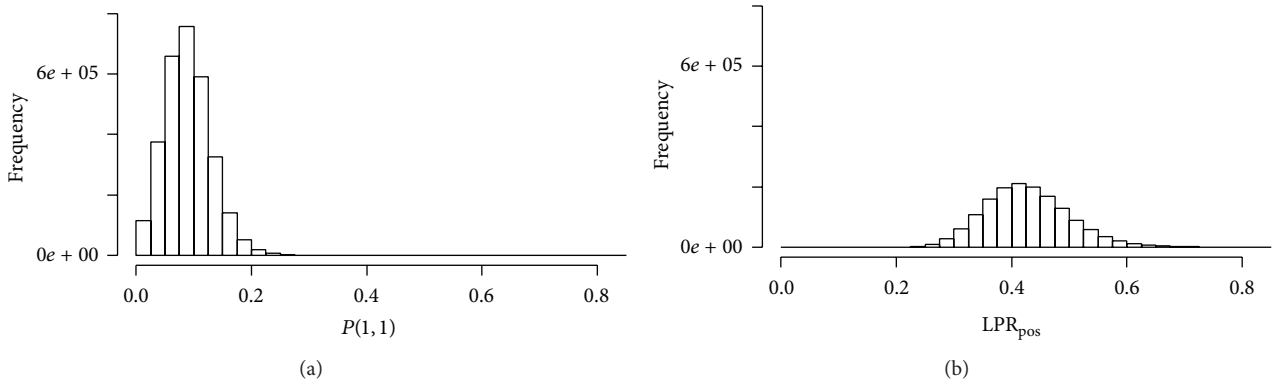


Figure 2: Distribution of gene pairs in the context of (a) $P(1,1)$ and (b) $\mathrm{LPR}_{\mathrm{pos}}$.

probabilities $P(b = 1, a = 1) + P(b = 0, a = 0) + P(b = -1, a = -1)$ actually indicates an integrated measure of both linear and probabilistic relations between the profiles of two genes and this term will be referred to as positive linear and probabilistic relation ($\mathrm{LPR}_{\mathrm{pos}}$) in the following. To our knowledge this is the first approach to measure similarity between two multivariate entities based on joint probability density masses in selected points giving emphasis on both linear and probabilistic relations.

3. Results

3.1. Effectiveness of LPR_{pos}. The distribution of all gene pairs in the context of $P(1,1)$ is shown in Figure 2(a). The average value of $P(1,1)$ is 0.0819. We calculated $\mathrm{LPR}_{\mathrm{pos}}$ for the gene pairs for which $P(1,1)$ is larger than the average value. The distribution of those gene pairs with respect to $\mathrm{LPR}_{\mathrm{pos}}$ is shown in Figure 2(b). The average value of $\mathrm{LPR}_{\mathrm{pos}}$ is 0.429. Initially we selected the highest 1%, 2%, 3%, 4%, and 5% gene pairs from the distribution of Figure 2(b), that is, gene pairs with higher $\mathrm{LPR}_{\mathrm{pos}}$ values, and determined the number of GO terms [23] shared by both the genes of each pair.

Figure 3(a) shows the percentage of selected gene pairs that share at least 1, 2, and, 3 GO terms and also that of equal number of randomly selected gene pairs. In the context of minimum number of shared GO terms the percentage of selected gene pairs is always much higher compared to that

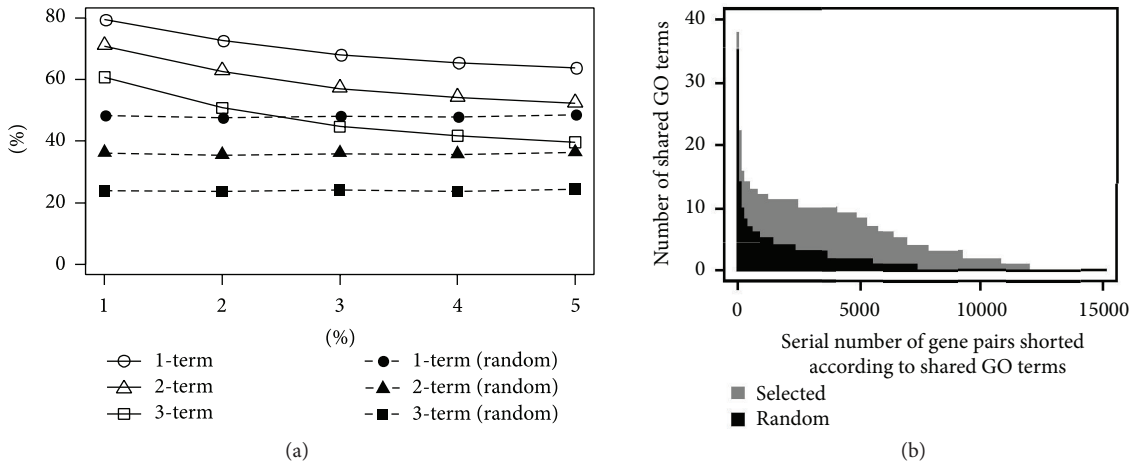


FIGURE 3: (a) x-axis is percentage of gene pairs of the distribution of Figure 2(b) selected based on higher LPR_{pos} values and y-axis is percentage of selected gene pairs that share at least 1, 2, or 3 GO terms. Empty markers correspond to gene pairs selected by the proposed method and filled markers corresponding to equal number of randomly selected gene pairs. (b) Actual number of GO terms shared by selected and random gene pairs corresponding to the 1% point of (a).

of randomly selected pairs. Figure 3(a) further shows that the higher the lower cutoff value of LPR_{pos} for a group of gene pairs is, the higher proportion of the gene pairs share common GO terms. To further illustrate the result we show in Figure 3(b) the actual number of shared GO terms for the highest 1% selected gene pairs and the equal number of random gene pairs which implies that the gene pairs selected based on LPR_{pos} share much more GO terms. Thus LPR_{pos} is a good measure to determine functional relation between genes.

3.2. FDR Analysis. We conducted FDR (false discovery rate) [24, 25] analysis to statistically assess the false positive rates among the selected gene pairs based on LPR_{pos}. For each pair of genes for which $P(1, 1)$ is above average we did the following.

(i) The numbers of 1 s, 0 s, and −1 s in the digital profile of both genes are counted.

(ii) Random profiles of both the genes are constructed by randomly imputing the same numbers of 1 s, 0 s, and −1 s. This process is repeated 100 times.

(iii) Then, $C(1, 1)$, $C(0, 0)$, and $C(-1, -1)$ are calculated for both real and random profile pairs. $C(k, k)\{k \in 1, 0, -1\}$ is the total number of profile points for which the expression level of both genes is k. In case of random profiles the average values corresponding 100 random profile pairs were considered.

(iv) A chi-square value is calculated as follows where N is the width of the expression matrix:

$$\chi^2 = \left[\sum_{k=1,0,-1} \frac{\{C(k,k)_{\text{real}} - C(k,k)_{\text{random}}\}^2}{C(k,k)_{\text{random}}}\right] + \frac{\{\sum_{k=1,0,-1} C(k,k)_{\text{real}} - \sum_{k=1,0,-1} C(k,k)_{\text{random}}\}^2}{N - \sum_{k=1,0,-1} C(k,k)_{\text{random}}}. \quad (5)$$

(v) Based on the chi-square value, a P-value for the gene pair is determined using R statistical software. Note that LPR_{pos} is directly proportional to $\sum_{k=1,0,-1} C(k, k)_{\text{real}}$.

Figure 4(a) shows the distribution of the gene pairs with respect to the P-values with a P-value interval of 0.05. For any given cutoff P-value the FDR is calculated as follows:

$$\text{FDR} = \frac{(\text{Total \# of gene pairs}) \times (P\text{-value})_{\text{cut-off}}}{\text{\# of gene pairs with } P\text{-value less than } (P\text{-value})_{\text{cut-off}}}. \quad (6)$$

Figure 4(b) shows the plot of FDR with respect to cutoff P-values. As the cutoff P-value decreases, FDR decreases rapidly and becomes roughly constant at P-value of 0.001. There are 25559 gene pairs for which the P-value is less than 0.001.

3.3. Network and Modules of the Selected Gene Pairs. Based on the FDR analysis of the above section, we selected 25559 gene pairs having highest LPR_{pos} values. Such selected gene pairs make a network consisting of 2131 nodes. We determined high density modules in that network using the network clustering algorithm DPClusO [26] and found 1154 modules of size 3 or more (see Supplementary File 1 in supplementary material available online at http://dx.doi.org/10.1155/2014/154594).

3.3.1. Richness of Similar Function Genes. To evaluate the richness of similar function genes in the modules we calculated their hypergeometric P-values by using the R package GOstats [27] in the context of all three types of GO terms: biological process (BP), cellular compartment (CC), and molecular function (MF). Figures 5(a), 5(b), and 5(c) show the distribution of the modules with respect to

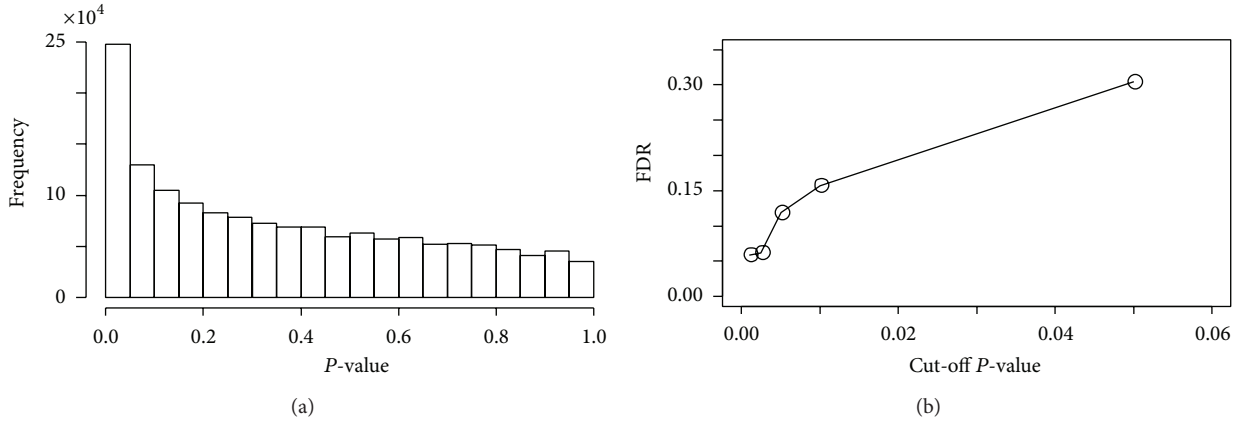


Figure 4: (a) Distribution of the gene pairs with respect to the χ-square P-values. (b) Plot of FDR with respect to cutoff P-values.

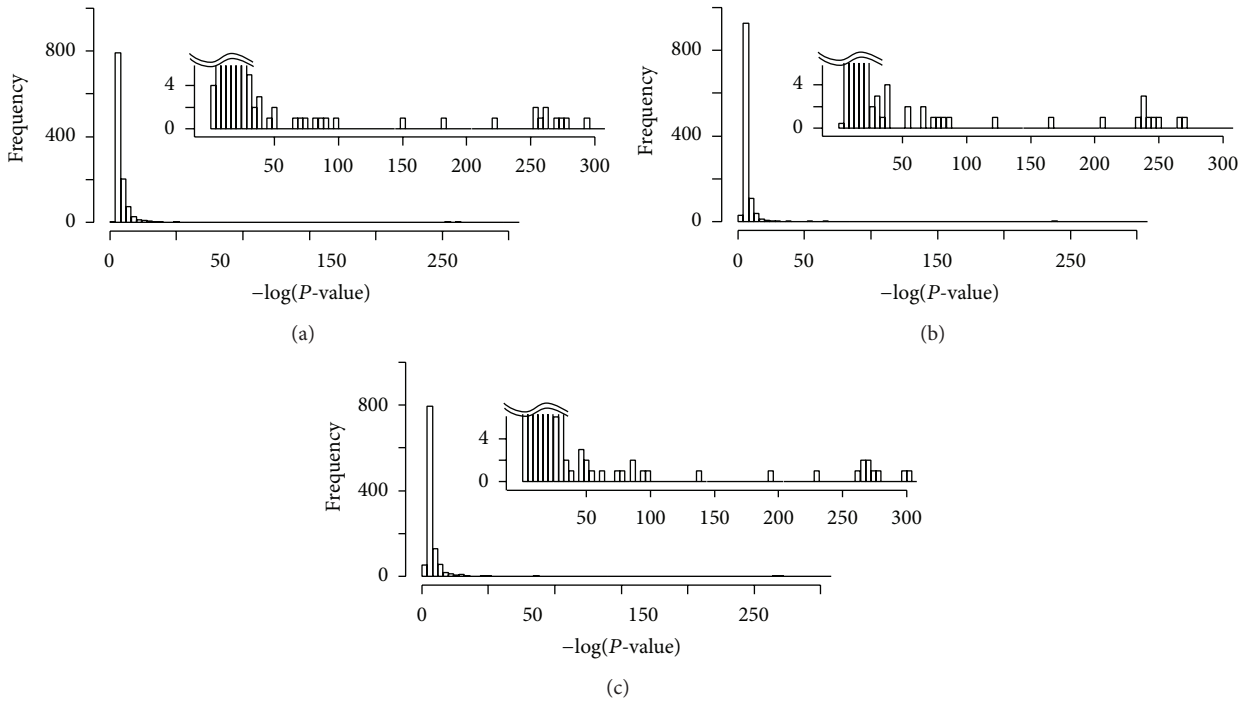


Figure 5: Distribution of the modules with respect to −log(P-value). P-values determined in the context of all three types of GO terms (a) biological process (BP), (b) molecular function (MF), and (c) cellular compartment (CC). The lower part of each graph is enlarged in the insets.

−log(P-value) which implies that almost all the modules are statistically significant. We selected 10 lowest P-value clusters corresponding to different GO terms from each of the three distributions of Figure 5 and their set union resulted in 22 clusters. Some biological information from the SGD database [28] about those 22 clusters is summarized in Table 2. Column 3 in Table 2 shows the P-values and corresponding GO terms determined by GOstats. Column 4 in Table 2 shows other GO terms retrieved from SGD database associated to the clusters covering many genes which implies that almost all the genes of each of the clusters could be associated to important GO terms which confirms the fact that the proposed method is a promising way to establish functional relation between genes based on expression data.

3.3.2. Richness of Similar Binding Sites. Furthermore to verify the presence of similar binding sites in the promoters of the genes included in individual modules we used the tool PRIMA (PRomoter Integration in Microarray Analysis) [29] from the software package EXPANDER [30]. Total 180 modules were found to have P-values less than 10^{-3} in the context of binding site enrichment of 57 various transcription

Table 2: Richness of similar function genes in selected clusters. For each cluster, hypergeometric *P*-values, corresponding GO terms, and also the actual number of genes of a particular function are indicated.

CID	Total number of genes	*P*-value/GO ID (From GOstats result)	Some relevant GO terms (corresponding number of genes) (From SGD database)
4	97	$1.20E-131$/GO:0022626 (CC) $2.62E-117$/GO:0002181 (BP) $4.80E-129$/GO:0003735 (MF)	Cytosolic ribosome (94), structural constituent of ribosome (94), cytoplasmic translation (93), ribosome (96)
16	76	$6.42E-24$/GO:0044391 (CC) $7.23E-17$/GO:0006412 (BP)	Ribosomal subunit (37), structural molecule activity (38)
19	73	$3.29E-23$/GO:0030529 (CC)	Ribonucleoprotein complex (47), intracellular part (73)
226	8	$1.50E-20$/GO:0000788 (CC) $1.93E-14$/GO:0006333 (BP)	Nuclear nucleosome (8), DNA bending complex (8)
1	113	$1.42E-17$/GO:0042254 (BP)	Cellular metabolic process (104), intracellular part (109)
44	34	$2.89E-16$/GO:0005840 (CC)	Cytosolic part (21), cytoplasm (34)
35	44	$3.35E-16$/GO:0010467 (BP)	Gene expression (41), primary metabolic process (43)
85	17	$4.76E-14$/GO:0044429 (CC)	Mitochondrial part (14), mitochondrion (16)
155	11	$6.28E-14$/GO:0051082 (MF) $4.97E-13$/GO:0006457 (BP)	Protein folding (9), protein binding (11), cellular protein metabolic process (10)
278	7	$3.00E-13$/GO:0000502 (CC)	Proteasome complex (7), proteasome storage granule (5)
87	16	$5.26E-13$/GO:0005730 (CC)	Nucleolus (12), non-membrane-bounded organelle (14)
107	14	$1.97E-12$/GO:0007005 (BP)	Mitochondrion organization (12), cellular component organization (13)
121	13	$5.32E-12$/GO:0006094 (BP)	Glycolysis (7), generation of precursor metabolites and energy (9)
442	5	$1.55E-11$/GO:0022904 (BP)	Mitochondrial respiratory chain (5), oxidoreductase complex (5)
173	10	$1.56E-11$/GO:0006457 (BP) $2.58E-08$/GO:0051082 (MF)	Protein folding (7), unfolded protein binding (5), protein binding (8)
282	7	$5.58E-11$/GO:0004298 (MF)	Modification-dependent protein catabolic process (7), roteasomal ubiquitin-independent protein catabolic process (5)
71	15	$5.90E-11$/GO:0005840 (CC)	Ribosome (13), ribonucleoprotein complex (14)
725	3	$1.61E-09$/GO:0003993 (MF)	Acid phosphatase activity (2)
214	9	$2.88E-09$/GO:0008121 (MF)	Hydrogen ion transmembrane transporter activity (5), single-organism metabolic process (7)
736	3	$4.03E-09$/GO:0004067 (MF)	Asparaginase activity (3)
1092	3	$2.26E-08$/GO:0015002 (MF)	Heme-copper terminal oxidase activity (3)
270	7	$2.32E-08$/GO:0015078 (MF)	Ion transmembrane transporter activity (6)

Table 3: Richness of binding sites in the promoters of the module genes corresponding to 10 different transcription factors.

CID	Size	TF	Number of Promo. (PRIMA)	*P*-value	Known regulatory relations (YEASTRACT)
3	98	YP00066 [SFP1]	58	$2.82E-42$	98
5	95	M00213 [RAP1]	55	$3.82E-28$	93
72	18	YP00036 [MBP1]	10	$4.40E-12$	12
155	11	M00169 [HSF]	7	$2.38E-09$	11
230	8	YP00068 [SIP4]	5	$7.89E-09$	4
227	8	YP00064 [RPN4]	8	$1.01E-08$	8
725	3	M00064 [PHO4]	3	$1.08E-08$	3
259	7	YP00076 [STB1]	5	$8.97E-08$	2
736	3	YP00013 [DAL82]	3	$3.65E-07$	0
233	8	YP00043 [MSN4]	8	$1.03E-06$	7

factors. The enrichment table generated by EXPANDER is in supplementary material (Supplementary Table 1). Table 3 shows information about 10 modules corresponding to lowest *P*-values involving 10 different transcription factors. We downloaded a list of known regulatory relations from the YEASTRACT database [31] and verified whether the genes in a module have regulatory relation with the associated transcription factor. Column 6 of Table 3 shows that a large number of genes in individual modules are already reported to be regulated by the corresponding transcription factor. Only in case of CID736, though all 3 genes contain in their promoters the binding site of the transcription factor DAL82, no regulatory relation between those genes is reported in the YEASTRACT database presently. However based on our analysis regulatory relations between DAL82 and those three genes may be predicted. Thus the proposed measure can also be integrated to other types of information for developing a method to predict regulatory relations between genes which is one of our future works.

4. Conclusions

In this work we propose a novel measure to determine functional relation between genes based on gene expression data. The present approach first digitizes the log-ratio type gene expression data to a matrix consisting of 1, 0, and −1 indicating highly expressed, no major change and highly suppressed conditions for genes, respectively. Then a probability density mass function table is constructed indicating nine joint probabilities for each pair of genes. Those pairs of genes were considered as functionally related for which the sum of probability density masses in selected points are statistically significant. We applied the method to a sample gene expression data of *S. cerevisiae*. It was found that substantial majority of the selected gene pairs share many GO terms. Also the network consisting of the selected gene pairs contains high density modules. It was shown that those modules were rich with similar function genes. Furthermore, it was verified that for many modules many of the genes contain similar binding sites in their promoters corresponding to known transcription factors of yeast and those transcription factors are known regulators of many of the genes in the corresponding module. Above all this work introduces a new approach for simultaneously measuring both linear and probabilistic relations between multivariate entities which is useful for handling multivariate data and big data biology.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is supported by the National Bioscience Database Center in Japan, the Ministry of Education, Culture, Sports, science, and Technology of Japan (Grant-in-Aid for Scientific Research on Innovation Areas "Biosynthetic Machinery. Deciphering and Regulating the System for Creating Structural Diversity of Bioactivity Metabolites (2007)").

References

[1] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal, "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues," *Bioinformatics*, vol. 18, no. 1, pp. S71–S77, 2002.

[2] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.

[3] A.-C. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[4] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[5] P. Uetz, L. Glot, G. Cagney et al., "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[6] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[7] A. H. Y. Tong, M. Evangelista, A. B. Parsons et al., "Systematic genetic analysis with ordered arrays of yeast deletion mutants," *Science*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[8] I. Miko, "Epistasis: gene interaction and phenotype effects," *Nature Education*, vol. 1, article 197, 2008.

[9] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, no. 5701, pp. 1555–1558, 2004.

[10] Z. Bar-Joseph, G. K. Gerber, T. I. Lee et al., "Computational discovery of gene modules and regulatory networks," *Nature Biotechnology*, vol. 21, no. 11, pp. 1337–1342, 2003.

[11] R. Nagarajan, S. Datta, M. Scutari, M. L. Beggs, G. T. Nolen, and C. a Peterson, "Functional relationships between genes associated with differentiation potential of aged myogenic progenitors," *Frontiers in Physiology*, vol. 1, article 21, 2010.

[12] Q. Wang, J. Sun, M. Zhou et al., "A novel network-based method for measuring the functional relationship between gene sets," *Bioinformatics*, vol. 27, no. 11, pp. 1521–1528, 2011.

[13] A. A. Margolin, I. Nemenman, K. Basso et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, article S7, 2006.

[14] A. Kundaje, M. Middendorf, M. Shah, C. H. Wiggins, Y. Freund, and C. Leslie, "A classification-based framework for predicting and analyzing gene regulatory response," *BMC Bioinformatics*, vol. 7, supplement 1, article S5, 2006.

[15] E. Segal, M. Shapira, A. Regev et al., "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[16] M. P. S. Brown, W. N. Grundy, D. Lin et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

[17] A. Mateos, J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky, "Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons," *Genome Research*, vol. 12, no. 11, pp. 1703–1715, 2002.

[18] A. Lægreid, T. R. Hvidsten, H. Midelfart, J. Komorowski, and A. K. Sandvik, "Predicting gene ontology biological process from temporal gene expression patterns," *Genome Research*, vol. 13, no. 5, pp. 965–979, 2003.

[19] G. P. S. Raghava and J. H. Han, "Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein," *BMC Bioinformatics*, vol. 6, article 59, 2005.

[20] J. Tuikkala, L. Elo, O. S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in microarray data with gene ontology," *Bioinformatics*, vol. 22, no. 5, pp. 566–572, 2006.

[21] M. Ouyang, W. J. Welsh, and P. Georgopoulos, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, vol. 20, no. 6, pp. 917–923, 2004.

[22] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, "pcaMethods—a bioconductor package providing PCA methods for incomplete data," *Bioinformatics*, vol. 23, no. 9, pp. 1164–1167, 2007.

[23] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[24] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[25] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[26] M. Altaf-Ul-Amin, M. Wada, and S. Kanaya, "Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking," *ISRN Biomathematics*, vol. 2012, Article ID 726429, 11 pages, 2012.

[27] S. Falcon and R. Gentleman, "Using GO stats to test gene lists for GO term association," *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.

[28] J. M. Cherry, C. Adler, C. Ball et al., "SGD: saccharomyces genome database," *Nucleic Acids Research*, vol. 26, no. 1, pp. 73–79, 1998.

[29] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh, "Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells," *Genome Research*, vol. 13, no. 5, pp. 773–780, 2003.

[30] R. Shamir, A. Maron-Katz, A. Tanay et al., "EXPANDER—an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, article 232, 2005.

[31] D. Abdulrehman, P. T. Monteiro, M. C. Teixeira et al., "YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface," *Nucleic Acids Research*, vol. 39, no. 1, pp. D136–D140, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 527029, 10 pages
http://dx.doi.org/10.1155/2014/527029

Review Article

Survey of Network-Based Approaches to Research of Cardiovascular Diseases

Anida Sarajlić and Nataša Pržulj

Department of Computing, Imperial College London, 180 Queen's Gate, South Kensington Campus, London SW72AZ, UK

Correspondence should be addressed to Nataša Pržulj; natasha@imperial.ac.uk

Received 26 November 2013; Accepted 7 February 2014; Published 20 March 2014

Academic Editor: Altaf-Ul- Amin

Cardiovascular diseases (CVDs) are the leading health problem worldwide. Investigating causes and mechanisms of CVDs calls for an integrative approach that would take into account its complex etiology. Biological networks generated from available data on biomolecular interactions are an excellent platform for understanding interconnectedness of all processes within a living cell, including processes that underlie diseases. Consequently, topology of biological networks has successfully been used for identifying genes, pathways, and modules that govern molecular actions underlying various complex diseases. Here, we review approaches that explore and use relationships between topological properties of biological networks and mechanisms underlying CVDs.

1. Introduction

Cardiovascular diseases (CVDs) cover a broad range of disorders which affect different parts of cardiovascular system and include coronary diseases, carotid diseases, peripheral arterial diseases, and aneurysms. They remain the leading health problem which affects more than 80 million individuals in the United States alone [1]. Based on the data from 2009, in the United States, on average one person dies of CVDs every 40 seconds. Coronary heart disease alone causes one out of every six deaths [1]. By year 2020 it is expected that Brazil, Russia, India, and China will contribute significantly to a global increase of additional 4% of deaths caused by CVDs [2].

Etiology of cardiovascular diseases is not simple. There are forms of CVDs that are Mendelian disorders resulting from a mutation on a single gene [3]. However, the majority are complex diseases occurring as a result of an interplay between multiple genes [3], as well as a variety of factors such as diet, dyslipidemia, hypertension, and body mass index [4]. For addressing this complexity, an integrative approach, that would take into account coaction between the multiple causes behind CVDs, seems to be the method of choice. This is because properties of a complex system as a whole cannot be completely discovered by simply observing properties of individual parts of the system without taking into account their interconnectedness [5]. Hence, different systems biology approaches have been used in CVD research, which has recently been reviewed elsewhere [6–9].

A mathematical concept of a *network* has been introduced in systems biology as it accurately captures the inner workings of many complex biological systems. For example, metabolic pathways are interconnected into a network, providing redundancy, adaptability, and robustness [10], thus enabling energy-efficient production of metabolites. Also, the fact that a specific network topology comes as a direct consequence of biological processes occurring between the elements of the underlying system highlights the importance of the topology as a valuable source of new biological knowledge.

In this survey, we focus on network-based systems biology approaches to CVD research. More specifically, we aim to investigate the extent to which network topology has contributed to novel medical insights into CVDs.

2. Topology of Biological Networks Reveals Disease Genes, Modules, and Pathways

2.1. Biological Data and Networks. Recent advances in high-throughput techniques have resulted in a number of large-scale biological data sets. In Table 1, we list commonly

Table 1: Databases of human molecular interaction and disease ontology data.

Database name	Type of data	URL
BioGRID	PPI and genetic interactions	http://thebiogrid.org/
HPRD	PPI, disease associations, posttranslational modifications, tissue expression, subcellular localization, and enzyme/substrate relationships	http://www.hprd.org/
DIP	Experimentally determined PPI	http://dip.doe-mbi.ucla.edu/dip/
HomoMINT	PPI	http://mint.bio.uniroma2.it/HomoMINT/
OPHID	PPI	http://ophid.utoronto.ca/ophidv2.204/
KEGG	Pathway maps, human diseases, drugs, orthology groups, genes, relations within genes, metabolites, biochemical reactions, and enzymes	http://www.genome.jp/kegg/
OMIM	Information on genes and genetic disorders	http://www.ncbi.nlm.nih.gov/omim

used databases of molecular interaction and disease ontology data for *H. sapiens*. These databases accumulate biological information, including interactions and relationships among biological macromolecules and metabolites, such as protein-protein interactions (PPI), genetic interactions, or enzyme-substrate relationships. The available data also include gene functional annotations, pathway maps, information on genetic disorders, and disease associations. As an example of the scale of available data, BioGRID currently lists 303,268 nonredundant physical interactions between 51,129 proteins across 48 organisms, while DRYGIN (http://drygin.ccbr.utoronto.ca/) contains 5,482,948 genetic interactions for *S. cerevisiae*.

A network is the same as a mathematical concept of a graph, denoted as a pair $G = \{V, E\}$, where V is a set of vertices (nodes) and E is a set of links (edges) that connect pairs of nodes [11]. When constructing a graph it is necessary to determine how biological elements and relations between them correspond to nodes and edges. For example, an edge in a protein network can be placed between two proteins if they bind together to perform their biological function; this results in a commonly used protein-protein interaction (PPI) network. Conversely, an edge between two proteins can also be placed if the two proteins share a common trait, such as being targeted by the same drug or causing the same disease. These associations are usually found by mining the scientific literature, resulting in an association network. Other highly exploited networks are genetic interaction networks, where genes correspond to nodes in the graph and edges represent functional associations between genes: an interaction between two genes occurs when the result of mutations in the genes is not just a combination of phenotypes of those mutations [12]. A metabolic network is a union of all metabolic pathways within a cell, where nodes correspond to metabolites and enzymes, and directed edges are metabolic reactions [10, 13, 14]. Regulator-gene interactions can be summed up into a transcriptional regulatory network [15]. Given various experimental limitations, up till date, only a handful of transcriptional regulatory networks for complex biological systems have been defined [16].

Graph theoretic approaches offer insight into the structure of these networks and allow us to single out properties of a network, or its parts, which are different from expected by random. Such findings can reveal the connection between a specific topological characteristic and related biological function or a process, such as a disease. Here, we will not provide details on global and local network properties nor specific algorithms commonly used in graph theory, such as algorithms for network clustering or alignment. For more details on these topics, see [17–20].

Note that a limiting factor regarding network analyses is the quality of data. Although large amounts of biological data are available, they are still noisy and incomplete. Techniques used for obtaining the data are often biased—they may not provide enough sensitivity to detect all changes in the system [21]. Outcomes of experiments depend not only on experimental design but also on the stringency of conditions of the experiments: for example, too stringent conditions can lead to false negative interactions, as opposed to false positive results obtained from experiments that were not stringent enough. Also, depending on the focus of the research and experimental design, some genes/proteins can be favoured and their possible interactions are explored more often, such as those of disease genes. This can impose a particular structure in the network, for example, false hubs, without reflecting the underlying network topology. In addition, not all biological processes can be accurately represented as interactions (edges in the network) between two elements. Often a process in a biological system requires more than two elements and involves different types of interactions. However, a benefit is that network representation gives an opportunity to reduce the complexity of biological data that is required for performing computational analyses. Different data sources offer various insights into underlying biological processes, and, only if integrated, they will yield the full meaning. Network analysis provides exactly insight into interconnectedness of the data that describe different processes within a living cell. Below we give a short overview of methods that use biological networks to extract new knowledge about diseases. Specifically, we focus on network biology in CVDs.

2.2. Exploring Disease through Network Topology. Topology of PPI networks has widely been explored and used for inferring involvement of proteins in biological functions and processes, including diseases. It has been shown that proteins

that are closer in the network are more likely to perform the same function [22]. In particular, *association by guilt* approach was used to infer functions of unannotated proteins: the direct neighbourhoods of proteins were examined looking for most common functions among annotated direct neighbours [23]. Similarly, the n neighbourhood of proteins [24] and shared neighbours of proteins [25] were analysed to annotate functions of unannotated proteins. These properties were used to associate genes with diseases using linkage methods (nomenclature adopted from [26]). In that sense, it has been shown that directly linked proteins in the human PPI network are more likely to cause similar diseases [27, 28] (simplified concept illustrated in Figure 1, panel (a)). A variant of linkage method was successfully applied to discover genes related to Alzheimer's disease [29].

Several other methods have shown that PPI network topology around proteins is a predictor of their function or their involvement in disease [30–32]. The local topology around a protein in a PPI network was summarized into a topological "signature" of a protein, *graphlet degree vector (GDV)* [30]. Proteins in the PPI network were grouped based on similarity of their "signatures," or GDV similarity, and it has been shown that proteins within those groups belong to same protein complexes, perform the same biological function, and are part of the same subcellular components [30]. Also, GDV similarity between proteins in the PPI network was used as a similarity measure for clustering proteins using series of clustering methods, resulting in clusters significantly enriched in cancer and disease related proteins. This leads to predictions of new melanogenesis related genes purely from the topology of the human PPI network and the predictions were phenotypically validated [31, 32].

Described methods used clustering of nodes in the network based on their topological properties (simplified example is illustrated in Figure 1, panel (b)). Note that this is different from clustering the network by identifying its topological modules: locally dense neighbourhoods in the network called graph clusters or network communities [17] (Figure 1, panel (c)). It is generally accepted that a subset of nodes is a good cluster, or community, if the induced subgraph is dense, with relatively few connections between the cluster nodes and nodes that are in the remaining part of the graph [33]. These topological modules often correspond to *functional modules*: aggregations of nodes similar in function, and to *disease modules*: sets of nodes that contribute to a specific disease phenotype [26]. Mitra et al. [34] thoroughly reviewed integrative approaches for identifying such functional modular structures in biological networks. Accordingly, module-based methods use assumption that nodes belonging to same topological or functional module are highly likely to be involved in the same disease. These methods have often been applied in studies related to cancer [35–37]. Another example of this principle is modules identified using community discovery algorithm [38], which resulted in the discovery of new links between Alzheimer's disease and CVDs at the coexpression and coregulation levels [39]. Several module-based methods have been applied to research of CVDs, which will be elaborated in more detail later in this survey.

An interesting survey on different methods that use network topology for predictions of disease genes [40] pointed out that many of the methods that rely on clustering algorithms, or linkage-based inference, are outperformed by random walk-based methods. Random walkers diffuse along the network starting from disease involved nodes with the same probability of visiting any neighbouring node—most visited genes are considered to be on the disease pathway and potentially involved in a particular disease. A method for prioritization of candidate disease genes using random walk analysis was tested on 110 disease gene families and significantly outperformed methods based on distance measures such as linkage-based methods or methods based on shortest paths to disease proteins [41].

2.3. Disease Networks. We are currently witnessing an increase in using disease networks, networks of biomolecules involved in a particular disease or a group of diseases, for exploring relationships between different diseases. For example, Goh et al. [42] created a bipartite "diseasome" network, where one partition consists of a set of diseases and the other of a set of disease genes (and where, by definition of a bipartite network, all edges in the network are between the partitions). They used it to generate two network projections: disease gene network and human disease network (which they found is clustered according to major disorder classes). By exploring centrality and peripherality of genes in the gene network, they showed that contrary to essential human genes that encode hub proteins—highly linked proteins in network, the majority of disease genes do not encode hubs and are localized in the periphery of the network [42].

Janjić and Pržulj [43] demonstrated the existence of topologically and functionally homogeneous "core subnetwork" of the human PPI network, which is enriched in disease genes, drug targets, and a small number of genes that have theoretically been proposed to be required for tumour formation, referred to as "driver genes" [44]. They call this subnetwork the "Core Diseasome" [43] and postulate it is the key to disease onset and progression and hence should be the primary object of therapeutic intervention.

CVD networks have recently gained interest, serving as a basis for a better understanding of the complexity behind the disease [6, 7]. In the next section we focus on various CVD networks with emphasis on the use of network topology. Note that henceforth we will use terms gene and protein interchangeably, as topological properties of proteins, represented as nodes the in PPI network, are commonly used to gain new knowledge about genes that encode these proteins.

3. Using Biological Networks in Research of CVDs

3.1. CVD Networks. There were several attempts to create biological networks relevant to various cardiovascular disorders.

A combination of methods based on experimental cell culture and data mining was used to collect a comprehensive set of vascular and atherosclerosis related genes [45]. In particular, public databases such as PubMed

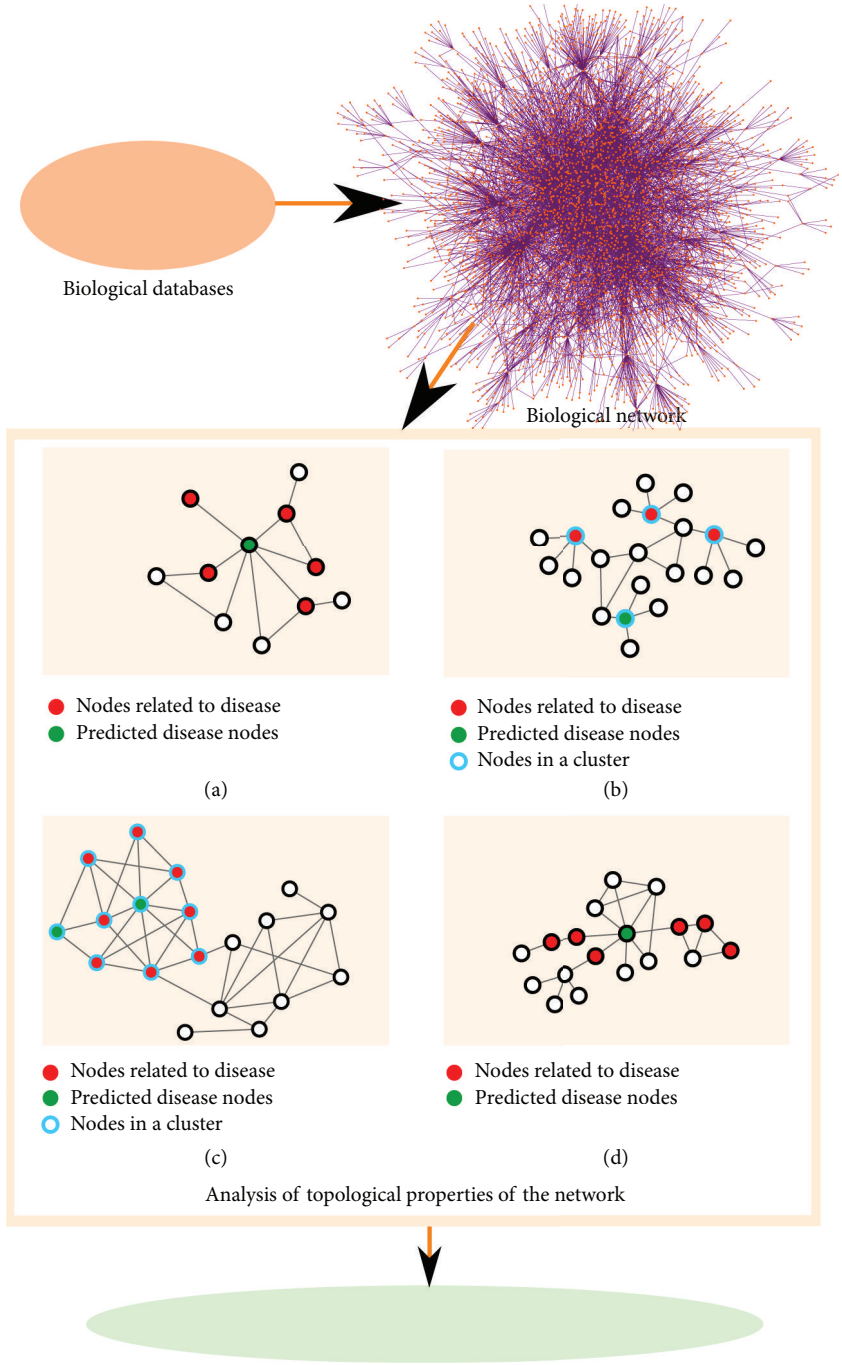


Figure 1: Using network topology to infer elements involved in disease. Panel (a): green node is associated with disease based on its neighbouring disease nodes (shown in red). Panel (b): nodes bordered in blue are part of the same cluster based on similar topology around them. Green node is associated with disease based on the cluster's enrichment in disease nodes (shown in red). Panel (c): nodes bordered in blue are part of the same graph cluster or community, in the network. Green nodes are associated with disease based on the community's enrichment in disease nodes (shown in red). Panel (d): node shown in green is associated with the disease, as a common node on shortest paths between nodes related to disease (shown in red).

(http://www.ncbi.nlm.nih.gov/pubmed) were searched for genes related to the terms *atherosclerosis, smooth muscle cell, endothelial cell, apoptosis, cytokine, and adhesion molecule*. This list of genes was then combined with genes obtained from sequencing clones from stimulated vascular cells in culture. Next, a large association network was constructed through semantic mining of published literature—an association between two genes was extracted from sentences in scientific literature that contained two gene names and a verb as defined by user context file. Also, coronary artery segments isolated from explanted hearts of 22 cardiac transplant patients were experimentally processed, resulting in

significant gene expression profiles obtained using significance analysis of microarrays (SAM) [46]. Then, for each gene from the large association network, a subnetwork was constructed. The subnetwork consisted of that gene and its neighbouring genes which were obtained using SAM analysis. A cumulative and average SAM scores were computed for each subnetwork and were used to identify subnetworks of high overall significance. Their central, "nexus," genes were singled out as potential regulators that may cause the disease phenotype [45].

A similar method was used for constructing an association network of human in-stent restenosis [47]. Genes relevant to the disease were collected using methods based on experimental cell culture and data mining, while associations between genes were obtained through text mining of MEDLINE (http://www.nlm.nih.gov/pubs/factsheets/medline.html) abstracts. Again, a subnetwork for each gene was constructed containing the gene and its direct neighbours in the network. Gene expressions were experimentally assessed from tissue samples of 89 patients using SAM analysis. Subnetworks were next compared based on the overall significance score calculated using SAM scores of the subnetwork members. Central nodes of these subnetworks were identified as successful targets for drug therapy.

Skogsberg et al. [48] revealed a regulatory gene network of cholesterol-responsive atherosclerosis genes that control formation of plaques in arteries, using analysis of gene expression in response to plasma cholesterol-lowering. They established a list of genes related to atherosclerosis, foam cells, smooth muscle cells, endothelial cells, and T cells using automated text mining of PubMed abstracts. The resulting network was proposed as a starting point for future research of novel atherosclerosis therapies.

Another PPI network of cardiovascular diseases was created from CVD related proteins that were identified using protein annotations from Uniprot database (searching for the keyword *cardiovascular*) and known protein-protein interactions from HPRD [49]. Only proteins with at least one known interaction in HPRD were taken into account. In addition to these proteins, their interacting partners in the PPI network, which also appear in the signalling pathways from KEGG database, were included in the network. The resulting CVD PPI network consisted of 55 proteins and 122 PPIs and was used to identify network CVD biomarkers as follows. (1) Single biomarker discovery was based on significantly different expressions between proteins in control patients and disease patients (significantly low P values); biomarkers were evaluated using not only P values but also support vector machine (SVM). (2) A candidate *pair biomarker* is composed of two single biomarkers and a PPI between them. Pair biomarkers were selected based on the best performance in SVM and significantly low P values. (3) Candidate *triple biomarker* is composed of three single biomarkers and PPIs between every pair among them. Again, triple biomarkers were selected based on the best performance in SVM and significantly low P values. (4) Multiple CVD biomarkers were identified in similar manner as combinations of different single ones, pair ones, and triple biomarkers.

As mentioned in Section 2.1, despite their important biological role, human transcriptional regulatory networks are still largely unexplored. Some of the reasons are experimental limitations and human cellular diversity [16]. However, there have been several attempts to construct a cardiac transcription network. For example, mRNA profiles were integrated with DNA-binding events of key cardiac transcription factors (TFs) [50]. Insights into combinatorial regulation by cardiac TFs showed that they compensate each other's functions. Cardiac transcription network was built based on findings from RNA knockdown experiments. Target genes that are important for the cardiovascular system were chosen based on their biological roles such as muscle contractility and cardiac growth. The network depicted the common regulation of several transcriptional factors and the impact of the post-transcriptional modulation of expression levels by miRNAs [50]. Another transcriptional network of cardiac TFs and genes important for cardiac function was constructed based on coexpression analysis involving TFs critical for hearth development. Coregulatory relationships between five such TFs were revealed [51]. These types of relationships can give a new perspective for understanding the complexity of CVDs.

The quality of biological data is crucial for constructing a reliable CVD network, as discussed in Section 2.1. New technologies, such as next generation sequencing platforms, have significantly increased DNA sequencing output [52] and as such will largely increase the size of available biological data. Therefore, next generation sequencing methods for gene expression profiling will change the approaches to studying many common complex disorders, including CVDs [53]. The resulting new insights into underlying mechanisms of CVDs will yield more complete CVD networks and open a window of opportunities for exploring the topology of these networks.

3.2. Correlating Network Topology with CVD Mechanisms. Several authors tried to explore whether basic topological information from a biological network, such as connectivity of the nodes, can be correlated with biological properties required for CVD onset and progression.

One example is a global PPI network in heart failure (HF) [54], created as a subnetwork of PPIs from HPRD that includes HF relevant genes. Next, differentially expressed genes in HF were identified from microarray data encoding molecular profiles of healthy versus HF subjects. Proteins encoded by these significantly differentially expressed genes were also included in the HF PPI subnetwork. This network was used to explore the relationship between gene coexpression levels and their connectivity in the HF network. It was discovered that hub proteins in the network are encoded by genes that display a significant diversity of coexpression patterns in comparison to peripheral proteins. However, hub proteins are not necessarily encoded by genes that are significantly differentially expressed. Analysis of gene ontology (GO) terms [55] revealed the relationship between connectivity of the proteins in this network and their involvement in specific biological processes, such as processes related to cardiac remodelling.

In their later work, the same authors explored dilated cardiomyopathy (DCM) genes [56], as DCM is recognised as a leading cause of HF. DCM genes were identified using gene expression profiles from three independent datasets, while associations with HF were identified using literature mining. Human HF PPI network was created using PPIs from HPRD by including genes known to be involved in HF and genes from the gene expression datasets along with biological pathways associated with them. Again, connectivities of nodes (proteins) in HF PPI network were compared to their gene expression patterns. Differential gene expression was measured using SAM analysis, resulting in *di*values representing genes' score of class differentiation. Focusing on significantly differentially expressed genes, it was found that superhubs and hubs in the network had a lower range of *di*values, while genes that encoded peripheral proteins in the network had a higher range of *di*values.

Several module-based approaches were applied to various CVD networks attempting to identify functional modules related to the disease or discover new associations between genes and disease. Diez et al. [57] created a combined gene association and correlation network, using data from 47 microarrays from a database of carotid endarterectomies (Biobank of Karolinska Endarterectomies, BiKE (http://ki.se/start)). The gene correlation network was constructed using statistical analysis of gene expression data. The association network was constructed using the list of differentially expressed genes, by performing literature search for each gene symbol and association keywords such as "*gene A activates gene B.*" The networks were then merged into an undirected network of atherosclerosis. This network was searched for active modules based on closeness centrality using *jActiveModules* Cytoscape plugin [58]. APOC1 gene was differentially expressed in atherosclerotic plaque and related to several important GO categories characteristic of the disease mechanism, so it was selected for a more detailed analysis. Hence, among detected modules, the one containing APOC1 gene was further inspected. This module was checked for GO enrichment. GO categories relevant to atherosclerosis mechanisms and etiology that were identified in this module were all characteristic of APOC1 gene, suggesting its importance in this disease.

Ischemic dilated cardiomyopathy (ICM) is one of the main pathological forms of DCM. A set of genes differentially expressed in ICM, downloaded from gene expression omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), and cardiac myocytes proteins retrieved from human protein atlas (HPA) [59] were merged to create another CVD relevant PPI network [60]. Information about PPIs was integrated from several public databases. The analysed largest connected component of this PPI network was divided in four layers, based on subcellular localization information. This revealed that the extracellular and plasma membrane layers contained more downregulated genes, while cytoplasm and nucleus contained more upregulated genes. Next, significantly overrepresented biological processes (BPs) were identified, and PPI network containing only proteins related to these GO BPs was then divided into 12 clusters according to BPs. It was shown that the number of PPIs between proteins involved in different BPs was associated with differential gene expression patterns.

Rende et al. [61] used topological features of PPI networks in search of genes common to CVDs and other diseases, by identifying functional modules of genes. They extended a core CVD network, consisting of proteins known to be associated with CVDs (manually curated from the literature), by including their direct interactors in PPI network, resulting in a cardiovascular disease "functional linkage network" (CFN). Hub proteins in this network were considered to be the key nodes that regulate molecular mechanisms of CVDs and interdependence between CVDs and other complex disorders. These hubs were identified using distributions of node degrees and betweenness centralities. Functional modules, highly connected subgraphs, were identified using a modularity measure based solely on topological properties, allowing modules to overlap. All hub proteins appeared in these functional modules. Presence of a protein in multiple functional modules in addition to its high connectivity implied that any changes regarding protein would affect all its functional modules. Next, proteins in functional modules were matched to diseases from OMIM database: 19 modules were associated with CVDs. Also, modules associated with at least two diseases were examined for functional GO term enrichment and were shown to be functionally linked. This approach revealed some significant complex disorders that cooccur with CVDs and identified relevant shared disease genes and shared disease functional modules.

Known causal congenital heart disease (CHD) genes and genes differentially expressed in this disease (named *target genes*) were mapped onto a PPI network with the aim of identifying gene modules relevant to CHD [62]. The network was modelled as an electrical circuit, where edges between nodes (genes) were used as a conductance of a resistor according to correlation of coexpression between the two end nodes. Shortest paths from one causal gene to all target genes were merged into a subnetwork, and the current flow for each gene in the subnetwork was computed to evaluate its importance. Genes were assigned to a subnetwork in which they scored best. This resulted in 12 disjoint modules for further analyses: relationships of individual modules with disease phenotypes, mutual coexpression among genes within the modules, functional enrichment, and pathway analysis. As a result, candidate disease genes and hub modules that regulate key pathways of CHD were identified.

Functional modules of gene coexpression networks were also explored in research of cardiac development, hypertrophy, and failure [63]. Datasets from microarray experiments involving myocardial tissue were collected from GEO and used for creating a weighted gene coexpression network, where edges represent adjacencies between genes based on weighted Pearson correlation between gene expression profiles. Gene modules were identified using agglomerative hierarchical clustering of adjacencies given by the topological overlap measure based on shared network neighbours. The modules were first identified in fetal tissue, followed by evaluating their reproducibility in normal adult, hypertrophied, and failing myocardial tissue. The analysis revealed specific gene coexpression modules that were present both in

Table 2: Methods that explored topology of biological networks in research of CVDs.

Network	Type of data/interactions in the network	Topological analysis performed on the data	Aims of topological analysis	Reference
Heart failure (HF) network	HF relevant genes, genes differentially expressed in HF and dilated cardiomyopathy (DCM), and PPI data	Connectivity of nodes	Relationship between gene connectivity and gene coexpression levels and their biological functions	[54, 56]
Network of atherosclerosis	Literature associations and gene expression data	Network modules identified based on closeness centrality	GO enrichment of network modules	[57]
Network of ischemic dilated cardiomyopathy (ICM)	Genes differentially expressed in ICM, cardiac myocytes proteins, and PPI data	Number of edges between network clusters	Correlation between number of edges between network clusters and differential gene expression patterns	[60]
Cardiovascular disease "functional linkage network" (CFN)	CVD proteins and PPI data	Degree distribution, betweenness centrality, and modularity measure	Associating functional modules (highly connected subgraphs) with diseases	[61]
Congenital heart disease (CHD) network	Known CHD genes, genes differentially expressed in CHD, and PPI data	Subnetworks based on shortest paths and current flow (network was modelled as an electrical circuit)	Functional subnetwork analysis in search of key pathways of CHD	[62]
Networks for analysis of cardiac development, hypertrophy, and failure	Gene coexpression data	Network modules based on hierarchical clustering and shared network neighbours	Identifying common modules in networks of different types of myocardial tissue	[63]
Human PPI network	PPI data	Node degree, neighbourhood enrichment, betweenness centrality, clustering coefficient, and shortest path length	Inferring coronary artery disease genes based on topological information	[65]
Human PPI network	PPI data	Clustering nodes based on graphlet degree vector similarity	Inferring new CVD genes based on clusters' enrichment in CVD genes	[66]

developing heart and in hypertrophied or failing myocardial tissue.

3.3. Methods for Utilizing Network Topology in CVD Research. In previous section, we described a variety of methods that used biological networks in search of genes, pathways, or functional modules that are significant for different types of CVDs.

We see that the majority of approaches focused on constructing biological networks of particular cardiovascular disorders. Several approaches further explored topological properties of these networks and use them in search of new CVD knowledge. In particular, modules in the network of atherosclerosis [57] were identified based on closeness centrality. Functional modules of a CVD network used for investigating relationships between CVD and other disorders were identified using modularity measure based solely on network topology [61]. The method for identifying modules in CHD utilized shortest paths in the network between genes of interest [62]. Also, some basic topological properties, such as node connectivity [54, 56], or the number of interactions between functional sets [60], were examined in correlation with disease. Note that the vast majority of the above-presented topological analyses focused on CVD subnetworks in isolation, rather than observing them as parts of a larger, more complete interaction network, such as the entire human PPI network.

This may be a limiting factor when exploring the interplay between genes involved in different CVD disorders or when targeting genes that have previously not been connected to CVDs. The importance of observing the neighbourhood of disease genes in the entire PPI network was emphasized in one of the studies related to atherosclerosis [64]. Functional enrichment test performed only on differentially expressed genes failed to detect biological processes related to the disease progression. However, the network that included both differentially expressed genes and genes that have high connectivity with them in the entire PPI network was functionally enriched in relevant biological processes. This analysis showed that the regulators of disease progression

should be looked for among genes that are not necessarily differentially expressed and within the context of the entire available PPI network.

We summarized the methods that used topology of biological networks in research of CVDs in Table 2. There are only few approaches that identified new genes relevant to CVDs relying solely on topological properties of entire PPI network. The example is the computational method based on six topological features (degree, neighbour count of disease genes, ratio of disease genes among neighbours, betweenness centrality, clustering coefficient, and mean shortest path length to disease gene) [65]. The constructed classifier was used on the PPI network to predict candidate genes for coronary artery disease.

The PPI network topology was also used for inferring proteins' involvement in CVDs as follows [66]. Proteins were clustered based on the similarity of topologies of their neighbourhoods in the PPI network, measured using *GDV similarity* [30]. The clusters were then checked for enrichment in CVD genes. The overlap of statistically significantly enriched clusters contained 10 *key CVD genes* and 17 predicted new CVD related genes. More than 70% of these predictions were validated in the literature. Also, both *key CVD genes* and predicted CVD genes were enriched in biological functions that CVD drug mechanisms rely on, showing that this approach may be successful in identifying potential drug targets.

4. Conclusion

The emerging interest in molecular interaction networks of various cardiovascular diseases has resulted in a number of association, gene expression, PPI, and transcriptional regulatory networks being examined to study atherosclerosis, in-stent restenosis, heart failure, dilated cardiomyopathy, ischemic dilated cardiomyopathy, and CVDs in general. Many of these networks were constructed using experimental data combined with literature mining, with the aim of identifying a broader set of genes involved in a particular cardiovascular disorder. These networks are a valuable platform for exploring the mechanisms of the disease. Nevertheless, their topologies have not been fully explored.

We surveyed studies that explored the link between some basic topological properties of CVD genes in networks and involvement of these genes in specific disease related processes. Several CVD networks were checked for enrichment in biological functions relevant to the disease, and functional modules in the networks were identified, in some cases using topological properties. However, topological analysis was usually limited to the disease specific subnetwork, without observing it in the context of a larger, more complete network. Such complete interaction networks were analysed only in few studies, which explored the topology around genes that were previously not associated with CVD and thus not present in the disease specific subnetwork. This resulted in predictions of novel CVD genes.

There is a huge potential in analysing CVD related molecular subnetworks and their topology in the context of the complete biomolecular interaction networks. Such approaches could give better insight into interconnectedness of different CVDs. They could help discover novel CVD genes and pathways responsible for the dependency between different disorders.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, the Serbian Ministry of Education and Science Project III44006, and ARRS Project J1-5454.

References

[1] S. Alan Go, D. Mozaffarian, V. L. Roger et al., "Executive summary: heart disease and stroke statistics—2013 update: a report from the american heart association," *Circulation*, vol. 127, no. 1, pp. 143–152, 2013.

[2] D. B. Mark, F. J. van de Werf, R. J. Simes et al., "Cardiovascular disease on a global scale: defining the path forward for research and practice," *European Heart Journal*, vol. 28, no. 21, pp. 2678–2684, 2007.

[3] S. Kathiresan and D. Srivastava, "Genetics of human cardiovascular disease," *Cell*, vol. 148, no. 6, pp. 1242–1257, 2012.

[4] C. E. Wheelock, Å. M. Wheelock, S. Kawashima et al., "Systems biology approaches and pathway tools for investigating cardiovascular disease," *Molecular BioSystems*, vol. 5, no. 6, pp. 588–602, 2009.

[5] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology," *Annual Review of Genomics and Human Genetics*, vol. 2, pp. 343–372, 2001.

[6] A. J. Lusis and J. N. Weiss, "Cardiovascular networks: systems-based approaches to cardiovascular disease," *Circulation*, vol. 121, no. 1, pp. 157–170, 2010.

[7] W. R. MacLellan, Y. Wang, and A. J. Lusis, "Systems-based approaches to cardiovascular disease," *Nature Reviews Cardiology*, vol. 9, no. 3, pp. 172–184, 2012.

[8] F. J. Azuaje, F. E. Dewey, D. L. Brutsaert, Y. Devaux, E. A. Ashley, and D. R. Wagner, "Systems-based approaches to cardiovascular biomarker discovery," *Circulation*, vol. 5, no. 3, pp. 360–367, 2012.

[9] S. Y. Chan, K. White, and J. Loscalzo, "Deciphering the molecular basis of human cardiovascular disease through network biology," *Current Opinion in Cardiology*, vol. 27, no. 3, pp. 202–209, 2012.

[10] H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, and A.-L. Barabásl, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[11] B. Bollobás, "Paul Erdös and probability theory," *Random Structures and Algorithms*, vol. 13, no. 3-4, pp. 521–533, 1998.

[12] R. Mani, R. P. St. Onge, J. L. Hartman IV, G. Giaever, and F. P. Roth, "Defining genetic interaction," *Proceedings of the National*

Academy of Sciences of the United States of America, vol. 105, no. 9, pp. 3461–3466, 2008.

[13] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761, pp. C47–C52, 1999.

[14] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási, "The implications of human metabolic network topology for disease comorbidity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 9880–9885, 2008.

[15] T. I. Lee, N. J. Rinaldi, F. Robert et al., "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.

[16] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos, "Circuitry and dynamics of human transcription factor regulatory networks," *Cell*, vol. 150, no. 6, pp. 1274–1286, 2012.

[17] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[18] M. Newman, *Networks: An Introduction*, Oxford University Press, New York, NY, USA, 2010.

[19] G. Valiente, *Algorithms on Trees and Graphs*, Springer, New York, NY, USA, 2002.

[20] N. Pržulj, "Protein-protein interactions: making sense of networks via graph-theoretic modeling," *BioEssays*, vol. 33, no. 2, pp. 115–123, 2011.

[21] D. K. Arrell and A. Terzic, "Network systems biology for drug discovery," *Clinical Pharmacology and Therapeutics*, vol. 88, no. 1, pp. 120–125, 2010.

[22] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, p. 88, 2007.

[23] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[24] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.

[25] M. P. Samanta and S. Liang, "Predicting protein functions from redundancies in large-scale protein interaction networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, pp. 12579–12583, 2003.

[26] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[27] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008.

[28] R. Aragues, C. Sander, and B. Oliva, "Predicting cancer involvement of genes from heterogeneous data," *BMC bioinformatics*, vol. 9, no. 172, p. 172, 2008.

[29] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's desease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15148–15153, 2004.

[30] T. Milenković and N. Pržulj, "Uncovering biological network function via graphlet degree signatures," *Cancer Informatics*, vol. 6, pp. 257–273, 2008.

[31] T. Milenković, V. Memišević, A. K. Ganesan, and N. Pržulj, "Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data," *Journal of the Royal Society Interface*, vol. 44, no. 7, pp. 353–350, 2010.

[32] H. Ho, T. Milenković, V. Memišević, J. Aruri, N. Pržulj, and A. K. Ganesan, "Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets," *BMC Systems Biology*, vol. 4, no. 1, p. 84, 2010.

[33] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.

[34] K. Mitra, A. R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[35] N. Bonifaci, A. Berenguer, J. Diez et al., "Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes," *BMC Medical Genomics*, vol. 1, p. 62, 2008.

[36] L. M. Heiser, N. J. Wang, C. L. Talcott et al., "Integrated analysis of breast cancer cell lines reveals unique signaling pathways," *Genome Biology*, vol. 10, no. 3, article R31, 2009.

[37] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[38] J. Ruan and W. Zhang, "Identifying network communities with a high resolution," *Physical Review E*, vol. 77, no. 1, Article ID 016104, pp. 1–12, 2008.

[39] M. Ray, J. Ruan, and W. Zhang, "Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases," *Genome Biology*, vol. 9, no. 10, article R148, 2008.

[40] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.

[41] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[42] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[43] V. Janjić and N. Pržulj, "The core diseasome," *Molecular Biosystems*, vol. 8, no. 10, pp. 2614–2625, 2012.

[44] A. Ashworth, C. J. Lord, and J. S. Reis-Filho, "Genetic interactions in cancer progression and treatment," *Cell*, vol. 145, no. 1, pp. 30–38, 2011.

[45] J. Y. King, R. Ferrara, R. Tabibiazar et al., "Pathway analysis of coronary atherosclerosis," *Physiological Genomics*, vol. 23, no. 1, pp. 103–118, 2005.

[46] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[47] E. A. Ashley, R. Ferrara, J. Y. King et al., "Network analysis of human in-stent restenosis," *Circulation*, vol. 114, no. 24, pp. 2644–2654, 2006.

[48] J. Skogsberg, J. Lundström, A. Kovacs et al., "Transcriptional profiling uncovers a network of cholesterol-responsive athe-

rosclerosis target genes," *PLoS Genetics*, vol. 4, no. 3, Article ID e1000036, 2008.

[49] G. Jin, X. Zhou, H. Wang et al., "The knowledge-integrated network biomarkers discovery for major adverse cardiac events," *Journal of Proteome Research*, vol. 7, no. 9, pp. 4013–4021, 2008.

[50] J. Schlesinger, M. Schueler, M. Grunert et al., "The cardiac transcription network modulated by gata4, mef2a, nkx2.5, srf, histone modifications, and microRNAs," *PLoS Genetics*, vol. 7, no. 2, Article ID e1001313, 2011.

[51] E. Poon, B. Yan, S. Zhang et al., "Transcriptome-guided functional analyses reveal novel biological properties and regulatory hierarchy of human embryonic stem cell-derived ventricular cardiomyocytes crucial for maturation," *PLoS ONE*, vol. 8, no. 10, Article ID e77784, 2013.

[52] A. J. Marian and J. Belmont, "Strategic approaches to unraveling genetic causes of cardiovascular diseases," *Circulation Research*, vol. 108, no. 10, pp. 1252–1269, 2011.

[53] B. Meder, J. Haas, A. Keller et al., "Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies," *Circulation*, vol. 4, no. 2, pp. 110–122, 2011.

[54] A. Camargo and F. Azuaje, "Linking gene expression and functional network data in human heart failure," *PLoS ONE*, vol. 2, no. 12, Article ID e1347, 2007.

[55] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[56] A. Camargo and F. Azuaje, "Identification of dilated cardiomyopathy signature genes through gene expression and network data integration," *Genomics*, vol. 92, no. 6, pp. 404–413, 2008.

[57] D. Diez, Å. M. Wheelock, S. Goto et al., "The use of network analyses for elucidating mechanisms in cardiovascular disease," *Molecular BioSystems*, vol. 6, no. 2, pp. 289–304, 2010.

[58] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software Environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[59] S. Hober and M. Uhlén, "Human protein atlas and the use of microarray technologies," *Current Opinion in Biotechnology*, vol. 19, no. 1, pp. 30–35, 2008.

[60] W. Zhu, L. Yang, and Z. Du, "Layered functional network analysis of gene expression in human heart failure," *PLoS ONE*, vol. 4, no. 7, Article ID e6288, 2009.

[61] D. Rende, N. Baysal, and B. Kirdar, "A novel integrative network approach to understand the interplay between cardiovascular disease and other complex disorders," *Molecular BioSystems*, vol. 7, no. 7, pp. 2205–2219, 2011.

[62] D. He, Z.-P. Liu, and L. Chen, "Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach," *BMC Genomics*, vol. 12, article 592, 2011.

[63] F. E. Dewey, M. V. Perez, M. T. Wheeler et al., "Gene coexpression network topology of cardiac development, hypertrophy, and failure," *Circulation*, vol. 4, no. 1, pp. 26–35, 2011.

[64] S. A. Ramsey, E. S. Gold, and A. Aderem, "A systems biology approach to understanding atherosclerosis," *EMBO Molecular Medicine*, vol. 2, no. 3, pp. 79–89, 2010.

[65] L. Zhang, X. Li, J. Tai, W. Li, and L. Chen, "Predicting candidate genes based on combined network topological features: a case study in coronary artery disease," *PLoS ONE*, vol. 7, no. 6, Article ID e39542, 2012.

[66] A. Sarajlić, V. Janjić, N. Stojković, Dj. Radak, and N. Pržulj, "Network topology reveals key cardiovascular disease genes," *PLoS ONE*, vol. 8, no. 8, Article ID e71537, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 136130, 15 pages
http://dx.doi.org/10.1155/2014/136130

Research Article

Essential Functional Modules for Pathogenic and Defensive Mechanisms in *Candida albicans* Infections

Yu-Chao Wang,[1] I-Chun Tsai,[2] Che Lin,[3] Wen-Ping Hsieh,[4] Chung-Yu Lan,[5] Yung-Jen Chuang,[6] and Bor-Sen Chen[2]

[1] *Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan*
[2] *Laboratory of Control and Systems Biology, Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan*
[3] *Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan*
[4] *Institute of Statistics, National Tsing Hua University, Hsinchu 30013, Taiwan*
[5] *Department of Life Science and Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu 30013, Taiwan*
[6] *Department of Medical Science and Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 30013, Taiwan*

Correspondence should be addressed to Yu-Chao Wang; yuchao@ym.edu.tw and Bor-Sen Chen; bschen@ee.nthu.edu.tw

Received 17 October 2013; Accepted 10 February 2014; Published 18 March 2014

Academic Editor: Shigehiko Kanaya

The clinical and biological significance of the study of fungal pathogen *Candida albicans* (*C. albicans*) has markedly increased. However, the explicit pathogenic and invasive mechanisms of such host-pathogen interactions have not yet been fully elucidated. Therefore, the essential functional modules involved in *C. albicans*-zebrafish interactions were investigated in this study. Adopting a systems biology approach, the early-stage and late-stage protein-protein interaction (PPI) networks for both *C. albicans* and zebrafish were constructed. By comparing PPI networks at the early and late stages of the infection process, several critical functional modules were identified in both pathogenic and defensive mechanisms. Functional modules in *C. albicans*, like those involved in hyphal morphogenesis, ion and small molecule transport, protein secretion, and shifts in carbon utilization, were seen to play important roles in pathogen invasion and damage caused to host cells. Moreover, the functional modules in zebrafish, such as those involved in immune response, apoptosis mechanisms, ion transport, protein secretion, and hemostasis-related processes, were found to be significant as defensive mechanisms during *C. albicans* infection. The essential functional modules thus determined could provide insights into the molecular mechanisms of host-pathogen interactions during the infection process and thereby devise potential therapeutic strategies to treat *C. albicans* infection.

1. Introduction

In daily life, human beings are exposed to environments containing a wide variety of microorganisms. It is inevitable that humans will sometimes face opportunistic threats posed by some of these microbes. Pathogens, microorganisms that cause their host disease, have evolved numerous strategies to invade their hosts, while hosts have also evolved corresponding defensive responses to these invading agents [1]. The result of such host-pathogen interactions can result in damage to or even death of the host. Therefore, investigating the mechanisms of host-pathogen interactions may help biologists and clinicians better understand the underlying biological scenario. Once the pathogenic mechanisms of pathogens and the corresponding defensive mechanisms employed by hosts are uncovered, novel strategies that assist hosts in responding to microbial infection may be developed.

Candida albicans, a fungal pathogen, is a kind of ubiquitous commensal yeast that inhabits the mouth, gastrointestinal tract, and the vagina in humans. Under normal conditions, *C. albicans* is harmless to humans. However, it can induce serious mucosal and life-threatening systemic infections in individuals who are immunocompromised due to such factors as infection with human immunodeficiency virus (HIV), organ transplantation, or cancer chemotherapy. In addition, *C. albicans* is a major cause of hospital-acquired

infection [2, 3]. *C. albicans* has many morphological forms including a yeast form, a pseudohyphal form, and a hyphal form. The ability to switch from the yeast to hyphal form has been proposed as one of the major factors accounting for the virulence of the organism, and other studies have demonstrated that nonfilamentous *C. albicans* mutants are avirulent [4–6].

Until recently, the mouse, the fruit fly, and the wax moth were the main model organisms for studies of *C. albicans* infection. However, there are certain disadvantages in using these organisms in such models. The fruit fly and the wax moth lack adaptive immunity [7, 8], and the mouse is too expensive for large-scale experiments. Therefore, Chao et al. [9] developed the zebrafish (*Danio rerio*) model as a minivertebrate host system for *C. albicans* infection studies. They showed that *C. albicans* can invade zebrafish and kill the host in a dose-dependent manner [9]. Brothers et al. also developed the zebrafish larva as a transparent vertebrate model of disseminated candidiasis, showing that the infection model reproduces many aspects of candidemia in mammalian hosts [10]. In addition, the zebrafish undergoes rapid embryonic development and requires relatively small spaces in which to breed, leading to low experimental costs and making it a suitable infection model organism. Furthermore, the zebrafish has both innate and adaptive immune systems [11] and therefore has become widely used in the study of human diseases [12].

Several studies have identified the virulence factors and the corresponding virulence-associated genes in *C. albicans* [13]. Other studies have investigated immune responses occurring during the infection process, especially pathogen recognition mechanisms [14]. However, these studies have mainly focused on specific genes and their particular roles in the infection process and have not investigated host-pathogen interaction from a systems point of view [15]. In the light of experimental observations in which about 50% of zebrafish were seen to die of extensive bleeding 18 hours after being infected with *C. albicans* (1×10^8 CFU) [9], we aimed to investigate both the functional modules of the activated pathogen essential in the invasion of zebrafish by *C. albicans* and the zebrafish functional modules likely to be responsible for defensive responses and the extensive bleeding. In other words, the goal of this study was to investigate the pathogenesis of *C. albicans* in fatal infections of zebrafish and the important defensive mechanisms employed by zebrafish against *C. albicans* infection from the network systems perspective. As a consequence, we simultaneously quantified the time-course gene expression profiles for both *C. albicans* and zebrafish during *C. albicans* infection. With the help of simultaneous host-pathogen interaction microarrays and other high-throughput omics data, the early-stage infection and late-stage infection protein interaction networks in both *C. albicans* and zebrafish were constructed. Protein-protein interactions are at the core of the intercellular interactions and control major biological functions. Differential interactions imply mechanistic changes that are a result of an organism's response to environmental conditions [16]. In the case of host-pathogen interaction, analyzing the differential interactions in a time-dependent manner can show how the host attempts to respond to the pathogen and how the pathogen responds within the host [16]. Therefore, changes in PPIs during infection may affect the pathogenesis of pathogens, while the reconfiguration of the protein-protein interactions in the host may reflect the activation of defensive mechanisms against pathogens. Using these constructed PPI networks, proteins with significant changes in their interaction profiles were considered to play important roles in infection pathology. Furthermore, from the identification of such significant proteins, the *C. albicans* functional modules associated with pathogenesis and the zebrafish functional modules involved in defense against *C. albicans* infection could be identified. It is hoped that by understanding the underlying pathogenic/defensive interaction mechanisms between the host and pathogen, biologists and clinicians may better appreciate how the pathogen infects its host and thereby devise effective therapeutic strategies to prevent loss of life in cases of *C. albicans* infection [17].

2. Materials and Methods

2.1. Omics Data Selection. In order to investigate important functional modules in *C. albicans*-zebrafish interactions, high-throughput omics data from many different sources were integrated, including simultaneous time-course gene expression profiles of *C. albicans* and zebrafish interactions obtained from microarray data, protein-protein interaction information from *Homo sapiens* and *Saccharomyces cerevisiae*, and ortholog data between humans and zebrafish and between *S. cerevisiae* and *C. albicans*. The time-course gene expression data were obtained from the GEO database (accession number: GSE32119). Experiments were performed to obtain *in vivo* genome-wide gene expression profiles simultaneously for both *C. albicans* and zebrafish during *C. albicans*-zebrafish interactions. Wild type AB strain zebrafish were intraperitoneally injected with *C. albicans* cell suspensions (SC5314 strain), and gene expression in both *C. albicans* and zebrafish was then assessed at nine subsequent times: 0.5, 1, 2, 4, 6, 8, 12, 16, and 18 hours after infection (hpi); this was performed three times in total [20].

Since there is little available in terms of protein interaction maps in either *C. albicans* or zebrafish, protein-protein interaction (PPI) information for these organisms was inferred from the interactome of *S. cerevisiae* and humans with the help of ortholog data [21]. Both the PPI data of *S. cerevisiae* and of humans were acquired from the Biological General Repository for Interaction Datasets (BioGRID) (http://thebiogrid.org/) [22]. The ortholog data pertaining to *C. albicans* and *S. cerevisiae* were retrieved from the *Candida* Genome Database (CGD) (http://www.candidagenome.org/) [23]; the ortholog data pertaining to zebrafish and humans were taken from the Zebrafish Model Organism Database (ZFIN) (http://zfin.org) [24] and the InParanoid database (http://InParanoid.sbc.su.se) [25]. In addition, gene annotations of *C. albicans* and zebrafish were obtained from CGD, the Gene Ontology (GO) database (http://www.geneontology.org/) [26], and the Database

for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) [27].

2.2. Selection of Protein Pool. The overall flowchart illustrating the approach adopted is shown in Figure 1. For both host data and pathogen data, gene expression profiles, ortholog data, and PPI data were used to construct dynamic PPI networks. In order to integrate gene expression profiles and PPI information, gene expression values were overlaid on the corresponding proteins as the protein expression levels [28]. Since the systems approach adopted in this study was based on dynamic protein-protein interaction networks, the extent of coverage of the interactome needed to be considered when selecting proteins of interest. For *C. albicans*, the PPI information was inferred from the interactions of *S. cerevisiae*, the best studied model system. However, the zebrafish has a much lower overall coverage in terms of protein interaction maps than *C. albicans*. Therefore, the selection of the protein pool was different for *C. albicans* and zebrafish. One-way analysis of variance (ANOVA) was employed to select differentially expressed proteins in *C. albicans*. The null hypothesis of ANOVA assumed that the average expression level of a protein would be the same at every time point [29]. Proteins with Bonferroni adjusted P values of less than 0.1 were selected in the protein pool as target proteins. For zebrafish, the protein pool included all proteins even though they were not differentially expressed. Since PPI networks were used in this study, those target proteins for which PPI information was not available were filtered out of the protein pool.

2.3. Protein Interaction Network Construction. Our strategy was to identify proteins significant in protein interaction network reconfiguration during the infection process and then to investigate the enriched functional modules composed of these significant proteins. For this reason, early- and late-stage PPI networks for both *C. albicans* and zebrafish were constructed for network configuration comparison. Previous histological analysis showed that the first zebrafish was seen to die 5 hours after being infected with *C. albicans* (1×10^8 CFU) and about 50% of zebrafish had died by 18 hpi [9]. Therefore, the gene expression data taken nine time points after infection were separated into two groups; one contained the 0.5–4 hpi data, the early stage of infection, and the other comprised of the 4–18 hpi data, the late stage of infection. Therefore, the PPI networks constructed from gene expression data within the 0.5–4 hpi period were designated as the early- stage PPI networks, and the gene expression data after 4 hpi were used to construct the late-stage PPI networks. With data pertaining to the proteins in the protein pool and the protein-protein interaction data obtained from the database mining, a rough PPI network for both post-infection stages was constructed for *C. albicans* and zebrafish by linking the proteins with the PPI information. However, under the specific conditions of the infection process, these rough PPI networks may be inappropriate because they were constructed from data obtained under all possible experimental conditions in the literature and databases. Therefore these rough PPI networks needed refining into a suitable network occurring specifically during infection with the help of gene expression profiles. In this study, a discrete dynamic model was employed to determine the PPI networks that occurred in the infection of zebrafish by *C. albicans* [30] (see Supplementary methods for details available online at http://dx.doi.org/10.1155/2014/136130). Based on the time-course microarray data, the system parameter estimation method and the model selection measurement Akaike Information Criterion (AIC) were then used to detect significant interactions [31, 32] (see Supplementary methods for details). In this way, with different sets of microarray data (0.5–4 hpi for the early stage and 4–18 hpi for the late stage), two refined PPI networks were constructed for the early and late stages of *C. albicans* infection of zebrafish for both organisms. These early and late stage PPI networks will be compared with each other to ascertain their network reconfigurations in the infection process.

2.4. Network Reconfiguration between the Early and Late Stages of the Infection Process. Living organisms take appropriate actions to respond to diverse environmental changes and internal perturbations. Through adjustment of their molecular interactions, organisms tend to maintain a proper, beneficial, or stable state in response to changes in such conditions [33]. Therefore the protein-protein network changes with these interaction variations over time to balance out the effects of environmental changes; that is, the PPI network reconfigures as corresponding protein interactions change to respond to the different conditions brought about by the infection process. A matrix indicating significant protein-protein interactions in the refined PPI network was constructed from those identified protein interaction abilities (see Supplementary methods). The established PPI interaction matrix of a refined PPI network can be thus represented:

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1K} \\ b_{21} & b_{22} & \cdots & b_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{K1} & b_{K2} & \cdots & b_{KK} \end{bmatrix}, \tag{1}$$

where b_{ij} denotes the identified interaction ability between proteins i and j and K represents the number of proteins in the refined PPI network. Hence, the interaction matrix of the differential PPI network contrasting the early-stage with late-stage PPI networks was expressed as follows:

$$\begin{aligned} D_l &= \begin{bmatrix} d_{11,l} & d_{12,l} & \cdots & d_{1K,l} \\ d_{21,l} & d_{22,l} & \cdots & d_{2K,l} \\ \vdots & \vdots & \ddots & \vdots \\ d_{K1,l} & d_{K2,l} & \cdots & d_{KK,l} \end{bmatrix} \\ &= \begin{bmatrix} b_{11,2l} - b_{11,1l} & b_{12,2l} - b_{12,1l} & \cdots & b_{1K,2l} - b_{1K,1l} \\ b_{21,2l} - b_{21,1l} & b_{22,2l} - b_{22,1l} & \cdots & b_{2K,2l} - b_{2K,1l} \\ \vdots & \vdots & \ddots & \vdots \\ b_{K1,2l} - b_{K1,1l} & b_{K2,2l} - b_{K2,1l} & \cdots & b_{KK,2l} - b_{KK,1l} \end{bmatrix}, \end{aligned} \tag{2}$$

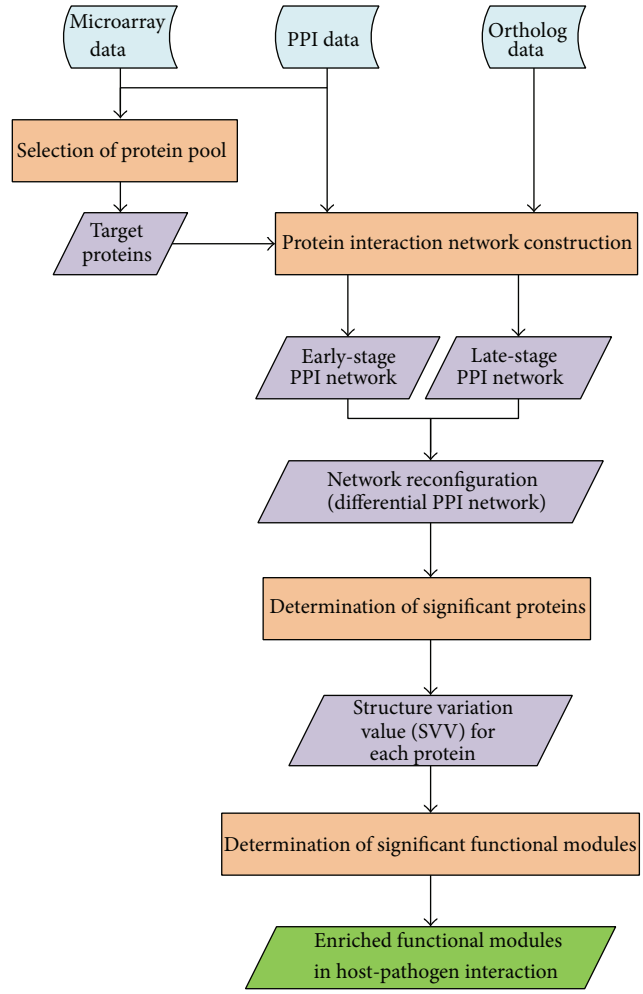


Figure 1: A flowchart for construction of protein-protein interaction networks and determination of enriched functional modules in host-pathogen interaction by comparing the early-stage and late-stage PPI networks. This figure shows the adopted approach in flowchart form. Blue boxes show the data sought in this study. Orange boxes indicate the steps used in the data-gathering process. Purple boxes represent the results of each processing step, and the green box denotes the final result of the whole approach.

where $d_{ij,l}$ indicates the change of protein interaction ability of the lth organism system between late-stage PPI network and early-stage PPI network for protein i and protein j, $b_{ij,1l}$ and $b_{ij,2l}$ represent the identified protein interaction ability between protein i and protein j for the early-stage PPI network and the late-stage PPI network of the lth organism, respectively, and l denotes the pathogen or host. Therefore, for each organism system, a matrix D_l was established to show the differential PPI network between the early-stage and late-stage PPI networks. Furthermore, the structural variations between the early-stage and late-stage PPI networks can be determined by the differential PPI network for each protein. The structure variation value (SVV) is considered to be an index to quantify the PPI network reconfiguration between these two stages in the infection process:

$$\mathrm{SVV}_l = \begin{bmatrix} \mathrm{SVV}_{1,l} \\ \mathrm{SVV}_{2,l} \\ \vdots \\ \mathrm{SVV}_{K,l} \end{bmatrix}, \tag{3}$$

where $\mathrm{SVV}_{i,l} = \sum_{j=1}^{K} |d_{ij,l}|$, l = host or pathogen, and $i = 1, \ldots, K$; that is, the reconfiguration of the protein i of the lth organism is obtained from the absolute sum of the ith row of D_l in (2) and the reconfiguration of the PPI network is obtained by the vector SVV_l for the lth organism. For a protein i of the lth organism system, $\mathrm{SVV}_{i,l}$ implies the extent of the structure change of the ith protein between the early stage and late stage of the infection process. SVV_l denotes the network reconfiguration of the lth organism.

2.5. Investigation of Significant Functional Modules in the Infection Process. During the host-pathogen interaction, the pathogen makes modifications to its PPI network for invasive purposes, while the host makes adjustments to its protein levels to defend itself against the pathogen. The participation of a protein in a specific biological process is correlated with the changes it undergoes in the PPI during that process. In this study, changes in the PPI structures in both organisms (i.e., the differential PPI networks) between the early and late stages of infection revealed proteins with significant

SVVs, which were then considered to play important roles in the infection process. Furthermore, functional modules made up of proteins with significant SVVs were regarded as important factors in the specific biological behavior of the infection process. Since no zebrafish died in the early stage (0.5–4 hpi) and the infected fish started to die in the late stage (4–18 hpi), these significant functional modules were considered to be important in conferring the virulence of the pathogen for *C. albicans*. In the case of zebrafish, these significant functional modules were possibly associated with defensive mechanisms by which certain biological processes were activated or inhibited in order to respond to *C. albicans* infection.

In order to determine the significance of the SVV of a given protein, an empirical P value was computed. A null distribution of SVVs was created based on the SVVs of random PPI networks. The random PPI networks were generated by permuting the network structure with the network size being constrained; that is, Erdős-Rényi random graph model was used to create the random PPI networks with the same number of protein interactions. Hence, the SVVs for each protein in the random PPI networks could be computed. With 100,000 iterations, the P value of a given SVV was calculated as the fraction of the number of random PPI networks in which that SVV was at least as large as the SVV of the real PPI network. SVVs with P values $\leq$ 0.05 were considered to be significant and the corresponding proteins were assumed to undergo significant changes in their interaction characteristics during the infection process.

Once proteins with significant SVVs were thus found in *C. albicans* interactions with zebrafish, gene annotations from the GO and DAVID databases were used to form enriched functional modules composed of the proteins identified as having significant SVVs. As time and environment change, the activated functional modules in an organism also change. We believe that the functional modules composed of proteins with significantly elevated SVVs in *C. albicans* are responsible for its pathogenesis. Likewise, enriched functional modules in infected zebrafish might be the major functional modules for defensive response induced by *C. albicans* infection.

3. Results

3.1. Construction of Dynamic PPI Networks and Identification of Significant Proteins in C. albicans Infection. Using the methods outlined above, several functional modules tentatively accounting for *C. albicans* pathogenicity and the associated defensive responses of zebrafish were investigated for their possible roles in the infection process. For *C. albicans*, the protein-protein interaction network was made up of 1,369 differentially expressed proteins selected in the protein pool. For zebrafish, a total of 7,861 proteins were all selected in the protein pool for PPI network construction without considering whether they were differentially expressed.

Although a rough PPI network can be constructed from target proteins along with PPI information obtained from database mining, it was more appropriate to refine the rough PPI network by microarray data of *C. albicans* infections. In this way, the refined early-stage and late-stage PPI networks were established for both *C. albicans* and zebrafish. For *C. albicans*, the early-stage PPI network was constructed from 1,318 proteins and 2,902 PPIs; the late-stage PPI network was comprised of 1,301 proteins and 4,045 PPIs. For zebrafish, there were 13,399 PPIs and 6,689 proteins in the early-stage PPI network and 18,807 interactions among 7,023 proteins in the late-stage PPI network. The extent of which a protein was considered significant in the infection process was based on the changes in the PPIs, that is, in the differential PPI networks (Supplementary Figure S1 and Figure S2 for *C. albicans* and zebrafish, resp.). In other words, significant proteins were identified by the comparison between edge variations of the PPI networks in the two stages of the infection process, as revealed by the SVVs in the differential PPI networks with P values $\leq$ 0.05 (distributions of SVVs for *C. albicans* and zebrafish were shown in Supplementary Figure S3). In this way, 139 *C. albicans* proteins were found to be of significance during the course of infection, and 380 zebrafish proteins were identified as playing important roles in defensive processes against *C. albicans*. Some functionally enriched modules comprised of these SVV-significant proteins are discussed in the following sections.

3.2. Investigation of Essential Functional Modules for Pathogenic and Defensive Mechanisms in C. albicans Infection

3.2.1. Functionally Enriched C. albicans Modules for Pathogenic Mechanism in the Infection Process. The 139 SVV-significant proteins seen in *C. albicans* can be divided into nine functional modules using annotations from the GO database. The nine modules are associated with hyphal morphogenesis, ion and small molecule transport, protein secretion, shifts in carbon utilization, stress responses, protein metabolism and catabolism, signal transduction, transcription-related processes, and other processes (proteins not belonging to the above-mentioned functional modules or lacking GO annotation, Figure 2). Eight out of nine functional modules are statistically enriched beyond what is expected by chance ($P < 0.05$, Fisher's exact test, except for other processes). Some of these nine functional modules are associated with general biological processes. Therefore, in this study, we have focused only on four of these modules: those playing a role in hyphal morphogenesis, ion and small molecule transport, protein secretion, and shifts in carbon utilization (Figure 2). We infer that these four enriched functional modules account for pathogenesis of *C. albicans* in infecting zebrafish, and further discussion of them is given below (Table 1).

(1) Hyphal morphogenesis: the fungal pathogen *C. albicans* grows in various morphogenic forms. The pathogen can exist in yeast form or undergo a process of morphogenesis to develop pseudohyphae or polarized hyphae. It has been demonstrated that the morphogenetic plasticity of *C. albicans* correlates closely with its pathogenicity and the phenotypic switch of the yeast to its hyphal form is a crucial factor in *C. albicans* pathogenesis [6]. Generally, mutant strains

Table 1: Enriched functional modules for pathogenic mechanisms in *C. albicans* and the corresponding significant proteins shown in Figure 2 during the host-pathogen interaction.

Functional module	Protein symbol
Hyphal morphogenesis	Cas4, Als3, Ifd6, Kis1, Mac4, Ndt80, orf19.6705
Ion and small molecule transport	Tna1, Agp2, Can1, Mac1, Als3, orf19.3769, Git1, Hut1, Mcd4, Mep1, Mrs7, orf19.1403, orf19.1427, orf19.2322.3, orf19.3132, orf19.3558, orf19.4897, Seo1, Sfc1, Vcx1
Protein secretion	Sap4, Sap5, Sap6, Sro77, Sys3, Ddi1, orf19.3247, orf19.7261, orf19.7604, orf19.841, Plb2, Prd1, Sec20, Spc3
Shifts in carbon utilization	Lat1, orf19.3782, Tes15, orf19.4121, Tes1, Pyc2, Acc1, Agp2, Pox1-3

This table lists four of the functional modules considered to be significant in *C. albicans* pathogenesis and the corresponding proteins with associated GO annotation.

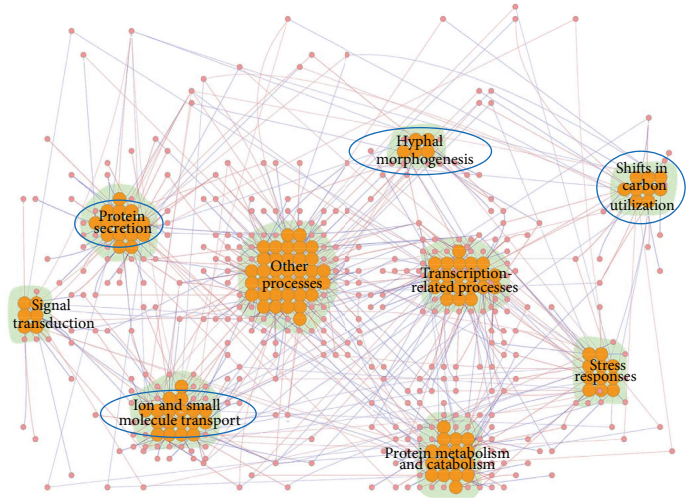


Figure 2: The functional modules composed of 139 *C. albicans* proteins found to be significant for pathogenic mechanisms in the infection process. This figure shows the differential PPI network constructed from 139 *C. albicans* proteins significant in the infection process and the interactions among them. Red and blue edges indicate positive and negative $d_{ij,1}$ values, respectively, as calculated using (2). The orange nodes represent the significant proteins, that is, proteins with SVV P values ≤ 0.05. There were nine enriched *C. albicans* functional modules occurring in host-pathogen interactions, playing roles in such processes as hyphal morphogenesis, ion and small molecule transport, protein secretion, shifts in carbon utilization, stress responses, protein metabolism and catabolism, signal transduction, transcription-related processes, and other processes. The functional modules marked with blue circles were investigated in this study. The figure was created using Cytoscape plugin Cerebral [18, 19]. The names of the proteins have been omitted for simplicity.

of *C. albicans* defective in the ability to form hyphae are less virulent in animal models [5, 34]. From a systematic level, considering the change of network structure between the two infection stages, several proteins with significant SVVs were included in the functional module, such as Cas4 and Als3 (Table 1). *PAG1*, alternatively known as *CAS4*, is a gene in the RAM network, a conserved signaling network that regulates polarized morphogenesis. The *CAS4* mutant showed hypersensitivity to cell wall-perturbing agents and loss of cell polarity [35]. Additionally, the hyphal form can interact with the yeast and pseudohyphal forms to produce biofilms, which can act as a source of recurrent infection and play an important role in resistance to antifungal agents [36]. Als3 is involved in biofilm formation and an *ALS3* mutant has been shown to be biofilm-defective *in vitro* [37]. In short, during *C. albicans* infection, morphogenesis was a crucial virulence-determining factor observed in the network structure variations between the early and late stages of infection.

(2) Ion and small molecule transport: *C. albicans* can thrive in various niches within its host, such as mouth, gastrointestinal tract, and the vagina. These niches are characterized by their diverse environments, with extreme variation in pH and nutrient composition. As a consequence, *C. albicans* must either adapt or, more likely, alter its niche in order to survive, possibly resulting in damage to the host tissue. Several proteins with significant SVVs were included in this functional module, such as Tna1, Agp2, and Can1, proteins responsible for nicotinic acid, carnitine, and amino acid transport, respectively (Table 1). These transporters, all showing significant interaction variations in network structures between the early and late stages of *C. albicans* infection, seemed to indicate that a reorganization of substrate uptake and utilization occurs. This reorganization could also enable *C. albicans* to assimilate available nutrients from its host allowing it to survive in a hostile microenvironment. Furthermore, using GO annotations, proteins like Als3, Mac1, and orf19.3769, which are involved in the transport

of such metal ions as iron, copper, and zinc, were also identified. Copper, zinc, and iron are all examples of nutritionally essential trace elements, also referred to as micronutrients. These minerals are required for growth and the optimal function of many organisms. Both excess and deficiency will have adverse effects in such organisms. Thus maintaining an adequate supply of micronutrients is critical for the viability of *C. albicans*. Previous research has shown that calprotectin from the cytoplasm of neutrophils can inhibit *C. albicans* growth through competition for zinc [38]. Therefore adequate zinc levels are required for *C. albicans* growth. In addition, superoxide dismutases cofactored with copper and zinc (Cu/ZnSOD) are found in *C. albicans* [39]. These enzymes play important roles in antioxidant defense when cells are exposed to oxygen. Antioxidants can inhibit oxidation reactions, which can produce free radicals, leading to cell damage or death. A previous study demonstrated that *C. albicans* lacking Cu/ZnSOD was more susceptible to macrophages and its virulence was seen to attenuate in mice [40]. Hence it would seem that the uptake of copper and zinc appears to be crucial in *C. albicans* growth and progression of infection. It is well known that *C. albicans* possesses iron acquisition mechanisms that are recognized to be essential for hyphal growth in the infection process [41]. Such mechanisms could also deprive the host of iron thereby exerting a harmful effect. Als3 is a hyphal-associated adhesin and invasin in *C. albicans* and is essential in ferritin-binding to the external hyphal layer [42]. Ferritin is an iron-containing host protein and therefore a potential iron source for pathogens. Previous studies have indicated that *C. albicans* mutants lacking *ALS3* displayed defective ferritin-binding abilities and that this also attenuated the pathogenic damage done to oral epithelial cells [41]. This distinctive iron-utilization characteristic contributes to the survival of the organism and pathogenesis in the host and also seems to play a role in hyphal formation, adhesion, and invasion during host-pathogen interactions. Taken together, ion and small molecule uptake and utilization enable the pathogen to adapt, invade, or even damage the host in *C. albicans* infection.

(3) Protein secretion: in the light of data from the differential PPI networks between the early and late infection stages, proteins having a role in protein secretion, such as Sap4 to Sap6, Sro77, and Sys3, showed significant network structure variations (Table 1). It is known that every cell is contained within a membrane that separates its interior from the external environment. Hence, the protein secretion system, which transports or extrudes molecules from the interior of microbes to the external environment, is an important mechanism by which microbes adapt and survive in their environments [43]. Several adhesins and extracellularly secreted hydrolases, such as secreted aspartyl proteinases (SAPs), secreted lipases (LIPs), and phospholipase B (PLB) [44], were identified as contributing to the virulence of *C. albicans*, and it is known that these proteins facilitate nutrient supplies, adhesion to host cells, tissue invasion, and even host cell damage. Dynamic changes to cell surface components and proteins released from the pathogen into the host cell environment can play important roles in host-pathogen interactions. That is to say, the virulence characteristics of *C. albicans* are closely related to its protein secretion mechanisms during infection. The Golgi apparatus is known to be involved in the secretion mechanism. Studies have demonstrated that in *C. albicans*, the Golgi complex, consisting of puncta, is redistributed to the distal portion of the extending hyphae whereas it is randomly distributed throughout the cytoplasm in the yeast form [45]. Since the Golgi apparatus is considered a locus of biomolecule manufacture, the relocation of the Golgi to the distal hyphal tip during hyphal formation means that post-Golgi secretory vesicles do not need to undergo long-distance transport from the cell body to the growing apical tip. Therefore, such Golgi redistribution would provide more rapid apical growth and result in further efficiency of the infection process [43, 45]. It also seems that clustered Golgi puncta give *C. albicans* the ability to secrete virulence-related proteins that adhere to, invade, or damage host tissue during host-pathogen interactions. However, there is no explicit evidence, except the well-known example of Sap4 to Sap6 [46], showing that mutant strains deficient in these identified protein secretion-related genes exhibit reduced virulence. Additional studies are needed to fully investigate the relationship between protein secretion and pathogenesis. Despite this, protein secretion mechanisms certainly enable *C. albicans* to adapt, survive, and invade the host, and therefore play an important role in host-pathogen interaction.

(4) Shifts in carbon utilization: once host immune cells recognize and attach to the pathogen-associated molecular patterns (PAMPs) of pathogens, phagocytosis is activated, in which pathogens are engulfed by the cell membranes of phagocytes to form an internal phagosome. Phagocytosis is a major cellular process used by hosts to destroy and remove pathogens. Some studies have shown that the phagosome of the host is a nutrient-poor environment for pathogens [47]. Moreover, *C. albicans* has been shown to undergo carbon starvation and glucose deprivation after internalization by macrophages [47]. Although glucose generally serves as the preferred source of energy and precursor for the synthesis of several other substances, the glucose-deficient environment of the macrophage requires carbon metabolism by the pathogen to be modulated after phagocytosis. Previous studies have shown that genes responsible for controlling the glyoxylate cycle and gluconeogenesis, which are used for the assimilation of two carbon compounds, are activated after *C. albicans* is exposed to macrophages. Additionally, acetyl-CoA is a precursor which can drive the glyoxylate cycle or gluconeogenesis and which could be derived from fatty acids from either *C. albicans* or the macrophage [47]. In this study, by looking at proteins with significant SVVs, a functional module of carbon utilization-related genes including Lat1, orf19.3782, and Tes15 was determined (Table 1). Annotation data from the *Candida* Genome Database revealed that Lat1 and orf19.3782 are associated with acetyl-CoA biosynthesis or transport. Tes15, orf19.4121, and Tes1 are involved in acyl-CoA metabolic processes; acyl-CoA is a coenzyme involved in the metabolism of fatty acids. Pyc2 is involved in the process of gluconeogenesis. Acc1, Agp2, and Pox1-3 are involved in fatty

acid biosynthesis, metabolic processes, and oxidation. It is speculated that during the early stage of infection, *C. albicans* is internalized by macrophages and some acetyl-CoA-, acyl-CoA-, and fatty acid-associated proteins, such as Lat1, Tes15, and Agp2 are activated (Figure 3). Nevertheless, during the late-stage of *C. albicans* infection, macrophages are killed and *C. albicans* begins to disseminate. When this occurs, glucose is used as the chief carbon source, causing the acetyl-CoA-associated and fatty acid-associated proteins to downregulate (Figure 3). However, the trends in the gene expression profiles of Tes15, orf19.4121, Tes1, and Pox1-3 are different from the other five proteins (Figure 3). This gradually increasing and subsequently decreasing expression profile suggests the induction of acetyl-CoA synthesis during infection. A possible explanation is that there is still some *C. albicans* that have been phagocytosed and therefore fatty acid-associated proteins are needed to produce glucose. Taken together, the rapid adaptation to ever-changing host environments, such as the reorganization of carbon utilization, enables *C. albicans* to survive and infect the host.

3.2.2. Functionally Enriched Zebrafish Modules for Defensive Mechanism during C. albicans Infection. To investigate what functional modules of zebrafish were induced by *C. albicans* infection, the bioinformatics database DAVID [27] and GO annotations were used for the analysis of zebrafish proteins. By applying functional annotation clustering in DAVID, 380 proteins identified as having significant SVVs could be classified into ten functional modules. These functional modules represent immune response, apoptosis mechanism, ion transport, protein secretion, hemostasis-related processes, signal transduction, transcription-related processes, embryonic morphogenesis and development, metabolism and catabolism, and other processes (proteins not belonging to the above-mentioned functional modules or without GO annotation, Figure 4). Among these ten modules, nine are statistically enriched ($P < 0.05$, Fisher's exact test, except for other processes). Since some functional modules represent general biological processes, in this study we have only focused on five functional modules: immune response, apoptosis mechanism, ion transport, protein secretion, and hemostasis-related process (Figure 4). The first four functional modules are considered to play defensive roles in the battle of host against pathogen. The functional module of the hemostasis-related process can be used to explain the pathological outcome, that is, the fatal bleeding of zebrafish seen in the *C. albicans*-zebrafish infection model in this study. These defensive functional modules are now further discussed (Table 2).

(1) Immune response: it is well known that *C. albicans* can be both a harmless commensal organism and a fatal pathogen. The transition from commensal organism to pathogen is dependent on the interaction between *C. albicans* and the host innate immune system. Several proteins with significant SVVs identified from the network structure variations during infection are involved in the immune response, such as Tlr2 and B2m (Table 2). During the interaction between host and pathogen, when the pattern recognition receptors (PRRs) of the host cells recognize fungal pathogen-associated molecular patterns (PAMPs), an innate response is usually triggered to combat the pathogen. These PRRs include various toll-like receptors (TLRs) which are expressed by different cell types, such as macrophages, monocytes, and dendritic cells, and are the primary immune sensors for the detection of invading pathogens [48]. One of the mechanisms involved in the innate immune recognitions of fungal pathogens is mediated by the Dectin-1/Tlr2 receptor complex that recognizes β-glucan, a major component of the cell wall of *C. albicans* [49]. In addition, the activation of macrophages and dendritic cells expressing TLRs will also initiate the adaptive immune response. Beta-2 microglobulin (B2m), a protein found to be with significant SVV in this study, is associated with MHC (major histocompatibility complex) class I molecules, which helps T cells to recognize antigens. It has been shown that B2m-knockout mice do not express MHC class I molecules and they lack CD8+ and natural killer T cells [50]. If dysfunction of the host immune system were to occur, pathogens would invade easily resulting in severe damage or even life-threatening systemic infection. The identification of the enrichment of the functional module underlying the immune response reinforces the important role the immune system plays in defensive mechanisms against invasive *C. albicans* in the encounter between *C. albicans* and zebrafish in the infection process.

(2) Apoptosis mechanism: apoptosis is a biological process which results in cell death and the proper modulation of apoptosis is essential for the survival of the host. However, hosts and pathogens induce apoptosis in different ways so as to gain optimal benefit. Pathogens have evolved diverse strategies to induce or inhibit host cell apoptosis, allowing the pathogen to evade the innate response and favoring further dissemination of the pathogen into the host tissues [51, 52]. In contrast to pathogens, hosts defend against infection by inducing apoptosis in infected cells and inhibiting apoptosis in immune cells [53, 54]. It has been demonstrated that the outer surface of the cell wall of *C. albicans* is coated with phospholipomannan (PLM) and that PLM binds to the membranes of macrophages and stimulates Tlr2-mediated apoptosis [55]. This results in macrophages apoptosis and enhanced survival of *C. albicans* [55, 56]. In contrast, studies have shown that resistance of monocytes to *C. albicans*-induced apoptosis may limit pathogen replication, protect monocyte viability, and therefore enhance the host defense response [53, 57]. In addition, although there is no explicit evidence indicating direct apoptotic effects on the nonphagocytic cells of the host in *C. albicans* infection, it has been shown that apoptosis affords the infected intestinal epithelial cells a mechanism for defense against invasive enteric pathogens by destroying infected or damaged epithelia [58]. Thus, we suggest that zebrafish induces apoptosis in *C. albicans*-infected cells to prevent pathogen dissemination or to kill pathogens. Taken together, during the *C. albicans*-zebrafish interaction, it seems that *C. albicans* interferes with the apoptosis mechanism of zebrafish to elude host defense and infect host cells, while zebrafish manipulate apoptosis to help eliminate the threat posed by *C. albicans*. No evidence

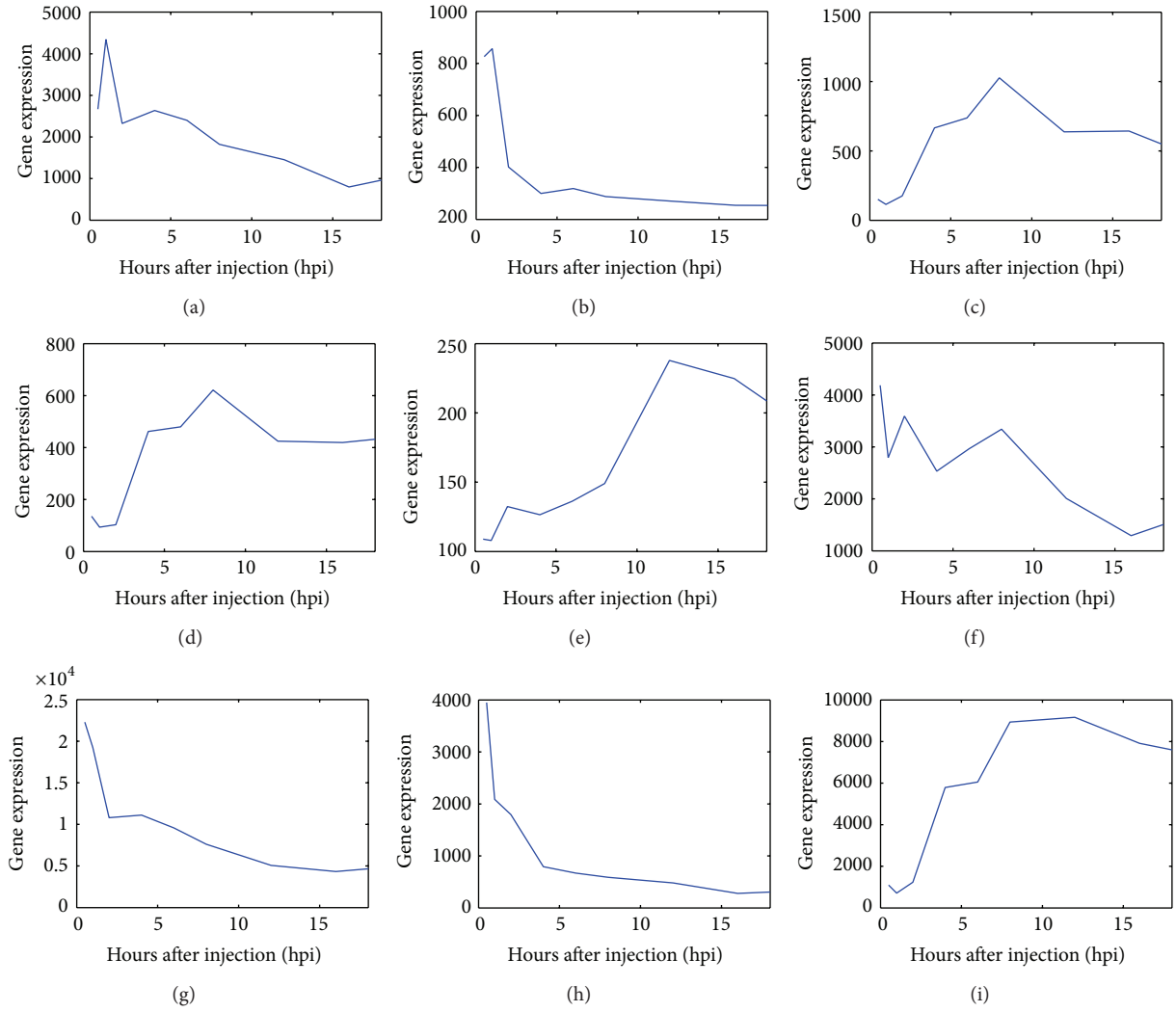


Figure 3: Gene expression profiles of *C. albicans* proteins in the functional module underlying shifts in carbon utilization: (a) Lat1, (b) orf19.3782, (c) Tes15, (d) orf19.4121, (e) Tes1, (f) Pyc2, (g) Acc1, (h) Agp2, and (i) Pox1-3.

exists to indicate that mutants with these identified apoptosis-related genes knocked out have higher mortality rates or are more highly susceptible to pathogen infection, but it is reasonable to speculate that dysfunctions in zebrafish apoptosis mechanisms would weaken the ability of the fish to defend against *C. albicans*, and zebrafish lacking such genes would be infected more severely and eventually die. Thus, the enriched functional module pertaining to apoptosis plays an important role in the complex host-pathogen interaction.

(3) Ion transport: from observations of significant SVVs, several proteins involved in calcium, potassium, iron, and zinc ion transport have been identified as significant in infection (Table 2). Previous studies have shown that the regulation of Ca^{2+} and K^{+}-signaling pathways is involved in T lymphocyte activation [59]. Additionally, intracellular calcium and potassium ion homeostasis influences apoptosis [60]. Therefore, proper calcium and potassium ion transport assists immune responses and apoptosis in zebrafish in response to *C. albicans* infection. In addition to calcium and potassium ions, there are also transporters that participate in micronutrient transport, such as iron and zinc. Micronutrients are nutrients required in small quantities to support normal physiological function. Pathogens have developed certain strategies to deprive the host of these micronutrients in order to promote growth and pathogenesis. Conversely, hosts sequester micronutrients from invading pathogens, that is, making these micronutrients unavailable to the pathogens, a concept termed nutritional immunity. Iron is an essential cofactor for several proteins and enzymes and, therefore, involved in numerous cellular functions and metabolic pathways [50]. A well-studied form of nutritional immunity is the iron-withholding defense system [61]. Using the approach adopted throughout this study, it was shown that transferrin-a (Tfa), a protein related to iron transport, undergoes a significant protein interaction change. Hosts have several iron-withholding mechanisms, and one of them acts through the host iron-binding proteins, the transferrins [41]. Thus, transferrin, responsible for iron scavenging in

Table 2: Enriched functional modules for defensive mechanisms in zebrafish and the corresponding significant proteins shown in Figure 4 during the host-pathogen interaction.

Functional module	Protein symbol
Immune response	Tlr2, B2m, Akt2, Akt2l, Apaf1, Cxcr3.2, Pik3r3a, Sigirr, Ticam1, Tlr20a, Vtna
Apoptosis mechanism	Akt2, Akt2l, Apaf1, Cdk5, Gdnfa, Nras, Phlda3, Pik3r3a, Plcg2, Prkar2ab, Sgk1, Tax1bp1a
Ion transport	Tfa, Abcc9, Cacnb3a, Cacnb4b, Clk2a, Cox5ab, Grid2, Grin1a, Grin1b, Kcnh1, Kcnq1, Sfxn1, si:ch211-12e13.7, si:ch211-258f14.5, Slc12a3, Slc26a6l, Slc39a6, Trpc6a, Trpm7, zgc:109934, zgc:162160
Protein secretion	Rab2a, Rab3da, Rab3db, Rab6ba, Rab8a, Rab10, Rab11a, Rab35, Ap1m2, Ap3m1, Atg4c, Bcap31, Naca, Nup85, Ramp2, Scamp2, Snx17, Tnpo2, Trpc4apa, zgc:113338
Hemostasis-related processes	Calcrla, Lama4, Acvrl1, Bmp4, Cdh2, Csrp1a, Ell, Gata2a, Hapln1b, Hopx, Nr2f2, Nrxn3a, Nrxn3b, Plxnb2a, Rab11a

The table lists five of the functional modules considered to be essential in the *C. albicans*-zebrafish interaction and the corresponding proteins with associated GO annotation.

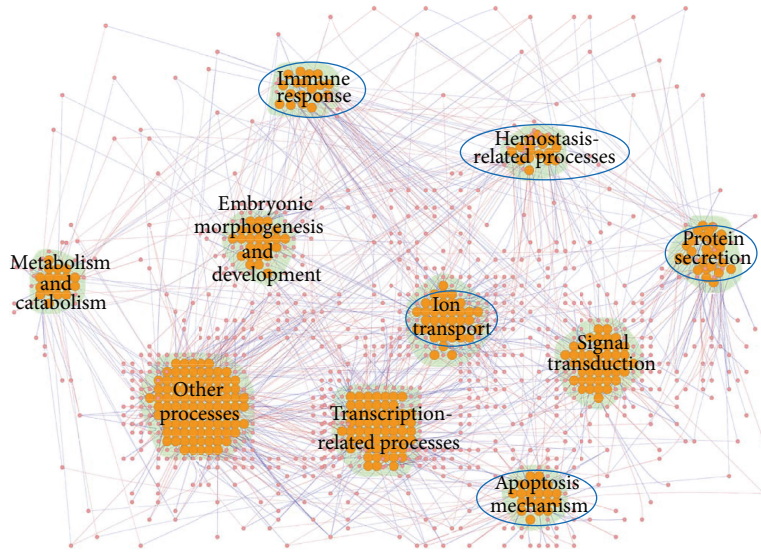


Figure 4: The functional modules composed of 380 zebrafish proteins found to be significant for defensive mechanisms in the infection process. This figure shows a differential PPI network constructed from 380 infection-significant zebrafish proteins and their interactions. Red and blue edges indicate positive and negative $d_{ij,l}$ values, respectively, calculated using (2). The orange nodes represent significant proteins, that is, proteins with SVV P values ≤ 0.05. There were ten enriched zebrafish functional modules occurring in host-pathogen interactions, including those underlying immune response, apoptosis mechanism, ion transport, protein secretion, hemostasis-related processes, signal transduction, transcription-related processes, embryonic morphogenesis and development, metabolism and catabolism, and other processes. The functional modules marked with blue circles were investigated in this study. The figure was created using Cytoscape plugin Cerebral [18, 19]. The protein names have been omitted for simplicity.

plasma and lymph, has antimicrobial activity in the host-pathogen interaction. Recent work has shown that there is also competition for micronutrients other than iron (zinc, e.g.) during the host-pathogen interaction [62]. Zinc, also essential for living organisms, plays a crucial role in the immune system, and zinc deficiency induces broad-spectrum defects in both innate and adaptive immunity [63]. Zinc sequestration by the host might inhibit microbial growth and protect against infection. A previous study has shown that calprotectin, a neutrophil-derived protein, competes with *C. albicans* for zinc which is needed for growth [64]. Taken together, it can be seen from the ion transport functional module investigated that it is reasonable to infer that normal ion transport systems, which limit micronutrient availability and are required for optimal immune or apoptotic function, assist zebrafish in defending against *C. albicans* infection.

(4) Protein secretion: like the situation in *C. albicans*, protein secretion or protein transport was identified as the enriched functional module during infection in zebrafish. Several proteins, especially those belonging to Rab family, showed significant network structure variations (Table 2). The Rab family is part of the Ras superfamily of small GTPases and functions in the regulation of intracellular vesicle trafficking and protein transport between different organelles and various secretory vesicles [65, 66]. Although there is no direct evidence linking zebrafish Rab family proteins with fungal infections, Rab GTPases have been found to be involved in the process of pathogen infection

in many other organisms. In *C. elegans*, the small GTPase Rab1 was shown to control innate immunity by mediating antimicrobial peptide gene expression [67]. In red drum fish (*S. ocellatus*), it has recently been reported that Rab1 regulates intracellular bacterial infection and thus is likely to play a role in bacteria-induced host immune defense [66]. In mammals, Rab5 and Rab7 have been shown to govern the early events of HIV-1 infection in human placental cells [68]. Additionally, Rab5 and Rab7 were demonstrated to affect the entry and transport of some viruses and bacteria [66]. These studies have indicated that Rab proteins are functionally associated with endocytosis and trafficking of intracellular pathogens, and pathogens evolve corresponding strategies to modulate Rab functions [69]. Based on the findings from other organisms and the fact that many Rab proteins were identified as SVV-significance in this study, we infer that Rab proteins in zebrafish also play important roles in host resistance against microbial infections. Further studies are needed to characterize the functions of zebrafish Rab proteins, which would help understand the relationship between protein secretion and host response during infection.

(5) Hemostasis-related processes: in this study, the pathological outcome of the host-pathogen interaction in zebrafish was found to be massive fatal bleeding. Based on the pathological outcome and the fact that *C. albicans* may cause deep-seated infection and disruption of endothelial surfaces [70], it seems that damage of host endothelial cells or blood vessels might occur during the interaction of *C. albicans* with zebrafish. Conversely, from the molecular perspective, several proteins important in hemostasis-related processes were found to have significant SVVs (Table 2). The zebrafish protein Calcrla, previously named Crlr, and protein Lama4 belong to this functional module. The calcitonin receptor-like receptor (Crlr) is a main endothelial cell receptor involved in cardiovascular homeostasis. In zebrafish, it has been demonstrated that mutation of *crlr*, which is associated with vascular development and angiogenesis, leads to atrophy of the trunk dorsal aorta or lack of blood circulation [71]. In this study, from the decreased Calcrla edge numbers in the protein interaction network dynamics, we may infer that Calcrla-employing biological processes were attenuated in the late stage of infection. Blood vessels are composed of two major cell types: endothelial cells and periendothelial cells. In addition to these cellular components, there are certain structural elements involved in the preservation of vascular integrity, such as adherens junctions, basement membranes, and the extracellular matrix [72]. Laminins are components of the basement membrane. Zebrafish with morpholino knockdown of *lama4* have been demonstrated to undergo cardiac dysfunction and embryonic hemorrhage [73]. In this study, the gene expression of the identified zebrafish laminin, alpha 4 (Lama4), declined as infection advanced (Figure 5), indicating that vascular integrity may not be maintained. Therefore, from the behavior of the hemostasis-related functional module identified, we speculate that *C. albicans* can penetrate endothelial cells and invade deeper tissues in zebrafish. In addition, the blood vessels of zebrafish were found to be damaged and vascular homeostasis could not be maintained during the late stage of *C. albicans* infection.

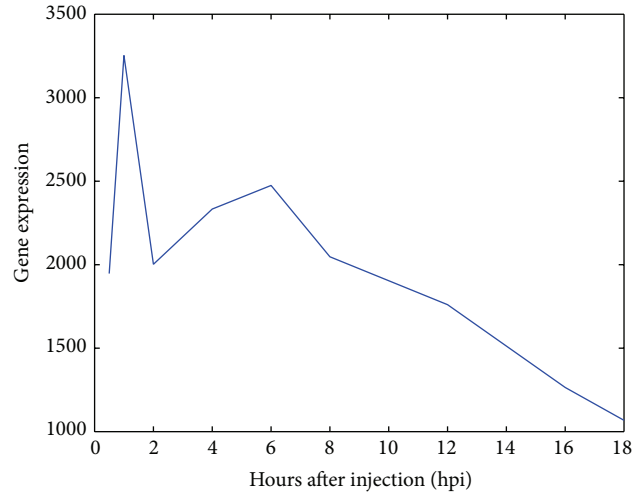


Figure 5: The gene expression profile of zebrafish laminin, alpha 4 (Lama 4).

4. Discussion

The importance of host-pathogen interactions has long been apparent to biologists and clinicians. It is vital to understand the possible factors determining the virulence of pathogens during an infection process. Simultaneously, the host defensive mechanisms and pathogenic mechanisms of damage, disease, and even mortality of the host are also of interest to researchers. Once the underlying molecular mechanisms are unveiled, it should become possible to develop various therapeutic strategies to prevent tissue damage and death caused by *C. albicans* infection. In this study, the pathogenic functional modules of *C. albicans* active during infection and the corresponding defensive functional modules of zebrafish occurring in response to the pathogenic threat were investigated by simultaneous host-pathogen interaction microarray data from both the systematic and molecular viewpoints. Through gene expression profiles, protein-protein interaction information obtained from database mining, and discrete dynamic interaction models, PPI networks of pathogen and host were constructed at two different infection stages. By comparing the refined PPI networks at the early and late infection stages to generate a differential PPI network, the PPI network reconfiguration and those proteins showing significant interaction changes during the infection period were determined. Furthermore, enriched functional modules among those proteins identified as playing significant roles were investigated in great detail by GO annotation. Hyphal morphogenesis, ion and small molecule transport, protein secretion, and shifts in carbon utilization were found to be the most important molecular mechanisms of pathogenesis in *C. albicans* infection. At the same time, immune responses, apoptosis, ion transport, and protein secretion were found

to be crucial molecular defensive mechanisms occurring in zebrafish in response to the pathogen. Additionally, we speculate from the functional module of hemostasis that *C. albicans* can damage the blood vessels of zebrafish, resulting in irreparable vascular destruction, which is consistent with the pathological outcome of fatal hemorrhage.

Biological systems are highly dynamic entities that continuously respond to environmental changes. However, few studies have investigated network reconfiguration or network rewiring to elucidate cellular responses [16]. The method employed in this study has been shown to be useful in constructing PPI networks and identifying the essential functional modules for pathogenic and defensive mechanisms in an infection process based on differential network analysis. Such an approach could highlight those interactions that changed dramatically across different conditions and potentially be suitable for the study of network comparisons with different cellular responses. Nevertheless, there are still some drawbacks to be addressed. First, the protein-protein interaction data for *C. albicans* and zebrafish used in this study were inferred from the interactomes of *S. cerevisiae* and humans with the help of corresponding ortholog data. Even though the imprecision in PPI information could lead to deviations between the constructed PPI network and the real situation, AIC was used to detect significant interactions under the specific condition of infection process with the help of gene expression profiles. In other words, the potential false positive PPIs that arose from ortholog-based inference could be pruned by AIC. In this case, the effect of the imprecise PPI information would be minimized. However, high coverage and reliable protein interaction maps for *C. albicans* and zebrafish would still benefit the construction of PPI networks and the investigation of essential functional modules in *C. albicans* infection in the future. Second, gene expression profiles were overlaid to estimate the expression of their corresponding proteins. However, there are several steps involved in the synthesis of proteins from mRNAs [74, 75]. The overlay of protein expression levels using gene expression values without any modification may result in inaccuracies in the identified PPI networks. Once high-throughput protein expression data are available, a great improvement in PPI network construction will be made. Third, the interactions of *C. albicans* with epithelial cells during the infection process can be roughly divided into three major steps: adhesion, invasion, and damage [76, 77]. Conversely, the host employs corresponding defensive mechanisms in response to invasive fungal infections, perhaps beginning with recognition, followed by defensive responses, and eventually a victor emerges from the competition of infection. However, due to limitations in the number of time points in the microarray data, the infection process was divided into only early and late stages in this study. If more time points in the microarray data could be obtained during the infection process, especially targeting the adhesion, invasion, and damage stages, more detailed stage-specific host-pathogen interactions could be investigated. In this way, the invasive and defensive strategies taken by *C. albicans* and zebrafish, respectively, could be more specifically elucidated.

Humans have to face a large number of challenges presented by pathogens during the course of a lifetime. Therefore, the investigation of molecular mechanisms of life-threatening infection is essential. For human fungal pathogens, the relevant research into infection could provide knowledge about network structures, infectious mechanisms, and bioecology [78]. Through a deeper understanding of pathogens, new therapeutic strategies against invasive microorganisms are also possible. However, few studies have explored the systems biology of pathogen infection and the host responses simultaneously. Hence, a comprehensive picture combining the pathogenic mechanisms of the pathogen and defensive mechanisms of its host could provide a novel antifungal drug discovery strategy to treat and prevent serious infectious disease, even mortality. In terms of the pathogen, the result of the analysis of *C. albicans* pathogenesis in this study highlights some functional modules associated with hyphal morphogenesis, ion and small molecule transport, protein secretion, and shifts in carbon utilization, in which all play crucial roles in invasion and damage to host cells. The molecules involved in these processes might be considered as potential targets for drug discovery. For hosts, the immune response, apoptosis, ion transport, protein secretion, and hemostasis-related processes were considered to be essential molecular mechanisms for defense and survival in *C. albicans* infections. Therefore, proteins involved in these functional modules could be therapeutically protected to prevent the irreparable damage caused by the infection. Most recently, we developed a computational framework to construct interspecies PPI network [79], which can be integrated with the current methodology to investigate the interspecies functional modules in the future. It is hoped, with the help of more detailed biological functional modules, that the treatment of life-threatening infection can be developed further and the mortality rates due to infection can ultimately be decreased.

5. Conclusions

In this study, with the help of simultaneous host-pathogen interaction microarrays for both *C. albicans* and zebrafish, we investigated essential functional modules for pathogenic and defensive mechanisms in *C. albicans* infections using differential network analysis. The early- and late-stage protein interaction networks for both organisms were first constructed. We then determined the network reconfiguration to identify the proteins with significant interaction variations during infection and to extract the enriched functional modules among these proteins. The hyphal morphogenesis, ion and small molecule transport, protein secretion, and shifts in carbon utilization functional modules in *C. albicans* were seen to play crucial roles in pathogen invasion and damage caused to host cells. The zebrafish functional modules like those involved in immune response, apoptosis mechanism, ion transport, protein secretion, and hemostasis-related processes were found to be significant as defensive mechanisms during *C. albicans* infection. The essential functional modules thus determined could provide insights into the molecular

mechanisms during the infection process and thereby help to devise potential therapeutic strategies to treat *C. albicans* infection.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Yu-Chao Wang and I-Chun Tsai contributed equally to this work.

Acknowledgment

This work was supported by the National Science Council of Taiwan under Grants NSC 102-2218-E-010-005-MY2 (to Yu-Chao Wang) and NSC 102-2745-E-007-001-ASP (to Bor-Sen Chen).

References

[1] F. P. Davis, D. T. Barkan, N. Eswar, J. H. Mckerrow, and A. Sali, "Host-pathogen protein interactions predicted by comparative modeling," *Protein Science*, vol. 16, no. 12, pp. 2585–2596, 2007.

[2] J. Naglik, A. Albrecht, O. Bader, and B. Hube, "*Candida albicans* proteinases and host/pathogen interactions," *Cellular Microbiology*, vol. 6, no. 10, pp. 915–926, 2004.

[3] J. Berman and P. E. Sudbery, "*Candida albicans*: a molecular revolution built on lessons from budding yeast," *Nature Reviews Genetics*, vol. 3, no. 12, pp. 918–930, 2002.

[4] R. A. Calderone and W. A. Fonzi, "Virulence factors of *Candida albicans*," *Trends in Microbiology*, vol. 9, no. 7, pp. 327–335, 2001.

[5] H.-J. Lo, J. R. Köhler, B. DiDomenico, D. Loebenberg, A. Cacciapuoti, and G. R. Fink, "Nonfilamentous *C. albicans* mutants are avirulent," *Cell*, vol. 90, no. 5, pp. 939–949, 1997.

[6] M. Whiteway and U. Oberholzer, "Candida morphogenesis and host-pathogen interactions," *Current Opinion in Microbiology*, vol. 7, no. 4, pp. 350–357, 2004.

[7] G. Cotter, S. Doyle, and K. Kavanagh, "Development of an insect model for the in vivo pathogenicity testing of yeasts," *FEMS Immunology & Medical Microbiology*, vol. 27, no. 2, pp. 163–169, 2000.

[8] A.-M. Alarco, A. Marcil, J. Chen, B. Suter, D. Thomas, and M. Whiteway, "Immune-deficient drosophila melanogaster: a model for the innate immune response to human fungal pathogens," *Journal of Immunology*, vol. 172, no. 9, pp. 5622–5628, 2004.

[9] C.-C. Chao, P.-C. Hsu, C.-F. Jen et al., "Zebrafish as a model host for *Candida albicans* infection," *Infection and Immunity*, vol. 78, no. 6, pp. 2512–2521, 2010.

[10] K. M. Brothers, Z. R. Newman, and R. T. Wheeler, "Live imaging of disseminated candidiasis in zebrafish reveals role of phagocyte oxidase in limiting filamentous growth," *Eukaryotic Cell*, vol. 10, no. 7, pp. 932–944, 2011.

[11] N. S. Trede, D. M. Langenau, D. Traver, A. T. Look, and L. I. Zon, "The use of zebrafish to understand immunity," *Immunity*, vol. 20, no. 4, pp. 367–379, 2004.

[12] G. J. Lieschke and P. D. Currie, "Animal models of human disease: zebrafish swim into view," *Nature Reviews Genetics*, vol. 8, no. 5, pp. 353–367, 2007.

[13] R. Martin, B. Wächtler, M. Schaller, D. Wilson, and B. Hube, "Host-pathogen interactions and virulence-associated genes during *Candida albicans* oral infections," *International Journal of Medical Microbiology*, vol. 301, no. 5, pp. 417–422, 2011.

[14] H. Kumar, T. Kawai, and S. Akira, "Pathogen recognition by the innate immune system," *International Reviews of Immunology*, vol. 30, no. 1, pp. 16–34, 2011.

[15] L. Rizzetto and D. Cavalieri, "Friend or foe: using systems biology to elucidate interactions between fungi and their hosts," *Trends in Microbiology*, vol. 19, no. 10, pp. 509–515, 2011.

[16] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, article 565, 2012.

[17] S. D. Durmuş Tekir and K. Ö. Ülgen, "Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era," *Biotechnology Journal*, vol. 8, no. 1, pp. 85–96, 2013.

[18] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[19] A. Barsky, J. L. Gardy, R. E. W. Hancock, and T. Munzner, "Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation," *Bioinformatics*, vol. 23, no. 8, pp. 1040–1042, 2007.

[20] Y. Y. Chen, C. C. Chao, F. C. Liu et al., "Dynamic transcript profiling of *Candida albicans* infection in zebrafish: a pathogen-host interaction study," *PLoS ONE*, vol. 8, no. 9, Article ID e72483, 2013.

[21] Y.-C. Wang, C.-Y. Lan, W.-P. Hsieh, L. A. Murillo, N. Agabian, and B.-S. Chen, "Global screening of potential *Candida albicans* biofilm-related transcription factors via network comparison," *BMC Bioinformatics*, vol. 11, article 53, 2010.

[22] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D698–D704, 2011.

[23] M. B. Arnaud, M. C. Costanzo, M. S. Skrzypek et al., "The *Candida* Genome Database (CGD), a community resource for *Candida albicans* gene and protein information," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D358–D363, 2005.

[24] Y. Bradford, T. Conlin, N. Dunn et al., "ZFIN: enhancements and updates to the zebrafish model organism database," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D822–D829, 2011.

[25] G. Östlund, T. Schmitt, K. Forslund et al., "Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D196–D203, 2010.

[26] M. A. Harris, J. Clark, A. Ireland et al., "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, supplement 1, pp. D258–D261, 2004.

[27] G. Dennis Jr., B. T. Sherman, D. A. Hosack et al., "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biology*, vol. 4, no. 5, article P3, 2003.

[28] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[29] M. Pagano and K. Gauvreau, *Principles of Biostatistics*, Duxbury Press, Pacific Grove, Calif, USA, 2nd edition, 2000.

[30] Y.-C. Wang and B.-S. Chen, "Integrated cellular network of transcription regulations and protein-protein interactions," *BMC Systems Biology*, vol. 4, article 20, 2010.

[31] R. Johansson, *System Modeling and Identification*, Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[32] H. Akaike, "New look at statistical-model Identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[33] D. R. Hyduke and B. O. Palsson, "Towards genome-scale signalling-network reconstructions," *Nature Reviews Genetics*, vol. 11, no. 4, pp. 297–307, 2010.

[34] R. Pukkila-Worley, A. Y. Peleg, E. Tampakakis, and E. Mylonakis, "*Candida albicans* hyphal formation and virulence assessed using a caenorhabditis elegans infection model," *Eukaryotic Cell*, vol. 8, no. 11, pp. 1750–1758, 2009.

[35] Y. Song, A. C. Seon, E. L. Kyung et al., "Role of the RAM network in cell polarity and hyphal morphogenesis in *Candida albicans*," *Molecular Biology of the Cell*, vol. 19, no. 12, pp. 5456–5477, 2008.

[36] J. Chandra, D. M. Kuhn, P. K. Mukherjee, L. L. Hoyer, T. McCormick, and M. A. Ghannoum, "Biofilm formation by the fungal pathogen *Candida albicans*: development, architecture, and drug resistance," *Journal of Bacteriology*, vol. 183, no. 18, pp. 5385–5394, 2001.

[37] C. J. Nobile, D. R. Andes, J. E. Nett et al., "Critical role of Bcr1-dependent adhesins in *C. albicans* biofilm formation in vitro and in vivo," *PLoS Pathogens*, vol. 2, no. 7, article e63, 2006.

[38] P. G. Sohnle, B. L. Hahn, and V. Santhanagopalan, "Inhibition of *Candida albicans* growth by calprotectin in the absence of direct contact with the organisms," *The Journal of Infectious Diseases*, vol. 174, no. 6, pp. 1369–1372, 1996.

[39] C.-S. Hwang, G.-E. Rhie, S.-T. Kim et al., "Copper- and zinc-containing superoxide dismutase and its gene from *Candida albicans*," *Biochimica et Biophysica Acta*, vol. 1427, no. 2, pp. 245–255, 1999.

[40] C.-S. Hwang, G.-E. Rhie, J.-H. Oh, W.-K. Huh, H.-S. Yim, and S.-O. Kang, "Copper- and zinc-containing superoxide dismutase (Cu/ZnSOD) is required for the protection of *Candida albicans* against oxidative stresses and the expression of its full virulence," *Microbiology*, vol. 148, no. 11, pp. 3705–3713, 2002.

[41] R. S. Almeida, D. Wilson, and B. Hube, "*Candida albicans* iron acquisition within the host," *FEMS Yeast Research*, vol. 9, no. 7, pp. 1000–1012, 2009.

[42] R. S. Almeida, S. Brunke, A. Albrecht et al., "The hyphal-associated adhesin and invasin Als3 of *Candida albicans* mediates iron acquisition from host ferritin," *PLoS Pathogens*, vol. 4, no. 11, Article ID e1000217, 2008.

[43] W. A. Fonzi, "The protein secretory pathway of *Candida albicans*," *Mycoses*, vol. 52, no. 4, pp. 291–303, 2009.

[44] M. Schaller, C. Borelli, H. C. Korting, and B. Hube, "Hydrolytic enzymes as virulence factors of *Candida albicans*," *Mycoses*, vol. 48, no. 6, pp. 365–377, 2005.

[45] P. C. G. Rida, A. Nishikawa, G. Y. Won, and N. Dean, "Yeast-to-hyphal transition triggers formin-dependent Golgi localization to the growing tip in *Candida albicans*," *Molecular Biology of the Cell*, vol. 17, no. 10, pp. 4364–4378, 2006.

[46] D. Sanglard, B. Hube, M. Monod, F. C. Odds, and N. A. R. Gow, "A triple deletion of the secreted aspartyl proteinase genes SAP4, SAP5, and SAP6 of *Candida albicans* causes attenuated virulence," *Infection and Immunity*, vol. 65, no. 9, pp. 3539–3546, 1997.

[47] M. C. Lorenz, J. A. Bender, and G. R. Fink, "Transcriptional response of *Candida albicans* upon internalization by macrophages," *Eukaryotic Cell*, vol. 3, no. 5, pp. 1076–1087, 2004.

[48] R. Medzhitov, "Recognition of microorganisms and activation of the immune response," *Nature*, vol. 449, no. 7164, pp. 819–826, 2007.

[49] M. G. Netea, N. A. R. Gow, C. A. Munro et al., "Immune sensing of *Candida albicans* requires cooperative recognition of mannans and glucans by lectin and Toll-like receptors," *The Journal of Clinical Investigation*, vol. 116, no. 6, pp. 1642–1650, 2006.

[50] U. E. Schaible and S. H. E. Kaufmann, "Iron and microbial infection," *Nature Reviews Microbiology*, vol. 2, no. 12, pp. 946–953, 2004.

[51] W. W. Navarre and A. Zychlinsky, "Pathogen-induced apoptosis of macrophages: a common end for different pathogenic strategies," *Cellular Microbiology*, vol. 2, no. 4, pp. 265–273, 2000.

[52] V. T. Heussler, P. Küenzi, and S. Rottenberg, "Inhibition of apoptosis by intracellular protozoan parasites," *International Journal for Parasitology*, vol. 31, no. 11, pp. 1166–1176, 2001.

[53] H. S. Kim, E. H. Choi, J. Khan et al., "Expression of genes encoding innate host defense molecules in normal human monocytes in response to *Candida albicans*," *Infection and Immunity*, vol. 73, no. 6, pp. 3714–3724, 2005.

[54] D. H. Dockrell, "The multiple roles of Fas ligand in the pathogenesis of infectious diseases," *Clinical Microbiology and Infection*, vol. 9, no. 8, pp. 766–779, 2003.

[55] D. Poulain and T. Jouault, "*Candida albicans* cell wall glycans, host receptors and responses: elements for a decisive crosstalk," *Current Opinion in Microbiology*, vol. 7, no. 4, pp. 342–349, 2004.

[56] S. Ibata-Ombetta, T. Idziorek, P.-A. Trinel, D. Poulain, and T. Jouault, "*Candida albicans* phospholipomannan promotes survival of phagocytosed yeasts through modulation of bad phosphorylation and macrophage apoptosis," *The Journal of Biological Chemistry*, vol. 278, no. 15, pp. 13086–13093, 2003.

[57] S. Heidenreich, B. Otte, D. Lang, and M. Schmidt, "Infection by *Candida albicans* inhibits apoptosis of human monocytes and monocytic U937 cells," *Journal of Leukocyte Biology*, vol. 60, no. 6, pp. 737–743, 1996.

[58] J. M. Kim, L. Eckmann, T. C. Savidge, D. C. Lowe, T. Witthöft, and M. F. Kagnoff, "Apoptosis of human intestinal epithelial cells after bacterial invasion," *The Journal of Clinical Investigation*, vol. 102, no. 10, pp. 1815–1823, 1998.

[59] M. D. Cahalan and K. G. Chandy, "Ion channels in the immune system as targets for immunosuppression," *Current Opinion in Biotechnology*, vol. 8, no. 6, pp. 749–756, 1997.

[60] S. P. Yu, L. M. T. Canzoniero, and D. W. Choi, "Ion homeostasis and apoptosis," *Current Opinion in Cell Biology*, vol. 13, no. 4, pp. 405–411, 2001.

[61] E. D. Weinberg, "Iron availability and infection," *Biochimica et Biophysica Acta*, vol. 1790, no. 7, pp. 600–605, 2009.

[62] T. E. Kehl-Fie and E. P. Skaar, "Nutritional immunity beyond iron: a role for manganese and zinc," *Current Opinion in Chemical Biology*, vol. 14, no. 2, pp. 218–224, 2010.

[63] A. H. Shankar and A. S. Prasad, "Zinc and immune function: the biological basis of altered resistance to infection," *The American Journal of Clinical Nutrition*, vol. 68, no. 2, supplement, pp. 447S–463S, 1998.

[64] S. J. Lulloff, B. L. Hahn, and P. G. Sohnle, "Fungal susceptibility to zinc deprivation," *The Journal of Laboratory and Clinical Medicine*, vol. 144, no. 4, pp. 208–214, 2004.

[65] H. Stenmark and V. M. Olkkonen, "The Rab GTPase family," *Genome Biology*, vol. 2, no. 5, article REVIEWS3007, 2001.

[66] Y.-H. Hu, T. Deng, and L. Sun, "The Rab1 GTPase of *Sciaenops ocellatus* modulates intracellular bacterial infection," *Fish & Shellfish Immunology*, vol. 31, no. 6, pp. 1005–1012, 2011.

[67] C. Couillault, N. Pujol, J. Reboul et al., "TLR-independent control of innate immunity in *Caenorhabditis elegans* by the TIR domain adaptor protein TIR-1, an ortholog of human SARM," *Nature Immunology*, vol. 5, no. 5, pp. 488–494, 2004.

[68] G. Vidricaire and M. J. Tremblay, "Rab5 and Rab7, but not ARF6, govern the early events of HIV-1 infection in polarized human placental cells," *Journal of Immunology*, vol. 175, no. 10, pp. 6517–6530, 2005.

[69] J. H. Brumell and M. A. Scidmore, "Manipulation of Rab GTPase function by intracellular bacterial pathogens," *Microbiology and Molecular Biology Reviews*, vol. 71, no. 4, pp. 636–652, 2007.

[70] J. R. Naglik, S. J. Challacombe, and B. Hube, "*Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis," *Microbiology and Molecular Biology Reviews*, vol. 67, no. 3, pp. 400–428, 2003.

[71] S. Nicoli, C. Tobia, L. Gualandi, G. de Sena, and M. Presta, "Calcitonin receptor-like receptor guides arterial differentiation in zebrafish," *Blood*, vol. 111, no. 10, pp. 4965–4972, 2008.

[72] A. Luttun and P. Verhamme, "Keeping your vascular integrity: what can we learn from fish?" *BioEssays*, vol. 30, no. 5, pp. 418–422, 2008.

[73] R. Knöll, R. Postel, J. Wang et al., "Laminin-α4 and integrin-linked kinase mutations cause human cardiomyopathy via simultaneous defects in cardiomyocytes and endothelial cells," *Circulation*, vol. 116, no. 5, pp. 515–525, 2007.

[74] T. Shiraishi, S. Matsuyama, and H. Kitano, "Large-scale analysis of network bistability for human cancers," *PLoS Computational Biology*, vol. 6, no. 7, Article ID e1000851, 2010.

[75] S. M. Friedman, R. Berezney, and I. B. Weinstein, "Fidelity in protein synthesis. The role of the ribosome," *The Journal of Biological Chemistry*, vol. 243, no. 19, pp. 5044–5048, 1968.

[76] W. Zhu and S. G. Filler, "Interactions of *Candida albicans* with epithelial cells," *Cellular Microbiology*, vol. 12, no. 3, pp. 273–282, 2010.

[77] K. Zakikhany, J. R. Naglik, A. Schmidt-westhausen, G. Holland, M. Schaller, and B. Hube, "*In vivo* transcript profiling of *Candida albicans* identifies a gene essential for interepithelial dissemination," *Cellular Microbiology*, vol. 9, no. 12, pp. 2938–2954, 2007.

[78] A. Dobson, "Population dynamics of pathogens with multiple host species," *The American Naturalist*, vol. 164, supplement 5, pp. S64–S78, 2004.

[79] Y.-C. Wang, C. Lin, M.-T. Chuang et al., "Interspecies protein-protein interaction network construction for characterization of host-pathogen interactions: a *Candida albicans*-zebrafish interaction study," *BMC Systems Biology*, vol. 7, article 79, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 985706, 8 pages
http://dx.doi.org/10.1155/2014/985706

Research Article

Visualization of Genome Signatures of Eukaryote Genomes by Batch-Learning Self-Organizing Map with a Special Emphasis on *Drosophila* Genomes

Takashi Abe,[1] Yuta Hamano,[2,3] and Toshimichi Ikemura[3]

[1] *Information Engineering, Niigata University, Niigata-shi, Niigata-ken 950-2181, Japan*
[2] *Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0101, Japan*
[3] *Nagahama Institute of Bio-Science and Technology, Nagahama-shi, Shiga-ken 526-0829, Japan*

Correspondence should be addressed to Toshimichi Ikemura; t_ikemura@nagahama-i-bio.ac.jp

Received 1 November 2013; Accepted 4 February 2014; Published 11 March 2014

Academic Editor: Altaf-Ul- Amin

A strategy of evolutionary studies that can compare vast numbers of genome sequences is becoming increasingly important with the remarkable progress of high-throughput DNA sequencing methods. We previously established a sequence alignment-free clustering method "BLSOM" for di-, tri-, and tetranucleotide compositions in genome sequences, which can characterize sequence characteristics (genome signatures) of a wide range of species. In the present study, we generated BLSOMs for tetra- and pentanucleotide compositions in approximately one million sequence fragments derived from 101 eukaryotes, for which almost complete genome sequences were available. BLSOM recognized phylotype-specific characteristics (e.g., key combinations of oligonucleotide frequencies) in the genome sequences, permitting phylotype-specific clustering of the sequences without any information regarding the species. In our detailed examination of 12 *Drosophila* species, the correlation between their phylogenetic classification and the classification on the BLSOMs was observed to visualize oligonucleotides diagnostic for species-specific clustering.

1. Introduction

Genome sequences, even protein-noncoding sequences, contain a wealth of information. The G + C content (%GC) is a fundamental characteristic of individual genomes and is used for a long period as a basic phylogenetic parameter to characterize individual genomes and genomic portions. The %GC, however, is too simple a parameter to differentiate wide varieties of genomes. Many groups have reported that oligonucleotide composition varies significantly among genomes and can be used to study genome diversity [1–8]. Because oligonucleotide composition can be used to distinguish species even with the same %GC, it has been called a "genome signature" [4, 5].

The unsupervised neural network algorithm known as Kohonen's Self-Organizing Map (SOM) is a powerful tool for clustering and visualizing high-dimensional vectorial data on a two-dimensional map [9–11]; oligonucleotide composition is an example of such high-dimensional data. We have previously developed a modified type SOM (batch-learning SOM: BLSOM) that depends on neither the order of data input nor the initial conditions, for codon frequencies in gene sequences [12] and oligonucleotide frequencies in genome sequences [13, 14]. BLSOM could recognize and visualize species-specific characteristics of codon or oligonucleotide frequencies in individual genomes, permitting clustering of genes or genome fragments according to species without the need for species information during BLSOM learning. Various high-performance supercomputers are now available for biological studies, and the BLSOM is suitable for actualizing high-performance parallel-computing with high-performance supercomputers. Therefore, this alignment-free clustering method was successfully applied to the phylogenetic classification of genome sequence fragments [15] and

to the analysis of a large number of microbial sequences obtained by metagenome studies of environmental and clinical samples [16].

To test the power of BLSOM to detect differences among eukaryote genomes and particularly among closely related species, the present study examined *Drosophila* in detail, for which the genomes of many closely related species have been sequenced. We constructed BLSOM with tetra- and pentanucleotide compositions in most (if not all) of the *Drosophila* genomes available and focused on species-specific characteristics of oligonucleotide frequencies in each *Drosophila* genome (genome signature), in connection with their phylogenetic classification.

2. Materials and Methods

2.1. Batch-Learning Self-Organizing Map (BLSOM). SOM is an unsupervised neural network algorithm that implements a characteristic nonlinear projection from the high-dimensional space of input data onto a two-dimensional array of weight vectors [9–11]. We modified the conventional SOM for genome informatics to make the learning process and resulting map independent of the order of data input, on the basis of batch learning SOM (BLSOM) [12, 13]. The initial weight vectors were defined by Principal Component Analysis (PCA) instead of random values. BLSOM learning was conducted as described previously [13], and the BLSOM program was obtained from UNTROD Inc. (takaabe@ie.niigata-u.ac.jp or y_wada@nagahama-i-bio.ac.jp).

2.2. Genome Sequences. Genome DNA sequences were obtained from http://hgdownload.cse.ucsc.edu/downloads.html. When the number of undetermined nucleotides (Ns) in a sequence exceeded 10% of the window size, the sequence was omitted from the analysis. When the number of Ns was less than 10%, the oligonucleotide frequencies were normalized to the length without Ns and included in the analysis.

3. Results

3.1. BLSOMs for 101 Eukaryote Genomes. A large number of genomes, including a wide variety of eukaryotes, have been sequenced with the remarkable progress of currently available DNA sequencing technologies. To investigate the clustering capacity of BLSOM for sequences derived from a wide range of eukaryotes, we first analyzed tetra- and pentanucleotide frequencies in ca. 1,800,000 nonoverlapping 5 kb sequences as well as ca. 900,000 nonoverlapping 10 kb sequences and overlapping 100 kb sequences with a 10 kb sliding step from 101 eukaryotic genomes, most of which were completely sequenced. To analyze sequences derived even from lower eukaryotes with small genome sizes in accurate detail, excess representation of higher eukaryotes' sequences derived from their large genomes had to be avoided. Therefore, for higher eukaryotes, 100 kb sequences were selected randomly from each large genome up to 200 Mb and used for the BLSOM analyses. In DNA databases, only one strand of a pair of complementary sequences is registered, and the choice between the two complementary sequences is often arbitrary in the registration. When global characteristics of oligonucleotide composition in the genome are considered, the distinction of frequencies between the complementary oligonucleotides (e.g., AAAC versus GTTT) is not important in most cases. In the present study, BLSOM was constructed with frequencies for degenerate sets in which the frequencies of a pair of complementary tetra- or pentanucleotides were added (DegeTetra- or DegePenta-BLSOM in Figure 1); this process roughly halved the computation time. The oligonucleotide frequencies were initially analyzed by PCA, and the resulting first and second principal components were used to set the initial weight vectors for the successive BLSOM. After 145 cycles of BLSOM learning, oligonucleotide frequencies in the fragment sequences were represented by the final weight vectors in the two-dimensional array. The resulting BLSOM revealed the clear separation (self-organization) of genomic fragments according to phylotypes (Figure 1). The computational time of DegeTetra-BLSOM was approximately six hours using high performance parallel computers; if PC servers were used, the time required for computation was 100 or more times as long.

In this figure, a node that included sequences from a single phylogenetic family was indicated in the color representing the family, while a node that included sequences from more than one family was indicated in black. Sequences from each family were clustered tightly on these tetra- and pentanucleotide BLSOMs. The accuracy level of separation by family on the 5-, 10-, and 100 kb DegeTetra-BLSOM was approximately 74, 90, and 99%, respectively, showing that longer sequences gave the higher accuracy of clustering according to phylotype.

In the 100 kb BLSOMs, the phylogenetic family territories were surrounded by contiguous white nodes, which contained no genome sequences in the final map. In other words, family borders could be drawn automatically on the basis of the contiguous white nodes. This is because the representative vectors of the family-specific nodes were very distinctive between different families even near territory borders. Much narrower white borders were observed within a certain family territory and primarily represented genus/species separation.

3.2. Diagnostic Oligonucleotides for Phylotype-Specific Clustering. %GC has long been used as a fundamental parameter for the phylogenetic characterization of species. We previously found that the %GC from the weight vector representing each node on a BLSOM was reflected in the horizontal axis [13]. Supporting the previous finding, the %GC increased globally from left to right on the present BLSOMs (the DegePenta-%GC panel in Figure 1(c)); therefore, sequences with high %GC (red in Figure 1(c)) were located on the right side. Importantly, sequences even with the same %GC were clearly separated on BLSOMs by a complex combination of oligonucleotide frequencies, resulting in accurate phylotype separation.

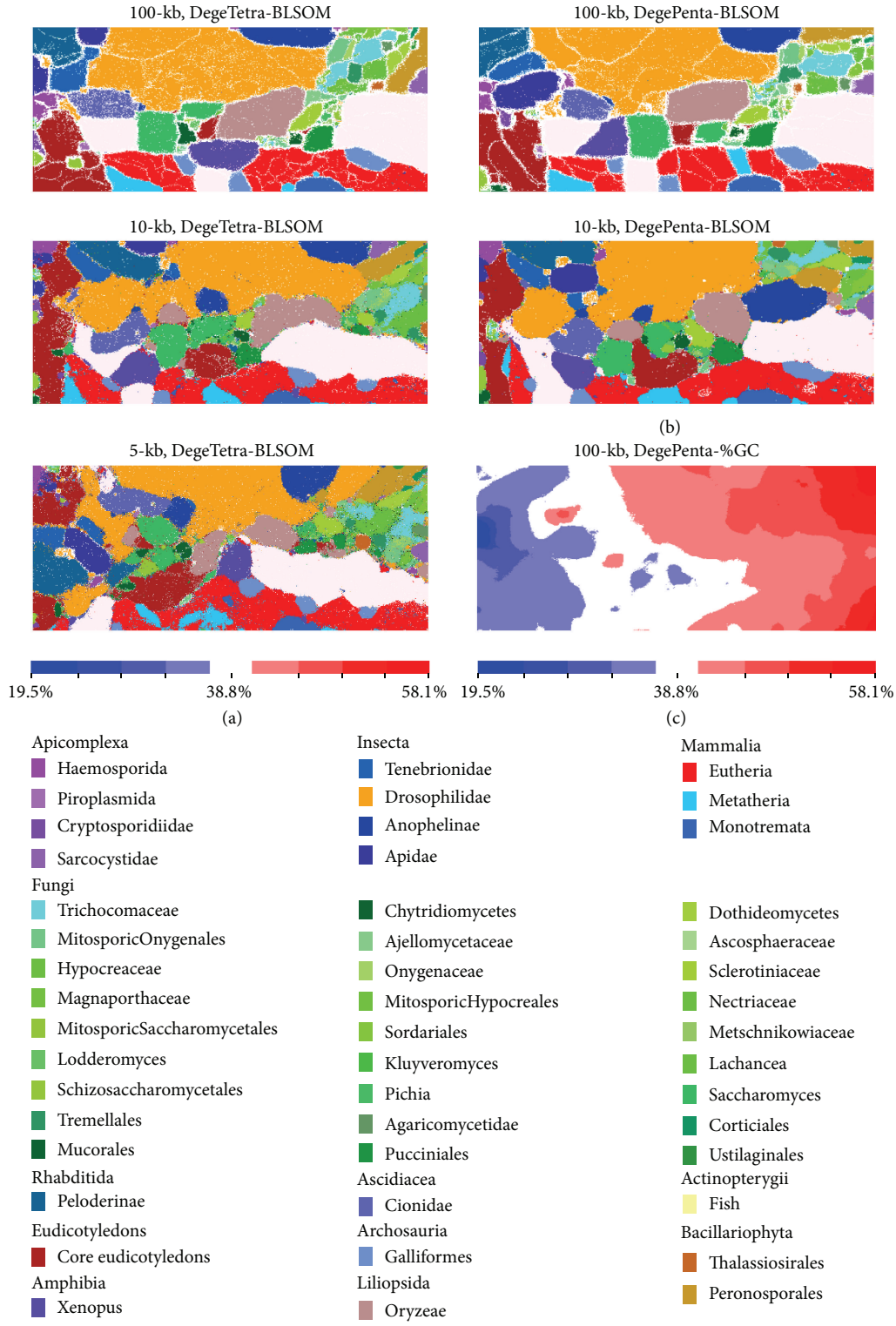


Figure 1: BLSOMs for the overlapping 100 kb with a 10 kb sliding step and the nonoverlapping 10- and 5 kb sequences from 101 eukaryotic genomes. (a) DegeTetra- and (b) DegePenta-BLSOMs. BLSOM was constructed with frequencies for degenerate sets in which the frequencies of a pair of complimentary tetra- or pentanucleotides were added. Nodes that include sequences from more than one phylogenetic family are indicated in black, those that contain no genomic sequences are indicated in white, and those containing sequences from a single family are indicated in colors. Differences in color were difficult to distinguish individual phylogenetic families because 47 families were analyzed simultaneously, but the observation that the back nodes were very rare showed the proper clustering of sequences according to phylotype. (c) 100 kb DegePenta-%GC; the %GC was calculated from the vectorial data representing each node in the 100 kb DegePenta-BLSOM and divided into nine categories with an equal number of nodes, as listed at the bottom of this panel; the %GC ranged from 19.5 to 58.1 and the midvalue was 38.8.

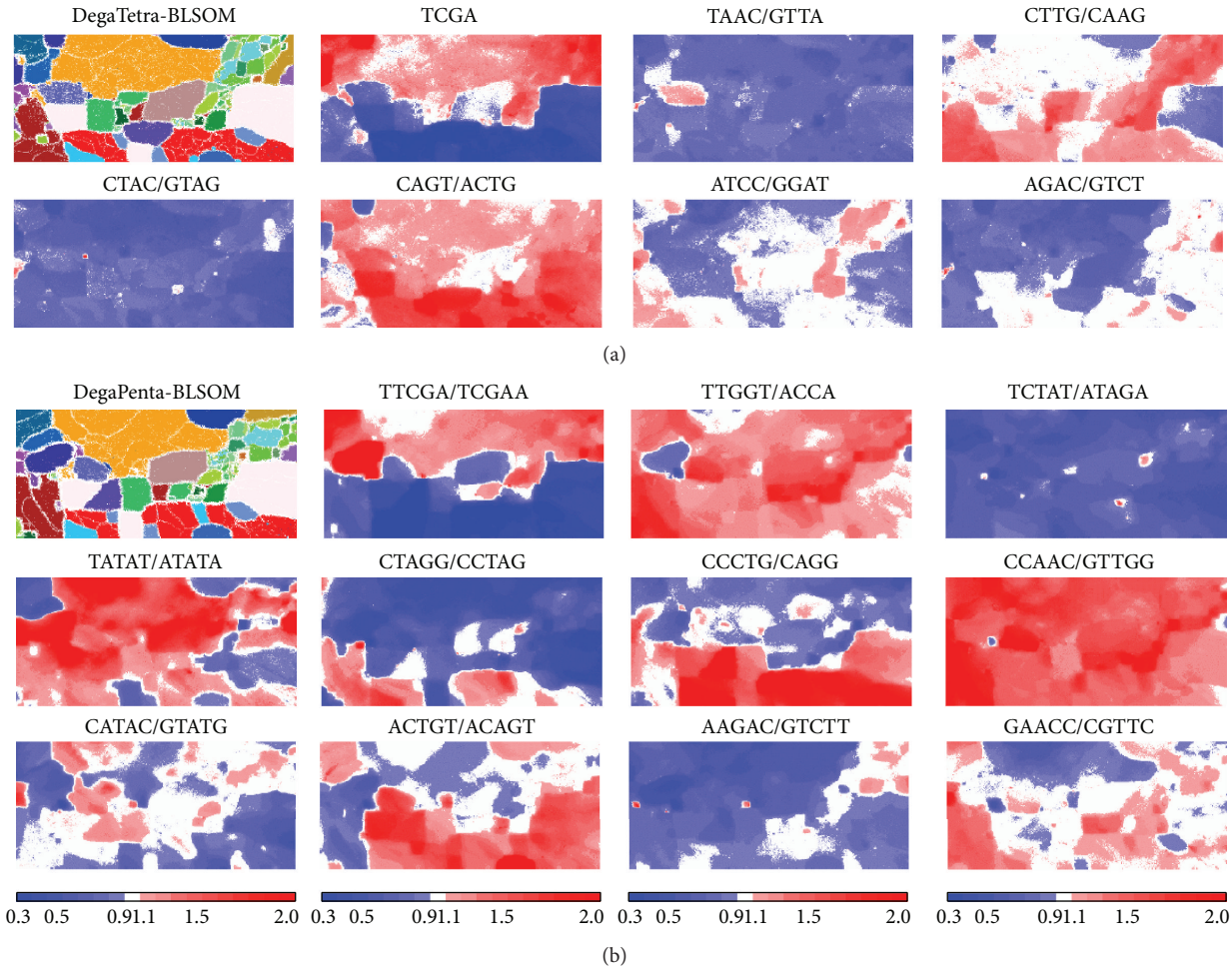


Figure 2: The level of each pair of complimentary tetra- or pentanucleotides on 100 kb BLSOMs. (a) DegeTetra- and (b) DegePenta-BLSOMs were those listed in Figure 1. Diagnostic examples of phylotype separations are presented. The level of each pair of complimentary tetra- or pentanucleotides in each node on the 100 kb DegeTetra- and DegePenta-BLSOMs in Figure 1 was calculated and normalized with the level expected from the mononucleotide composition of the node. The observed/expected ratio is indicated in colors shown at the bottom of the figure.

BLSOM provides a powerful ability for visualizing diagnostic oligonucleotides that contribute to the self-organization of sequences according to phylotype [14, 15]. The frequency of each tetra- or pentanucleotide obtained from the weight vector for each node in the 100 kb DegeTetra- or DegePenta-BLSOM listed in Figure 1 was calculated and normalized with the level expected from the mononucleotide composition calculated from the vectorial data representing each node. The observed/expected ratios were illustrated in red (overrepresented), blue (underrepresented), and white (moderately represented) in Figure 2. This normalization allowed oligonucleotide frequencies in each node to be studied independent of the %GC of the sequences [15]. Transitions between red (overrepresentation) and blue (underrepresentation) for various tetra- and pentanucleotides often coincided exactly with phylotype territory borders, indicating that BLSOM recognized the phylotype-specific combination of oligonucleotide frequencies that was the representative signature of each genome. Seven tetranucleotide and eleven pentanucleotide examples, which were diagnostic for phylotype territory formation, were presented in Figure 2; distribution patterns for all tetra- and pentanucleotides were listed in Supplementary Figures S1 and S2 (see Figures S1 and S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/985706). It should be stressed that complex combinations of many oligonucleotides contributed to the self-organization of sequence fragments according to phylotype and that BLSOM could visualize the diagnostic oligonucleotides in an easy-to-understand manner.

3.3. Visualization of 12 Drosophila Genome Sequences. In Figure 1, we analyzed 101 eukaryotic genomes that covered a wide range of phylogenetically distant eukaryotes. Then, to investigate the clustering capacity of BLSOM for the genomes of phylogenetically closely related species, nodes containing sequences derived only from 12 *Drosophila* genomes were

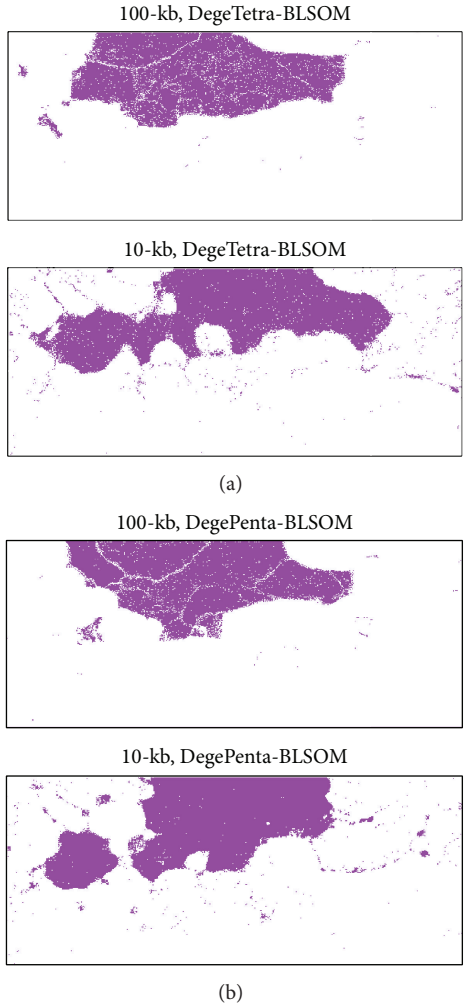


Figure 3: Distribution patterns of nodes containing sequences derived only from 12 *Drosophila* genomes on 100 kb DegeTetra- (a) and DegePenta- (b) BLSOMs, which were listed in Figure 1.

specifically marked in pink on the BLSOMs (Figure 3). In Figure 4, we examined the correlation between the classification pattern on the DegePenta-BLSOM and the phylogenetic classification according to *Drosophila* genomes by referring to a phylogenetic tree for the 12 *Drosophila* species, which was obtained from Flybase [17]. Distribution patterns for species belonging to one of the five *Drosophila* groups, which were specified by five boxes in the phylogenetic tree, were similar to each other, but patterns for species belonging to different groups were clearly distinct; the same result was obtained in DegeTetra-BLSOM (data not shown). This showed that BLSOM could properly extract sequence characteristics of *Drosophila* genomes with phylogenetic clustering through sequence homology searching. The diagnostic oligonucleotides contributing to the clustering according to the *Drosophila* group on BLSOMs could be assigned as shown in Figure 2. For example, the frequency of CTTCG was low only for the melanogaster group, while those of ATTCX, TGGTC, and TTCGY were low only for the virilis group; see the distribution patterns for tetra- and pentanucleotides listed in Supplementary Figures S1 and S2.

3.4. BLSOMs Constructed with 12 Drosophila Genomes. The results presented in Figures 1 and 2 showed that BLSOM could analyze almost all eukaryotic genome sequences available from the current DNA databanks simultaneously on a single map and visualize their genome signatures. This comprehensive, panoramic view of a huge number of genome sequences will become increasingly important because a large number of genomes (closely or distantly related with each other) have been sequenced intensively with next-generation DNA sequencers. This is one capacity of BLSOM. The results presented in Figures 3 and 4 also showed that BLSOM had a good capacity for distinguishing closely related species. When focusing only on closely related species, such as those belonging to one genus, BLSOM constructed only for these species may provide much detailed information. To examine this possibility, we constructed DegeTetra- and DegePenta-BLSOMs for 100 kb sequences with a 10 kb sliding step derived from the 12 *Drosophila* genomes. Nodes containing sequences only from one *Drosophila* group listed in Figure 4 were marked in the color representing the group (Figures 5(a) and 5(b)). In Figure 5(c), nodes containing sequences only from one species on the DegePenta-BLSOM were marked in the color representing the species. Species belonging to one group had similar patterns or their territories were adjacently located; for their phylogenetic closeness, refer to the phylogenetic tree listed in Figure 4. It should also be noted that there was segmentation of the territory of one species and/or that there were minor satellite territories apart from its major territory. Similar results were also obtained from DegeTetra-BLSOM (data not shown). These characteristics of individual species on BLSOMs may presumably represent genome characteristics of these species.

In Figure 5(d), examples of the pentanucleotides diagnostic for separation by group/species were presented; the results of all pentanucleotides were presented in Supplementary Figure S3. Although the functions of the oligonucleotides diagnostic for separation by group/species have not yet been studied, some of them may relate to the biological functions and/or evolutionary processes leading to the construction of the present genomes [18–20].

4. Discussion and Conclusion

When characteristic oligonucleotides, both underrepresented and overrepresented in each genome, are considered (Figures 2 and 4), various molecular mechanisms, including context-dependent mutation, repair, and modification, may be responsible [1–8]. For overrepresented sequences, preferences for sequences recognized by ubiquitous DNA-binding proteins and ubiquitous repetitive elements must be considered. It should also be mentioned that oligonucleotides, such as tetra-heptanucleotides, often represent motif sequences responsible for sequence-specific protein binding (e.g., transcription factor binding). Occurrences of such motif oligonucleotides should differ from the occurrences expected from

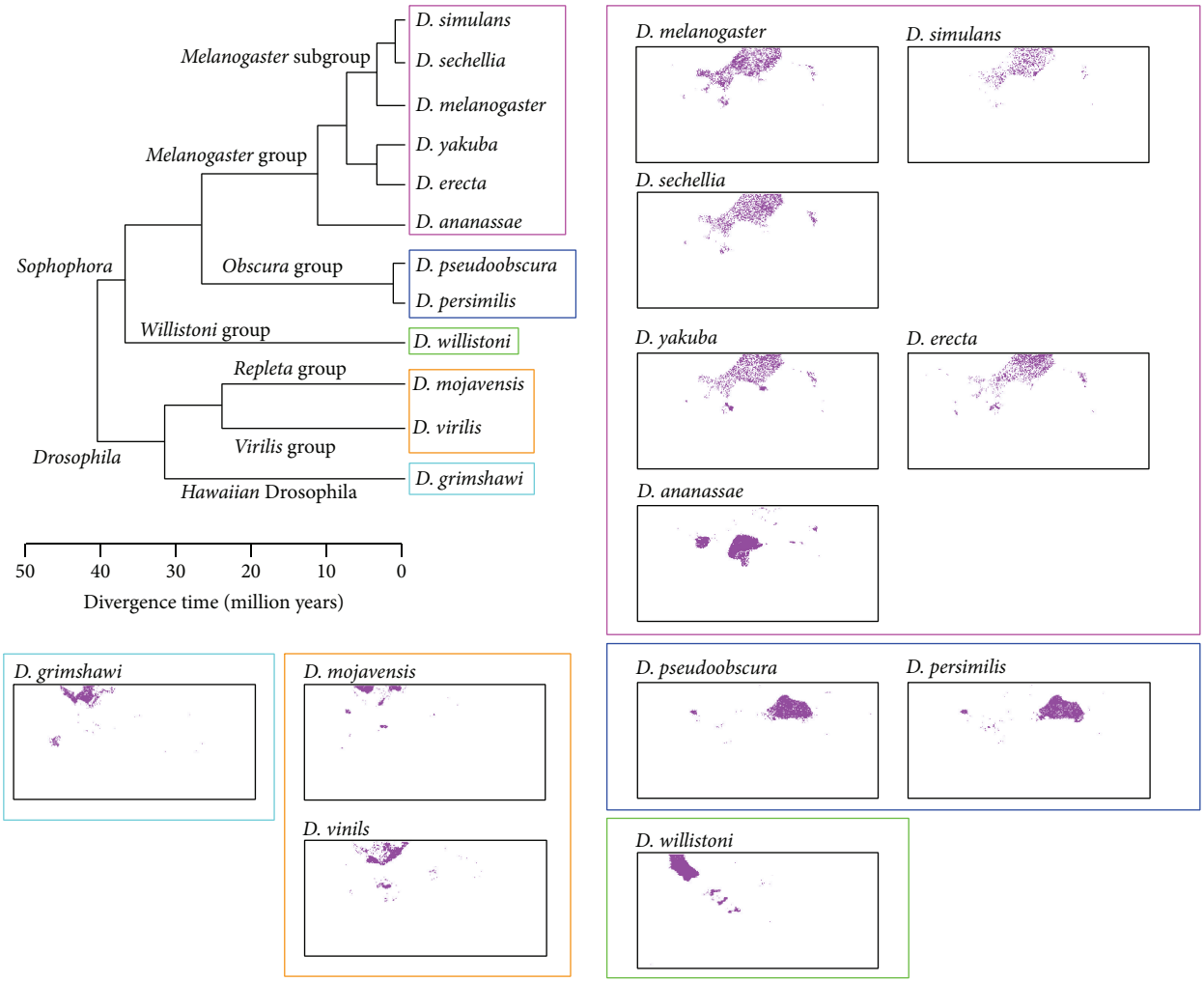


Figure 4: Distribution pattern for each of the 12 *Drosophila* genomes on 100 kb DegePenta-BLSOM listed in Figure 1. The phylogenetic tree for the 12 *Drosophila* species was obtained from Flybase [17]. The species belonging to one of the five *Drosophila* groups were separately specified by five boxes with different colors in the phylogenetic tree. A distribution pattern of nodes containing sequences derived from one *Drosophila* species was presented in the box marked with the color representing the respective *Drosophila* group, which was specified in the phylogenetic tree.

the mononucleotide composition in the respective genome and may differ among genomes and among genomic portions within a single genome. Actually, we have recently found that a pentanucleotide-BLSOM for the human genome can detect the characteristic enrichment of many transcription-factor-binding motifs in pericentric heterochromatin regions [21]. Functional signals, such as transcription-regulatory signals, are typically longer than pentanucleotides; therefore, analyses of longer oligonucleotides will become important. To conduct BLSOM with longer oligonucleotides, such as hexa- and heptanucleotides (4,096- and 16,384-dimensional data), for a large number of currently available genome sequences, large-scale computation using a high-performance supercomputer is essential, and the BLSOM algorithm is suitable for high-performance parallel computing.

For almost half of genes from the novel genomes sequenced, it has become clear that protein functions cannot always be estimated through sequence homology searching. We have applied BLSOM to protein sequence studies to analyze the frequency of oligopeptides and found the separation (self-organization) of proteins according to their functions [22]. This finding indicates that the BLSOM can be used for protein function estimation that does not rely on sequence homology searching and troublesome and confusing sequence alignment. Large-scale BLSOM analyses of a large amount and a wide variety of genome and protein sequences facilitate efficient extraction of fundamental information that supports research and development in a broad range of life sciences and industrial fields.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

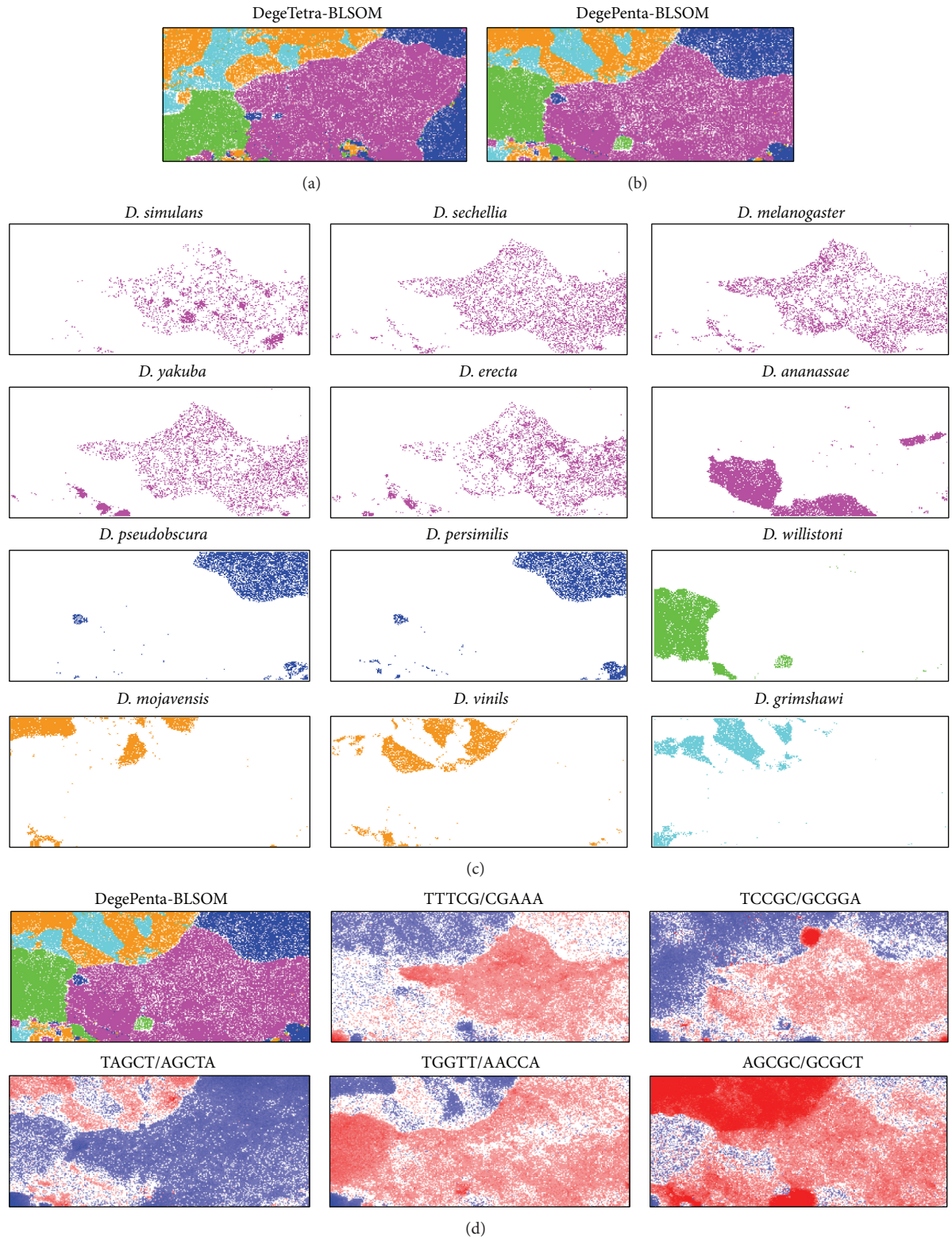


FIGURE 5: DegeTetra- (a) and DegePenta- (b) BLSOMs for the overlapping 100 kb sequences with a 10 kb sliding step derived from the 12 *Drosophila* genomes. Nodes containing sequences derived from genomes belonging to more than one group were indicated in black and those belonging to one group were indicated in colors, which were used to distinguish the boxes representing the five *Drosophila* groups. (c) The distribution pattern for each *Drosophila* genome on the DegePenta-BLSOM, which was listed in (b). Nodes were indicated in the color representing the group. (d) The level of each pair of complimentary pentanucleotides on the DegePenta-BLSOM listed in (b) was shown as described in Figure 2.

Acknowledgments

This work was supported by the Integrated Database Project and Grant-in-Aid for Scientific Research (C) and for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computation was done in part with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

[1] R. Nussinov, "Doublet frequencies in evolutionary distinct groups," *Nucleic Acids Research*, vol. 12, no. 3, pp. 1749–1763, 1984.

[2] G. J. Phillips, J. Arnold, and R. Ivarie, "Mono-through hexanucleotide composition of the *Escherichia coli* genome: a markov chain analysis," *Nucleic Acids Research*, vol. 15, no. 6, pp. 2611–2626, 1987.

[3] S. Karlin, J. Mrázek, and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *Journal of Bacteriology*, vol. 179, no. 12, pp. 3899–3913, 1997.

[4] S. Karlin, "Global dinucleotide signatures and analysis of genomic heterogeneity," *Current Opinion in Microbiology*, vol. 1, no. 5, pp. 598–610, 1998.

[5] S. Karlin, A. M. Campbell, and J. Mrázek, "Comparative DNA analysis across diverse genomes," *Annual Review of Genetics*, vol. 32, pp. 185–225, 1998.

[6] E. P. C. Rocha, A. Viari, and A. Danchin, "Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons," *Nucleic Acids Research*, vol. 26, no. 12, pp. 2971–2980, 1998.

[7] A. J. Gentles and S. Karlin, "Genome-scale compositional comparisons in Eukaryotes," *Genome Research*, vol. 11, no. 4, pp. 540–546, 2001.

[8] Bernardi, *Structural and Evolutionary Genomics: Natural Selection in Genome Evolution*, Elsevier, 2004.

[9] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[10] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[11] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceedings of the IEEE*, vol. 84, no. 10, pp. 1358–1383, 1996.

[12] S. Kanaya, M. Kinouchi, T. Abe et al., "Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome," *Gene*, vol. 276, no. 1-2, pp. 89–99, 2001.

[13] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for unveiling hidden genome signatures," *Genome Research*, vol. 13, no. 4, pp. 693–702, 2003.

[14] T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura, "Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes," *Gene*, vol. 365, no. 1-2, pp. 27–34, 2006.

[15] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples," *DNA Research*, vol. 12, no. 5, pp. 281–290, 2005.

[16] R. Nakao, T. Abe, A. M. Nijhof et al., "A novel approach, based on BLSOMs (Batch Learning Self-Organizing Maps), to the microbiome analysis of ticks," *ISME Journal*, vol. 7, no. 5, pp. 1003–1015, 2013.

[17] S. Tweedie, M. Ashburner, K. Falls et al., "FlyBase: enhancing *Drosophila* gene ontology annotations," *Nucleic Acids Research*, vol. 37, no. 1, pp. D555–D559, 2009.

[18] Drosophila 12 Genomes Consortium, "Evolution of genes and genomes on the *Drosophila* phaylogeny," *Nature*, vol. 450, no. 7167, pp. 203–218, 2007.

[19] A. Startk, M. F. Lin, P. Kheradpour et al., "Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures," *Nature*, vol. 450, no. 7167, pp. 219–232, 2007.

[20] M. W. Hahn, M. V. Han, and S.-G. Han, "Gene family evolution across 12 *Drosophila* genomes," *PLoS genetics*, vol. 3, no. 11, p. e197, 2007.

[21] Y. Iwasaki, K. Wada, Y. Wada, T. Abe, and T. Ikemura, "Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance," *Chromosome Research*, vol. 21, pp. 461–474, 2013.

[22] T. Abe, S. Kanaya, H. Uehara, and T. Ikemura, "A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses," *DNA Research*, vol. 16, no. 5, pp. 287–298, 2009.