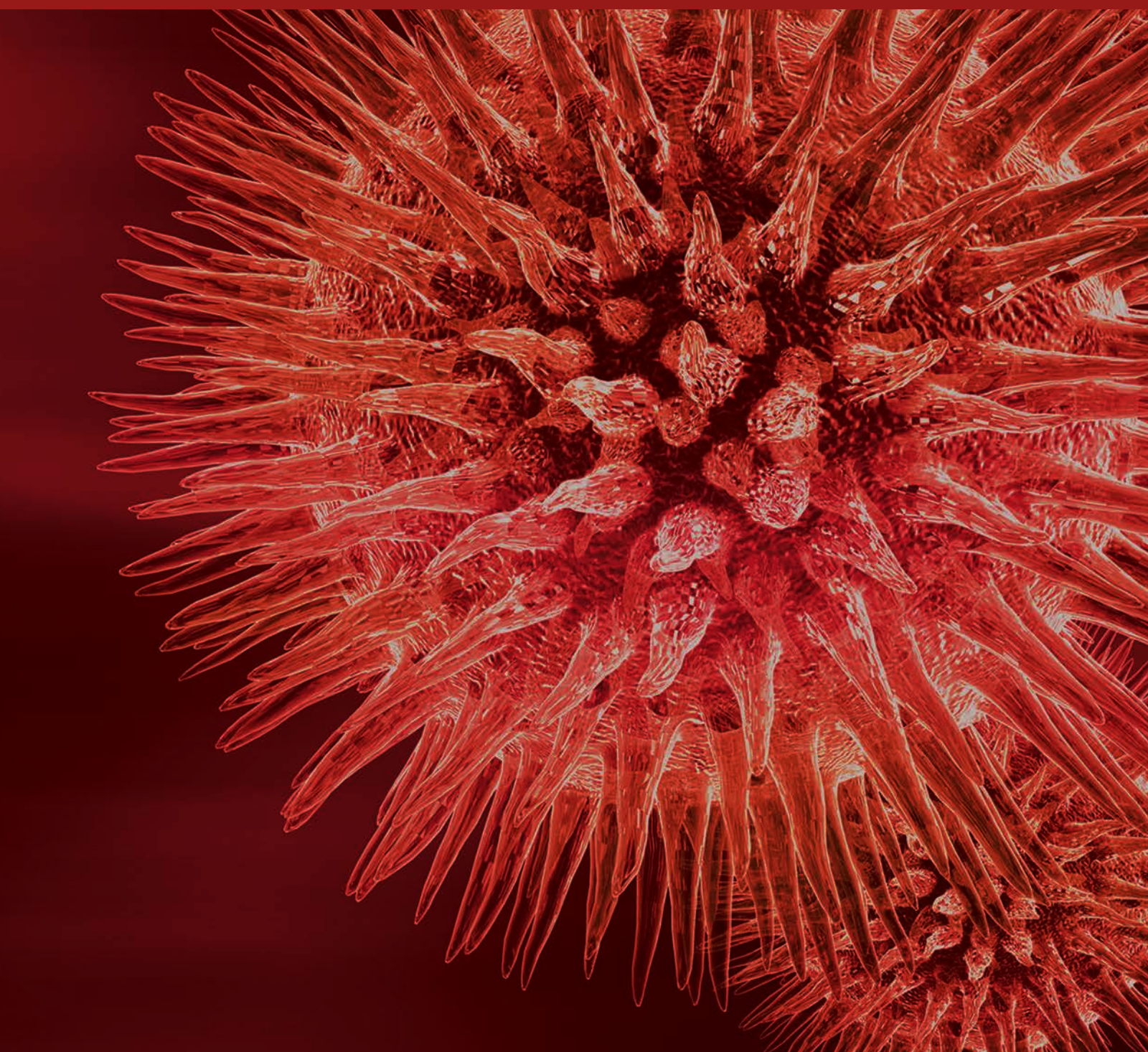
BioMed Research International

Big Data and Network Biology 2016

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



Big Data and Network Biology 2016

BioMed Research International

Big Data and Network Biology 2016

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

Copyright © 2017 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "BioMed Research International." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 9432460, 2 pages
http://dx.doi.org/10.1155/2017/9432460

Editorial
Big Data and Network Biology 2016

Shigehiko Kanaya,[1] Md. Altaf-Ul-Amin,[1] Samuel K. 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, West Java 16680, Indonesia*

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

Received 22 December 2016; Accepted 26 December 2016; Published 26 January 2017

Copyright © 2017 Shigehiko Kanaya et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Big data science towards biology has emerged as an important research and application field. The scope of big data bioinformatics and the need for integrated systems biology applications have posed significant challenges to computing systems. With the growing deluge of molecular biological data, big data methodologies, including data mining algorithms and processing techniques, are becoming particularly relevant. These encompass data accumulation, storage, retrieval, classification, and visualization. These have become increasingly important themes of research in the past decade. Network analysis has become popular for systems level analysis of biological phenomena. Both micro- and macrolevel biological networks of various types facilitate the development of theory and methodology for solving core scientific issues. The application ranges from studies of ecological structure, biodiversity and environment, and evolution and extinction of species. It is also more widely used in studies on metabolic regulation and biomarker identification particularly for noncommunicable diseases such as cancer, Alzheimer's disease, and diabetes. The versatility of these topics, theory and methodology of big data and network biology, justifies the aims and scope of this special issues. The present special issue is however not an exhaustive representation of these topics.

This special issue contains six papers. One article presents a biomedical text mining approach and another article talks about algorithms for inferring gene regulatory networks. Four other papers present methodology related to genomics, transcriptomics, metabolomics, and drug repositioning.

The paper "Novel Approach to Classify Plants Based on Metabolite-Content Similarity" proposed an unsupervised approach to classify plants based on their known metabolite content data. Plants were classified based on structurally similar metabolite groups to reduce the influence of incomplete data. The resulting plant clusters were found to be consistent with known evolutional relations of plants and reveal the significance of metabolite content as a taxonomic marker.

The paper "A Systematic Framework for Drug Repositioning from Integrated Omics and Drug Phenotype Profiles Using Pathway-Drug Network" proposes a systematic framework that employs experimental genomic knowledge and pharmaceutical knowledge to reposition drugs for a specific disease. The experimental results showed that the proposed framework is a useful approach to discover promising candidates for breast cancer treatment.

The paper "MapReduce Algorithms for Inferring Gene Regulatory Networks from Time-Series Microarray Data Using an Information-Theoretic Approach" proposes new MapReduce algorithms for inferring gene regulatory networks (GRNs) on a Hadoop cluster in a cloud environment. These algorithms employ an information-theoretic approach to infer GRNs using time-series gene expression profiles. Experimental results show that their MapReduce program is much faster and achieves slightly better prediction accuracy than a state-of-the-art R program.

The paper "Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism" introduces one of a series of methods for correlation-based network

generation and analysis using freely available software. The pipeline used published metabolomics data of a population of human breast carcinoma cell lines MDA-MB-231 under normal and hypoxia conditions. The analysis revealed significant differences between the metabolic networks in response to the tested conditions.

The paper "Semisupervised Learning Based Disease-Symptom and Symptom-Therapeutic Substance Relation Extraction from Biomedical Literature" presents a method of constructing two models for extracting the relations between the disease and symptom, and symptom and therapeutic substance from biomedical texts, respectively. The authors apply two semisupervised learning algorithms, Co-Training and Tri-Training, to boost the relation extraction performance.

Horizontal gene transfer (HGT) has had an important role in eukaryotic genome evolution. The paper "Horizontally Transferred Genetic Elements in the Tsetse Fly Genome: An Alignment-Free Clustering Approach Using Batch Learning Self-Organising Map (BLSOM)" employs BLSOM to explore the genome of *Glossina morsitans* for evidence of HGT from microorganisms. The predicted donors of HGT candidate include diverse bacteria that have not previously been associated with the tsetse fly. These findings provide a basis for understanding the coevolutionary history of the tsetse fly and its microbes and establish the effectiveness of BLSOM for the detection of HGT.

Acknowledgments

We thank the authors of the articles in this special issue for their 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 which helped a lot in the improvement of the manuscripts.

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

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 6261802, 8 pages
http://dx.doi.org/10.1155/2017/6261802

Research Article

MapReduce Algorithms for Inferring Gene Regulatory Networks from Time-Series Microarray Data Using an Information-Theoretic Approach

Yasser Abduallah,[1] Turki Turki,[2] Kevin Byron,[1,3] Zongxuan Du,[1] Miguel Cervantes-Cervantes,[4] and Jason T. L. Wang[1,3]

[1]*Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102, USA*
[2]*Computer Science Department, King Abdulaziz University, P.O. Box 80221, Jeddah 21589, Saudi Arabia*
[3]*Bioinformatics Program, New Jersey Institute of Technology, Newark, NJ 07102, USA*
[4]*Department of Biological Sciences, Rutgers University, Newark, NJ 07102, USA*

Correspondence should be addressed to Turki Turki; tturki@kau.edu.sa and Jason T. L. Wang; wangj@njit.edu

Received 7 June 2016; Revised 14 November 2016; Accepted 13 December 2016; Published 22 January 2017

Academic Editor: Farit M. Afendi

Copyright © 2017 Yasser Abduallah et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Gene regulation is a series of processes that control gene expression and its extent. The connections among genes and their regulatory molecules, usually transcription factors, and a descriptive model of such connections are known as gene regulatory networks (GRNs). Elucidating GRNs is crucial to understand the inner workings of the cell and the complexity of gene interactions. To date, numerous algorithms have been developed to infer gene regulatory networks. However, as the number of identified genes increases and the complexity of their interactions is uncovered, networks and their regulatory mechanisms become cumbersome to test. Furthermore, prodding through experimental results requires an enormous amount of computation, resulting in slow data processing. Therefore, new approaches are needed to expeditiously analyze copious amounts of experimental data resulting from cellular GRNs. To meet this need, cloud computing is promising as reported in the literature. Here, we propose new MapReduce algorithms for inferring gene regulatory networks on a Hadoop cluster in a cloud environment. These algorithms employ an information-theoretic approach to infer GRNs using time-series microarray data. Experimental results show that our MapReduce program is much faster than an existing tool while achieving slightly better prediction accuracy than the existing tool.

1. Introduction

Current biotechnology has allowed researchers in various fields to obtain immense amounts of experimental information, ranging from macromolecular sequences and gene expression data to proteomics and metabolomics. In addition to large-scale genomic information obtained through such methods as third-generation DNA sequencing, newer technology, such as RNA-seq and ChIP-seq, has allowed researchers to fine-tune the analysis of gene expression patterns [1–3]. More information on interactions between transcription factors and DNA, both qualitative and quantitative, is increasingly emerging from microarray data [4–6]. Although microarrays alone do not provide direct evidence of functional connections among genes, the attachment of transcription factors (TFs) and their binding sites (TFBSs), located at specific gene promoters, influences transcription and modulates RNA production from a particular gene, thus establishing a first level of functional interaction. Since the TFs are gene-encoded polypeptides and the target TFBSs belong to different genes, analyses of TFs-TFBSs interactions could reveal gene networks and may even contribute to the elucidation of unknown GRNs [7]. Besides contributing to inferring and understanding these interactions, determining GRNs also provides models of such connections [8]. GRNs could be the basis to infer more complex networks, encompassing gene, protein, and metabolic spaces, as well as

the entangled and often overlooked signaling pathways that interconnect them [9–13].

The core GRN apparatus consists of the sum of *cis*-regulatory modular DNA sequence elements that interact with TFs. These sequences read and process information incoming from the cell, transducing it into the formation of gene products while modulating their abundance [14]. To make them easier to understand, GRN models must be genome-oriented and viewable at different levels, from global patterns of gene expression, down to *cis*-regulatory DNA and nucleotide sequences [15].

Interactions among genes are mediated by gene products such as DNA-binding proteins (including TFs) and miRNAs. The analyses of gene interactions can be difficult if time-series data are part of the experimental design [16]. Analysis of genes, gene products, and metabolism (the Three Spaces of gene networks) would require additional computing resources. Among the previously ignored components of gene networking are miRNAs [17, 18]. In addition to their importance as regulatory elements in gene expression, the capacity of miRNAs to be transported from cell to cell implicates them in a panoply of pathophysiological processes that include antiviral defense, tumorigenesis, lipometabolism, and glucose metabolism [16]. This role in disease complicates our understanding of translational regulation via endogenous miRNAs. In addition, miRNAs seem to be present in different types of foods [19] with potential implications in human health and disease. Understanding the biogenesis, transport, and mechanisms of action of miRNAs on their target RNA would result in possible therapies, requiring large amounts of computational power, which can be attained by cloud computing and process parallelizing.

Detailed experimental analysis of several functional regulatory elements has revealed that they consist of dense clusters of unique, short DNA sequences specifically recognized by a range of TFs. Biochemically, protein-binding to these sequences controls the regulatory output of the clusters and, from an informational perspective, clustered specific target sites determine the type of regulatory outcome and the cellular functions that will be performed. GRNs are encoded in the DNA and can be thought of as a sequence-dependent regulatory genome, given that TFs recognize specific short DNA motifs. The small length of these motifs means that they will occur frequently but randomly within the enormity of the total genome of a particular organism [20–22]. Therefore, to parse functional regulatory elements using bioinformatics requires the analysis of copious amounts of genomic data.

Analyses of time-series data from microarrays can show the chronological expression of specific genes or groups of genes. These temporal patterns can be used to infer or propose causal relationships in gene regulation [23]. Thus, genes in logical networks can modulate the extent of each other's gene expression over the life span of a cell or a whole organism. Time-series microarray data shed light on a complex but measurable regulatory system, allowing for a more precise inference of gene interaction.

Numerous algorithms have been developed for inferring GRNs [24–27]. In this paper, we present a new approach, tailored to cloud computing, to infer GRNs using time-series microarray data. Using time as a variable in the analysis of GRNs permits the study of changes in cellular phenotype, as opposed to a snapshot in a limited time frame that may reveal static interactions but not progression of gene expression phenomena. The time-series datasets used in this work come from DREAM4 challenges [28–31] and an Affymetrix Yeast Genome 2.0 Array downloaded from NCBI's Gene Expression Omnibus. The array contains 5,744 probe sets and includes 10,928 *Saccharomyces cerevisiae* genes with 49 time points and transcriptional oscillations of about 4 hours. These oscillations reveal cell redox states, which in turn result from changes in metabolic fluxes and cell cycle phases [32].

As knowledge in several biological fields leads to an ever-expanding accumulation in gene expression data, the main consideration in data processing is that analysis of information becomes increasingly time-consuming, thus creating a demand to speed up the analytical process. In order to obtain results more expeditiously, we develop information-theoretic algorithms using MapReduce that run on a distributed, multinode Apache Hadoop cluster in a cloud environment. Cloud resources are increasingly more flexible and affordable compared with local traditional computing resources. Cloud computing advantages in the field of bioinformatics research are well known [33–39].

2. Materials and Methods

2.1. Framework. Previous information-theoretic algorithms for network inference were implemented in the R programming language using steady state data [40] and time-series data [23]. The tool using steady state data is named ARACNE [40] and the tool using time-series data is named TimeDelay-ARACNE [23]. ARACNE infers an undirected network, which basically shows whether two genes are mutually dependent rather than the regulatory relationship between the genes. In contrast, TimeDelay-ARACNE infers a directed network, in which an edge from gene A to gene B indicates that A regulates the expression of B.

In contrast to the R-based ARACNE and TimeDelay-ARACNE, our proposed information-theoretic framework is tailored to the MapReduce programming paradigm. Like TimeDelay-ARACNE [23], the input of our framework is a set of time-series gene expression data and the output is an inferred gene regulatory network. The input dataset contains m genes, and each gene has n expression values recorded at n different time points, respectively. Our framework consists of three steps. Step 1 aims to detect, for each gene g, the first time point t ($t > 1$) at which a substantial change in the gene expression of g with respect to the gene expression of g at time point 1 takes place. This t is referred to as the time point of Substantial Change of Expression (ScE) of gene g, denoted as ScE(g). Step 2 calculates, for two genes g_x and g_y, the influence of g_x on g_y, denoted as influence(g_x, g_y), based on the ScE values of the genes. Step 3 determines the edges between genes using their influence values. Below we present details of the proposed framework.

Step 1 (calculation of ScE). Let $g(t)$ be the expression value of gene g at time point t. We say g is activated (or induced)

at time point t ($t > 1$) if $g(t)/g(1) > \tau$, where $\tau > 1$ is a threshold. We say g is inhibited (or repressed) at time point t ($t > 1$) if $g(t)/g(1) < 1/\tau$. For each gene g, we maintain two sets of time points: $g^+(t)$ and $g^-(t)$; $g^+(t)$ contains all time points at which g is induced and $g^-(t)$ contains all time points at which g is repressed. Initially, $g^+(t) = \varnothing$ and $g^-(t) = \varnothing$. The two sets of time points are then updated as follows. For each time point t ($t > 1$),

$$\begin{aligned} &\text{if } \frac{g(t)}{g(1)} > \tau, \quad \text{then } g^+(t) = g^+(t) \cup \{t\}, \\ &\text{if } \frac{g(t)}{g(1)} < \frac{1}{\tau}, \quad \text{then } g^-(t) = g^-(t) \cup \{t\}. \end{aligned} \tag{1}$$

If $1/\tau \leq g(t)/g(1) \leq \tau$, then g is neither induced nor repressed at time point t. In this case, we simply ignore this time point t without adding t to $g^+(t)$ or $g^-(t)$. The value of τ used in this study is set to 1.2. With this threshold value and datasets used in the study (DREAM4 [28–31]), there is a significant difference between the mean of the gene expression values of the time points at which g is induced and the mean of the gene expression values of the time points at which g is repressed according to Student's t-test ($p < 0.05$).

Let ScE(g) represent the first time point t ($t > 1$) at which g is either induced or repressed; that is,

$$\text{ScE}(g) = \min\{g^+(t) \cup g^-(t)\}. \tag{2}$$

For any two genes g_a and g_b, there are three cases to be considered.

Case 1 (ScE(g_a) < ScE(g_b)). We send the ordered pair (g_a, g_b) and the expression values of the two genes to Step 2.

Case 2 (ScE(g_b) < ScE(g_a)). We send the ordered pair (g_b, g_a) and the expression values of the two genes to Step 2.

Case 3 (ScE(g_a) = ScE(g_b)). We send g_a and g_b with a tag indicating that both of the ordered pairs (g_a, g_b) and (g_b, g_a) should be considered, together with their gene expression values, to Step 2.

Step 2 (calculation of influence values). For each pair of genes (g_x, g_y) received from Step 1, we calculate the time-delayed mutual information [41] between the genes as follows:

$$I^k\left(g_x, g_y^{(k)}\right) = \sum_{1 \leq i \leq n-k} p\left(g_x^i, g_y^{i+k}\right) \log \frac{p\left(g_x^i, g_y^{i+k}\right)}{p\left(g_x^i\right) p\left(g_y^{i+k}\right)}, \tag{3}$$

where n is the total number of time points, $p(g_x^i)$ is the marginal distribution of g_x^i, and $p(g_x^i, g_y^{i+k})$ is the joint distribution of g_x^i and g_y^{i+k}. (In our implementation, a hash table is used to calculate the joint distribution to save time and space.) The parameter k, $1 \leq k \leq h$, represents the length of delayed time and h is the maximum length of delayed time. (In the study presented here, h is set to 3.) The notation g_x^i denotes the gene expression of g_x at time point i and g_y^{i+k} is the gene expression of g_y at time point $i+k$. Figure 1 illustrates how to calculate time-delayed mutual information. There are 21 time points in Figure 1. The length of delayed time is 2 (i.e., $k = 2$). Each rectangle represents the gene expression value obtained at some time point. Mutual information of rectangles with the same color is computed. Then, the influence of g_x on g_y is calculated as follows:

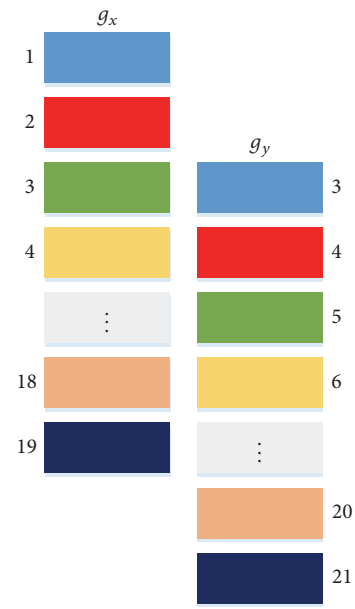


Figure 1: Illustration of how to calculate time-delayed mutual information.

$$\text{influence}\left(g_x, g_y\right) = \max_{1 \leq k \leq h} \left\{I^k\left(g_x, g_y^{(k)}\right)\right\}. \tag{4}$$

Referring to the three cases in Step 1, for Case 1, we calculate influence(g_a, g_b) and send (g_a, g_b) and influence(g_a, g_b) to Step 3. For Case 2, we calculate influence(g_b, g_a) and send (g_b, g_a) and influence(g_b, g_a) to Step 3. For Case 3, if influence(g_a, g_b) ≥ influence(g_b, g_a), then we send (g_a, g_b) and influence(g_a, g_b) to Step 3; otherwise, we send (g_b, g_a) and influence(g_b, g_a) to Step 3.

Step 3 (determination of edges between genes). Let ε be a threshold. For each pair (g_x, g_y) received from Step 2, if influence(g_x, g_y) > ε, then we create an edge from g_x to g_y indicating that g_x substantially influences g_y or g_x regulates the expression of g_y; that is, there is a predicted present edge from g_x to g_y. If influence(g_x, g_y) ≤ ε, then we do not create an edge from g_x to g_y; that is, there is a predicted absent edge from g_x to g_y. With the predicted present and absent edges, we are able to infer or reconstruct a gene regulatory network. The value of ε used in this study is set to 0.96. With this threshold value and datasets used in the study (DREAM4 [28–31]), there is a significant difference between the mean of the influence values of the predicted present edges and the mean of the influence values of the predicted absent edges according to Student's t-test ($p < 0.05$).

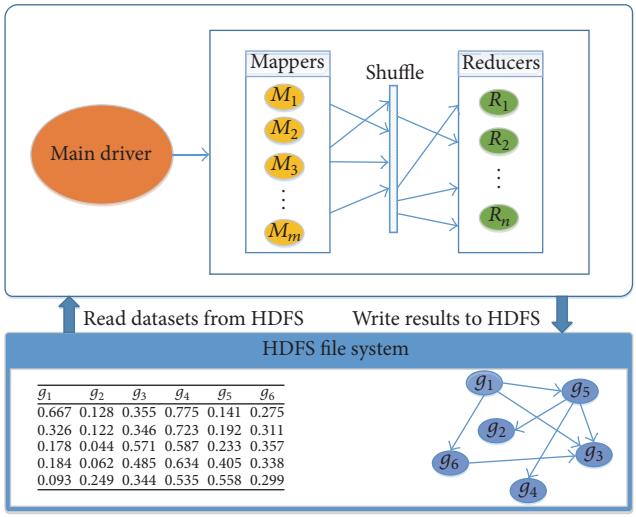


FIGURE 2: Conceptual description of the Hadoop MapReduce implementation of our proposed algorithms.

2.2. MapReduce Algorithms. Figure 2 presents a high-level conceptual description of the Hadoop MapReduce implementation of our proposed framework. The system includes a driver and one or more mappers and reducers. The driver takes the input from the user, including the locations of input and output files, as well as algorithm parameters and thresholds. The driver prepares a job with the required configuration, sends the job to Hadoop to start it, and calculates the time that the job takes to be completed. The mappers are user-defined programs (UDPs), which prepare data and perform calculations, if needed, and then send the processed data (key-value pairs) to the reducers. The reducers are also UDPs, which perform the final processing and write the results into the output file. Hadoop optimizes the number of mappers for a job. The user can control the number of reducers needed for completing the job. The Hadoop distributed file system (HDFS) is a global repository for storage of the input flat file (in plain text format) with gene expression data and the output file with an inferred gene regulatory network.

Each gene has an identifier. (We use g_x to represent both a gene and its identifier when the context is clear.) Each line in the input file contains a pair of genes and their expression values. Genes are sorted based on their identifiers. Each pair of genes g_x, g_y occurs in the input file only once; specifically, the gene pair in which the identifier of g_x is less than the identifier of g_y occurs in the input file. Suppose there are m genes. There are $(m \times m - 1)/2$ lines in the input file.

We develop four MapReduce algorithms, named M0, M1, M2, and M3, respectively. These algorithms differ in which steps, as described in Section 2.1, are performed by mappers.

Algorithm M0. In this algorithm, mappers perform zero steps. Reducers have to do Steps 1, 2, and 3. In the key-value pairs transmitted between mappers and reducers, the key is a pair of genes, and the value contains the expression profiles of the genes.

Algorithm M1. In this algorithm, mappers perform Step 1. Reducers have to do Steps 2 and 3. In the key-value pairs transmitted between mappers and reducers, the key is a pair of genes and the value contains the expression profiles of the genes. Part of the value is a tag indicating which case in Step 1 applies to the pair of genes.

Algorithm M2. In this algorithm, mappers have to do Steps 1 and 2. Reducers perform Step 3. In the key-value pairs transmitted between mappers and reducers, the key is an ordered pair of genes (g_x, g_y) and the value is influence(g_x, g_y).

Algorithm M3. In this algorithm, mappers have to do Steps 1, 2, and 3. Reducers perform zero steps. In the key-value pairs transmitted between mappers and reducers, the key is the edge $g_x \rightarrow g_y$ and the value is the influence of g_x on g_y that exceeds the threshold ε.

The time needed by mappers is bounded by $O(m^2/M)$ and the time needed by reducers is bounded by $O(m^2/R)$, where m is the number of genes, M is the number of mappers, and R is the number of reducers. Thus, the time complexity of our MapReduce algorithms is $O(m^2/M) + O(m^2/R)$. Note that this is a very pessimistic upper bound since reducers often work in parallel with mappers, and hence the actual time needed by the algorithms is much less. Note also that, in practice, $M > R$, and hence the time complexity of our algorithms is bounded by $O(m^2/R)$.

3. Results and Discussion

3.1. Experimental Results. The four algorithms described in Section 2.2 were implemented in MapReduce and Java on a Hadoop infrastructure (cloud), which is a virtual environment based on VMware Big Data Extensions (BDE). The infrastructure hardware cluster associated with BDE is comprised of two IBM iDataPlex dx360 M4 servers. Each dx360 M4 server is comprised of two Intel Xeon 2.7 GHz E5-2680 (8-Core) CPUs for a total of 16 cores per server. With the enabling of hyperthreading, the number of logical processors is doubled to provide 32 logical processors per server. Each server has 128 GB RAM.

The dataset used in the experiments was GSE30052 [32], downloaded from the Gene Expression Omnibus (GEO) at http://www.ncbi.nlm.nih.gov/geo/. This dataset was created using an Affymetrix Yeast Genome 2.0 Array containing 5,744 probe sets for *S. cerevisiae* gene expression analysis. The dataset contains 10,928 genes with 49 time points. The dataset is split into key-value pairs as described in Section 2.2 and the input file has $(10{,}928 \times 10{,}927)/2$ lines, taking up 26.8 GB of disk space. Hadoop assigns 254 mappers to this dataset. The default value for the number of reducers is set to 20.

We divided GSE30052 into smaller datasets that were subsets of GSE30052 with varying numbers of genes. Figure 3 compares the running times of the four MapReduce algorithms described in Section 2.2 for varying dataset sizes. It can be seen from the figure that all the four algorithms scale well when the dataset size becomes large (i.e., the number of genes increases). Algorithm M2 performs the best. This happens because, with M2, the mappers, working

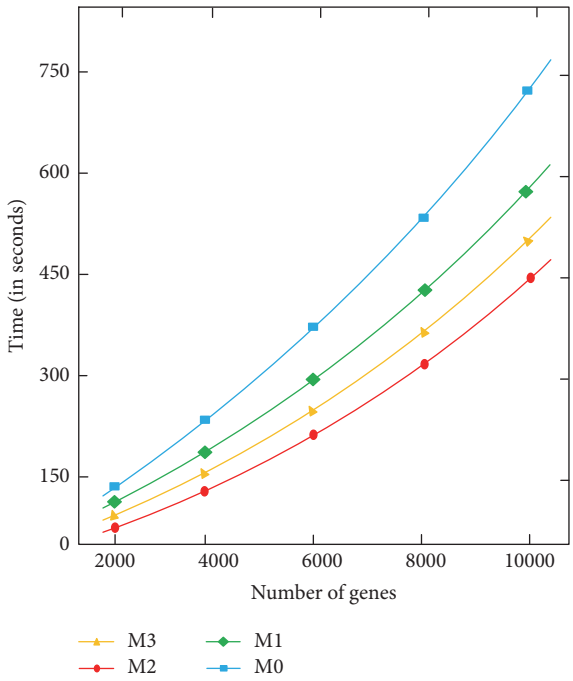


Figure 3: Performance comparison of the four MapReduce algorithms.

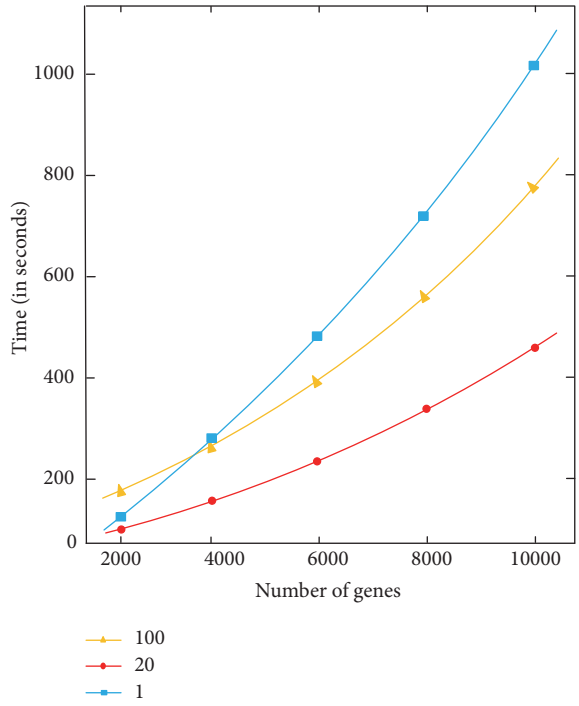


Figure 4: Effect of the number of reducers on the performance of the M2 algorithm.

in parallel, share some workload with the reducers, which perform a relatively smaller amount of computation while writing results into the output file. It is worth noting that M2 is better than M3, in which the mappers have to do all the computation. Algorithm M0 performs the worst. With M0, the reducers have to do all the calculations and become too busy to quickly complete the job.

We then fixed the algorithm and used M2 in all subsequent experiments. Figure 4 shows running times of the M2 algorithm with 1, 20, and 100 reducers, respectively. It can be seen that the optimal number of reducers is 20. With this configuration, the reducers work at their maximum limit. When there are too many (e.g., 100) reducers, the overhead is too large, and as a consequence the system is slowed down. On the other hand, when only one reducer is employed, the reducer is overloaded and the overall performance of the system degrades.

Finally, we conducted experiments to compare the MapReduce implementation of the M2 algorithm running on the Hadoop cluster (denoted MRC), the MapReduce implementation of the M2 algorithm running on a standalone single-node server (denoted as MRS), the Java implementation of the M2 algorithm running on a single-node server, and the R implementation of the related time-delayed mutual information algorithm, TimeDelay-ARACNE [23]. In Figure 5, it can be seen that MRC is highly scalable and that it outperforms the other three programs. Notably, due to Hadoop's overhead, MRS is even slower than the Java program. The R program is not scalable; its running time dramatically increases as the dataset becomes large.

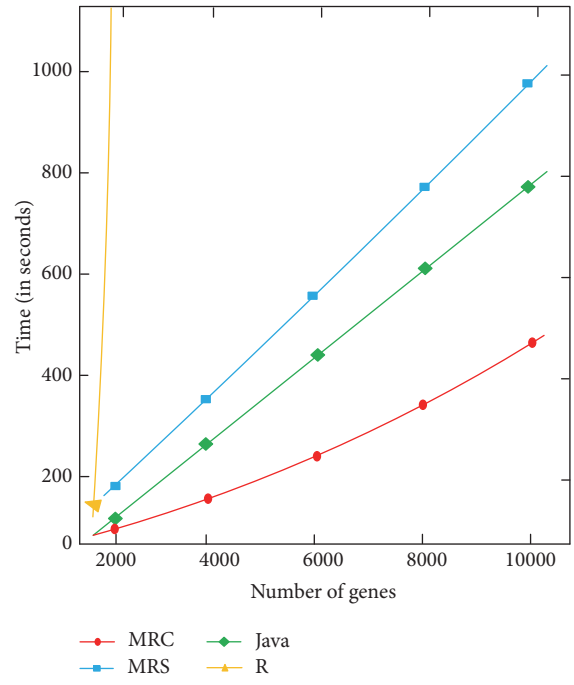


Figure 5: Performance comparison of the MapReduce algorithm (M2) and related programs.

3.2. Discussion. Our information-theoretic algorithms for inferring gene regulatory networks are implemented in MapReduce and run on a Hadoop cluster. A tool that is closely related to our work is the TimeDelay-ARACNE program in R [23], which also infers gene regulatory networks from time-series gene expression profiles using an information-theoretic

approach. As shown in Figure 5, the TimeDelay-ARACNE program in R does not scale well whereas our MapReduce program is highly scalable when running on the Hadoop cluster. Furthermore, our MapReduce program differs from the TimeDelay-ARACNE R program in that our algorithm is deterministic whereas the R program is implemented based on a nondeterministic algorithm, specifically Markov random fields. For the same dataset and parameter values, the R program produces different results in different executions. In contrast, our MapReduce program always produces the same result, in different executions, for the same dataset.

To evaluate the accuracy of these programs, we adopted the five time-series gene expression datasets available in the DREAM4 100-gene in silico network inference challenge [28–31]. Each dataset contains 10 times series, where each time series has 21 time points, for 100 genes. Thus, in each time series, each gene has 21 gene expression values. Totally, there are 50 time series in the five datasets. Each time-series dataset is associated with a gold standard file, where the gold standard represents the ground truth of the network structure for the time-series data. Each edge in the gold standard represents a true regulatory relationship between two genes. Our experimental results showed that the average accuracy of TimeDelay-ARACNE on the five datasets is approximately 92.4%. The average accuracy of our deterministic algorithm (M2) is about 93.1%, which is slightly better than TimeDelay-ARACNE. It was observed that, in a dataset, when M2 is better than TimeDelay-ARACNE, it has a higher accuracy than TimeDelay-ARACNE for every execution of TimeDelay-ARACNE on the dataset. Furthermore, our MapReduce program (M2) takes, on average, 20 seconds to infer a gene regulatory network on a DREAM4 dataset while the average time used by the TimeDelay-ARACNE R program is 3,500 seconds.

We also tested different values for the parameters τ, ε, and the maximum length of delayed time, h, used in the proposed algorithms. Experimental results showed that the default values for these parameters ($\tau = 1.2$, $\varepsilon = 0.96$, and $h = 3$) achieve the highest accuracy. When compared with other parameter values (e.g., the maximum length of delayed time $h = 6, \varepsilon = 0.35$, or $\tau = 2$), the accuracy achieved by the default parameter values is significantly higher than the accuracy achieved by the other parameter values according to Wilcoxon signed rank tests [42] ($p < 0.05$).

4. Conclusions

We have presented four MapReduce algorithms for reconstructing gene regulatory networks from time-series microarray data using an information-theoretic approach. Our experimental results showed that the algorithm (M2) that uses mappers to perform a large portion of work and reducers to perform a relatively small amount of computation achieves the best performance. This M2 algorithm is faster than an algorithm in which the mappers have to do all the computation. Moreover, the M2 algorithm is much faster than another algorithm in which the reducers have to do all the computation and become too busy to quickly complete the job.

When tested on DREAM4 datasets with 100 genes in each dataset, our MapReduce program (M2) is slightly better than a closely related R program (TimeDelay-ARACNE [23]) in terms of accuracy; furthermore, our MapReduce program is much faster than the existing R program. When tested on a big dataset (GSE30052 [32]) with 10,928 genes, our MapReduce program was found to be highly scalable whereas the R program was not (cf. Figure 5). It should be pointed out, however, that the comparison with the TimeDelay-ARACNE R program is not completely fair. Our MapReduce program is based on a parallel algorithm whereas the TimeDelay-ARACNE R program is based on a sequential algorithm. Further study would be needed to investigate parallel versions of the R program or a new TimeDelay-ARACNE R package that supports parallelization.

The work presented here shows that distributing highly parallel tasks in a cloud environment achieves higher performance than running the tasks in a standalone or noncloud environment. In general, cloud computing can provide the power to integrate the ever-increasing information about the Three Spaces of gene networks [8] as well as the multipronged signal transduction pathways traversing these spaces. Comprehending systems biology and functional genomics could eventually contribute to a better grasp of organismal physiology. Thus, the cloud would provide computing power that is needed as the analysis of multilevel processes becomes more complicated. Cloud computing will enable genome-scale network inference as demonstrated in this study.

Epigenetics [43] is an emerging aspect of gene regulation whose study would require enormous computing capacity. This type of posttranslational regulation cross talk involves chemical modifications of DNA and histones in a process known as chromatin remodeling. The role of genetic and epigenetic networks in a variety of health conditions is now coming into view. For example, there are at least 450 different genes associated with intellectual disability and related cognitive disorders. Some of these genes are involved in synaptic plasticity and cell signaling whereas others are epigenetic genes involved in chromatin modifications [44]. Analysis of the interactions across these genes and networks, as well as finding new mutations, will require the development of highly expeditious bioinformatics tools to mine the anticipated high amounts of data.

Genome-scale metabolic models are becoming essential in biomedical applications, and researchers are moving towards building such models [45]. MapReduce algorithms could become a powerful tool in the analyses of all aspects of gene networking in the Three Spaces paradigm. In general, cloud computing could facilitate the handling of the vast amounts of information (big data) that such analyses require.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this article.

Acknowledgments

The authors thank N. Patel and L. Zhong for useful conversations concerning gene network inference.

References

[1] E. E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing," *Human Molecular Genetics*, vol. 19, no. 2, pp. R227–R240, 2010.

[2] C. Angelini and V. Costa, "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems," *Frontiers in Cell and Developmental Biology*, vol. 2, article no. 51, 2014.

[3] J. Qin, Y. Hu, F. Xu, H. K. Yalamanchili, and J. Wang, "Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods," *Methods*, vol. 67, no. 3, pp. 294–303, 2014.

[4] J. B. Brown and S. E. Celniker, "Lessons from modENCODE," *Annual Review of Genomics and Human Genetics*, vol. 16, pp. 31–53, 2015.

[5] V. Filkov, S. Skiena, and J. Zhi, "Analysis techniques for microarray time-series data," *Journal of Computational Biology*, vol. 9, no. 2, pp. 317–330, 2002.

[6] F. Zhu, L. Shi, J. D. Engel, and Y. Guan, "Regulatory network inferred using expression data of small sample size: application and validation in erythroid system," *Bioinformatics*, vol. 31, no. 15, pp. 2537–2544, 2015.

[7] T. Werner, S. M. Dombrowski, C. Zgheib et al., "Elucidating functional context within microarray data by integrated transcription factor-focused gene-interaction and regulatory network analysis," *European Cytokine Network*, vol. 24, no. 2, pp. 75–90, 2013.

[8] P. Brazhnik, A. De La Fuente, and P. Mendes, "Gene networks: how to put the function in genomics," *Trends in Biotechnology*, vol. 20, no. 11, pp. 467–472, 2002.

[9] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, article 565, 2012.

[10] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano, "Reverse engineering cellular networks," *Nature Protocols*, vol. 1, no. 2, pp. 662–671, 2006.

[11] N. Kibinge, N. Ono, M. Horie et al., "Integrated pathway-based transcription regulation network mining and visualization based on gene expression profiles," *Journal of Biomedical Informatics*, vol. 61, pp. 194–202, 2016.

[12] M. Altaf-Ul-Amin, F. M. Afendi, S. K. Kiboi, and S. Kanaya, "Systems biology in the context of big data and networks," *BioMed Research International*, vol. 2014, Article ID 428570, 11 pages, 2014.

[13] M. Altaf-Ul-Amin, T. Katsuragi, T. Sato, and S. Kanaya, "A glimpse to background and characteristics of major molecular biological networks," *BioMed Research International*, vol. 2015, Article ID 540297, 14 pages, 2015.

[14] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[15] W. J. R. Longabaugh, E. H. Davidson, and H. Bolouri, "Visualization, documentation, analysis, and communication of large-scale gene regulatory networks," *Biochimica et Biophysica Acta*, vol. 1789, no. 4, pp. 363–374, 2009.

[16] L. Xu, B.-F. Yang, and J. Ai, "MicroRNA transport: a new way in cell communication," *Journal of Cellular Physiology*, vol. 228, no. 8, pp. 1713–1719, 2013.

[17] K. Chen and N. Rajewsky, "The evolution of gene regulation by transcription factors and microRNAs," *Nature Reviews Genetics*, vol. 8, no. 2, pp. 93–103, 2007.

[18] L. Zhong, J. T. L. Wang, D. Wen, V. Aris, P. Soteropoulos, and B. A. Shapiro, "Effective classification of microRNA precursors using feature mining and AdaBoost algorithms," *OMICS*, vol. 17, no. 9, pp. 486–493, 2013.

[19] A. E. Wagner, S. Piegholdt, M. Ferraro, K. Pallauf, and G. Rimbach, "Food derived microRNAs," *Food and Function*, vol. 6, no. 3, pp. 714–718, 2015.

[20] W. J. R. Longabaugh, E. H. Davidson, and H. Bolouri, "Computational representation of developmental genetic regulatory networks," *Developmental Biology*, vol. 283, no. 1, pp. 1–16, 2005.

[21] M. Levine and E. H. Davidson, "Gene regulatory networks for development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4936–4942, 2005.

[22] E. H. Davidson and D. H. Erwin, "Gene regulatory networks and the evolution of animal body plans," *Science*, vol. 311, no. 5762, pp. 796–797, 2006.

[23] P. Zoppoli, S. Morganella, and M. Ceccarelli, "TimeDelay-ARACNE: reverse engineering of gene networks from time-course data by an information theoretic approach," *BMC Bioinformatics*, vol. 11, article no. 154, 2010.

[24] S. R. Maetschke, P. B. Madhamshettiwar, M. J. Davis, and M. A. Ragan, "Supervised, semi-supervised and unsupervised inference of gene regulatory networks," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 195–211, 2014.

[25] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 14, pp. 6286–6291, 2010.

[26] N. Patel and J. T. L. Wang, "Semi-supervised prediction of gene regulatory networks using machine learning algorithms," *Journal of Biosciences*, vol. 40, no. 4, pp. 731–740, 2015.

[27] T. Turki and J. T. L. Wang, "A new approach to link prediction in gene regulatory networks," in *Intelligent Data Engineering and Automated Learning—IDEAL 2015: 16th International Conference, Wroclaw, Poland, October 14–16, 2015, Proceedings*, vol. 9375 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2015.

[28] A. Greenfield, A. Madar, H. Ostrer, and R. Bonneau, "DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models," *PLoS ONE*, vol. 5, no. 10, Article ID e13397, 2010.

[29] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano, "Generating realistic in silico gene networks for performance assessment of reverse engineering methods," *Journal of Computational Biology*, vol. 16, no. 2, pp. 229–239, 2009.

[30] R. J. Prill, D. Marbach, J. Saez-Rodriguez et al., "Towards a rigorous assessment of systems biology models: the DREAM3 challenges," *PLoS ONE*, vol. 5, no. 2, Article ID e9202, 2010.

[31] T. Schaffter, D. Marbach, and D. Floreano, "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods," *Bioinformatics*, vol. 27, no. 16, pp. 2263–2270, 2011.

[32] S. L. Chin, I. M. Marcus, R. R. Klevecz, and C. M. Li, "Dynamics of oscillatory phenotypes in Saccharomyces cerevisiae reveal a network of genome-wide transcriptional oscillators," *FEBS Journal*, vol. 279, no. 6, pp. 1119–1130, 2012.

[33] B. Langmead, K. D. Hansen, and J. T. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna," *Genome Biology*, vol. 11, no. 8, p. R83, 2010.

[34] T. Nguyen, W. Shi, and D. Ruden, "CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping," *BMC Research Notes*, vol. 4, article no. 171, 2011.

[35] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[36] S. Zhao, K. Prenger, L. Smith et al., "Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing," *BMC Genomics*, vol. 14, article 425, 2013.

[37] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends," *BioData Mining*, vol. 7, no. 1, article 22, 2014.

[38] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, vol. 15, no. 4, pp. 637–647, 2014.

[39] A. O'Driscoll, V. Belogrudov, J. Carroll et al., "HBLAST: parallelised sequence similarity—a Hadoop MapReducable basic local alignment search tool," *Journal of Biomedical Informatics*, vol. 54, pp. 58–64, 2015.

[40] 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.

[41] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review. A. Third Series*, vol. 33, no. 2, article 1134, 1986.

[42] Y. Song, L. Hua, B. A. Shapiro, and J. T. L. Wang, "Effective alignment of RNA pseudoknot structures using partition function posterior log-odds scores," *BMC Bioinformatics*, vol. 16, no. 1, article no. 39, 2015.

[43] E. Li, "Chromatin modification and epigenetic reprogramming in mammalian development," *Nature Reviews Genetics*, vol. 3, no. 9, pp. 662–673, 2002.

[44] H. Van Bokhoven, "Genetic and epigenetic networks in intellectual disabilities," *Annual Review of Genetics*, vol. 45, pp. 81–104, 2011.

[45] C. Zhang, B. Ji, A. Mardinoglu, J. Nielsen, and Q. Hua, "Logical transformation of genome-scale metabolic models for gene level applications and analysis," *Bioinformatics*, vol. 31, no. 14, pp. 2324–2331, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 5296729, 12 pages
http://dx.doi.org/10.1155/2017/5296729

Research Article

Novel Approach to Classify Plants Based on Metabolite-Content Similarity

Kang Liu, Azian Azamimi Abdullah, Ming Huang, Takaaki Nishioka, 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 6 August 2016; Revised 14 November 2016; Accepted 30 November 2016; Published 9 January 2017

Academic Editor: Yudong Cai

Copyright © 2017 Kang Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Secondary metabolites are bioactive substances with diverse chemical structures. Depending on the ecological environment within which they are living, higher plants use different combinations of secondary metabolites for adaptation (e.g., defense against attacks by herbivores or pathogenic microbes). This suggests that the similarity in metabolite content is applicable to assess phylogenic similarity of higher plants. However, such a chemical taxonomic approach has limitations of incomplete metabolomics data. We propose an approach for successfully classifying 216 plants based on their known incomplete metabolite content. Structurally similar metabolites have been clustered using the network clustering algorithm DPClus. Plants have been represented as binary vectors, implying relations with structurally similar metabolite groups, and classified using Ward's method of hierarchical clustering. Despite incomplete data, the resulting plant clusters are consistent with the known evolutional relations of plants. This finding reveals the significance of metabolite content as a taxonomic marker. We also discuss the predictive power of metabolite content in exploring nutritional and medicinal properties in plants. As a byproduct of our analysis, we could predict some currently unknown species-metabolite relations.

1. Introduction

Plant taxonomy is the science that explores, describes, names, and classifies plants. The systematic and phylogenetic analysis of plants is traditionally based on macroscopic and microscopic morphological characteristics and is known to be turbulent [1]. The study of DNA and to a certain extent m-RNA and proteins has led to the immense subject of molecular biology, which has been increasingly applied to reconstruct the phylogeny of higher and lower plants [2]. The use of molecular data in plant taxonomy has been highly successful in many instances but has the following two limitations. First, current technologies that use genomic compartments instead of the entire genome data usually only partially reveal the evolutional relations among plants. The number of organisms with completely known genomes in Kyoto Encyclopedia of Genes and Genomes (KEGG) has now reached 4505 but includes only 65 plants (November 2016). This indicates that it is still impractical to reconstruct plant taxonomy using the entire genome information. Second, recent research has indicated that horizontal gene transfer occurs in multicellular eukaryotes, especially in plants, and has an important role in their eukaryotic evolution. This suggests that phylogenetic reconstruction cannot be determined conclusively from sequence data [3, 4]. Paralleled with molecular biology, exploration of the phylogenetic distance between species based on metabolites, either alone or in combination with sequence features, has also begun. Clemente et al. (2007) presented a method for assessing the structural similarity of metabolic pathways for several organisms and reconstructed phylogenies that were very similar to the National Center for Biotechnology Information (NCBI) taxonomy [5]. Borenstein et al. (2008) predicted the phylogenetic tree by comparing seed metabolite compound content [6]. Mano et al. (2010) considered the topology of pathways as chains and used a pathway-alignment method to classify species [7]. Chang et al. (2011) proposed an approach from the perspective of enzyme substrates and corresponding products in which each organism is represented as a vector of substrate-product pairs. The vectors were then compared to

reconstruct a phylogenetic tree [8]. Ma et al. (2013) demonstrated the usefulness of the global alignment of multiple metabolic networks to infer the phylogenetic relationships between species [9]. However, most of these studies have focused on microorganisms, such as archaea, rather than multicellular eukaryotes.

Plants are the major contributors of natural products and are usually rich in nutritional or medicinal properties. Many natural products are biologically active and have been used for thousands of years as traditional medicines. Classifying plants on the basis of their chemical constituents, which is also known as plant chemosystematics, could be helpful in discovering new edible and medicinal plants and solving selected taxonomical problems [2, 10, 11]. Traditional chemosystematics of plants is based on the presence or absence of selected secondary metabolites, which is far from the holistic approach involving metabolite content [10, 11]. The incomplete data of metabolite constituents of plants limits the ability to solve taxonomical problems and discovery of new natural products or medicinal properties of plants.

With the rapid development of metabolomics, metabolite-related databases (DBs) have been created, including KNApSAcK, which contains accumulated information about species-metabolite relations including information about many secondary metabolites of plants [12]. Such information can be used in the systems-biological studies on the interactions between plants, including the activities of medicinal plants as well as interactions between plants and their environments [13]. Metabolite content refers to all small molecules that are the products or intermediates of metabolism (metabolites) that are present within a biological organism. The metabolite content of plants is dominated by secondary metabolites [14], which are usually of high structural diversity [15]. As a rule, secondary metabolites are often similar within members of a clade, and plants within a taxon often represent similar metabolite content and bioactive properties. Therefore, the metabolite content of plants can be used as a taxonomy marker to distinguish plants and other organisms [11]. However, the expression of secondary metabolites of a given structural type has frequently arisen on a number of occasions in different parts of the plant kingdom. This discrepancy could be due either to convergent evolution or to differential gene expression [11]. This suggests that the metabolite content of plants may reveal more information of the interaction and bioactive pattern of plants rather than morphology characteristics. Such metabolite-content-based classification not only reveals the phylogenetic relationship of plants but also can be used for studying the relationship of plants in terms of their bioactive properties, guiding prediction of medicinal properties in bioprospecting, exploring new nutritional or economic uses of plants, and solving taxonomical problems. Previously, microorganism species have been classified based on the volatile metabolites emitted by them, and the results have been well explained in terms of their pathogenicity [16]. This finding indicates that it is possible to classify other species, such as plants, based on metabolite-content similarity. With the development of plants metabolomics and big data biology, it is now possible to investigate the metabolite content of plants on a cross-class level [17, 18].

The KNApSAcK Core DB is an extensive plant-metabolite relation DB that can be applied in multifaceted plant research, such as identification of metabolites, construction of integrated DBs, and bioinformatics and systems biology [19, 20] and can be considered an advanced source of metabolite content of plants. The KNApSAcK Core DB contains 109,976 species-metabolite relationships that encompass 22,399 species and 50,897 metabolites, and these numbers are still growing [13]. In this paper, we propose an approach to classify plants based on metabolite-content similarity. The metabolite-content data of plants and structure data of compounds are mainly obtained from the KNApSAcK Core DB and partially from PubChem DB [21, 22]. We measure the structural similarity between two metabolites by using the concept of the Tanimoto coefficient [23, 24], construct a network by selecting highly structurally similar metabolite pairs, and determine structurally similar groups of metabolites by using the DPClus algorithm [25]. We then link plants to such metabolite groups instead of individual metabolites to represent the plants as binary vectors. Several structurally similar metabolites are generally involved in a metabolic pathway. Thus, the use of structurally similar metabolite groups in this study can help to reduce the effect of missing data. Next, the metabolite-content similarity between plants is calculated based on binary similarity coefficients which then transformed into metabolite-content distances. Plants are finally classified using the hierarchical clustering method, and the resulting classification is evaluated by comparing it with the NCBI taxonomy [26]. Our classification results reveal both the phylogeny- and bioactivity-based relations among plants. We also use a support vector machine (SVM) algorithm to classify the plants by their economic uses [27, 28]. The classification performance reveals the predictive power of metabolite content in exploring nutritional and medicinal properties of plants. As a byproduct of our analysis, we can predict some currently unknown species-metabolite relations. To the best of our knowledge, we are the first to classify plants based on metabolite content.

2. Materials and Methods

2.1. Dataset and Preliminaries. The major input data are species-metabolite relationships obtained from the KNApSAcK Core DB, which is a part of the KNApSAcK Family DB [13]. The KNApSAcK Core DB contains most of the published information about species-metabolite relations, but this is obviously far from complete regarding plants and other living organisms. In the preprocessing step, we removed the plants with inadequate plant-metabolite relations to guarantee that the amount of metabolite content of selected plants is sufficient enough to reveal their interrelations.

We collected the molecular structure description files for the metabolites in our dataset as additional input data. The KNApSAcK Core DB provides MOL molecular structure files for most of the metabolites. For metabolite compounds with structure files that cannot be obtained from the KNApSAcK Core DB, we downloaded the SDF files directly from the

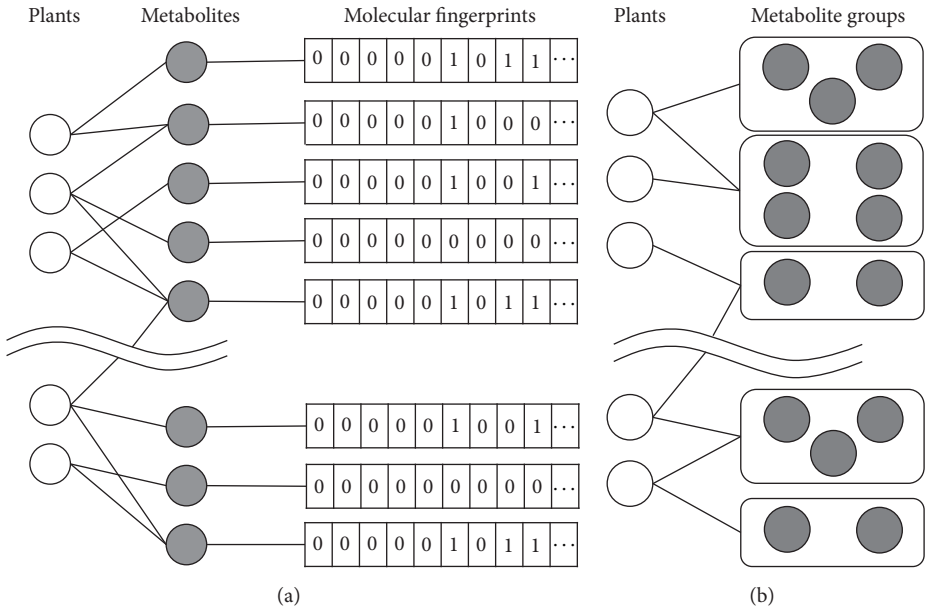


Figure 1: (a) Bipartite graph of plant-metabolite relations. Molecular structures of metabolites are described by 166-bit atom pair fingerprints, which are used to calculate Tanimoto structure similarity score for each metabolite pair. (b) Bipartite graph of plant versus metabolite-group relations. Each plant has been associated with metabolite groups instead of single metabolites to reduce effect of incomplete data.

PubChem DB [21, 22]. We used R package ChemmineR (v2.26.0) to generate atom pair fingerprints from molecular structure description files for all the metabolite compounds [29]. These molecular fingerprints were used to measure the structural similarity for all the metabolite pairs. Figure 1(a) illustrates the binary plant-metabolite relations and corresponding molecular fingerprints.

2.2. Network Construction of Metabolites Based on Chemical Structure Similarity. Very little is known of the complete set of metabolite content of plants. Therefore, for classifying plants based on currently available metabolite-content data, an approach that can compensate for the limitations of missing data is needed. Adjacent metabolites along a metabolic pathway are often related to similar substructures; therefore, it can be assumed that structurally similar metabolites are involved in the same or similar pathway. Therefore, plants that share highly structurally similar metabolites are likely to have common pathways; thus, they are likely to be within the same category and represent similar bioactivity. To compensate for the gap in missing data, we primarily linked plants to structurally similar metabolite groups instead of individual metabolites for this study.

For the purpose of determining structurally similar metabolite groups, we initially constructed a network of metabolites based on chemical structure similarity. We used the Tanimoto coefficient to measure the structural similarity between two metabolites [23]. Willett (2014) investigated different structural similarity measures and concluded that chemoinformatics research on structural similarity would continue to be largely based on the use of 2D fingerprints, and the Tanimoto coefficient has been established as the standard for similarity searching [30]. The Tanimoto coefficient between two metabolites A and B is defined as follows, which is the proportion of the features shared by two compounds divided by their union:

$$\text{Tanimoto}\,(A, B) = \frac{AB}{A + B - AB}. \quad (1)$$

The variable AB is the number of features common in both compounds, while A and B are the number of features that are related to the respective individual compounds. The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. The Tanimoto coefficient can be calculated from molecular fingerprints using the R package ChemmineR [29]. Empirically, a Tanimoto coefficient value larger than 0.85 indicates that the compared compounds represent highly similar bioactive features [31]. We used 0.85 as the threshold to insert an edge between two metabolites and constructed a network of metabolites.

2.3. Clustering of Metabolites Based on DPClus. The DPClus algorithm is a graph-clustering algorithm that can be used to extract densely connected nodes as a cluster [25, 32]. This algorithm can be applied to an undirected simple graph $G = (N, E)$ that consists of a finite set of nodes N and a finite set of edges E. Two important parameters are used in this algorithm (i.e., density d and cluster property cp). Density d_k of any cluster k is the ratio of the number of edges present in the cluster ($|E|$) to the maximum possible number of edges in the cluster ($|E|_{\max}$). The cluster property of a node n with respect to cluster k is represented as

$$\text{cp}_{nk} = \frac{E_{nk}}{d_k \times N_k}, \quad (2)$$

where N_k is the number of nodes in k and E_{nk} is the total number of edges between n and each node of k.

In this study, we applied the DPClus algorithm to the structural similarity network of metabolites. The metabolites were divided into many groups such that each group contains structurally similar compounds and can be treated as a distinctive pattern of structure. Each metabolite group might be related to a certain pathway, which is related to the phylogeny and ecology of plants. A plant is related to a metabolite group if it is related to any metabolite in the group. Thus, the original plant-metabolite relations are transformed into plant versus metabolite-group relations, as shown in Figure 1(b). We used such groups to measure the similarity between plants, thus reducing the effects of incomplete metabolite-content data.

2.4. Clustering of Plants Based on Metabolite Groups. The relations between plants and structurally similar metabolite groups can be expressed with a sparse binary matrix, which is defined as M. Element $M_{ij} = 1$ means that plant i contains at least one metabolite of group j, and $M_{ij} = 0$ means that plant i contains no metabolite of group j. Therefore, for each plant, we obtain a binary vector such that each bit corresponds to the presence or absence of a metabolite group.

Let two plants be described by the binary vectors x and y, each comprised of p variables with values either 1 or 0 ("1" indicates presence while "0" indicates absence), and p is the total number of metabolite groups. The Simpson similarity coefficient between plants can be calculated as

$$S_s = \frac{a}{\min\{(a+b),(a+c)\}}. \quad (3)$$

Here, a, b, and c are the frequencies of the events $x\&y$, $x\&\overline{y}$, and $\overline{x}\&y$, respectively [33–35].

To strengthen our finding with more support, we also used the Jaccard coefficient, which was previously considered as a similarity measure between different organisms in different contexts [33, 36]. The Jaccard similarity coefficient can be calculated as

$$S_j = \frac{a}{a+b+c}. \quad (4)$$

We transformed a similarity coefficient, s, to a distance coefficient, d, by the transformation $d = 1-s$ and classified the plants by using Ward's hierarchical clustering method using R.

2.5. Classification of Plants by SVMs. Support vector machines are supervised machine learning models for classification and regression analysis [27, 28]. An SVM training algorithm builds a model by constructing decision boundaries in feature space. Examples are predicted to belong to a category based on the boundaries.

To study the relationship between metabolite groups and economic uses of plants and evaluate the predictive power of metabolite content in guiding the discovery of natural products or medicinal properties in plants, we used an SVM algorithm, which was implemented by the function svm in R package e1071 v1.6-7, to classify plants by using default parameters [37–39]. We used economic uses as labels and corresponding metabolite groups as features. The classification performance is evaluated by using a confusion matrix. In a confusion matrix, the sum of a column represents the instances in a predicted class, while the sum of a row represents the instances in an actual class. All programs in this research were run in R v3.3.1.

3. Results and Discussion

3.1. Data Preprocessing. The KNApSAcK Core DB contains a total of 111199 species-metabolite binary relations that encompass 25658 species and 50899 metabolites. This DB was developed by collecting information on numerous metabolites of various organisms from published literature and several DBs, including PubChem [21, 22]. The species-metabolite relations in the KNApSAcK Core DB can be represented as a bipartite graph, as shown in Figure 1(a). The degree distribution of species in a species-metabolite bipartite graph follows a power law trend (see Supplementary Figure 1 of the Supplementary Material available online on https://doi.org/10.1155/2017/5296729) [40]. The metabolite-content data of plants in the KNApSAcK Core DB is unbalanced, i.e., many plants are associated with only a few metabolites and a few plants are associated with many metabolites, while other plants are in a between situation. One of the reasons behind this is that different plants have metabolic pathways of varying complexity. Medicinal plants usually contain more metabolites compared to edible plants because the former have gone through less artificial selections and preserved more secondary metabolites during evolution. Another reason is that the metabolomics of some important plants have been studied more systematically. The recorded metabolite content of such plants is more comprehensive compared to wild plants. Therefore, in our current research, we selected 216 plants from a total of 25658 plants in the KNApSAcK Core DB, such that each of the 216 plants is reported to be associated with no less than 30 metabolites, with 135 being the maximum number and 31 being the minimum. There are a total of 6522 metabolites related to the 216 plants in our input dataset.

3.2. Plant Representation Based on Metabolite-Content Similarity. We dealt with 6522 metabolites involving 216 plants. We determined the Tanimoto coefficients between all possible metabolite pairs (21264981 pairs). We selected 54528 metabolite pairs with Tanimoto values greater than 0.85, which are 0.25% of all the metabolite pairs. On average, each metabolite is related to about eight different metabolites. We connected all the selected metabolite pairs and constructed a network of metabolites, as shown in Figure 2(a). This network involves 5085 metabolites and the other 1437 metabolites are not included in the network; that is, each of these metabolites is not structurally similar to any other metabolites. The 5085 metabolites included in the network are divided into 669 connected components. The degree distribution of the network also follows a power law trend (Figure 2(b)) [40].

To compensate for the gap in incomplete data regarding species-metabolite relations, we associated plants with

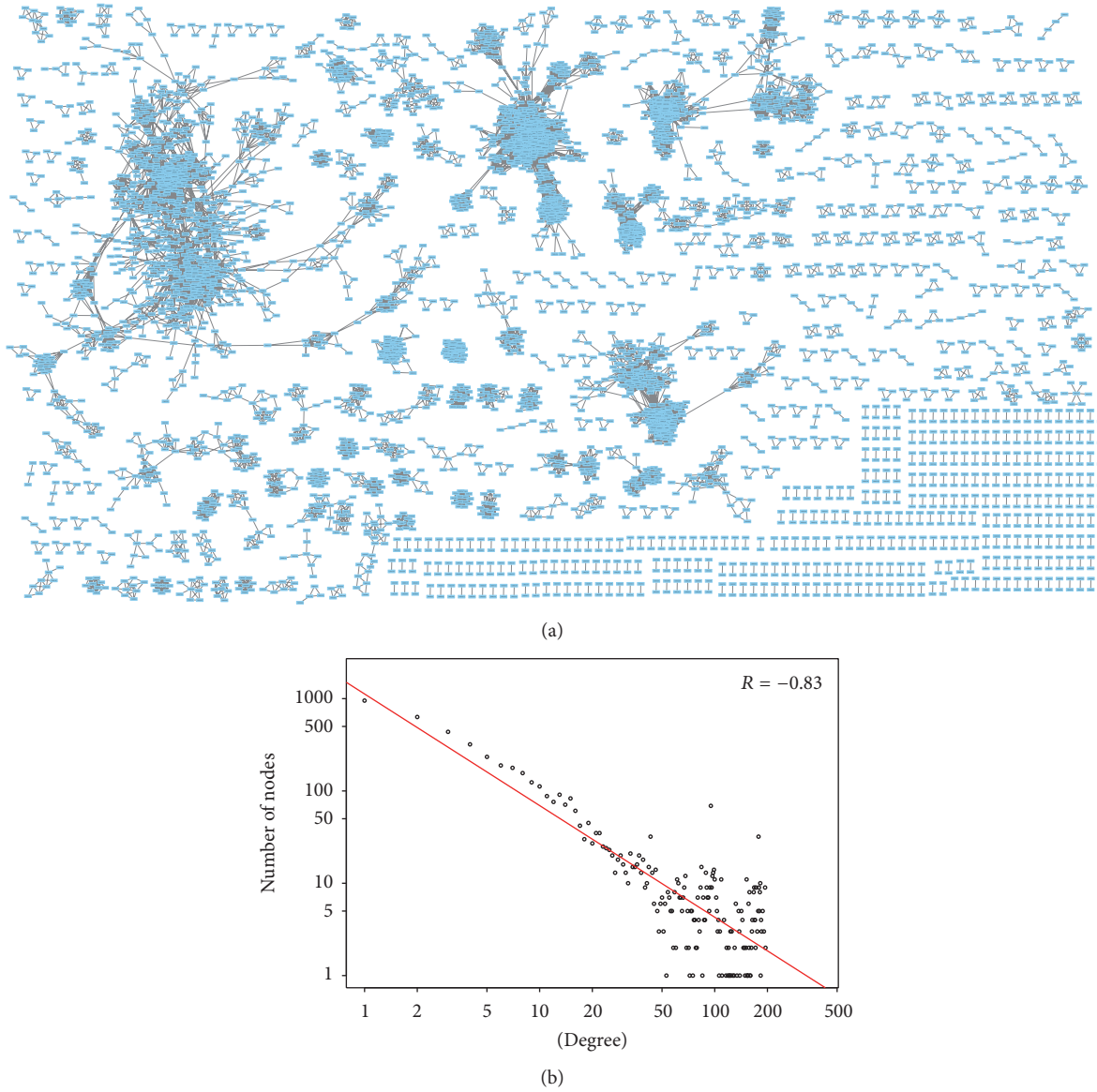


Figure 2: (a) Structural-similarity-based network of metabolites (plotted using network analysis tool *Cytoscape* v3.3.0). This network is composed of many isolated components, and each component contains different number of nodes. (b) Degree distribution of the network in log scale.

structurally similar metabolite groups instead of individual metabolites. To achieve this, we applied the DPClus algorithm to the network of metabolites we developed, as discussed in the previous section. We did DPClus clustering with the following settings: cluster property cp was set to 0.5, density value d was set to 0.9, minimum cluster size was set to 2, and we used the overlapping mode.

The DPClus algorithm generated 1150 clusters (i.e., metabolite groups, involving 4700 metabolites). The largest group contained 174 metabolites, and there were 510 metabolite groups containing only 2 metabolites. Figure 3 shows the frequency of metabolite groups with respect to size (the count of metabolites) in both normal scale and log-log scale (inset), and this distribution also follows a power law trend [40]. A total of 1822 metabolites not included in any cluster are considered as groups consisting of a single metabolite.

All clusters, large or small, contained structurally similar metabolites. Large clusters might be related to different metabolic pathways, but small clusters are likely related to specific metabolic pathways. A plant is related to a metabolite group if it is reported to contain any metabolite in the group. A plant can be represented as a binary vector such that each bit of the vector corresponds to the presence or absence of a metabolite group.

3.3. Clustering of Plants Based on Metabolite-Content Similarity. We calculated the plant-plant similarity by using two commonly used binary similarity coefficients Simpson

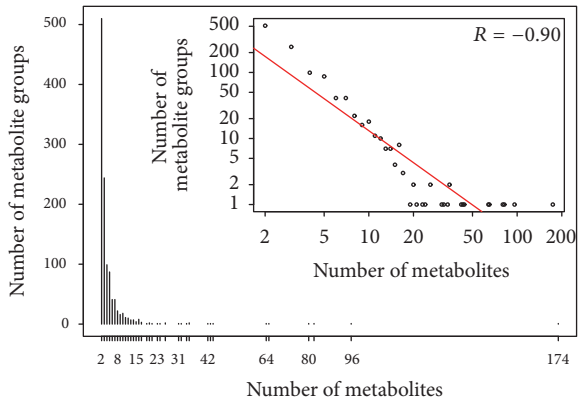


Figure 3: Frequency of metabolite groups with respect to group size. x-axes represent number of metabolites belonging to one metabolite group, and y-axes represent frequency of such metabolite groups. Frequency of metabolite groups in log scale is shown in inset figure.

and Jaccard [33]. The Jaccard coefficient has been used as a similarity measure to compare the enzyme content of metabolic networks in each pair of organisms [36]. The Simpson coefficient was devised to minimize the effect of the unequal size of two faunas being compared and having in the denominator only the number of taxa in a sample having the smaller number [34, 35].

We transformed a similarity score into a distance score d using $d = 1 - s$ and then conducted Ward's hierarchical clustering analysis. Thus, we determined two dendrograms corresponding to two types of coefficients with our approach.

We used the NCBI taxonomy of the 216 plants generated using a web-based tool from the NCBI homepage (http://www.ncbi.nlm.nih.gov/taxonomy) as the reference classification [26]. The NCBI classification reflects the phylogenetic patterns within a plant group primarily based on morphology. According to the NCBI taxonomy, the 216 plants spread over 52 families with the largest family *Fabaceae* containing 42 plants.

We compared the dendrogram trees generated with our approach with the NCBI taxonomy based on a similarity score called Baker's Gamma correlation coefficient using R package *dendextend* v1.3.0 [41, 42]. Baker's Gamma correlation coefficient ranges from −1 to +1, with positive values, meaning that the two trees are statistically similar. The results show that both Simpson- and Jaccard-coefficient-based trees produced similar scores (i.e., 0.062 and 0.059, resp.), indicating that both trees are statistically similar with the NCBI taxonomy. We can also extract phylogeny relations from the trees by referring to the NCBI taxonomy.

Overall, we found that the Simpson coefficient performed better than the Jaccard coefficient. In the Simpson coefficient tree, more plants from the same genus or family appeared nearer to each other compared to the Jaccard coefficient tree. We illustrate this fact by pointing out some examples in Supplementary Figure 2. The better performance of the Simpson coefficient is also reflected with the Baker's Gamma correlation coefficient. Therefore, for further explanation, we selected the Simpson coefficient tree and classified the plants into 48 groups by cutting the dendrogram at variable threshold heights empirically chosen to enrich the clusters with plants of the same genus or family. Supplementary Figure 3 shows the dendrogram together with group IDs produced by our classification method.

The main defined ranks in the NCBI taxonomic hierarchy are as follows: *superkingdom*, *kingdom*, *phylum*, *subclass*, *order*, *family*, *subfamily*, *tribe*, *genus*, and *species* (from high to low). We collected the taxonomy information of 216 plants that we considered in this study and annotated each plant with ranks of *family* and *genus* (we used the scientific names of plants where the first word of a plant name represents the *genus* to which the plant belongs). Table 1 lists the 48 groups of plants based on our clustering result with their taxonomic and usage information. The plants are arranged by different groups, and for each group plants within the same *family* or *genus* are arranged together to highlight the internal phylogeny relations. In the dendrogram of Supplementary Figure 3, neighboring plants belonging to the same *genus* or *family* are indicated by horizontal bold colored lines. Each *genus* or *family* is indicated by a specific color. It is evident that many clusters are rich with plants from the same *genus* or *family*. Thus, our results imply that plants in the same taxon correspond to similar metabolite content. Taking into account the inadequate amount of metabolite data and limited number of plants we considered for certain families, the results from our approach are very promising. These indicate that the proposed approach was designed to compensate for the shortcomings of limited data. Some deviations in our classification from the NCBI taxonomy can be explained in terms of ecological relationships or bioactive similarity. This implies that, compared to morphology-based taxonomy, metabolite-content-based classification reveals more information about the bioactive similarity among plants, which is related to the nutritional and medicinal properties of plants. Therefore, metabolite-content-based classification can be used as a time-efficient predictive tool for guiding discovery of edible and medicinal properties in wild plants.

3.4. Predicting Currently Unknown Plant-Metabolite Relations. The species-metabolite relation data in the KNApSAcK Core DB were collected from previously published papers. Many more plant-metabolite relations will inevitably be discovered in the future. However, based on our study, we can predict some not yet known plant-metabolite relations. When several plants are included in the same cluster with our approach, it implies that those plants contain many metabolites that are either the same or different but structurally very similar. When several plants contain a different subset of a group of structurally similar metabolites and they are very close according to morphological taxonomy, we can assume that all those plants contain the union of the metabolites currently detected in them. The basis of this assumption is that similar metabolic pathways are expected to be active in plants within a given taxon group.

In our experiments, we found structurally similar metabolite groups of different sizes, large and small. However, the metabolites belonging to a smaller group are likely to be closely related along a certain metabolic pathway. Therefore,

Table 1: Taxonomic and use information of 216 plants. Group ID, plant names, taxonomic ranks (*family*), and economic uses are mentioned in consecutive columns. Economic uses of plants are represented as the following abbreviations: E (edible), M (medicinal), L (landscaping,), T (timber), P (poisonous), and W (wild plant). Some plants are both edible and medicinal and are annotated as E/M.

Group	Plant	Family	Use
1	*Citrus limon*	Rutaceae	E
	Citrus aurantifolia	Rutaceae	E/M
	Citrus paradisi	Rutaceae	E
	Citrus sinensis	Rutaceae	E
	Citrus reticulata	Rutaceae	E
	Citrus aurantium	Rutaceae	E
2	*Houttuynia cordata*	Saururaceae	E/M
	Houttuynia emeiensis	Saururaceae	W
	Rhodiola rosea	Crassulaceae	M
3	*Artemisia annua*	Asteraceae	M
	Artemisia capillaris	Asteraceae	M
	Rhaponticum carthamoides	Asteraceae	W
	Solanum lycopersicum	Solanaceae	E
4	*Anthemis aciphylla*	Asteraceae	W
	Artemisia annua L.	Asteraceae	M
	Centaurea sessilis	Asteraceae	W
	Valeriana officinalis	Caprifoliaceae	M
	Persicaria minus	Polygonaceae	M
	Mentha arvensis	Lamiaceae	M
	Peucedanum paniculatum	Apiaceae	W
5	*Zingiber officinale*	Zingiberaceae	E/M
	Alpinia galanga	Zingiberaceae	E/M
	Rosmarinus officinalis	Lamiaceae	M
	Cistus albidus	Cistaceae	W
	Pinus halepensis	Pinaceae	L
6	*Myrtus communis*	Myrtaceae	M
	Leptospermum scoparium	Myrtaceae	M
	Santolina corsica	Asteraceae	W
7	*Curcuma amada*	Zingiberaceae	E/M
	Curcuma aeruginosa	Zingiberaceae	W
	Cistus creticus	Cistaceae	W
	Melaleuca leucadendra	Myrtaceae	M
	Piper arboreum	Piperaceae	W
	Piper fimbriulatum	Piperaceae	W
	Cedrus libani	Pinaceae	L
	Cyperus rotundus	Cyperaceae	M
8	*Pseudotsuga menziesii*	Pinaceae	T
	Pinus sylvestris	Pinaceae	T
	Picea abies	Pinaceae	T
	Citrus unshiu	Rutaceae	E
9	*Prunus persica*	Rosaceae	E
	Prunus avium	Rosaceae	E
	Prunus cerasus	Rosaceae	E
10	*Pisum sativum*	Fabaceae	E
	Lathyrus odoratus	Fabaceae	L
	Allium cepa	Amaryllidaceae	E
11	*Linum usitatissimum*	Linaceae	T
	Vicia faba	Fabaceae	E
	Carthamus tinctorius	Asteraceae	M
12	*Phaseolus lunatus*	Fabaceae	E
	Phaseolus vulgaris	Fabaceae	E
	Phaseolus coccineus	Fabaceae	E
13	*Triticum aestivum*	Poaceae	E
	Zea mays	Poaceae	E
	Spinacia oleracea	*Amaranthaceae*	E
14	*Raphanus sativus*	Brassicaceae	E
	Brassica napus	Brassicaceae	P
	Malus domestica	Rosaceae	E
15	*Hordeum vulgare*	Poaceae	E
	Oryza sativa	Poaceae	E
	Cucumis sativus	Cucurbitaceae	E
	Glycine max	Fabaceae	E
	Helianthus annuus	Asteraceae	E
16	*Eriobotrya japonica*	Rosaceae	E
	Cassia fistula	Fabaceae	M
	Aesculus hippocastanum	Hippocastanaceae	P
	Camellia sinensis	Theaceae	E
	Rheum sp.	Polygonaceae	W
17	*Robinia pseudoacacia*	Fabaceae	L
	Colophospermum mopane	Fabaceae	T
	Acacia mearnsii	Fabaceae	W
18	*Sinocrassula indica*	Crassulaceae	M
	Sedum sarmentosum	Crassulaceae	M
	Rhodiola sachalinensis	Crassulaceae	M
	Phyllanthus emblica	Phyllanthaceae	E/M
	Psidium guajava	Myrtaceae	E
	Phellodendron amurense	Rutaceae	M
	Epimedium sagittatum	Berberidaceae	M
19	*Solanum lycopersicum*	Solanaceae	E
	Solanum tuberosum	Solanaceae	E
	Nicotiana tabacum	Solanaceae	M
20	*Capsicum annuum*	Solanaceae	E
	Petunia x hybrida	Solanaceae	L
	Daucus carota	Apiaceae	W
	Asclepias curassavica	Apocynaceae	L
	Humulus lupulus	Cannabaceae	M
	Cyperus rotundus	Cyperaceae	M
21	*Glycyrrhiza uralensis*	Fabaceae	M
	Glycyrrhiza aspera	Fabaceae	W
	Glycyrrhiza glabra	Fabaceae	E/M
	Glycyrrhiza inflata	Fabaceae	M
22	*Lupinus luteus*	Fabaceae	W
	Lupinus albus	Fabaceae	E
	Derris scandens	Fabaceae	W
	Erythrina variegata	Fabaceae	L
	Erythrina senegalensis	Fabaceae	M
23	*Euchresta japonica*	Fabaceae	W
	Euchresta formosana	Fabaceae	W
	Sophora flavescens	Fabaceae	M
	Maackia amurensis	Fabaceae	L
	Sophora secundiflora	Fabaceae	W
	Daphniphyllum oldhamii	Daphniphyllaceae	M
24	*Medicago sativa*	Fabaceae	E
	Clitoria ternatea	Fabaceae	E
	Trifolium pratense	Fabaceae	M
	Sophora japonica	Fabaceae	T
	Lespedeza homoloba	Fabaceae	W
	Melilotus messanensis	Fabaceae	W
	Glycyrrhiza pallidiflora	Fabaceae	W

Table 1: Continued.

Group	Plant	Family	Use
	Dalbergia odorifera	Fabaceae	T
25	*Corydalis claviculata*	Papaveraceae	W
	Papaver somniferum	Papaveraceae	M
	Corydalis solida	Papaveraceae	W
	Cocculus laurifolius	Menispermaceae	W
	Stephania cepharantha	Menispermaceae	W
	Stephania cepharantha	Menispermaceae	W
	Cocculus pendulus	Menispermaceae	W
	Annona cherimola	Annonaceae	E
	Xylopia parviflora	Annonaceae	W
26	*Brassica oleracea*	Brassicaceae	E
	Brassica rapa	Brassicaceae	E
	Armoracia lapathifolia	Brassicaceae	E
	Hesperis matronalis	Brassicaceae	L
27	*Alstonia macrophylla*	Apocynaceae	T
	Alstonia angustifolia	Apocynaceae	M
	Alstonia angustifolia var. *latifolia*	Apocynaceae	M
28	*Millettia pinnata*	Fabaceae	L
	Millettia pinnata	Fabaceae	L
	Neorautanenia amboensis	Fabaceae	W
	Tephrosia purpurea	Fabaceae	P
	Amorpha fruticosa	Fabaceae	L
	Piscidia erythrina	Fabaceae	T
29	*Gymnadenia conopsea*	Orchidaceae	M
	Bletilla striata	Orchidaceae	M
30	*Taiwania cryptomerioides*	Cupressaceae	T
	Chamaecyparis formosensis	Cupressaceae	T
	Cryptomeria japonica	Cupressaceae	T
31	*Gutierrezia microcephala*	Asteraceae	P
	Saussurea lappa	Asteraceae	M
	Artemisia spp.	Asteraceae	W
	Citrus spp.	Rutaceae	E
	Citrus sudachi	Rutaceae	M
	Murraya paniculata	Rutaceae	M
	Cannabis sativa	Cannabaceae	M
	Iris domestica	Iridaceae	M
32	*Tabernaemontana coffeoides*	Apocynaceae	W
	Kopsia dasyrachis	Apocynaceae	W
	Catharanthus roseus	Apocynaceae	M
	Rauvolfia vomitoria	Apocynaceae	W
33	*Nardostachys chinensis*	Caprifoliaceae	W
	Acritopappus confertus	Asteraceae	W
	Isodon xerophilus	Lamiaceae	W
	Cynanchum sublanceolatum	Apocynaceae	W
	Caesalpinia crista	Fabaceae	T
	Murraya euchrestifolia	Rutaceae	W
	Curcuma zedoaria	Zingiberaceae	E
34	*Garcinia mangostana*	Clusiaceae	E/M
	Garcinia dulcis	Clusiaceae	W
35	*Atalantia buxifolia*	Rutaceae	W
	Ruta graveolens	Rutaceae	E/M
	Clausena excavata	Rutaceae	W
	Angelica furcijuga	Apiaceae	E/M
36	*Andrographis paniculata*	Acanthaceae	M
	Scutellaria baicalensis	Lamiaceae	M

Table 1: Continued.

Group	Plant	Family	Use
37	*Zanthoxylum simulans*	Rutaceae	M
	Zanthoxylum integrifolium	Rutaceae	W
38	*Magnolia denudata*	Magnoliaceae	M
	Magnolia officinalis	Magnoliaceae	M
	Aeschynanthus bracteatus	Gesneriaceae	W
39	*Broussonetia papyrifera*	Moraceae	E
	Morus alba	Moraceae	E/M
	Artocarpus communis	Moraceae	E
40	*Sinapis alba*	Brassicaceae	E
	Vachellia rigidula	Fabaceae	E
41	*Lycium chinense*	Solanaceae	M
	Mandragora autumnalis	Solanaceae	M
	Angelica sinensis	Apiaceae	M
42	*Cullen corylifolium*	Fabaceae	M
	Calophyllum inophyllum	Calophyllaceae	T
	Juniperus phoenicea	Cupressaceae	W
43	*Taxus cuspidata*	Taxaceae	P
	Taxus brevifolia	Taxaceae	M
	Taxus baccata	Taxaceae	M
	Taxus wallichiana	Taxaceae	M
	Taxus chinensis	Taxaceae	M
	Taxus mairei	Taxaceae	M
	Taxus yunnanensis	Taxaceae	M
44	*Panax notoginseng*	Araliaceae	M
	Panax ginseng	Araliaceae	M
	Panax pseudoginseng var. *notoginseng*	Araliaceae	M
	Panax ginseng C.A. Meyer	Araliaceae	M
	Bupleurum rotundifolium	Apiaceae	M
	Beta vulgaris	Amaranthaceae	E
	Bellis perennis	Asteraceae	E/M
45	*Xylocarpus granatum*	Meliaceae	W
	Spiraea formosana	Rosaceae	W
	Hibiscus taiwanensis	Malvaceae	W
	Begonia nantoensis	Begoniaceae	W
	Alpinia blepharocalyx	Zingiberaceae	W
	Taraxacum formosanum	Asteraceae	W
46	*Aristolochia elegans*	Aristolochiaceae	L
	Aristolochia heterophylla	Aristolochiaceae	M
47	*Artabotrys uncinatus*	Annonaceae	W
	Annona purpurea	Annonaceae	E
	Rubia yunnanensis	Rubiaceae	M
	Withania somnifera	Solanaceae	M
48	*Salvia officinalis*	Lamiaceae	E/M
	Orthosiphon stamineus	Lamiaceae	W
	Plantago major	Plantaginaceae	M
	Rehmannia glutinosa	Rehmanniaceae	M
	Olea europaea	Oleaceae	E/M
	Lonicera japonica	Caprifoliaceae	M
	Eleutherococcus senticosus	Araliaceae	M
	Diospyros kaki	Ebenaceae	E
	Punica granatum	Lythraceae	E
	Curcuma domestica	Zingiberaceae	E/M

Table 2: Reported plant-metabolite relations of 6 plants of genus *Citrus* with a given metabolite group (including 2 metabolites: *Limonene* and *Cyclohexane*). 1/0 indicates presence/absence of a metabolite in a plant.

	Citrus limon	*Citrus aurantifolia*	*Citrus paradisi*	*Citrus sinensis*	*Citrus reticulata*	*Citrus aurantium*
Limonene	1	1	1	1	1	1
Cyclohexane	0	1	1	1	1	0

Table 3: Predicted unrecorded metabolites for 6 *Citrus* plants, encompassing 38 plant-metabolite relations.

Species	Predicted unrecorded metabolites
Citrus limon	Gibberellin A4; methyl salicylate; cyclohexane; *o*-isopropenyl toluene; jasmonic acid; 10′-apoviolaxanthal; alpha-trans-bergamotene
Citrus aurantifolia	Methyl salicylate; citral; benzeneacetaldehyde; *o*-isopropenyl toluene; methyl epijasmonate; salvigenin
Citrus paradisi	Rhoifolin; isopropanol; methyl salicylate; citral; benzeneacetaldehyde; *o*-isopropenyl toluene
Citrus sinensis	Isoscutellarein 7,8-dimethyl ether; isoscutellarein 7,8,4′-trimethyl ether; *o*-isopropenyl toluene; methyl epijasmonate; salvigenin; gibberellin A53; violaxanthin
Citrus reticulata	Gibberellin A81; gibberellin A9; isopropanol; citral; 6-demethoxytangeritin; tetramethylscutellarein
Citrus aurantium	Apigenin 7-rutinoside; methyl salicylate; salvigenin; cyclohexane; benzeneacetaldehyde; *o*-isopropenyl toluene

for predicting currently unknown plant-metabolite relations, we focused on only smaller metabolite groups and empirically considered the metabolite groups of size no more than eight.

In summary, we follow the following steps to improve prediction accuracy.

Step 1. We select a group of plants that are in the same cluster according to our approach and at the same time belong to the same genus or family. Let us call such a group S.

Step 2. We determine the set (K) of structurally similar metabolite groups of size no more than eight such that each metabolite group is associated with at least two plants in S.

Step 3. All the metabolites of a metabolite group in K are assigned to the plants in S which are associated with the group. This process is repeated for each group in K.

Based on known information, however, we exclude some metabolites that are mainly structure isomers from this prediction process because some isomers are usually produced by different pathways [43, 44]. We discuss this method with an example as follows.

Predicting Metabolites for Citrus Plants. Six *Citrus* plants *(Citrus limon, Citrus aurantifolia*, *Citrus paradisi*, *Citrus sinensis*, *Citrus reticulata*, and *Citrus aurantium)* are considered an excellent group in our classification (Group 1 in Table 1, we call it group S) and belong to the same genus *(Citrus)*. We extract the set K of metabolite groups (with size no more than eight) in which each metabolite group is associated with at least two plants in S. There is a total of 58 such metabolite groups in K. For each metabolite group in K which is related to multiple plants, we can construct a plant-metabolite table. Table 2 is a plant-metabolite table for a given metabolite group that contains two metabolites, *Limonene* and *Cyclohexane*, and their association to six plants in S. In Table 2, "1" means that the metabolite is reported in the corresponding plant and "0" means that the metabolite is unreported in that plant. We treat all these unreported plant-metabolite relations as currently unknown but actual relations. We repeat this process for all 58 metabolite groups in K and obtain a list of unrecorded metabolites for the plants in S, which we show in Table 3. Following this method, we can predict some currently unrecorded metabolites and find some widespread medicinal species that can be substitutions of more endangered relatives currently being used [45].

Not all the predicted metabolites might actually be produced in given plants because of the complexity of metabolic pathway evolution. On the contrary, many true relations could not be predicted due to the limitation of the incomplete data source. However, with developments in plant metabolomics, we may be able to add more plant-metabolite relations in our analysis in the future and produce better results. For other plant groups, we can also predict numerous unrecorded metabolites. We list all the predicted plant-metabolite relations in Supplementary Table 1.

3.5. Relationship between Metabolite Content and Uses of Plants. Our unsupervised approach for classifying plants is based on metabolite-content similarity using hierarchical clustering. Our results substantially match those of traditional morphology-based taxonomy. However, our results further reflect the usage patterns of plants.

The metabolite content of plants is always related to their bioactive properties, and the similarity of the metabolite content of plants can reveal their bioactive similarity. Generally, medicinal properties are not randomly distributed in different classes of plants. Some plant classes are represented by more medicinal plants than others. It is suggested that there is a phylogenetic pattern in medicinal properties even within one genus [45–47]. A similar distribution could also be observed in our classification that plants with certain uses are concentrated in the same group. Many plant groups in our classification are of similar usage patterns. A plant

Table 4: Resulting confusion matrix from support vector machine (SVM) algorithm. 162 plants are labeled as edible (E), medicinal (M), timber (T), landscaping (L), and poisonous (P), and SVM model was constructed to classify them.

	M	E	T	L	P	Recognition rate [%]
M	81	0	0	0	0	100
E	1	47	0	0	0	97.9
T	6	0	8	0	0	57.2
L	8	1	0	5	0	35.7
P	4	1	0	0	0	0

Total: 162 plants. Accuracy: 87.0%.

is frequently related to multiple uses, but we only consider the most common use in this paper. We collected all the plant resource information from various data sources, including Wikipedia (https://www.wikipedia.org), and annotated plants by their uses such as medicinal, edible, ornamental, forestry, poisonous, and timber. Table 1 lists the usage patterns of 216 plants. The economic uses of plants are represented by different letters (E: edible, M: medicinal, L: landscaping, including forestry and ornamental plants, T: timber, P: poisonous, and W: wild plants that are not yet widely used by humans). Eleven groups (ID: 1, 9, 10, 12, 13, 14, 15, 19, 26, 39, and 40) involving 38 plants mostly consist of edible plants, and 14 groups (ID: 2, 4, 6, 18, 21, 27, 29, 31, 36, 38, 41, 43, 44, and 48) involving 69 plants mostly consist of medicinal plants. Moreover, 3 groups (ID: 8, 17, and 30) involving 10 plants mostly consist of landscaping or timber plants. This implies that the proposed classification approach of plants is consistent with their economic uses.

In this section, we investigate the relations between usage patterns and metabolite content of plants using a supervised classification technique. We considered every metabolite group as a pathway pattern such that each group can be used as a feature for classifying plants by their uses. For this analysis, we considered 48 edible plants (E), 81 medicinal plants (M), 14 timber plants (T), 14 landscaping plants (including forestry and ornamental plants), and 5 poisonous plants (P). We considered the plants that have both edible and medicinal uses (plants with "E/M" in Table 1) as medicinal plants. We applied an SVM algorithm to classify the plants, using economic uses of plants as labels and corresponding metabolite groups as features. Classification performance was evaluated from the resulting confusion matrix, as shown in Table 4. The rows of the confusion matrix indicate documented uses of plants and columns indicate the predicted uses from the SVM algorithm. *Recognition rate* is the proportion of correctly predicted plants corresponding to a class.

We found that all the medicinal plants and all but one edible plant were classified correctly. This implies that the metabolite content of medicinal and edible plants substantially differs. However, half the timber and landscaping plants were classified as medicinal plants. Therefore, timber and landscaping plants are somewhat related to medicinal plants in terms of metabolite content. All the poisonous plants were classified incorrectly: four plants were classified as medicinal plants and one as edible. This implies that poisonous plants are more similar to medicinal plants. Many poisonous plants can be used in treating specific diseases if the doses are carefully controlled [48]. In summary, edible plants represent exclusive metabolite content and can be differently classified from inedible plants. Furthermore, metabolite-content-based classification also reveals the predictive power of medicinal properties in bioprospecting. This indicates that our proposed approach can be used for exploring nutritional or medicinal properties of plants.

4. Conclusion

We proposed an approach for comparing the metabolite content of plants and classifying plants by their metabolite content. We showed that with this approach we can classify plants similar to the traditional morphology-based plant taxonomy. Naturally, this work can be generalized from various perspectives. First, our approach can be regarded as a novel chemosystematics method that can be used to consider the global metabolite content of plants instead of a group of metabolites as done in previous research. The resulting classification is consistent with natural phylogenetic and chemosystematics patterns of plants. Some deviations in our classification from the NCBI taxonomy can be explained in terms of bioactive similarity. Moreover, the complexity and known extent of metabolite content vary for different plants. We found that the Simpson coefficient can minimize the effect of the unequal size of the metabolite content of organisms and performs better in comparing metabolite content of plants than the Jaccard coefficient, which has been widely used as a similarity measure in various biological studies.

We also described a method for predicting unrecorded metabolites by structurally similar metabolite groups and phylogenetic relation of plants. With this method, we can predict some unrecorded metabolites and find new edible/medicinal plants from wild plants that have not been used by humans. Moreover, we studied the relation between the metabolite content of plants and their economic uses. We found that edible and medicinal plants represent unique metabolic pathway patterns and can be classified with an SVM algorithm with our integrated metabolite-content data. Our proposed metabolite-content-based plant-classification approach reveals the predictive power of medicinal properties in bioprospecting. The performance of this approach depends on the completeness of the metabolite-content data we use because metabolite groups, which were regarded as metabolic pathway patterns in our research, have been extracted from the background network of metabolites by using the DPClus algorithm. Therefore, if we can add more plant-metabolite relations, we can classify metabolites and species more accurately. Also, metabolites along identical pathways always correspond to high structural similarity. Our approach will be useful for predicting metabolic pathways in plants.

Competing Interests

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

Acknowledgments

This work was supported by JSPS KAKENHI (Grant no. JP16K07223), National Bioscience Database Center in Japan, and NAIST Big Data Project.

References

[1] P. Besse, Ed., *Molecular Plant Taxonomy: Methods and Protocols*, Humana Press, 2014.

[2] T. Reynolds, "The evolution of chemosystematics," *Phytochemistry*, vol. 68, no. 22–24, pp. 2887–2895, 2007.

[3] P. J. Keeling and J. D. Palmer, "Horizontal gene transfer in eukaryotic evolution," *Nature Reviews Genetics*, vol. 9, no. 8, pp. 605–618, 2008.

[4] C. Gao, X. Ren, A. S. Mason et al., "Horizontal gene transfer in plants," *Functional & Integrative Genomics*, vol. 14, no. 1, pp. 23–29, 2014.

[5] J. C. Clemente, K. Satou, and G. Valiente, "Phylogenetic reconstruction from non-genomic data," *Bioinformatics*, vol. 23, no. 2, pp. e110–e115, 2007.

[6] E. Borenstein, M. Kupiec, M. W. Feldman, and E. Ruppin, "Large-scale reconstruction and phylogenetic analysis of metabolic environments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 38, pp. 14482–14487, 2008.

[7] A. Mano, T. Tuller, O. Béjà, and R. Y. Pinter, "Comparative classification of species and the study of pathway evolution based on the alignment of metabolic pathways," *BMC Bioinformatics*, vol. 11, no. 1, article no. S38, 2010.

[8] C.-W. Chang, P.-C. Lyu, and M. Arita, "Reconstructing phylogeny from metabolic substrate-product relationships," *BMC Bioinformatics*, vol. 12, supplement, p. S27, 2011.

[9] C.-Y. Ma, S.-H. Lin, C.-C. Lee, C. Y. Tang, B. Berger, and C.-S. Liao, "Reconstruction of phyletic trees by global alignment of multiple metabolic networks," *BMC Bioinformatics*, vol. 14, supplement 2, article S12, 2013.

[10] R. Singh, "Chemotaxonomy: a tool for plant classification," *Journal of Medicinal Plants*, vol. 4, no. 2, pp. 90–93, 2016.

[11] M. Wink, "Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective," *Phytochemistry*, vol. 64, no. 1, pp. 3–19, 2003.

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

[13] 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.

[14] R. Hegnauer, "Chemical characters in plant taxonomy: some possibilities and limitations," *Pure and Applied Chemistry*, vol. 14, no. 1, pp. 173–187, 1967.

[15] E. Pichersky and D. R. Gang, "Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective," *Trends in Plant Science*, vol. 5, no. 10, pp. 439–445, 2000.

[16] A. A. Abdullah, M. Altaf-Ul-Amin, N. Ono et al., "Development and mining of a volatile organic compound database," *BioMed Research International*, vol. 2015, Article ID 139254, 13 pages, 2015.

[17] V. Marx, "Biology: the big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[18] M. Altaf-Ul-Amin, F. M. Afendi, S. K. Kiboi, and S. Kanaya, "Systems biology in the context of big data and networks," *BioMed Research International*, vol. 2014, Article ID 428570, 11 pages, 2014.

[19] 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.

[20] Y. Nakamura, F. Mochamad Afendi, A. Kawsar 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.

[21] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "PubChem: integrated platform of small molecules and biological activities," *Annual Reports in Computational Chemistry*, vol. 4, pp. 217–241, 2008.

[22] 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, supplement 2, no. 2, pp. W623–W633, 2009.

[23] J. W. Godden, L. Xue, and J. Bajorath, "Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 1, pp. 163–166, 2000.

[24] X. Chen and C. H. Reynolds, "Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1407–1414, 2002.

[25] 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.

[26] S. Federhen, "The NCBI taxonomy database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D136–D143, 2012.

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

[28] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Tech. Rep., Department of Computer Science, National Taiwan University, 2003, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[29] Y. Cao, A. Charisi, L.-C. Cheng, T. Jiang, and T. Girke, "ChemmineR: a compound mining framework for R," *Bioinformatics*, vol. 24, no. 15, pp. 1733–1734, 2008.

[30] P. Willett, "The calculation of molecular structural similarity: principles and practice," *Molecular Informatics*, vol. 33, no. 6-7, pp. 403–413, 2014.

[31] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, "Do structurally similar molecules have similar biological activity?" *Journal of Medicinal Chemistry*, vol. 45, no. 19, pp. 4350–4358, 2002.

[32] Md. 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.

[33] S. S. Choi, S. H. Cha, and C. C. Tappert, "A survey of binary similarity and distance measures," *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[34] W. C. Fallaw, "A test of the Simpson coefficient and other binary coefficients of faunal similarity," *Journal of Paleontology*, pp. 1029–1034, 1979.

[35] H.-W. Ma and A.-P. Zeng, "Phylogenetic comparison of metabolic capacities of organisms at genome level," *Molecular Phylogenetics and Evolution*, vol. 31, no. 1, pp. 204–213, 2004.

[36] K. Deyasi, A. Banerjee, and B. Deb, "Phylogeny of metabolic networks: a spectral graph theoretical approach," *Journal of Biosciences*, vol. 40, no. 4, pp. 799–808, 2015.

[37] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[38] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.

[39] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, *Misc Functions of the Department of Statistics (e1071)*, TU Wien, Vienna, Austria, 2005.

[40] 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.

[41] F. B. Baker, "Stability of two hierarchical grouping techniques Case I: sensitivity to data errors," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 440–445, 1974.

[42] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering," *Bioinformatics*, vol. 31, no. 22, pp. 3718–3720, 2015.

[43] P. M. Dewick, *Medicinal Natural Products. A Biosynthetic Approach*, John Wiley and Sons, Chichester, UK, 3rd edition, 2009.

[44] J. McMurry and T. Begley, *The Organic Chemistry of Biological Pathways*, chapter 3, Roberts and Company Publishers, Englewood, Colo, USA, 2005.

[45] C. H. Saslis-Lagoudakis, B. B. Klitgaard, F. Forest et al., "The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: an example from *pterocarpus* (*leguminosae*)," *PLoS ONE*, vol. 6, no. 7, Article ID e22275, 2011.

[46] N. Rønsted, M. R. E. Symonds, T. Birkholm et al., "Can phylogeny predict chemical diversity and potential medicinal activity of plants? A Case Study of Amaryllidaceae," *BMC Evolutionary Biology*, vol. 12, no. 1, article 182, 2012.

[47] M. Ernst, C. H. Saslis-Lagoudakis, O. M. Grace et al., "Evolutionary prediction of medicinal properties in the genus *Euphorbia L.*," Scientific Reports 6, 2016.

[48] N. Tamilselvan, T. Thirumalai, P. Shyamala, and E. David, "A review on some poisonous plants and their medicinal values," *Journal of Acute Disease*, vol. 3, no. 2, pp. 85–89, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7147039, 17 pages
http://dx.doi.org/10.1155/2016/7147039

Research Article

A Systematic Framework for Drug Repositioning from Integrated Omics and Drug Phenotype Profiles Using Pathway-Drug Network

Erkhembayar Jadamba[1] and Miyoung Shin[2]

[1]*Bio-Intelligence & Data Mining Laboratory, Graduate School of Electrical Engineering and Computer Science, Kyungpook National University, 1370 Sangyeok-dong, Buk-gu, Daegu 702-701, Republic of Korea*
[2]*School of Electronics Engineering, Kyungpook National University, 1370 Sangyeok-dong, Buk-gu, Daegu 702-701, Republic of Korea*

Correspondence should be addressed to Miyoung Shin; shinmy@knu.ac.kr

Received 17 June 2016; Revised 12 October 2016; Accepted 20 October 2016

Academic Editor: Md. Altaf-Ul-Amin

Copyright © 2016 E. Jadamba and M. Shin. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Drug repositioning offers new clinical indications for old drugs. Recently, many computational approaches have been developed to repurpose marketed drugs in human diseases by mining various of biological data including disease expression profiles, pathways, drug phenotype expression profiles, and chemical structure data. However, despite encouraging results, a comprehensive and efficient computational drug repositioning approach is needed that includes the high-level integration of available resources. In this study, we propose a systematic framework employing experimental genomic knowledge and pharmaceutical knowledge to reposition drugs for a specific disease. Specifically, we first obtain experimental genomic knowledge from disease gene expression profiles and pharmaceutical knowledge from drug phenotype expression profiles and construct a pathway-drug network representing a priori known associations between drugs and pathways. To discover promising candidates for drug repositioning, we initialize node labels for the pathway-drug network using identified disease pathways and known drugs associated with the phenotype of interest and perform network propagation in a semisupervised manner. To evaluate our method, we conducted some experiments to reposition 1309 drugs based on four different breast cancer datasets and verified the results of promising candidate drugs for breast cancer by a two-step validation procedure. Consequently, our experimental results showed that the proposed framework is quite useful approach to discover promising candidates for breast cancer treatment.

1. Introduction

Developing and discovering a new drug is a very costly and time consuming process, which can take 10–17 years with a cost of 1.3 billion dollars. Despite large investments in research and development each year, there are still only a small number of new drugs approved successfully by the Food and Drug Administration (FDA) each year. Increasing failure rates, high costs, and the lengthy testing process for drug development have led to a process called drug repositioning [1], which refers to identifying and developing new uses for existing drugs to reduce the risk and cost.

Traditional drug repositioning methods primarily use information on chemical structure, side effects, and drug phenotypes and explore similar drugs based on the assumption that structurally similar drugs tend to share common indications [2–4]. In other words, the key idea behind these approaches is that molecularly similar drug structures often affect proteins and biological systems in similar ways [4]. For example, Swamidass [5] used chemical structure data to identify unexpected connections between a known drug and a disease and explored the hypothesis that if a drug has the same target as a known drug, then this new drug would also have activity against the disease. As another approach, Keiser et al. used 3665 US FDA-approved and investigational drugs that together had hundreds of targets, defining each target by its ligands. The chemical similarities between the drugs and ligand sets predicted thousands of unanticipated associations,

which have been used to develop new indications for many drugs.

Alternatively, some approaches use a drug phenotype, which is the expression profile of patients undergoing treatment with a drug. For example, the Connectivity Map (CMap) [6, 7] project is exploring the effects of a large number of FDA-approved chemicals (1309 drugs) on gene expression, and these effects are measured in four different cell lines, allowing researchers to analyze the different expression patterns of drug's target genes. Many computational approaches have been introduced to reposition drugs using CMap by analyzing drug-associated expression signatures to match a repositioned drug's effect with a shared perturbed gene expression profile for another disease, under the assumption that drugs that share similar CMap expression signatures have similar therapeutic applications. Using the CMap data, Iorio et al. [8] developed a drug repositioning method by constructing a drug-drug similarity network using gene set enrichment analysis (GSEA) [9] that could compute the similarity between pairs of drugs. Several different studies [3, 10–13] showed that using CMap expression profiles with a combination of various data sources such as drug target databases, drug chemical structures, and drug side effects was an improvement over the current drug target identification methods.

Moreover, the rapid developments in genomics and high-throughput technologies have produced a large volume of disease gene expression profiles, protein-protein interactions, and pathways. The high-level integration of these resources using network-based approaches is reported to have great potential for discovering novel drug indications for existing drugs [14]. For example, Chen et al. [15] introduced two different inference methods for predicting drug-disease associations based on basic network topology using a bipartite graph constructed from DrugBank [16] and Online Mendelian Inheritance in Man (OMIM) [17]. Emig et al. [18] integrated gene expression profiles, drug targets, disease information, and interactions for drug repositioning. Hu and Agarwal [19] created a disease-drug network using disease microarray datasets and predicted new indications for existing drugs using their disease-drug network.

Although many of the above methods have shown encouraging results for finding new indications for old drugs, there are still some limitations. For example, Yildirim et al. [20] concluded that most drugs with distinct chemical structures target the same proteins, and Keiser et al. [21] reported that structurally similar drugs may also target proteins with dissimilar functions, stating that using chemical structure alone is insufficient for successful drug repositioning [22]. In addition, care should be taken when using only the drug phenotype (drug treated) expression profile (such as CMAP) for drug repositioning because some portion of the genes or pathways that show statistically significant expression differences in cell lines treated with the drug may be expressed only because of the drug's side effects or toxicity. Furthermore, the genes expressed in the drug treated profiles for specific disease cell line or tissue only represent a small subset of the biological pathways, whereas the cooperation of genes plays an important role in complex diseases such as cancer. Pathway-based drug repositioning may be a better alternative for drug repositioning for specific diseases such as cancer.

To overcome the above limitations, the current drug repositioning methods require a comprehensive and efficient computational drug repositioning approach that incorporates powerful machine learning approaches using the high-level integration of available data such as disease gene expression profiles (disease profile), drug treated expression profiles (drug phenotype profile), and drug databases (STITCH [23], DrugBank [16], therapeutic target database (TTD) [24]) to discover new drugs for a human diseases. In this study, we aim to develop a systematic computation framework that repositions drugs by employing disease profile and drug phenotype profiles on the drug network along with integrated omics data.

2. Materials and Methods

In the framework as shown in Figure 1, we firstly identify *disease-specific pathways* by using an integrative analysis of multiple disease gene expression profiles and construct a *pathway-drug network* structure using pathway-drug associations derived from the CMap *drug phenotype profile*. Then to discover promising candidates, for drug repositioning, we initialize node labels for the pathway-drug network using identified disease pathways and known drugs associated with breast cancer and perform network propagation in a semisupervised manner.

In the following, the detailed explanations of our proposed framework for repositioning and evaluation method are described.

2.1. Finding Disease-Specific Pathways from Multiple Disease Expression Profiles. To identify disease pathways related to a specific disease, conventional approaches have usually focused on identifying enriched pathways between cases and controls using data from a single experiment. Specifically, when using real experimental data such as microarray gene expression data, it is possible for different studies to report different results for disease-specific pathways. That is, the results are often not reproducible or not robust even to the mildest data perturbation, so the integrated analysis of multiple existing studies can increase the reliability and generalizability of results [25]. To address these issues, our approach identifies a disease-specific pathway based on disease pathway enrichment using multiple gene expression profiles for a given phenotype, in which the disease pathway enrichment results are integrated. Each disease expression profile is preprocessed, and the pathways that show significant differences between case and control samples are identified by GSEA [9], which returns the enrichment score (ES) and nominal p value for each pathway. These scores are used for comparison analysis across pathways to detect significant pathways.

Here, we considered that the integration of pathways significantly enriched for each expression profile could better represent "disease-specific pathways" for the phenotype of interest. To integrate, the pathways with a nominal p value

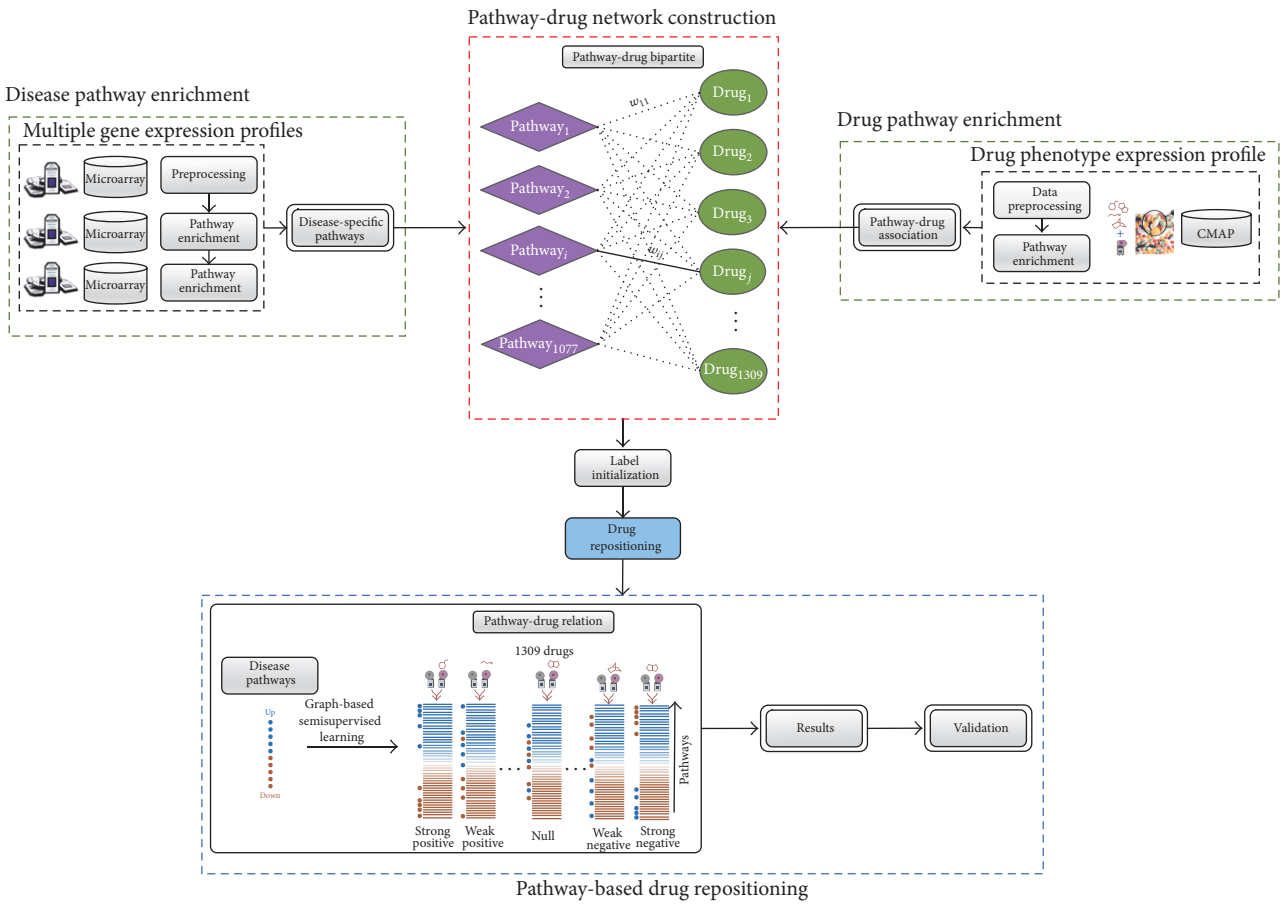


FIGURE 1: The proposed framework for drug repositioning. The proposed framework consists of several steps. First, disease-specific pathways are identified by disease pathway enrichment of multiple expression profiles for the disease of interest. Second, the drug pathway network is constructed from the drug pathway associations obtained from the drug phenotype profiles. Once the network is constructed, initial labels are assigned using disease-specific pathways and known drugs associated with the given disease. Finally, pathway-based drug repositioning is performed using semisupervised network propagation. The identified drugs are evaluated, and the final results are obtained.

less than 0.01 ($p < 0.01$) are selected as significant pathways for each expression profile, and their union is defined as "disease-specific pathways." Figure 2 presents an illustration of the integration process.

2.2. Deriving Pathway-Drug Associations from CMap Drug Phenotype Profiles. To define a pathway-drug association, pathway-drug enrichment is established from the drug phenotype expression profile (CMap: Connectivity Map) [6, 7], which contains the gene expression profiles obtained from five different cancer cell lines treated with 1309 (v2) small drug molecules, most of which are FDA-approved drugs, for a total 6100 data points representing gene expression results with control vehicle samples. The CMap data are preprocessed, batch effects are removed, and pathway enrichments are estimated by GSEA as in previous studies [11, 26, 27]. As a result, each pathway (1077) has an ES for each drug molecule (1309). The strength of the ES indicates the association degree of a pathway with a drug. As shown in Figure 3, the pathway-drug association can be represented as a 1077×1309 matrix, where the columns list the drugs and the rows list the pathways.

2.3. Pathway-Drug Network Construction. A pathway-drug network was established from the drug pathway association profile. By using the pathway-drug enrichment matrix (Figure 3), the pathway-drug bipartite graph structure $G = (U, V, E, w)$ was constructed, whose vertices can be divided into two disjoint sets: U (pathways) and V (drugs) such that every edge $e \in E$ with weight w represents the enrichment of pathway $u_i \in U$ by drug $v_j \in V$. In other words, each node in the network corresponds to a drug or pathway, and each edge corresponds to the association between them. It can be observed that drugs tend to bind with disease-specific pathways. All nodes were initially unlabeled as 0. Semisupervised learning on a network requires a small amount of labeled data with a large amount of unlabeled data.

To use the constructed bipartite graph for drug repositioning, we made following assumption as in [4]: If pharmacologically different drugs induce the same phenotype of

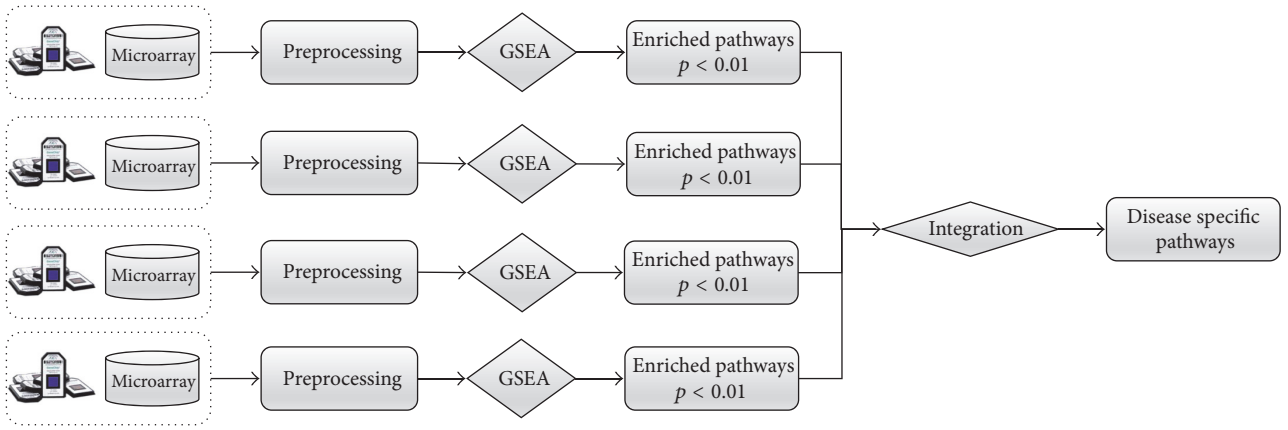


Figure 2: Disease pathway enrichment. Disease-specific pathways are identified from multiple gene expression profiles for the same disease. For each profile, enriched pathways with $p < 0.01$ are selected and integrated by taking their union. The resulting pathways are considered disease-specific pathways for the given disease.

interest, then most of molecular pathways they target must be shared. In other words, *drugs used to treat the same disease (phenotype) target similar pathways*. For example, if we have some prior knowledge on certain drugs that are used to treat a specific disease, then most of the molecular pathway they target should be similar. In Figure 4, the blue drugs (breast cancer treatment drugs) target pathway "B," and the green drugs (prostate cancer treatment drugs) target pathway "D." From this information, it is can be concluded that drug "K" can likely be used to treat prostate cancer, when the weight (ES) is high enough. This is main assumption that we make in our proposed framework for pathway-based drug repositioning. Defining the initial knowledge (or initial labels for nodes) is also one of the key steps in this work.

2.4. Label Initialization on a Pathway-Drug Network. To initialize the pathway-drug labels for the *U (pathways)* and *V (drugs)* disjoint sets, we used disease-specific pathways inferred from the multiple gene expression profiles and known treatment drugs for the given phenotype (breast cancer) were obtained from three different public resources: the Maya Clinic, Cancer Organization, and TTD. The identified disease-specific pathways were mapped to the *U (pathways) set* and labeled as 1, and the remaining pathways were labeled as 0.

For the *V (drugs)* set, a more accurate prediction is possible if we can set the labels for the drug set in the pathway-drug network using previously known information about the disease-related drugs prior to using network propagation to predict drugs associated with the disease. Therefore, we first verified known drugs used for the treatment of the disease of interest using public drug-related sources, including the Maya Clinic database, Cancer Organization database, and TTD, and then determined the labels for the drug set in the pathway-drug network. These drugs were mapped to the *V (drugs)* set and labeled as 1, and the remaining drugs were labeled as 0.

2.5. Drug Repositioning by Semisupervised Learning. Once the initial labeling of the pathway-drug network was completed, we predicted the repositioned drugs by learning the drug nodes and pathway nodes with the network propagation algorithm. The bipartite graph can be defined as $G = (V, U, E, w)$, where V and U are the node sets that are the disjoint node, in which the nodes of each node set are expressed as v and u, respectively. E is the set of edges between V and U, and w represents the weights of these edges. The weight of a specific edge is expressed as $w(v, u)$. The function for the sum of all weight values for a node can be defined as $d(v) = \sum_{(v,u)} w(v, u)$. Now, let us examine the network propagation algorithm based on the definition of the previously defined bipartite graph. First, the network propagation algorithm normalizes the weights of the bipartite graph using the following formula:

$$B = D_v^{-1/2} * W * D_u^{-1/2}. \quad (1)$$

Here, W is a matrix containing the weights of the bipartite graph, D_v and D_u are the diagonal matrices with the values of $D_{i_v i_v} = d(v)$ and $D_{i_u i_u} = d(u)$, respectively, and B is the matrix of the normalized weights. Second, network propagation is performed for the bipartite graph using formulae (2) and (3), iterating over the objective function of the graph-based semisupervised learning algorithm.

For each $v \in V$,

$$f(v)^t = (1 - \alpha)\, y(v) + \alpha \sum_{u \in U} B_{i_v i_u} f(u)^{t-1}. \quad (2)$$

For each $u \in U$,

$$f(u)^t = (1 - \alpha)\, y(u) + \alpha \sum_{v \in V} B_{i_v i_u} f(v)^{t-1}. \quad (3)$$

Here, t is the number of iterations and y is the initial label of the corresponding node. The parameter α has a value between 0 and 1 and acts to regulate the relative weight of the initial

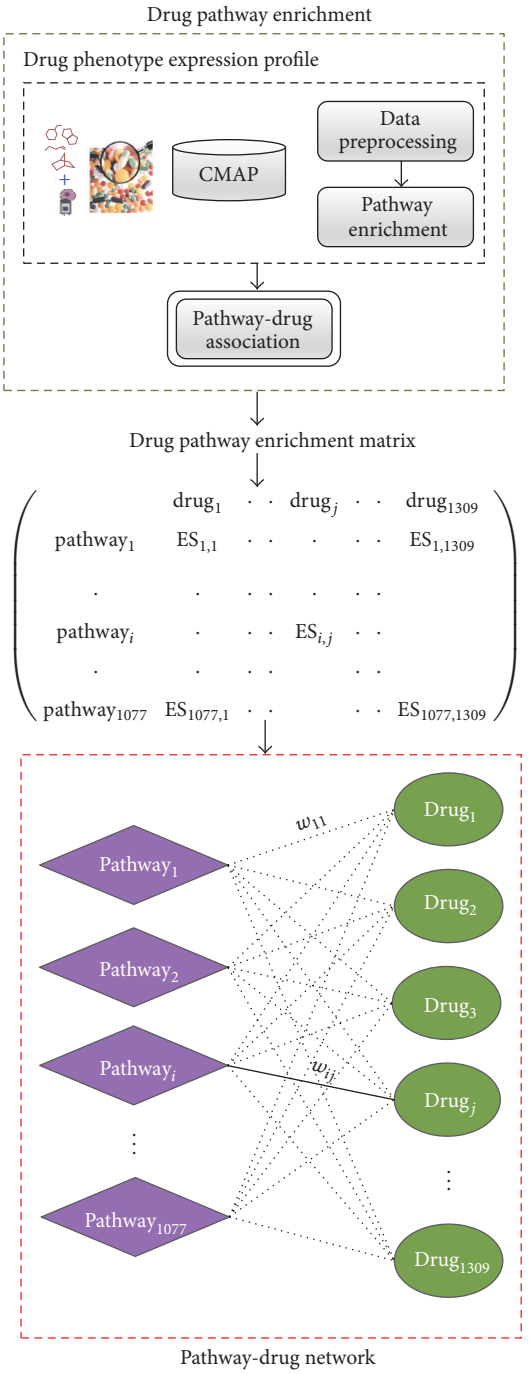


FIGURE 3: Drug pathway association and pathway-drug network. Associations between a drug and pathways are defined by drug pathway enrichment from drug phenotype expression profiles. The strength of $ES_{i,j}$ represents the enrichment of $pathway_j$ when treated with $drug_j$.

label and the learned label. $y(v)$ and $y(u)$ are the initial labels for the drugs and pathway, respectively, whereas $f(v)$ and $f(u)$ are the final label scores. Finally, network propagation is completed when the values of $f(v)$ and $f(u)$ converge.

If the network propagation algorithm is executed over the pathway-drug bipartite according to the above method, the learned drugs label scores can be obtained. As the label score of a drug increases, the drug can be considered a more promising candidate for drug repositioning for the given phenotype. Therefore, we define the values of the final drug label scores as the drug repositioning scores and use them to predict disease-associated drugs from the pathway-drug network. In addition, all obtained label scores are normalized by the Z-score using the following equation:

$$Z_{\text{drug}_i} = \frac{l_{\text{drug}_i} - \text{mean}(l)}{\text{std}(l)}, \tag{4}$$

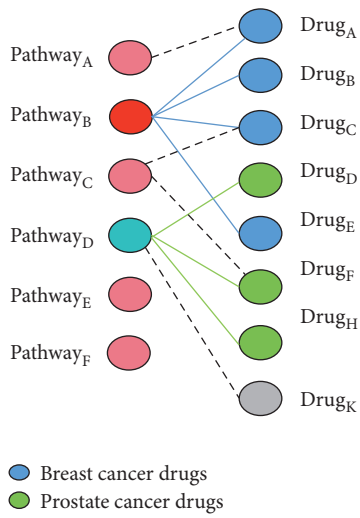


FIGURE 4: Similar drugs used for the same disease share most of the molecular pathway they target.

where l is the label score vector for all drugs and l_{drug_i} is the final label score for drug$_i$. For each drug, the corresponding p value was estimated based on the Z-score for Gaussian distribution. For more conservative results, we chose drugs with $p < 0.001$ as promising drug candidates for drug repositioning for the given disease. The selected promising drug candidates are evaluated by our validation methods and chosen for further investigation.

3. Results and Discussion

We tested our proposed framework to reposition 1309 drugs for breast cancer.

3.1. Finding Disease-Specific Pathways in Breast Cancer. To obtain breast cancer-specific pathways, we used publicly available breast cancer expression profiles (GSE15852 [28], GSE20437 [29], GSE2043 [30], and GSE2990 [31]) from the Gene Expression Omnibus (GEO) [32]. Table 1 shows the detailed characteristics of the expression profiles used in our study. Each dataset was preprocessed using RMA techniques [33] and implemented in R using the BioConductor package, which includes a large number of metadata packages appropriate for different types of microarrays. Supplementary Figure 1, in Supplementary Material available online at http://dx.doi.org/10.1155/2016/7147039, shows the results of preprocessing. For each dataset, the corresponding annotation databases were downloaded separately, and each probe was mapped to a HUGO [34] gene symbol; a probe was discarded if it did not match any symbol. In addition, if a gene had multiple probes (many-to-one), the gene expression values were averaged over the probes.

The human metabolic and signaling pathways were obtained from the Molecular Signature Database (MSigDB) [35]. As shown in Table 2, we chose the canonical pathways in the curated gene sets that contain 1077 pathways collected from KEGG [36], Reactome [37], and BioCarta (http://www.biocarta.com/).

For each dataset, a pathway was defined as breast cancer enriched by GSEA when $p < 0.01$. To integrate, the enriched pathways with nominal p values less than 0.01 ($p < 0.01$) were selected as significant pathways for each expression profile, and their union was defined as the "disease-specific pathways." Table 3 shows the number of enriched pathways for each dataset and the integrated pathways obtained by taking their union. Table 4 shows an example of enriched pathways in breast cancer by using experiment dataset (GSE2990). In the Supplementary Material, Tables 1–4 provide the GSEA analysis results for each cancer expression profile and list the identified disease-specific pathways that were used for label initialization on the pathway-drug network.

3.2. Breast Cancer Drug Repositioning Using the Proposed Approach. From the four different breast cancer expression profiles, 143 pathways were identified as significantly enriched. On the pathway-drug network, these pathways were mapped to the *U (pathways) set and* initially labeled as 1, and the remaining 934 pathways were labeled as 0. In addition, known drugs used for the treatment for breast cancer were obtained from three different public resources, the Maya Clinic, Cancer Organization, and TTD. Sixty-one drugs approved to treat breast cancer were obtained from the Maya Clinic, 49 drugs were obtained from the Cancer Organization, and 11 drugs were obtained from TTD. Next, after mapping these drugs to the drug pathway network only 10 drugs were successfully mapped. Moreover, the 10 mapped drugs (tamoxifen, letrozole, doxorubicin, vinblastine, exemestane, aminoglutethimide, methotrexate, paclitaxel, megestrol, and fulvestrant) were labeled as 1 on *V (drugs)*, whereas all remaining drugs (1299) were labeled as 0.

Once the initial labels of the pathway-drug network were chosen, we predicted promising candidates related to breast cancer using semisupervised network propagation, as shown in Figure 5. As a result, we considered 17 drugs with $p < 0.001$, as shown in Table 5, and found that 10 of them are already known drugs. The remaining seven drugs were considered as promising drug candidates for breast cancer and used for further validation to examine their association with breast cancer.

3.3. Validation of Promising Candidate Drugs. To validate the predicted drugs, we recommend the use of two different methods. Drugs that have been successfully validated by both methods are considered to be confirmed for repositioning for breast cancer.

3.3.1. Biological Validation. Biological validation was performed by manually checking the evidence in the biological literature on promising drug candidates. We manually searched for any possible indication of the repositioned drugs for breast cancer. As shown in Table 6, for each promising drug candidate, several different lines of evidence in the literature were found indicating its possible use for breast cancer. Based on these results, we concluded that six drugs of seven drugs were confirmed by biological validation for their new

Table 1: Breast cancer gene expression datasets.

#	Dataset id	Samples (control/case)	Probes	Platform	References
1	GSE15852	86 (43/43)	22283	GPL96 (HG-U133A)	Pau Ni et al. [28]
2	GSE20438	42 (18/26)	22283	GPL96 (HG-U133A)	Graham et al. [29]
3	GSE2043	286 (180/106)	22283	GPL96 (HG-U133A)	Wang et al. [30]
4	GSE2929	193 (64/129)	22283	GPL96 (HG-U133A)	Sotiriou et al. [31]

The gene expression datasets were downloaded from the NCBI Gene Expression Omnibus (GEO).

Table 2: Pathway data.

Database	Pathways	# of gene sets	URL
MSigDB (c2-canonical pathways)	KEGG	186	http://www.genome.jp/kegg/
	Reactome	674	http://www.reactome.org/
	BioCarta	217	http://www.biocarta.com/
Total		1077	

Table 3: Breast cancer disease-specific pathways for each dataset.

	GSE15852	GSE20438	GSE2043	GSE2929	Integrated pathways
# of pathways $p < 0.01$	109	7	17	21	**143**

Enriched pathways were identified by GSEA. For integration, the pathways with (p value < 0.01) were selected as significant pathways for each expression profile and their union was defined as "disease-specific pathways."

Table 4: Breast cancer pathways from GSE2990 ($p < 0.01$).

#	Name	ES	NES	p value
1	REACTOME_DEFENSINS	−0.767	−1.599	0.008230452
2	REACTOME_ORGANIC_CATION_ANION_ZWITTERION_TRANSPORT	0.771	1.668	0.004016064
3	REACTOME_G0_AND_EARLY_G1	−0.771	−1.601	0.008213553
4	REACTOME_CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	−0.601	−1.670	0
5	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	0.756	1.644	0.00617284
6	REACTOME_PECAM1_INTERACTIONS	−0.855	−1.685	0.001972387
7	REACTOME_GABA_SYNTHESIS_RELEASE_REUPTAKE_AND_DEGRADATION	0.773	1.558	0.003891051
8	REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	−0.773	−1.613	0.008438818
9	BIOCARTA_GATA3_PATHWAY	0.746	1.688	0.002004008
10	REACTOME_NEUROTRANSMITTER_RELEASE_CYCLE	0.709	1.747	0
11	BIOCARTA_G2_PATHWAY	−0.705	−1.704	0.00625
12	REACTOME_PASSIVE_TRANSPORT_BY_AQUAPORINS	−0.757	−1.718	0.001865672
13	REACTOME_PYRIMIDINE_METABOLISM	−0.702	−1.726	0.001964637
14	BIOCARTA_ACTINY_PATHWAY	−0.734	−1.749	0.001945525
15	REACTOME_SYNTHESIS_OF_GLYCOSYLPHOSPHATIDYLINOSITOL_GPI	0.703	1.699	0.001976285
16	REACTOME_ENDOGENOUS_STEROLS	−0.762	−1.694	0.005703422
17	REACTOME_GLYCOSPHINGOLIPID_METABOLISM	0.590	1.763	0.006048387

Supplementary files 2, 3, 4, and 5 provide the full pathway enrichment analysis results for the breast cancer expression profiles.

Table 5: Predicted drugs after pathway-based drug repositioning.

Drug name	Ranking score	Z-score	p value	Description
Doxorubicin*	0.999	8.935	0	It works by intercalating DNA, with the most serious adverse effect being life threatening heart damage.
Exemestane*	0.803	6.880	$2.99E-12$	Tyrosine kinase inhibitor which selectively inhibits HER2
Methotrexate*	0.708	5.892	$1.91E-09$	blocks the production of steroids derived from cholesterol and is clinically used in the treatment of Cushing's syndrome and metastatic breast cancer.
Megestrol*	0.668	5.464	$2.33E-08$	It binds to and inhibits the enzyme dihydrofolate reductase, resulting in inhibition of purine nucleotide and thymidylate synthesis and, subsequently, inhibition of DNA and RNA syntheses.
Paclitaxel*	0.646	5.235	$8.26E-08$	It binds to and stabilizes microtubules, preventing their depolymerization and so inhibiting cellular motility, mitosis, and replication.
Aminoglutethimide*	0.637	5.148	$1.32E-07$	It blocks the production of steroids derived from cholesterol and is clinically used in the treatment of Cushing's syndrome and metastatic breast cancer.
Tamoxifen*	0.634	5.113	$1.59E-07$	It is an antagonist of the estrogen receptor in breast tissue via its active metabolite, hydroxytamoxifen. In other tissues such as the endometrium, it behaves as an agonist and thus may be characterized as a mixed agonist/antagonist.
Vinblastine*	0.625	5.020	$2.59E-07$	It is an antimicrotubule drug used to treat certain kinds of cancer, including Hodgkin's lymphoma, non-small cell lung cancer, breast cancer, head and neck cancer, and testicular cancer.
Fulvestrant*	0.604	4.802	$7.86E-07$	It is drug treatment of hormone receptor-positive metastatic breast cancer in postmenopausal women with disease progression following antiestrogen therapy. It is an estrogen receptor antagonist with no agonist effects, which works by downregulating the estrogen receptor.
Letrozole*	0.579	4.539	$2.83E-06$	It is an oral nonsteroidal aromatase inhibitor for the treatment of hormonally responsive breast cancer.
MS-275**	0.530	4.023	$2.87E-05$	Entinostat, also known as SNDX-275 and MS-275, is a benzamide histone deacetylase inhibitor undergoing *clinical trials* for treatment of various cancers.
GW-8510**	0.477	3.467	0.000263332	Cyclin-dependent kinase 5 inhibitors: inhibition of dopamine transporter activity.
Camptothecin**	0.475	3.452	0.000278495	It is an alkaloid isolated from the stem wood of the Chinese tree, *Camptotheca acuminata*. This compound selectively inhibits the nuclear enzyme DNA topoisomerase.
Phenoxybenzamine**	0.461	3.303	0.000478379	It is an alpha-adrenergic antagonist with long duration of action. It has been used to treat hypertension and as a peripheral vasodilator.
Tyrphostin_AG-825**	0.447	3.159	0.000792504	It is tyrosine kinase inhibitor, which selectively inhibits HER2.
Alsterpaullone**	0.447	3.150	0.000815292	CDC2 protein kinase, antiangiogenic potential of small molecular inhibitors of cyclin-dependent kinases in vitro.
Celastrol**	0.442	3.100	0.000966191	Celastrol is a remedial ingredient isolated from the root extracts of "*Tripterygium wilfordii*" (Thunder of God vine) and "*Celastrus regelii*." In "in vitro" and "in vivo" animal experiments, celastrol exhibits antioxidant, anti-inflammatory, anticancer, and insecticidal activities.

*Known breast cancer drug. **Potential drug candidate for repositioning.

Table 6: Literature evidences for the promising drug candidates for breast cancer.

Promising drugs	Biological validation	Related literature for possible usage for breast cancer
MS-275**	**o**	(i) [45]: the potential anticancer activity of MS-275 in combination with pentoxifylline in panel of cell lines and human breast cancer xenograft model in vitro and in vivo. (ii) [46]: MS-275 sensitizes TRAIL-resistant breast cancer cells, inhibits angiogenesis and metastasis, and reverses epithelial-mesenchymal transition in vivo. (iii) [47]: HDAC inhibitors (MS-275) enhance the apoptosis-inducing potential of TRAIL in breast carcinoma.
GW-8510**	◇	(i) [48]: repositioning of a cyclin-dependent kinase inhibitor GW-8510 as a ribonucleotide reductase M2 inhibitor to treat human colorectal cancer. In addition, GW-8510 induced autophagic cell death. (ii) [49]: in cell viability tests, four candidate drugs, GW-8510, etacrynic acid, ginkgolide A, and 6-azathymine, are identified as having high inhibitory activities against cancer cells.
Camptothecin**	**o**	(i) [50]: the camptothecin targets WRN protein: mechanism and relevance in clinical breast cancer. (ii) [51]: CRLX101, an investigational camptothecin-containing nanoparticle-drug conjugate, targets cancer stem cells and impedes resistance to antiangiogenic therapy in mouse models of breast cancer. (iii) [52]: STI571 sensitizes breast cancer cells to 5-fluorouracil, cisplatin, and camptothecin in a cell type-specific manner. (iv) [53]: acquired camptothecin resistance of human breast cancer MCF-7/C4 cells with normal topoisomerase I and elevated DNA repair. (v) known as "Happy Tree" in Chinese traditional cancer treatment.
Phenoxybenzamine**	**x**	(i) Not enough evidence.
Tyrphostin_AG-825**	**o**	(i) [54]: C-Src activation by ErbB2 leads to attachment-independent growth of human breast epithelial cells. (ii) [55]: using in vivo mouse models of breast cancer: using gefitinib, ERBB1 inhibition rapidly inhibits tumor cell motility and invasion but not intravasation, whereas ERBB2 inhibition by AG825 rapidly blocks intravasation. (iii) [56]: tyrphostin AG 825 has been used in combination with hypericin-mediated photodynamic therapy (HY-PDT) for evaluating its therapeutic effects in HER2 overexpressing human breast cancer cells.
Alsterpaullone**	**o**	(i) [57]: the antitumor effects of ALP through induction of apoptosis in breast cancer and leukemia cells. Identification of alsterpaullone as a novel small molecule inhibitor to target group 3 medulloblastoma. (ii) [58]: baicalein blocked survivin expression in lung and breast cancer cells. Alsterpaullone is a CDC2 kinase inhibitor (43). Both CDC25 phosphatase and CDC2 kinase inhibitors enhanced the baicalein-induced cancer cell death.
Celastrol**	**o**	(i) [59]: anticancer effect of celastrol on human triple negative breast cancer: possible involvement of oxidative stress, mitochondrial dysfunction, apoptosis, and PI3K/Akt pathways. (ii) [60]: celastrol induces that apoptosis of breast cancer cells and inhibits invasion via downregulation of MMP-9.

** Potential drug candidate for repositioning.

usage in breast cancer treatment, with phenoxybenzamine not being confirmed.

3.3.2. Computational Evaluation on the Validation Network. In drug repositioning, it is difficult to compare and evaluate the performances of computational methods. To address this issue, several recent studies have focused on curating a comprehensive and public catalog of existing drug indications using a manual process [4].

Therefore, to develop a better evaluation method using computational methods, a validation network was constructed using information on three different relationships, drug-drug, drug-gene, and gene-gene, from the STITCH and STRING databases [38]. The drug-drug relationship information was obtained from the STITCH (v4) [39] database, which contains data on the interactions between small molecules and the edges between two chemicals that are expressed using a score between 0 and 900 defined

Table 7: Degree centrality of promising drug candidates on the validation network.

Rank	Drug name	Degree centrality
1	Tamoxifen*	0.661
2	Doxorubicin*	0.554
3	Paclitaxel*	0.536
4	Fulvestrant*	0.268
5	Methotrexate*	0.268
6	Camptothecin**	0.232
7	Letrozole*	0.214
8	Vinblastine*	0.196
9	Exemestane*	0.179
10	Megestrol*	0.125
11	Aminoglutethimide*	0.107
12	MS-275**	0.089
13	Alsterpaullone**	0.071
14	GW-8510**	0.036
15	Phenoxybenzamine**	0.036
16	Celastrol**	0.036
17	Tyrphostin_AG-825**	0.018

*Known breast cancer drug. **Potential drug candidate for repositioning.

from the chemical similarity between drugs. The drug-gene network was constructed from STITCH (for human) protein-chemical interactions with the help of the STRING database which provides 4,523,609 relationships for humans with the correlations between proteins and chemicals recorded as scores using information obtained from experimental results, text-mining, or predicted correlations. The gene-gene network was constructed from the STRING database, where A PPI network can be described as a complex system of proteins linked by interactions. Two proteins or genes that physically interact are represented as adjacent nodes connected by an edge. Each protein id (unipro id) is converted to the corresponding gene symbols using annotation databases provided in the STRING protein-protein interaction database. For computational evaluation, we have selected a maximum of 40 neighbors of drugs (17 drugs) with a weight criterion of $r > 0.4$ from the validation network derived from STITCH. The constructed validation network is illustrated in Figure 6.

To investigate the node properties in a network, network topology measurements (degree centrality and betweenness) and linkage analysis (PageRank) are often used. Degree centrality represents the number of interactions/edges/connections for a node. Biological networks are mostly scale-free networks, in which most nodes have few edges and a small number of nodes (hub) have a very high degree centrality. Betweenness is measured by the shortest paths between all nodes in the network and nodes that have the "shortest path" going through them are called bottlenecks. These hub and bottleneck nodes are topologically important and are usually functionally essential nodes (genes and drugs that have significant biological roles). Nodes connected to the hub and bottleneck node directly can also be functionally important. In addition, link analysis is a technique used to evaluate relationships (connection weights). The PageRank is a popular link analysis algorithm based on idea that a node should be significant if other significant nodes contain links to it.

By answering the following biological questions for the promising drug candidates, we identified the most promising drugs among them.

(i) Which candidate drug has an interesting/important relationship (connections) with known drugs?

(ii) Which candidate drug has the hub/bottleneck property on the validation network?

(iii) Which candidate drugs are connected to known breast cancer target genes?

For this purpose, we checked the network properties of promising drug candidates on the validation network using degree centrality, betweenness, and PageRank. Among them, the network topology measurements (degree centrality and betweenness) are designed to produce a ranking which allows indication of the most important vertices and not designed to measure the influence of neighbor nodes in general. Therefore, for better validation of promising candidates on validation network, PageRank algorithm seems to be more preferable which evaluates the nodes by considering their connection weights to the influential neighbors nodes.

From the results shown in Table 7, the popular breast cancer drug "tamoxifen" was identified as the most important hub node with degree centrality of 0.661 on the validation network. Among the promising drug candidates, *camptothecin* showed the hub node property with the highest degree centrality (0.232) among the other five (MS-275, GW-8510, phenoxybenzamine, tyrphostin_AG-825, and alsterpaullone). Table 8 shows the neighbor nodes of the camptothecin on the validation network where it has a strong chemical similarity with the known drugs doxorubicin, paclitaxel, vinblastine, and methotrexate. A close look at this relationship is shown in Figure 7(a), and this evidence seems to point to the possibility of using the camptothecin for breast cancer treatment because structurally similar drugs usually bind the same disease targets. In addition, from Table 8 and Figures 7(a) and 7(b), it can be seen that camptothecin has a strong target relation with the genes that play active role in breast cancer including TOP1, ABCB1, TOP2A, CASP3, and TP53 (neighbors) and EGFR (second-degree neighbor). TOP1 and TOP2A were reported to inhibit the breast cancer resistant proteins [40]. ABCB1 is known as prognostic factor in breast cancer patients [41]. CASP3 expression loss represents an important cell survival mechanism in breast cancer patients [42] and it inhibits the growth of breast cancer cells. EGFR was one of the first identified important targets in breast cancer, and half of breast cancer cases overexpress EGFR.

The candidate drugs MS-257 and alsterpaullone showed relatively higher degree centrality values among the remaining drugs. Table 9 and Figure 8 show the neighbor nodes relationship of MS-257 on the validation network, where it

Table 8: The neighbors of candidate drug "camptothecin" on the validation network.

Nodes	Description	Weight
TOP1	Topoisomerase (DNA) I: the reaction catalyzed by topoisomerases leads to the conversion of one topological isomer of DNA to another.	0.999
CASP3	Caspase 3: apoptosis-related cysteine peptidase: it is involved in the activation cascade of caspases responsible for apoptosis execution.	0.965
TP53	Tumor protein p53: it acts as a tumor suppressor in many tumor types and induces growth arrest or apoptosis depending on the physiological circumstances and cell type.	0.965
Doxorubicin*	It is a drug used in cancer chemotherapy; it works by intercalating DNA, with the most serious adverse effect being life threatening heart damage.	0.890
ABCG2	ATP-binding cassette, subfamily G (WHITE), member 2; xenobiotic transporter that may play an important role in the exclusion of xenobiotics from the brain. It may be involved in brain-to-blood efflux. It appears to play a major role in the multidrug resistance phenotype of several cancer cell lines.	0.873
CDK1	Cyclin-dependent kinase 1: it plays a key role in the control of the eukaryotic cell cycle. It is required in higher cells for entry into S phase and mitosis. p34 is a component of the kinase complex that phosphorylates the repetitive C-terminus of RNA polymerase II.	0.846
ABCB1	ATP-binding cassette, subfamily B (MDR/TAP), member 1; energy-dependent efflux pump responsible for decreased drug accumulation in multidrug-resistant cells.	0.843
BCL2	B-cell CLL/lymphoma 2: it suppresses apoptosis in a variety of cell systems including factor-dependent lymphohematopoietic and neural cells. It regulates cell death by controlling the mitochondrial membrane permeability. It appears to function in a feedback loop system with caspases. It inhibits caspase activity either by preventing the release of cytochrome c from the mitochondria and/or by binding to the apoptosis-activating factor (APAF-1).	0.820
Paclitaxel*	It binds to and inhibits the enzyme dihydrofolate reductase, resulting in inhibition of purine nucleotide and thymidylate synthesis and, subsequently, inhibition of DNA and RNA syntheses.	0.812
CDK2	Cyclin-dependent kinase 2; involved in the control of the cell cycle; interacting with cyclins A, B1, B3, D, or E. Activity of CDK2 is maximal during S phase and G2.	0.754
Vinblastine*	An antimicrotubule drug used to treat certain kinds of cancer, including Hodgkin's lymphoma, non-small cell lung cancer, breast cancer, head and neck cancer, and testicular cancer.	0.560
Methotrexate*	It blocks the production of steroids derived from cholesterol and is clinically used in the treatment of Cushing's syndrome and metastatic breast cancer.	0.554
TOP2A	Topoisomerase (DNA) II alpha 170 kDa; control of topological states of DNA by transient breakage and subsequent rejoining of DNA strands.	0.431

*Known breast cancer drug.

Table 9: The neighbors of candidate drug "MS-257" on the validation network.

Genes	Description	Weight
HDAC1	Histone deacetylase 1; responsible for the deacetylation of lysine residues on the N-terminal part of the core histones (H2A, H2B, H3, and H4). Histone deacetylation gives a tag for epigenetic repression and plays an important role in transcriptional regulation, cell cycle progression, and developmental events.	0.987
TP53	Tumor protein p53: it acts as a tumor suppressor in many tumor types and induces growth arrest or apoptosis depending on the physiological circumstances and cell type.	0.831
CASP3	Caspase 3, apoptosis-related cysteine peptidase; involved in the activation cascade of caspases responsible for apoptosis execution.	0.827
CCND1	cyclin D1; essential for the control of the cell cycle at the G1/S (start) transition.	0.822
CYP3A4	Cytochrome P450, family 3, subfamily A, polypeptide 4; cytochromes P450 are a group of heme-thiolate monooxygenases.	0.433

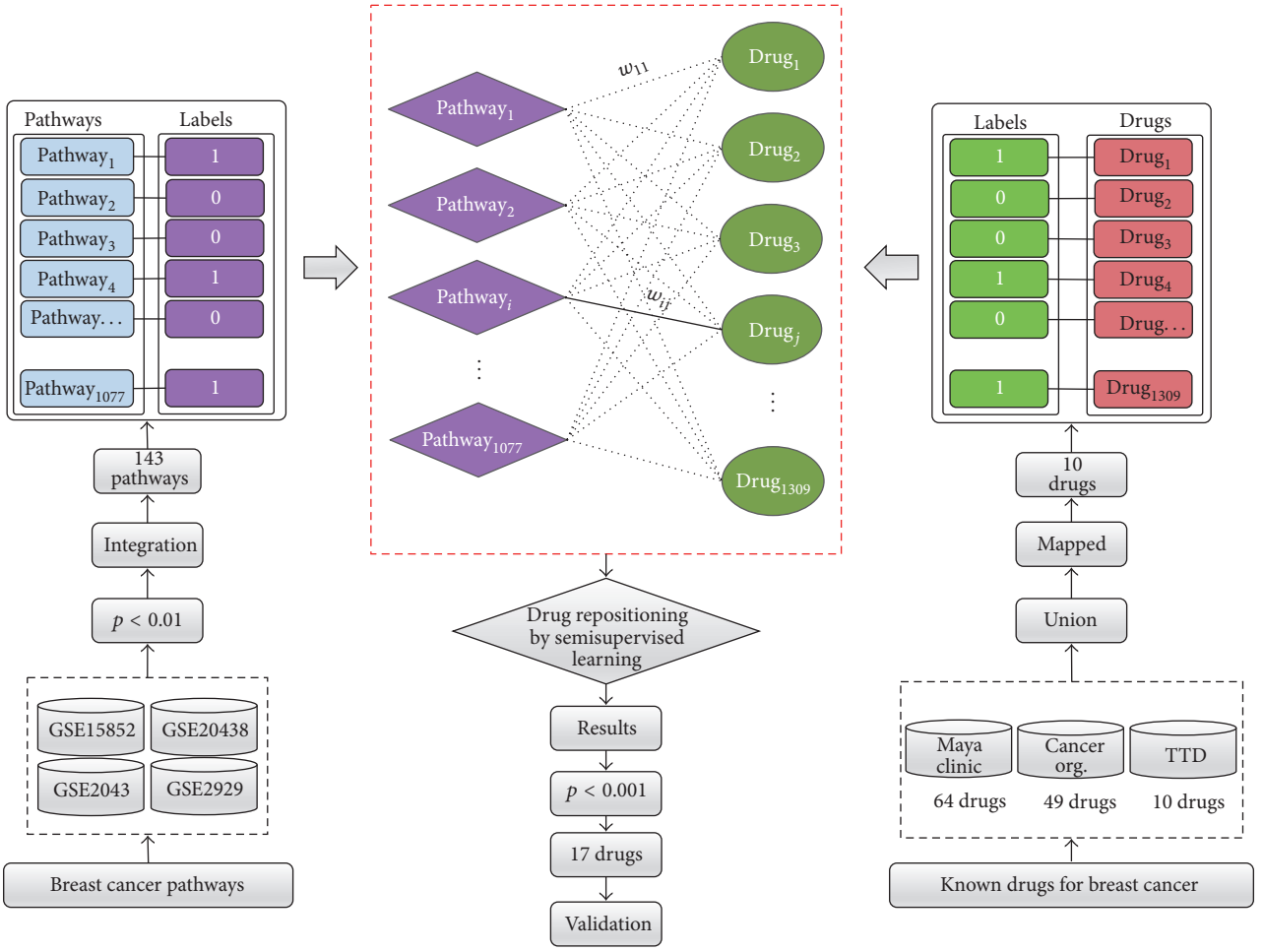


Figure 5: Breast cancer drug repositioning. Ten known drugs approved to treat breast cancer were obtained from the Maya Clinic, Cancer Org, and TTD. A total of 143 breast cancer-specific pathways were identified from multiple breast cancer expression profiles. Successfully mapped pathways and drugs were labeled as 1. Once labels were initialized on the pathway-drug network, we repositioned drugs for breast cancer using semisupervised learning. Predicted drugs with $p < 0.001$ were considered promising candidate drugs, and their associations with breast cancer were investigated using two different validation methods.

has strong target relationships with the genes HDAC1, TP53, CASP3, CCND1, and CYP3A4. Overexpression of HDCA1 represents clinicopathological indicators of disease progression in human breast cancer [43]. CCDN1 was reported to be a therapeutic target in breast cancer [44], and it has an indirect relationship with breast cancer susceptibility gene BRCA1. The betweenness results are summarized in Table 10. Among promising drug candidates only camptothecin and MS-275 showed some bottleneck node properties. Tamoxifen was defined as the most important bottleneck drug for breast cancer. Finally, we evaluated the connection weights of candidate drugs on the validation network using PageRank algorithm. We chose the alpha parameter as 0.85, which is the most commonly used value for this parameter with original Google PageRank algorithm. As shown in Table 11, camptothecin (0.257), alsterpaullone (0.102), and MS-275 (0,088) exhibited higher ranking scores than the other promising candidate drugs.

From the evidences shown above, we concluded that *camptothecin*, *MS-257*, and *alsterpaullone* exhibited the strongest network property evidences for breast cancer on the validation network. In general, all of the promising candidates successfully passed the computational evaluation on the network.

After performing biological and computational evaluations of the promising candidate drugs, we selected *camptothecin* as the most promising candidate because it was the most successful in both evaluation processes. For MS-278, GW-85, AG825, alsterpaullone, and celastrol, there was strong literature evidence with a reasonable network property. Thus, as shown in Figure 9, camptothecin, MS-278, alsterpaullone, GW-85, and AG825 and were validated as repositioned drugs and indicated for further investigation in breast cancer treatment.

4. Summary

We introduced a new systematic framework for disease-specific drug repositioning from integrated gene expression profiles on a pathway-drug network constructed from drug

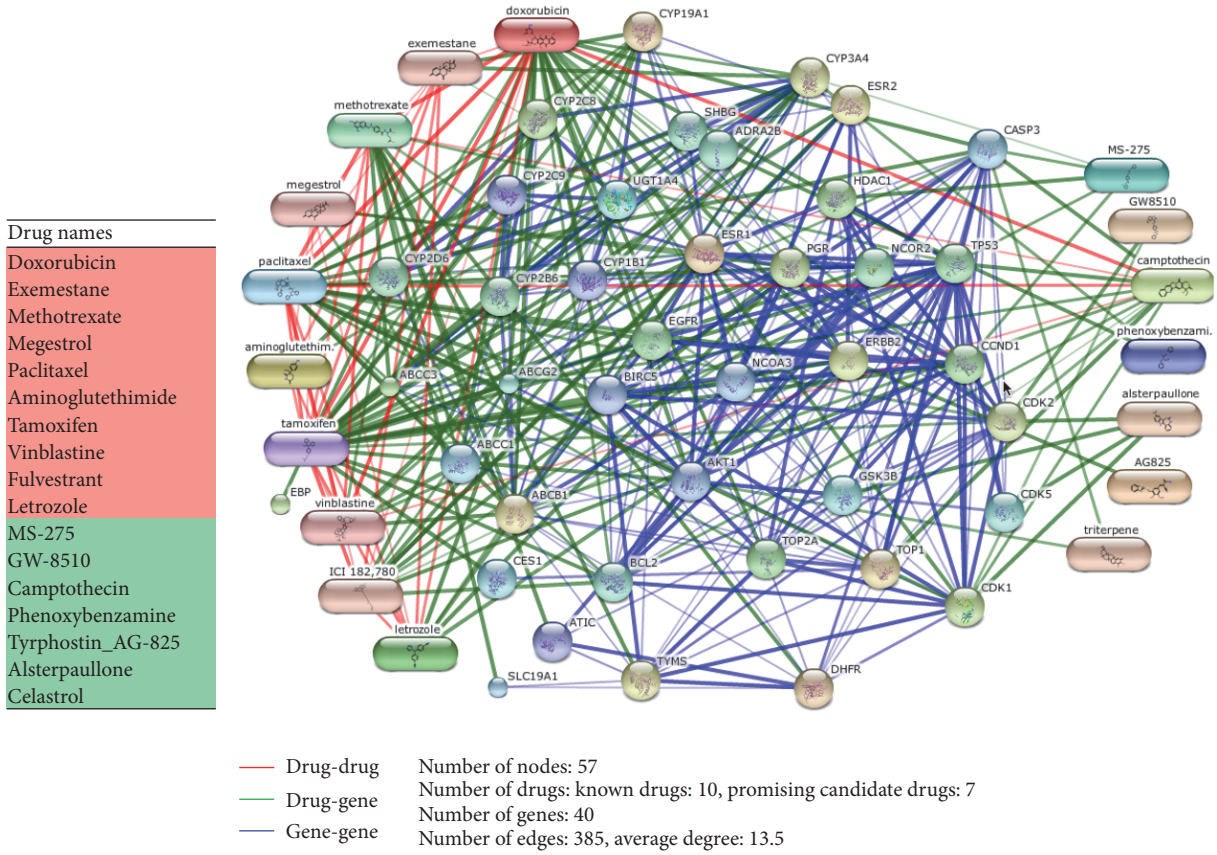
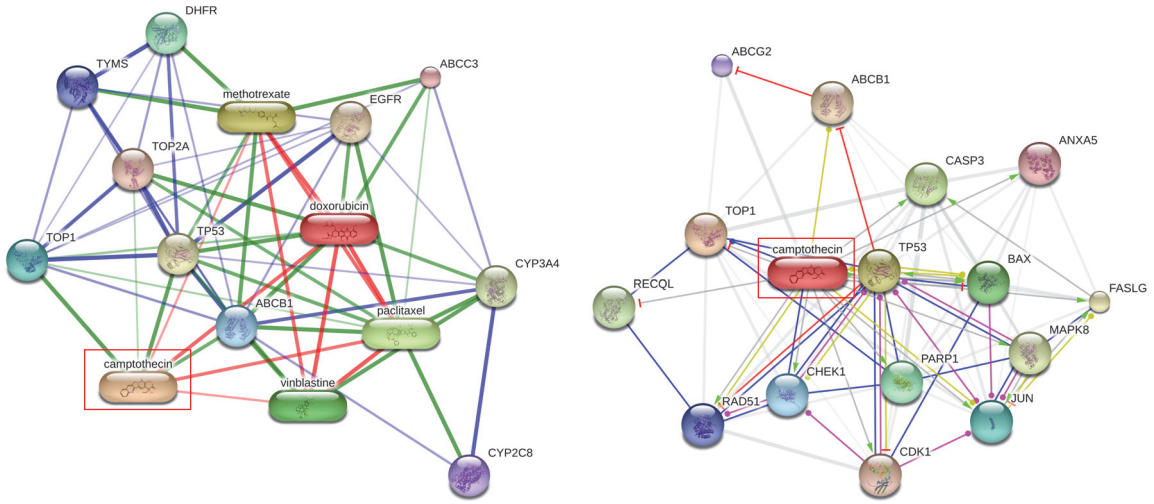
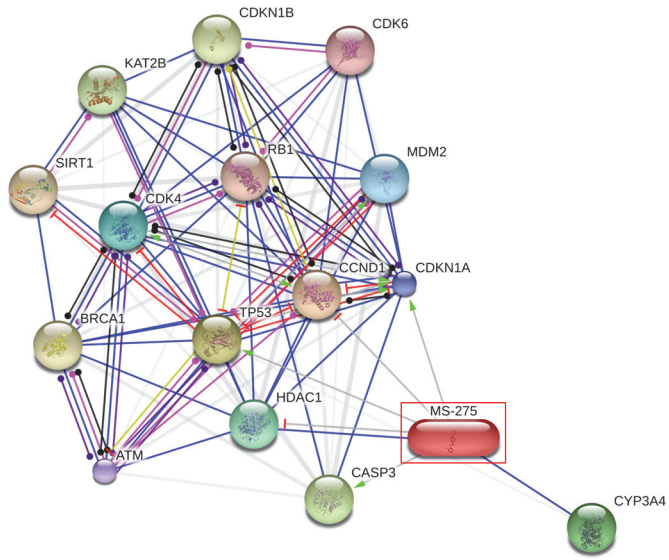


Figure 6: Known drugs and promising drug candidates on the validation network. The validation network for 16 drugs was constructed from STITCH. Each node is a drug or a gene. The green edges represent drug-gene interactions, and the red edges indicate drug-drug interactions; the blue edges represent gene-gene relationships obtained from STRING. Wider edges reflect stronger relationships between nodes. For easier implementation and visualization, a maximum of 40 neighbors of drugs (17 nodes) with a weight criterion of $r > 0.4$ were selected. As indicated in the figure, some drugs have significant topological features on the validation network.



(a) (b)

Figure 7: The candidate drug camptothecin on the validation network. (a) Camptothecin has a strong relationship (chemical similarity) with known breast cancer drugs: doxorubin, paclitaxel, vinblastine, and methotrexate. (b) Camptothecin has direct target relationship with the genes playing active roles in breast cancer including TOP1, ABCB1, TOP2A, CASP3, and TP53 (neighbors). Moreover, it has an indirect relationship with the breast cancer target gene EGFR.



Directly connected genes

Genes	Description	Weight
HDAC1	Histone deacetylase 1; responsible for the deacetylation of lysine residues on the N-terminal part of the core histones (H2A, H2B, H3, and H4). Histone deacetylation gives a tag for epigenetic repression and plays an important role in transcriptional regulation, cell cycle progression, and developmental events	0.987
TP53	Tumor protein p53; it acts as a tumor suppressor in many tumor types; it induces growth arrest or apoptosis depending on the physiological circumstances and cell type	0.831
CASP3	Caspase 3, apoptosis-related cysteine peptidase; it is involved in the activation cascade of caspases responsible for apoptosis execution	0.827
CCND1	Cyclin D1: it is essential for the control of the cell cycle at the G1/S (start) transition	0.822
CYP3A4	Cytochrome P450, family 3, subfamily A, polypeptide 4; cytochromes P450 are a group of heme-thiolate monooxygenases	0.433

Neighbor genes

Genes	Description	Weight	Activation	Inhibition	Binding	Phenotype	Catalysis	Posttrans. m	Reaction	Expression
CDK4	Cyclin-dependent kinase; probably involved in the control of the cell cycle	0.999	•		•				•	•
MDM2	mdm2 p53 binding protein homolog (mouse), inhibiting TP53/p53 and TP73/p73 cell cycle	0.999	•		•			•		•
CDKN1A	Cyclin-dependent kinase inhibitor 1A (p21, Cip1), whose role is mediated by p53/TP53 as an inhibitor of cellular proliferation in response to DNA damage	0.999	•		•				•	•
ATM	Ataxia telangiectasia mutated; serine/threonine protein kinase	0.999			•		•	•		•
CDK6	Cyclin-dependent kinase 6; probably involving the control of cell	0.999			•					
RB1	Retinoblastoma 1; key regulator of entry into cell division that acts as a tumor suppressor	0.999			•		•			•
SIRT1	Sirtuin (silent mating type information regulation 2 homolog) 1 (*S. cerevisiae*); NAD-dependent	0.999			•			•		•
BRCA1	Breast cancer 1, early onset; the BRCA1-BARD1 heterodimer coordinates a diverse range of cell pathways such as DNA damage repair, ubiquitination, and transcriptional regulation to maintain genomic stability	0.999			•					
CDKN1B	Cyclin-dependent kinase inhibitor 1B (p2, Kip1); important regulator of cell cycle progression	0.999			•				•	•
KAT2B	K(lysine) acetyltransferase 2B; it has histone acetyl transferase activity with core histones and nucleosome core particles	0.999			•			•		

Figure 8: The candidate drug MS-275 on the validation network. MS-275 has a strong target relationship with the breast cancer genes HDAC1, TP53, CASP3, CCND1, and CYP3A4. Furthermore, it has an indirect relationship with the well-known breast cancer gene BRCA1.

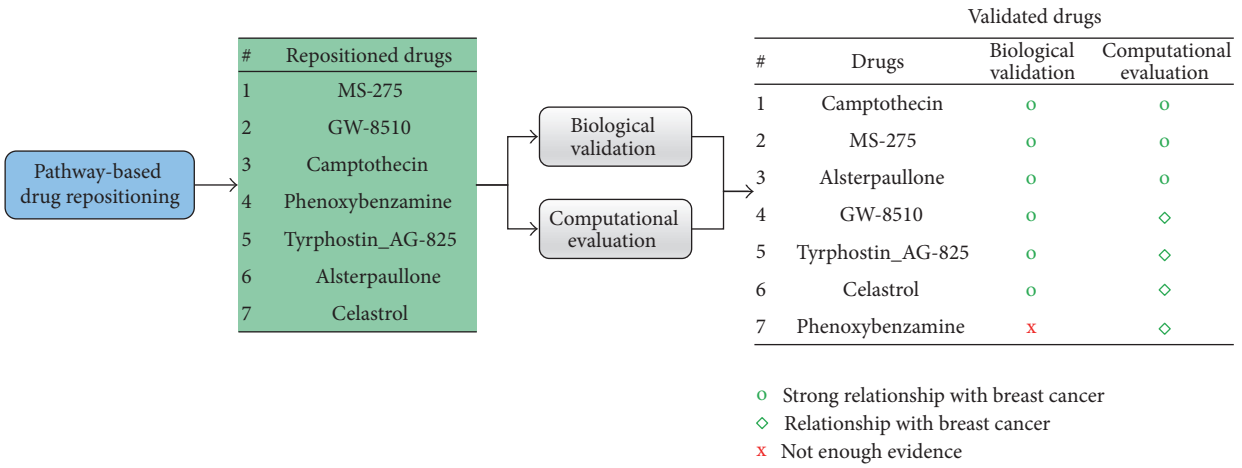


Figure 9: Validated drugs. Candidate drugs with successful results for both the biological validation and computational evaluation are considered repositioned drugs for breast cancer.

Table 10: Betweenness of promising drug candidates on the validation network.

Rank	Drug name	Betweenness
1	Tamoxifen*	172
2	Paclitaxel*	37
3	Doxorubicin*	32
4	Camptothecin**	13
5	Exemestane*	13
6	Fulvestrant*	12
7	Methotrexate*	10
8	Vinblastine*	6
9	Megestrol*	6
10	Aminoglutethimide*	2
11	Letrozole*	2
12	MS-275**	2
13	GW-8510**	0
14	Phenoxybenzamine**	0
15	Tyrphostin_AG-825**	0
16	Alsterpaullone**	0
17	Celastrol**	0

*Known breast cancer drug. **Potential drug candidate for repositioning.

Table 11: PageRank of promising drug candidates on the validation network ($\alpha = 0.85$).

Rank	Drug name	Ranking score
1	Tamoxifen*	0.990
2	Doxorubicin*	0.692
3	Paclitaxel*	0.663
4	Methotrexate*	0.373
5	Fulvestrant*	0.316
6	Camptothecin**	0.257
7	Letrozole*	0.252
8	Vinblastine*	0.235
9	Exemestane*	0.176
10	Aminoglutethimide*	0.108
11	Alsterpaullone**	0.102
12	Megestrol*	0.101
13	MS-275**	0.088
14	Phenoxybenzamine**	0.080
15	GW-8510**	0.026
16	Celastrol**	0.023
17	Tyrphostin_AG-825**	0.010

*Known breast cancer drug. **Potential drug candidate for repositioning.

phenotype expression profiles (CMap) using semisupervised learning. The proposed pathway-based drug repositioning process showed encouraging results when using four different disease expression profiles to predict candidate drugs for disease-specific repositioning.

Two different methods were employed to evaluate the repositioned drugs. The drugs that passed both evaluation methods successfully were considered the most promising drugs to target breast cancer. As a result, several drugs, including camptothecin, MS-275, alsterpaullone, GW-8510, AG 825, and celastrol were identified as possible drugs to be repositioned to treat breast cancer, and these results are supported by multiple lines of evidence in the public literature. Specifically, *camptothecin* was the most promising drug candidate because it showed a high network property on the validation network and was supported by evidence in the literature.

Despite the interesting results, our method for drug repositioning was developed and validated in only using integrated mRNA gene expression profiles. However, the strategy can be easily improved to include other experimental data types, such as RNA-seq, miRNA, DNA-methylation, and single nucleotide polymorphism (SNP) information. Finally, the increasing number of genomic and pharmaceutical databases necessitates the further development of the method to identify new drugs and targets for rare cancer

subtypes, develop personalized medicine, and design targeted cancer therapies.

Competing Interests

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

Acknowledgments

This study was supported by the BK21 Plus project funded by the Ministry of Education, Korea (21A20131600011).

References

[1] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[2] T. I. Oprea and J. Mestres, "Drug repurposing: far beyond new targets for old drugs," *The AAPS Journal*, vol. 14, no. 4, pp. 759–763, 2012.

[3] F. Napolitano, Y. Zhao, V. M. Moreira et al., "Drug repositioning: a machine-learning approach through data integration," *Journal of Cheminformatics*, vol. 5, no. 1, pp. 1–9, 2013.

[4] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings in Bioinformatics*, vol. 17, no. 1, pp. 2–12, 2016.

[5] S. J. Swamidass, "Mining small-molecule screens to repurpose drugs," *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 327–335, 2011.

[6] J. Lamb, E. D. Crawford, D. Peck et al., "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[7] J. Lamb, "The connectivity map: a new tool for biomedical research," *Nature Reviews Cancer*, vol. 7, no. 1, pp. 54–60, 2007.

[8] F. Iorio, R. Tagliaferri, and D. Di Bernardo, "Identifying network of drug mode of action by gene expression profiling," *Journal of Computational Biology*, vol. 16, no. 2, pp. 241–251, 2009.

[9] 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.

[10] Y. Silberberg, A. Gottlieb, M. Kupiec, E. Ruppin, and R. Sharan, "Large-scale elucidation of drug response pathways in humans," *Journal of Computational Biology*, vol. 19, no. 2, pp. 163–174, 2012.

[11] F. Iorio, R. Bosotti, E. Scacheri et al., "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 33, pp. 14621–14626, 2010.

[12] J. A. Parkkinen and S. Kaski, "Probabilistic drug connectivity mapping," *BMC Bioinformatics*, vol. 15, article 113, 2014.

[13] J. Yu, P. Putcha, and J. M. Silva, "Recovering drug-induced apoptosis subnetwork from connectivity map data," *BioMed Research International*, vol. 2015, Article ID 708563, 11 pages, 2015.

[14] A. Pujol, R. Mosca, J. Farrés, and P. Aloy, "Unveiling the role of network and systems biology in drug discovery," *Trends in Pharmacological Sciences*, vol. 31, no. 3, pp. 115–123, 2010.

[15] H. Chen, H. Zhang, Z. Zhang, Y. Cao, and W. Tang, "Network-based inference methods for drug repositioning," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 130620, 7 pages, 2015.

[16] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1091–D1097, 2014.

[17] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 2005.

[18] D. Emig, A. Ivliev, O. Pustovalova et al., "Drug target prediction and repositioning using an integrated network-based approach," *PLoS ONE*, vol. 8, no. 4, Article ID e60618, 2013.

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

[20] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network," *Nature Biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[21] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnology*, vol. 25, no. 2, pp. 197–206, 2007.

[22] F. Tan, R. Yang, X. Xu et al., "Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity," *Molecular BioSystems*, vol. 10, no. 5, pp. 1126–1138, 2014.

[23] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D684–D688, 2008.

[24] F. Zhu, Z. Shi, C. Qin et al., "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1128–D1136, 2012.

[25] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, "Key issues in conducting a meta-analysis of gene expression microarray datasets," *PLoS Medicine*, vol. 5, no. 9, p. e184, 2008.

[26] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort, and P. Bork, "Drug-induced regulation of target expression," *PLoS Computational Biology*, vol. 6, no. 9, Article ID e1000925, 2010.

[27] F. Napolitano, F. Sirci, D. Carrella, and D. di Bernardo, "Drug-set enrichment analysis: a novel tool to investigate drug mode of action," *Bioinformatics*, vol. 32, no. 2, pp. 235–241, 2016.

[28] I. B. Pau Ni, Z. Zakaria, R. Muhammad et al., "Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context," *Pathology Research and Practice*, vol. 206, no. 4, pp. 223–228, 2010.

[29] K. Graham, A. De Las Morenas, A. Tripathi et al., "Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile," *British Journal of Cancer*, vol. 102, no. 8, pp. 1284–1293, 2010.

[30] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[31] C. Sotiriou, P. Wirapati, S. Loi et al., "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 2006.

[32] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data

repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[33] 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.

[34] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, "Genenames.org: the HGNC resources in 2015," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1079–D1085, 2015.

[35] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[36] M. Kanehisa, M. Araki, S. Goto et al., "KEGG for linking genomes to life and the environment," *Nucleic Acids Research*, vol. 36, no. 1, pp. D480–D484, 2008.

[37] L. Matthews, G. Gopinath, M. Gillespie et al., "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Research*, vol. 37, no. 1, pp. D619–D622, 2009.

[38] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. 1, pp. D447–D452, 2015.

[39] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild et al., "STITCH 4: integration of protein-chemical interactions with user data," *Nucleic Acids Research*, vol. 42, no. D1, pp. D401–D407, 2014.

[40] H. Jandu, K. Aluzaite, L. Fogh et al., "Molecular characterization of irinotecan (SN-38) resistant human breast cancer cell lines," *BMC Cancer*, vol. 16, no. 1, article 34, 2016.

[41] H.-J. Kim, S.-A. Im, B. Keam et al., "ABCB1 polymorphism as prognostic factor in breast cancer patients treated with docetaxel and doxorubicin neoadjuvant chemotherapy," *Cancer Science*, vol. 106, no. 1, pp. 86–93, 2015.

[42] E. Devarajan, A. A. Sahin, J. S. Chen et al., "Down-regulation of caspase 3 in breast cancer: a possible mechanism for chemoresistance," *Oncogene*, vol. 21, no. 57, pp. 8843–8851, 2002.

[43] B. M. Müller, L. Jana, A. Kasajima et al., "Differential expression of histone deacetylases HDAC1, 2 and 3 in human breast cancer—overexpression of HDAC2 and HDAC3 is associated with clinicopathological indicators of disease progression," *BMC Cancer*, vol. 13, no. 1, article 215, pp. 1–8, 2013.

[44] E. A. Musgrove, C. E. Caldon, J. Barraclough, A. Stone, and R. L. Sutherland, "Cyclin D as a therapeutic target in cancer," *Nature Reviews Cancer*, vol. 11, no. 8, pp. 558–572, 2011.

[45] S. Nidhyanandan, T. S. Boreddy, K. B. Chandrasekhar, N. D. Reddy, N. M. Kulkarni, and S. Narayanan, "Phosphodiesterase inhibitor, pentoxifylline enhances anticancer activity of histone deacetylase inhibitor, MS-275 in human breast cancer in vitro and in vivo," *European Journal of Pharmacology*, vol. 764, pp. 508–519, 2015.

[46] R. K. Srivastava, R. Kurzrock, and S. Shankar, "MS-275 sensitizes TRAIL-resistant breast cancer cells, inhibits angiogenesis and metastasis, and reverses epithelial-mesenchymal transition in vivo," *Molecular Cancer Therapeutics*, vol. 9, no. 12, pp. 3254–3266, 2010.

[47] T. R. Singh, S. Shankar, and R. K. Srivastava, "HDAC inhibitors enhance the apoptosis-inducing potential of TRAIL in breast carcinoma," *Oncogene*, vol. 24, no. 29, pp. 4609–4623, 2005.

[48] Y. Y. Hsieh, C. J. Chou, H. L. Lo, and P. M. Yang, "Repositioning of a cyclin-dependent kinase inhibitor GW8510 as a ribonucleotide reductase M2 inhibitor to treat human colorectal cancer," *Cell Death Discovery*, vol. 2, Article ID 16027, 2016.

[49] F.-H. Chung, Y.-R. Chiang, A.-L. Tseng et al., "Functional Module Connectivity Map (FMCM): a framework for searching repurposed drug compounds for systems treatment of cancer and an application to colorectal adenocarcinoma," *PLoS ONE*, vol. 9, no. 1, Article ID e86299, 2014.

[50] R. A. Shamanna, H. Lu, D. L. Croteau et al., "Camptothecin targets WRN protein: mechanism and relevance in clinical breast cancer," *Oncotarget*, vol. 7, no. 12, pp. 13269–13284, 2016.

[51] S. J. Conley, T. L. Baker, J. P. Burnett et al., "CRLX101, an investigational camptothecin-containing nanoparticle-drug conjugate, targets cancer stem cells and impedes resistance to antiangiogenic therapy in mouse models of breast cancer," *Breast Cancer Research and Treatment*, vol. 150, no. 3, pp. 559–567, 2015.

[52] J. T. Sims, S. Ganguly, L. S. Fiore, C. J. Holler, E.-S. Park, and R. Plattner, "STI571 sensitizes breast cancer cells to 5-fluorouracil, cisplatin and camptothecin in a cell type-specific manner," *Biochemical Pharmacology*, vol. 78, no. 3, pp. 249–260, 2009.

[53] A. Fujimori, M. Gupta, Y. Hoki, and Y. Pommier, "Acquired camptothecin resistance of human breast cancer MCF-7/C4 cells with normal topoisomerase I and elevated DNA repair," *Molecular Pharmacology*, vol. 50, no. 6, pp. 1472–1478, 1996.

[54] L. G. Sheffield, "C-Src activation by ErbB2 leads to attachment-independent growth of human breast epithelial cells," *Biochemical and Biophysical Research Communications*, vol. 250, no. 1, pp. 27–31, 1998.

[55] D. Kedrin, J. Wyckoff, P. J. Boimel et al., "ERBB1 and ERBB2 have distinct functions in tumor cell invasion and intravasation," *Clinical Cancer Research*, vol. 15, no. 11, pp. 3733–3739, 2009.

[56] Sigma-Aldrich, *Tyrphostin AG 825: Highlights of Prescribing Information*, Sigma-Aldrich, 2016.

[57] C. C. Faria, S. Agnihotri, S. C. Mack et al., "Identification of alsterpaullone as a novel small molecule inhibitor to target group 3 medulloblastoma," *Oncotarget*, vol. 6, no. 25, pp. 21718–21729, 2015.

[58] J.-I. Chao, W.-C. Su, and H.-F. Liu, "Baicalein induces cancer cell death and proliferation retardation by the inhibition of CDC2 kinase and survivin associated with opposite role of p38 mitogen-activated protein kinase and AKT," *Molecular Cancer Therapeutics*, vol. 6, no. 11, pp. 3039–3048, 2007.

[59] S. Shrivastava, M. K. Jeengar, V. S. Reddy, G. B. Reddy, and V. G. M. Naidu, "Anticancer effect of celastrol on human triple negative breast cancer: possible involvement of oxidative stress, mitochondrial dysfunction, apoptosis and PI3K/Akt pathways," *Experimental and Molecular Pathology*, vol. 98, no. 3, pp. 313–327, 2015.

[60] C. Mi, H. Shi, J. Ma, L. Z. Han, J. J. Lee, and X. Jin, "Celastrol induces the apoptosis of breast cancer cells and inhibits their invasion via downregulation of MMP-9," *Oncology Reports*, vol. 32, no. 6, pp. 2527–2532, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 3164624, 8 pages
http://dx.doi.org/10.1155/2016/3164624

Research Article

Horizontally Transferred Genetic Elements in the Tsetse Fly Genome: An Alignment-Free Clustering Approach Using Batch Learning Self-Organising Map (BLSOM)

Ryo Nakao,[1,2] Takashi Abe,[3] Shunsuke Funayama,[3] and Chihiro Sugimoto[4,5,6]

[1] *Unit of Risk Analysis and Management, Hokkaido University Research Center for Zoonosis Control, Kita 20, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan*
[2] *Laboratory of Parasitology, Department of Disease Control, Graduate School of Veterinary Medicine, Hokkaido University, Kita 18, Nishi 9, Kita-Ku, Sapporo, Hokkaido 060-0818, Japan*
[3] *Graduate School of Science & Technology, Niigata University, No. 8050, Igarashi 2-no-cho, Nishi-ku, Niigata 950-2181, Japan*
[4] *Division of Collaboration and Education, Hokkaido University Research Center for Zoonosis Control, Kita 20, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan*
[5] *Global Station for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, Kita 20, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan*
[6] *Department of Disease Control, School of Veterinary Medicine, University of Zambia, P.O. Box 32379, Lusaka, Zambia*

Correspondence should be addressed to Chihiro Sugimoto; sugimoto@czc.hokudai.ac.jp

Received 30 June 2016; Revised 26 September 2016; Accepted 8 November 2016

Academic Editor: Farit M. Afendi

Copyright © 2016 Ryo Nakao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tsetse flies (*Glossina* spp.) are the primary vectors of trypanosomes, which can cause human and animal African trypanosomiasis in Sub-Saharan African countries. The objective of this study was to explore the genome of *Glossina morsitans morsitans* for evidence of horizontal gene transfer (HGT) from microorganisms. We employed an alignment-free clustering method, that is, batch learning self-organising map (BLSOM), in which sequence fragments are clustered based on the similarity of oligonucleotide frequencies independently of sequence homology. After an initial scan of HGT events using BLSOM, we identified 3.8% of the tsetse fly genome as HGT candidates. The predicted donors of these HGT candidates included known symbionts, such as *Wolbachia*, as well as bacteria that have not previously been associated with the tsetse fly. We detected HGT candidates from diverse bacteria such as *Bacillus* and Flavobacteria, suggesting a past association between these taxa. Functional annotation revealed that the HGT candidates encoded loci in various functional pathways, such as metabolic and antibiotic biosynthesis pathways. These findings provide a basis for understanding the coevolutionary history of the tsetse fly and its microbes and establish the effectiveness of BLSOM for the detection of HGT events.

1. Introduction

Tsetse flies (*Glossina* spp.) are the primary vectors of trypanosome parasites; they cause human African trypanosomiasis (or sleeping sickness) and animal African trypanosomiasis (or nagana) in Sub-Saharan African countries. The flies harbour three maternally transmitted endosymbionts, *Wigglesworthia glossinidia*, *Sodalis glossinidius,* and *Wolbachia pipientis*, which influence host physiology. For example, *Wigglesworthia* provides essential nutrients, such as vitamins, to the host [1, 2] and influences host immune maturation [3]. Although the precise role of *Sodalis* in the tsetse fly is not clear, it appears to influence various host properties, such as longevity and susceptibility to trypanosome infections [4–6]. In many arthropod species [7, 8], *Wolbachia* induces strong cytoplasmic incompatibility, which was also observed in the tsetse fly [9]. In addition to these common bacteria, a recent microbial population analysis using a deep-sequencing approach revealed other facultative microorganisms from diverse bacterial families in the guts of tsetse flies, though

their relative abundances were very low compared to that of the symbiont *Wigglesworthia* [10].

In addition to the parasitism of bacterial organisms themselves, partial genome sequences of *Wolbachia* are incorporated into the tsetse fly genome. Initially, Doudoumis et al. reported the incorporation of short fragments of three *Wolbachia* genes (16S rDNA, *fbpA,* and *wsp*) in the genomes of laboratory and natural *Glossina morsitans morsitans* (*Gmm*) populations [11]. Subsequently, a whole-genome sequencing project revealed large insertions of the *Wolbachia* genome in the *Gmm* genome via horizontal gene transfer (HGT) events [12, 13]. These insertions were identified by extracting *Wolbachia*-specific sequences from whole-genome Sanger sequencing reads and pyrosequencing data based on nucleotide homology with the complete genome sequences of three *Wolbachia* strains (wMel, wRi and wBm) [13]. Fluorescent *in situ* hybridisation analyses further confirmed the presence of these insertions in *Gmm* on the two sex chromosomes (X and Y) and the supernumerary B-chromosome [12, 13].

HGT elements can be detected by two main methods: phylogeny-based and composition-based methods [14]. The first method relies on sequence alignments; HGT is identified when the position of a query sequence in a tree does not match that of a reference phylogeny. Although this approach is robust, the frequency of HGT events may be underestimated, especially when there is a lack of information on donor sequences [15]. The second method relies on nucleotide compositional features, such as G+C content, nucleotide frequencies, or codon usage [16–18], and theoretically does not require sequence homology. Batch learning self-organising map (BLSOM) is an alignment-free clustering method that generates a map independently of the order in which data are input via a learning process [19, 20]. This method enables the clustering of genomic sequence fragments based on the similarity of oligonucleotide frequencies, without any other taxonomical information; it has been successfully applied in genomic and metagenomic studies [20–22].

The objective of this study was to characterise HGT from microorganisms in the genome of *Gmm* using the alignment-free clustering method BLSOM. Using BLSOM, we detected a number of HGT candidates from diverse origins. In a comparison of the results for HGT from *Wolbachia* between methods, there was a high level of agreement between BLSOM and BLASTn, a homology-based approach. Based on functional annotation, these potential HGT elements encoded loci in various functional pathways.

2. Materials and Methods

2.1. Genome Sequences. The tsetse fly (*Gmm*) genome (Accession number CCAG010000000) and all prokaryotic sequences identified to the species level (n = 5,600) were obtained from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). When the number of undetermined nucleotides (Ns) exceeded 10% of the window size (5 kb), the sequence was omitted from the analysis. When the number of Ns was less than 10%, the oligonucleotide frequencies were normalised to the length without Ns and included in the analysis.

2.2. Batch Learning Self-Organising Map. G+C% is a fundamental value for the phylogenetic classification of microbial genomes, including viral genomes, but it cannot differentiate a wide variety of genomes. Oligonucleotide composition can distinguish species, even those with the same G+C%, because it varies substantially among genomes; accordingly, it is referred to as a "genome signature" [23]. Multivariate analyses, such as factor correspondence analysis and principal component analysis (PCA), are useful to investigate variation in gene sequences [24]. However, the clustering power of conventional multivariate analyses is inadequate when massive quantities of sequence data from a wide variety of genomes are analysed collectively. Kohonen's self-organising map (SOM) is a powerful tool for clustering and visualising high-dimensional data vectors on a two-dimensional plane [25, 26]. To handle codon and oligonucleotide composition as high-dimensional data vectors, we modified the conventional SOM to develop the BLSOM [19, 20], which is suitable for genome sequence analyses and high-performance parallel computing. The initial weight vectors were defined by PCA, instead of random values, based on the finding that PCA can classify gene sequences into groups of known biological categories. Weight vectors ($\mathbf{w}_{ij}$) were arranged in the two-dimensional lattice denoted by i (=0, 1,. . ., I–1) and j (=0, 1,. . ., J–1). Weight vectors ($\mathbf{w}_{ij}$) were set and updated as described previously [19, 27]. A BLSOM program suitable for PC cluster systems is available on our website (http://bioinfo.ie.niigata-u.ac.jp/?BLSOM).

2.3. Detection of HGT Candidates in the Tsetse Fly Genome and Prediction of Their Origins Using BLSOM. To identify HGT candidates in the tsetse fly genome derived from prokaryotes, two types of large-scale BLSOM were used, that is, Tsetse+Prokaryotes- and Genus-BLSOM, using all genome sequences deposited in DDBJ/ENA/GenBank. A Tsetse+Prokaryotes-BLSOM was constructed with a degenerate tetranucleotide composition for all 5 kb sequences derived from tsetse fly genome sequences of longer than 5 kb plus 5,600 identified prokaryotes for which at least 10 kb of sequence was available from DDBJ/ENA/GenBank. The degenerate tetranucleotide composition was the composition of degenerate sets in which a pair of complementary tetranucleotides was added (e.g., ATGC and GCAT). To obtain more detailed phylotype information for the prokaryotic sequences, Genus-BLSOM was constructed for each phylum derived from 5,600 identified prokaryotes.

For tsetse fly contigs of longer than 5 kb (9,710 contigs), a 5 kb window with a 1 kb step was used to obtain 303,250 segments (Figure 1, Step 1), which were mapped to Tsetse+Prokaryotes-BLSOM by identifying the lattice point with the minimum Euclidian distances in the multidimensional space (Figure 1, Step 2). For every lattice point at which tsetse fly genomic segments were mapped to prokaryotic territories, the most abundant phylum was identified, and

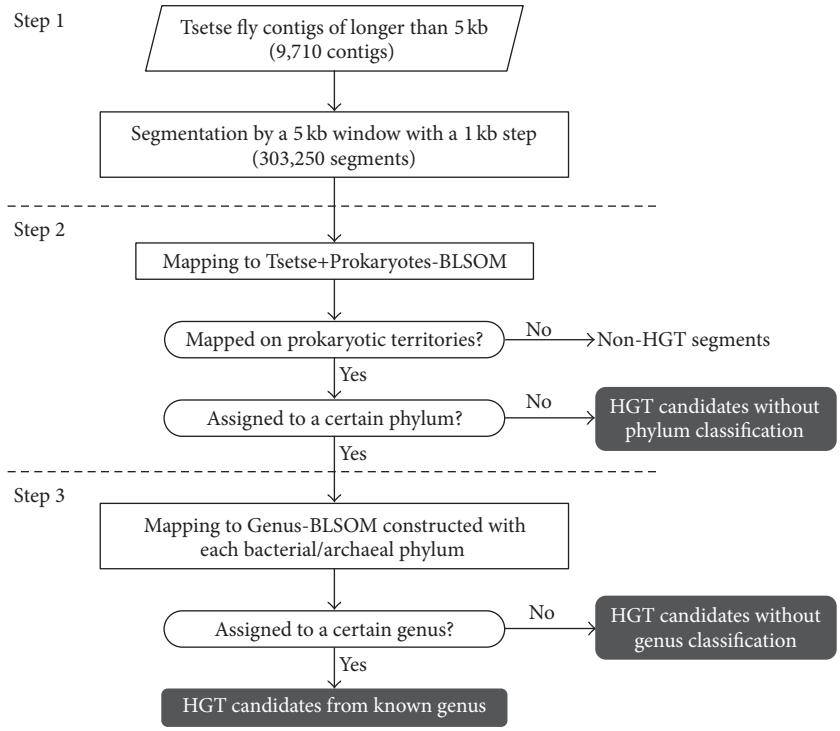


Figure 1: Workflow for data processing and BLSOM analysis.

the mapped tsetse fly genomic segments were tentatively assumed to belong to the phylum. Finally, when the most abundant phylum in more than 40% of the segments derived from a single tsetse fly contig was the same, the tsetse fly contig was assigned to this phylum by BLSOM. To identify the phylogenetic origin of the tsetse fly genomic segments that were mapped to the prokaryotic territories on Tsetse+Prokaryotes-BLSOM, they were successively mapped on Genus-BLSOM (Figure 1, Step 3). Similar stepwise mappings of tsetse fly genomic segments on BLSOMs constructed with sequences from more detailed phylogenetic categories (e.g., genera) were conducted.

2.4. Detection of HGT Candidates by BLASTn. To detect HGT candidates derived from *Wolbachia*, a local BLASTn search was conducted with the contig sequences of the tsetse fly that exceeded 5 kb against the NCBI nonredundant nucleotide database. When more than 1 kb of the sequence showed similarity with *Wolbachia* with top-hits and an *E*-value threshold of 1×10^{-5}, the contigs were considered HGT candidates derived from *Wolbachia*.

2.5. Functional Classification of HGT Candidates. The sequence fragments with prokaryote origins were functionally annotated using KEGG (Kyoto Encyclopedia of Genes and Genomes) mapping [28] with the KAAS web server (http://www.genome.jp/tools/kaas/) [29]. KEGG Orthology (KO) assignments were obtained using the single-directional best-hit method. The organisms included in the analysis were as follows (based on IDs): hsa, dme, ath, sce, pfa, eco, sty, hin, pae, nme, hpy, rpr, mlo, bsu, sau, lla, spn, cac, mge, mtu, ctr, bbu, syn, aae, mja, afu, pho, and ape. The organisms in the database are listed on the KAAS web server (http://www.genome.jp/kaas-bin/kaas_org).

3. Results and Discussion

3.1. Detection of HGT Candidates Using BLSOM. Of 303,250 sequence segments obtained from the tsetse genome, we found that 11,524 sequences (3.8%) clustered with reads from prokaryotes and thus were HGT candidates according to Tsetse+Prokaryotes-BLSOM (Figure 2). These sequences were distributed across 2,960 different contigs, corresponding to 30.48% of all contigs. We assigned the most sequences to the phylum Firmicutes (n = 758), followed by the phyla Bacteroidetes (n = 370), Alphaproteobacteria (n = 90), and Gammaproteobacteria (n = 23) (Table 1). We did not assign 1,671 contigs to phyla owing in part to the presence of HGTs from multiple phyla within the same contig. It is also possible that these candidates were introduced by ancient HGT events, and their oligonucleotide compositions drifted over time [30], limiting the use of composition-based methods for classification.

We performed further characterisation of donor sequences to the genus level using Genus-BLSOM (Figure 3). The results of four dominant phyla are summarised in Table 2. We assigned the most sequences to the genus *Bacillus* (n = 239). The second most highly represented origin was the class Flavobacteria (n = 187), which we were unable to classify to the genus level owing to coclustering with genome sequences that lacked genus information, followed by the genera

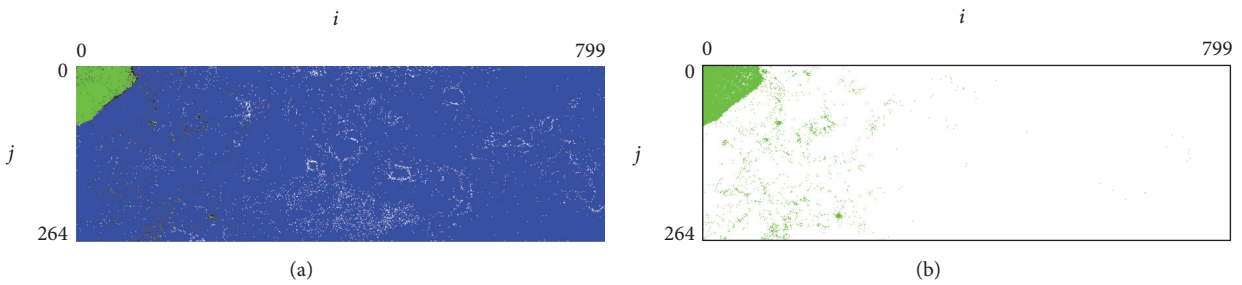


Figure 2: Tsetse+Prokaryotes-BLSOM. (a) BLSOM using the degenerate tetranucleotide set for the tsetse fly plus 5,600 identified prokaryotes. Lattice points that include the sequences from the tsetse fly are indicated in green, those that contain no genomic sequences are indicated in white, and those containing sequences from a prokaryote are indicated in blue. Lattice points that include both tsetse fly- and prokaryote-sequences are shown in black. (b) Distribution of tsetse fly genome sequences. Only green lattice points are shown.

Table 1: Origins of HGT candidates at the phylum level.

Phylum	Number of contigs
Actinobacteria	10
Alphaproteobacteria	90
Aquificae	2
Bacteroidetes	370
Betaproteobacteria	4
Crenarchaeota	1
Cyanobacteria	1
Epsilonproteobacteria	16
Euryarchaeota	12
Firmicutes	758
Fusobacteria	1
Gammaproteobacteria	23
Spirochetes	1
Unassigned	1,671
Total	2,960

Prediction was obtained using Tsetse+Prokaryotes-BLSOM.

Staphylococcus ($n = 134$), *Enterococcus* ($n = 83$), *Wolbachia* ($n = 56$), *Polaribacter* ($n = 37$), and *Listeria* ($n = 31$). The HGT candidates associated with the genera *Wigglesworthia* and *Sodalis*, which are common endosymbionts of the tsetse fly, were not detected using BLSOM. Most tsetse flies are heavily infected with *Wigglesworthia*; its abundance reaches over 99% in natural *Gmm* populations [10]. The lack of genome sequences associated with *Wigglesworthia* may support the high quality of tsetse fly genome sequences, since some genome data are contaminated by symbiont genomes, which can lead to the false-positive detection of HGT events [31]. This result also suggests that the symbiosis between *Wigglesworthia* and the tsetse fly was recent, as suggested by its genome features [32]. Nonetheless, we cannot exclude the possibility of bacterial genome contaminations in the tsetse fly genome since diverse bacteria exist in tsetse fly [10] and their sequences might not have been completely removed during the genome assembly process.

The high frequency of HGT candidates from the genus *Bacillus* suggests that there was a strong association between the tsetse fly and *Bacillus* in the past. Members of the genus *Bacillus* are ubiquitous in nature and have been isolated from diverse environments such as water, soil, plants, animals, and air [33]. Some species, such as *Bacillus thuringiensis*, have been well studied as agents of biological control of arthropods [34]. Kaaya and Darji infected several *Bacillus* species, including *B. thuringiensis*, to the adult *Gmm* and found that the mortality of *Gmm* was depending on the bacterial species [35], indicating that some *Bacillus* species may have infected *Gmm* persistently without adverse effect on the hosts and served as HGT donors. In fact, in a microbiota analysis of one tsetse fly species, *Glossina fuscipes fuscipes*, the bacteria belonging to the genus *Bacillus* were found dominant in a culture-dependent manner [36].

In contrast, there is no report on the relationship between the tsetse fly and Flavobacteria, which was identified as a second dominant donor of HGT candidates in this study (Table 2). Flavobacteria are symbionts in several arthropods [37–43], which indicates a high probability of the proliferation of this group of bacteria in arthropod hosts including tsetse fly. A comparative genome analysis of a flavobacterial symbiont (*Blattabacterium* strain Bge) in the omnivorous German cockroach (*Blattella germanica*) suggested that it plays roles in nutrient supply to the host, amino acid catabolism, and nitrogen excretion [41]. Flavobacterial symbionts in the ladybird (*Coleomegilla maculata*), and coccinellid beetle (*Adonia variegata*) induce male-killing [38, 39], in which male progeny in infected females die during embryogenesis. This phenomenon is widely recognised in other bacteria, such as *Wolbachia*, *Rickettsia*, *Arsenophonus*, *Spiroplasma,* and *Cardinium* [44]. Hurst et al. proposed that two male-killing symbionts cannot coexist at equilibrium in a single host species based on an observational study of the two-spot ladybird (*Adalia bipunctata*) infected with two symbionts, *Rickettsia* and *Spiroplasma* [39]. The presence of *Wolbachia* in the tsetse fly and HGT elements from *Wolbachia* in the tsetse fly genome [12, 13] may explain the absence of Flavobacteria in current tsetse fly populations.

3.2. HGT Candidates Derived from Wolbachia. We performed a BLASTn analysis to detect HGT candidates derived from *Wolbachia*. For 38 contigs, we detected sequence homology with *Wolbachia* sequences based on the criteria described earlier. Of these 38 contigs, we identified 36 as

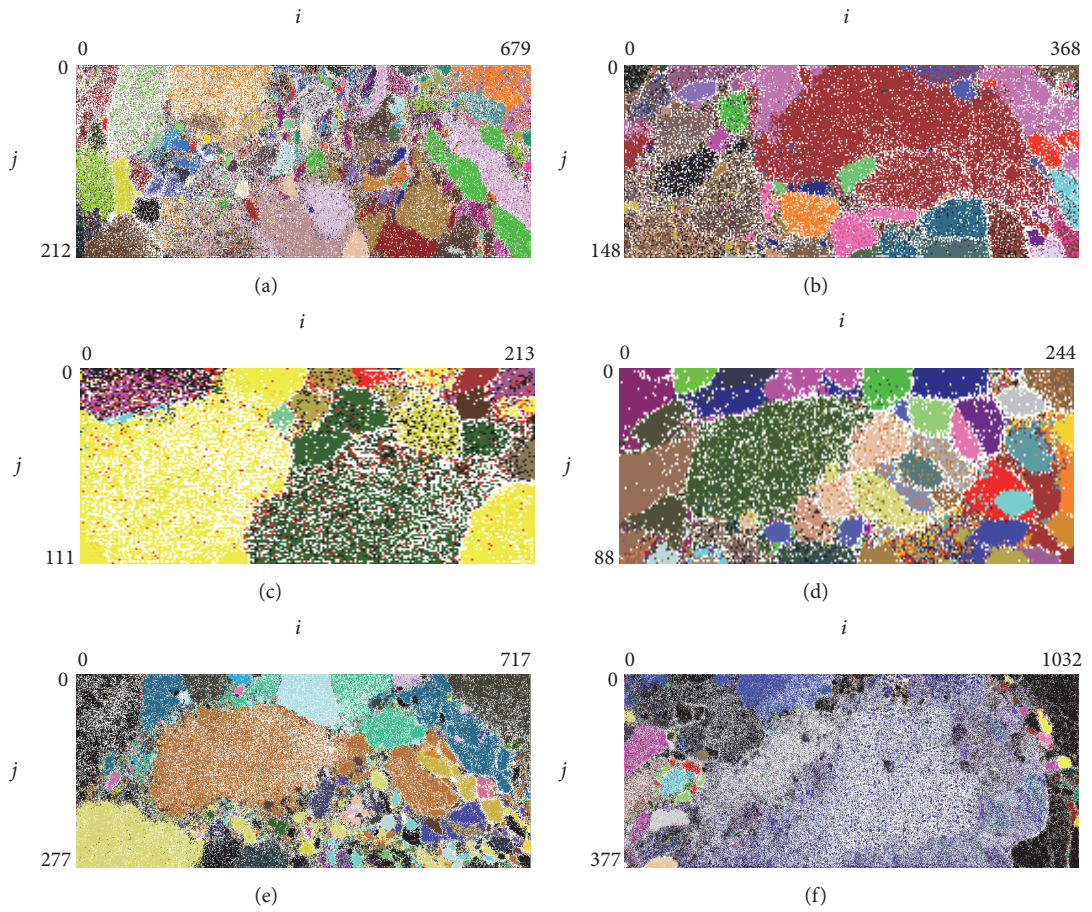


Figure 3: Genus-BLSOM. (a) Alphaproteobacteria. (b) Bacteroidetes. (c) Epsilonproteobacteria. (d) Euryarchaeota. (e) Firmicutes. (f) Gammaproteobacteria. Lattice points that include sequences from more than one genus are indicated in black, those including no sequences are indicated in white, and those including sequences from a single genus are indicated in individual color.

HGT candidates from *Wolbachia* using BLSOM, while we assigned the remaining two contigs to the genus *Rickettsia* using BLSOM. Accordingly, we observed a high level of agreement (36/38 = 94.7%) between the two approaches, indicating that the analytical sensitivity of BLSOM is at least comparable to that of BLASTn. These results also suggest that the genetic elements derived from *Wolbachia* were recently introduced to the tsetse fly genome. Based on a comparison between the BLSOM and BLASTn results, we observed 20 contigs that were only identified as HGT candidates using BLSOM. There may be as-yet unidentified *Wolbachia* strains, explaining the failure to identify HGT candidates based on sequence homology. In fact, for short regions (i.e., 338, 492, 497, and 497 bp) of four contigs identified as HGT candidates from *Wolbachia* only using BLSOM, we detected sequence homology with a *Wolbachia* endosymbiont of *Gmm* (Accession number AWUH01000121) with an *E*-value of $<1 \times 10^{-5}$.

Genetic elements related to *Wolbachia* that were presumably obtained by HGT have been detected in the genomes of multiple arthropod species, including the adzuki bean beetle (*Callosobruchus chinensis*) [45, 46], a fruit fly (*Drosophila ananassae*) [47], parasitoid wasps (*Nasonia* spp.) [47], the pea aphid (*Acyrthosiphon pisum*) [48], two mosquito species (*Aedes aegypti* and *Aedes mascarensis*) [49, 50], and the longicorn beetle (*Monochamus alternatus*) [51]. Some HGT events could be explained by nuclear-phage recombination, as proposed previously [49], but further studies are needed to determine the specific mechanisms of transfer. Since BLSOM can be used to detect *Wolbachia*-derived HGTs that exhibit low sequence similarity with known *Wolbachia* strains, it provides an alternative method with which to explore the mechanisms of HGT. The application of BLSOM to an increasing number of eukaryotic genomes will reveal the diversity and frequency of *Wolbachia*-derived HGTs in other arthropods, including vectors of medical and veterinary importance.

3.3. Functional Classification of HGT Candidates. We mapped all of the HGT candidates identified using Tsetse+Prokaryotes-BLSOM (n = 11,524) to the KEGG pathway. The KO included 317 biological pathways. The predicted pathways were mainly related to "Metabolic pathways" (193 molecules), "Biosynthesis of secondary metabolites" (75 molecules), and "Biosynthesis of antibiotics" (50 molecules). These results suggested that the HGT candidates have the

Table 2: Origins of HGT candidates at the genus level.

Phylum	Genus[1]	Number of contigs
Alphaproteobacteria	*Anaplasma*	4
	Bartonella	8
	Ehrlichia	1
	Neorickettsia	8
	Rickettsia	5
	Wolbachia	56
	Unassigned	8
Bacteroidetes	*Bacteroides*	4
	Cytophaga	1
	Dyadobacter	1
	Flavobacteria[2]	187
	Flavobacterium	12
	Kordia	18
	Leadbetterella	3
	Mucilaginibacter	1
	Paludibacter	9
	Pedobacter	1
	Polaribacter	37
	Prevotella	10
	Psychroflexus	4
	Spirosoma	4
	Unassigned	78
Firmicutes	*Clostridium*	4
	Bacillus	239
	Enterococcus	83
	Epulopiscium	4
	Erysipelothrix	1
	Geobacillus	3
	Lactobacillus	4
	Lactococcus	1
	Leuconostoc	17
	Listeria	31
	Lysinibacillus	7
	Oenococcus	1
	Paenibacillus	1
	Peptoniphilus	1
	Staphylococcus	134
	Streptococcus	9
	Thermoanaerobacter	3
	Turicibacter	2
	Veillonella	3
	Unassigned	210
Gammaproteobacteria	*Acinetobacter*	4
	Enterobacter	6
	Escherichia	1
	Haemophilus	1
	Shewanella	1
	Vibrio	1
	Xylella	1
	Unassigned	8

[1]Prediction was obtained using Genus-BLSOM. [2]Classification to the class level was obtained.

potential to affect a large number of metabolic activities; however, further analyses are essential to demonstrate the active transcription of HGT-acquired genes using transcriptomics or gene-specific reverse transcription-PCR; such analyses can provide initial evidence for the functional importance of HGT-acquired genes [52]. In general, genes transferred to host genomes are pseudogenised via the acquisition of mutations, including insertions and deletions [46, 52, 53]. Unfortunately, since we could not employ RNA sequencing data into our analysis, it is not clear to what extent the detected HGT candidates have been pseudogenised. Nonetheless, active transcription of HGT-acquired genes has been detected in recipient hosts, such as a *Wolbachia*-derived gene in the *Aedes albopictus* C6/36 cell line [54]. Moreover, an increasing number of studies suggests that HGT-acquired genes facilitate the establishment of obligate mutualistic relationships between arthropods and their symbionts [55]. Analyses of the functional roles of HGT-acquired genes may improve our understanding of the complex interactions between the tsetse fly, microbes, and pathogens.

4. Conclusions

We investigated the use of BLSOM to detect HGT candidates in the tsetse fly genome. Using BLSOM, we successfully detected a number of HGT candidates from diverse bacterial origins. The HGT candidates represented 3.8% of the tsetse fly genome. The predicted donors of these HGT elements included *Wolbachia*, a well-known symbiont of the tsetse fly. In addition, using BLSOM, we identified HGT candidates from bacteria that have not previously been associated with the tsetse fly. We observed the HGT candidates from diverse bacteria such as *Bacillus* and Flavobacteria, suggesting a strong past association between these taxa. In a comparison between BLASTn and BLSOM results for the detection of HGT candidates from *Wolbachia*, the analytical sensitivity of BLSOM was at least comparable to that of the sequence homology-based approach. Furthermore, BLSOM can be used to detect HGT elements from organisms with low similarity with currently available sequences. These data obtained using BLSOM provide a basis for understanding the coevolutionary history of the tsetse fly and its microbes.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Authors' Contributions

Ryo Nakao and Takashi Abe contributed equally to this work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) (25850195) and (A) (15H05633), for Scientific Research (C) (26330327), and for Scientific Research on Innovative Areas (16H06429, 16K21723, and 16H06431). The computation was done in part with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

References

[1] G. M. Attardo, C. Lohs, A. Heddi, U. H. Alam, S. Yildirim, and S. Aksoy, "Analysis of milk gland structure and function in *Glossina morsitans*: milk protein production, symbiont populations and fecundity," *Journal of Insect Physiology*, vol. 54, no. 8, pp. 1236–1242, 2008.

[2] R. Pais, C. Lohs, Y. Wu, J. Wang, and S. Aksoy, "The obligate mutualist *Wigglesworthia glossinidia* influences reproduction, digestion, and immunity processes of its host, the tsetse fly," *Applied and Environmental Microbiology*, vol. 74, no. 19, pp. 5965–5974, 2008.

[3] B. L. Weiss, R. Mouchotte, R. V. M. Rio et al., "Interspecific transfer of bacterial endosymbionts between tsetse fly species: infection establishment and effect on host fitness," *Applied and Environmental Microbiology*, vol. 72, no. 11, pp. 7013–7021, 2006.

[4] C. Dale and S. C. Welburn, "The endosymbionts of tsetse flies: manipulating host-parasite interactions," *International Journal for Parasitology*, vol. 31, no. 5-6, pp. 628–631, 2001.

[5] A. Geiger, S. Ravel, R. Frutos, and G. Cuny, "*Sodalis glossinidius* (Enterobacteriaceae) and vectorial competence of *Glossina palpalis gambiensis* and *Glossina morsitans morsitans* for *Trypanosoma congolense* Savannah type," *Current Microbiology*, vol. 51, no. 1, pp. 35–40, 2005.

[6] A. Geiger, S. Ravel, T. Mateille et al., "Vector competence of *Glossina palpalis gambiensis* for *Trypanosoma brucei* s.l. and genetic diversity of the symbiont *Sodalis glossinidius*," *Molecular Biology and Evolution*, vol. 24, no. 1, pp. 102–109, 2007.

[7] J. H. Werren, L. Baldo, and M. E. Clark, "*Wolbachia*: master manipulators of invertebrate biology," *Nature Reviews Microbiology*, vol. 6, no. 10, pp. 741–751, 2008.

[8] A. Saridaki and K. Bourtzis, "*Wolbachia*: more than just a bug in insects genitals," *Current Opinion in Microbiology*, vol. 13, no. 1, pp. 67–72, 2010.

[9] U. Alam, J. Medlock, C. Brelsfoard et al., "*Wolbachia* symbiont infections induce strong cytoplasmic incompatibility in the tsetse fly *glossina morsitans*," *PLoS Pathogens*, vol. 7, no. 12, Article ID e1002415, 2011.

[10] E. Aksoy, E. L. Telleria, R. Echodu et al., "Analysis of multiple tsetse fly populations in Uganda reveals limited diversity and species-specific gut microbiota," *Applied and Environmental Microbiology*, vol. 80, no. 14, pp. 4301–4312, 2014.

[11] V. Doudoumis, G. Tsiamis, F. Wamwiri et al., "Detection and characterization of *Wolbachia* infections in laboratory and natural populations of different species of tsetse flies (genus *Glossina*)," *BMC Microbiology*, vol. 12, supplement 1, article S3, 2012.

[12] International Glossina Genome Initiative, "Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis," *Science*, vol. 344, no. 6182, pp. 380–386, 2014.

[13] C. Brelsfoard, G. Tsiamis, M. Falchetto et al., "Presence of extensive *Wolbachia* symbiont insertions discovered in the genome of its host *Glossina morsitans morsitans*," *PLoS Neglected Tropical Diseases*, vol. 8, no. 4, Article ID e2728, 2014.

[14] O. Adato, N. Ninyo, U. Gophna, and S. Snir, "Detecting horizontal gene transfer between closely related taxa," *PLoS Computational Biology*, vol. 11, no. 10, Article ID e1004408, 2015.

[15] J. Tamames and A. Moya, "Estimating the extent of horizontal gene transfer in metagenomic sequences," *BMC Genomics*, vol. 9, article 136, 2008.

[16] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis," *Gene*, vol. 238, no. 1, pp. 143–155, 1999.

[17] S. Garcia-Vallvé, A. Romeu, and J. Palau, "Horizontal gene transfer in bacterial and archaeal complete genomes," *Genome Research*, vol. 10, no. 11, pp. 1719–1725, 2000.

[18] Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori, "Biased biological functions of horizontally transferred genes in prokaryotic genomes," *Nature Genetics*, vol. 36, no. 7, pp. 760–766, 2004.

[19] 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.

[20] 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.

[21] 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.

[22] 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, no. 5, pp. 1003–1015, 2013.

[23] S. Karlin, "Global dinucleotide signatures and analysis of genomic heterogeneity," *Current Opinion in Microbiology*, vol. 1, no. 5, pp. 598–610, 1998.

[24] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé, "Codon catalog usage and the genome hypothesis," *Nucleic Acids Research*, vol. 8, no. 1, pp. r49–r62, 1980.

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

[26] 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.

[27] S. Kanaya, Y. Kudo, T. Abe, T. Okazaki, C. D. Carpio, and T. Ikemura, "Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome," *Genome Informatics*, vol. 9, pp. 369–371, 1998.

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

[29] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, no. 2, pp. W182–W185, 2007.

[30] J. R. Brown, "Ancient horizontal gene transfer," *Nature Reviews Genetics*, vol. 4, no. 2, pp. 121–132, 2003.

[31] I. I. Artamonova and A. R. Mushegian, "Genome sequence analysis indicates that the model eukaryote *Nematostella vectensis* harbors bacterial consorts," *Applied and Environmental Microbiology*, vol. 79, no. 22, pp. 6868–6873, 2013.

[32] L. Akman, A. Yamashita, H. Watanabe et al., "Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*," *Nature Genetics*, vol. 32, no. 3, pp. 402–407, 2002.

[33] M. Pignatelli, A. Moya, and J. Tamames, "EnvDB, a database for describing the environmental distribution of prokaryotic taxa,"

Environmental Microbiology Reports, vol. 1, no. 3, pp. 191–197, 2009.

[34] A. Bravo, S. Likitvivatanavong, S. S. Gill, and M. Soberón, "*Bacillus thuringiensis*: a story of a successful bioinsecticide," *Insect Biochemistry and Molecular Biology*, vol. 41, no. 7, pp. 423–431, 2011.

[35] G. P. Kaaya and N. Darji, "Mortality in adult tsetse, *Glossina morsitans morsitans*, caused by entomopathogenic bacteria," *Journal of Invertebrate Pathology*, vol. 54, no. 1, pp. 32–38, 1989.

[36] J. M. Lindh and M. J. Lehane, "The tsetse fly *Glossina fuscipes fuscipes* (Diptera: Glossina) harbours a surprising diversity of bacteria other than symbionts," *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, vol. 99, no. 3, pp. 711–720, 2011.

[37] C. Bandi, G. Damiani, L. Magrassi, A. Grigolo, R. Fani, and L. Sacchi, "Flavobacteria as intracellular symbionts in cockroaches," *Proceedings of the Royal Society B: Biological Sciences*, vol. 257, no. 1348, pp. 43–48, 1994.

[38] G. D. D. Hurst, T. C. Hammarton, C. Bandi, T. M. O. Majerus, D. Bertrand, and M. E. N. Majerus, "The diversity of inherited parasites of insects: the male-killing agent of the ladybird beetle *Coleomegilla maculata* is a member of the Flavobacteria," *Genetical Research*, vol. 70, no. 1, pp. 1–6, 1997.

[39] G. D. D. Hurst, C. Bandi, L. Sacchi et al., "*Adonia variegata* (Coleoptera: Coccinellidae) bears maternally inherited Flavobacteria that kill males only," *Parasitology*, vol. 118, no. 2, pp. 125–134, 1999.

[40] S. Van Borm, A. Buschinger, J. J. Boomsma, and J. Billen, "Tetraponera ants have gut symbionts related to nitrogen-fixing root-nodule bacteria," *Proceedings of the Royal Society B: Biological Sciences*, vol. 269, no. 1504, pp. 2023–2027, 2002.

[41] M. J. López-Sánchez, A. Neef, J. Peretó et al., "Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*," *PLoS Genetics*, vol. 5, no. 11, Article ID e1000721, 2009.

[42] Y. Matsuura, R. Koga, N. Nikoh, X.-Y. Meng, S. Hanada, and T. Fukatsu, "Huge symbiotic organs in giant scale insects of the genus *Drosicha* (Coccoidea: Monophlebidae) harbor flavobacterial and enterobacterial endosymbionts," *Zoological Science*, vol. 26, no. 7, pp. 448–456, 2009.

[43] M. Rosenblueth, L. Sayavedra, H. Sámano-Sánchez, A. Roth, and E. Martínez-Romero, "Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (Hemiptera: Coccoidea)," *Journal of Evolutionary Biology*, vol. 25, no. 11, pp. 2357–2368, 2012.

[44] D. Kageyama, S. Narita, and M. Watanabe, "Insect sex determination manipulated by their endosymbionts: incidences, mechanisms and implications," *Insects*, vol. 3, no. 1, pp. 161–199, 2012.

[45] N. Kondo, N. Nikoh, N. Ijichi, M. Shimada, and T. Fukatsu, "Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14280–14285, 2002.

[46] N. Nikoh, K. Tanaka, F. Shibata et al., "*Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes," *Genome Research*, vol. 18, no. 2, pp. 272–280, 2008.

[47] J. C. Dunning Hotopp, M. E. Clark, D. C. S. G. Oliveira et al., "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes," *Science*, vol. 317, no. 5845, pp. 1753–1756, 2007.

[48] N. Nikoh and A. Nakabachi, "Aphids acquired symbiotic genes via lateral gene transfer," *BMC Biology*, vol. 7, article no. 12, 2009.

[49] L. Klasson, Z. Kambris, P. E. Cook, T. Walker, and S. P. Sinkins, "Horizontal gene transfer between *Wolbachia* and the mosquito *Aedes aegypti*," *BMC Genomics*, vol. 10, article no. 33, 2009.

[50] M. Woolfit, I. Iturbe-Ormaetxe, E. A. McGraw, and S. L. O'Neill, "An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*," *Molecular Biology and Evolution*, vol. 26, no. 2, pp. 367–374, 2009.

[51] T. Aikawa, H. Anbutsu, N. Nikoh, T. Kikuchi, F. Shibata, and T. Fukatsu, "Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome," *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1674, pp. 3791–3798, 2009.

[52] S. M. Degnan, "Think laterally: horizontal gene transfer from symbiotic microbes may extend the phenotype of marine sessile hosts," *Frontiers in Microbiology*, vol. 5, article 638, 2014.

[53] V. Doudoumis, U. Alam, E. Aksoy et al., "Tsetse-*Wolbachia* symbiosis: comes of age and has great potential for pest and disease control," *Journal of Invertebrate Pathology*, vol. 112, no. 1, pp. S94–S103, 2013.

[54] Q. Hou, J. He, J. Yu et al., "A case of horizontal gene transfer from *Wolbachia* to *Aedes albopictus* C6/36 cell line," *Mobile Genetic Elements*, vol. 4, no. 2, Article ID e28914, 2014.

[55] A. Nakabachi, "Horizontal gene transfers in insects," *Current Opinion in Insect Science*, vol. 7, article no. 113, pp. 24–29, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8313272, 9 pages
http://dx.doi.org/10.1155/2016/8313272

Research Article

Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism

Albert Batushansky,[1] David Toubiana,[2] and Aaron Fait[1]

[1]*The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, 84990 Midreshet Ben-Gurion, Israel*
[2]*Telekom Innovation Laboratories, Department of Information Systems Engineering, Ben-Gurion University of the Negev, 84105 Beer Sheva, Israel*

Correspondence should be addressed to Albert Batushansky; batushanskya@missouri.edu

Received 11 May 2016; Revised 3 August 2016; Accepted 18 August 2016

Academic Editor: Shigehiko Kanaya

Copyright © 2016 Albert Batushansky et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the last decade vast data sets are being generated in biological and medical studies. The challenge lies in their summary, complexity reduction, and interpretation. Correlation-based networks and graph-theory based properties of this type of networks can be successfully used during this process. However, the procedure has its pitfalls and requires specific knowledge that often lays beyond classical biology and includes many computational tools and software. Here we introduce one of a series of methods for correlation-based network generation and analysis using freely available software. The pipeline allows the user to control each step of the network generation and provides flexibility in selection of correlation methods and thresholds. The pipeline was implemented on published metabolomics data of a population of human breast carcinoma cell lines MDA-MB-231 under two conditions: normal and hypoxia. The analysis revealed significant differences between the metabolic networks in response to the tested conditions. The network under hypoxia had 1.7 times more significant correlations between metabolites, compared to normal conditions. Unique metabolic interactions were identified which could lead to the identification of improved markers or aid in elucidating the mechanism of regulation between distantly related metabolites induced by the cancer growth.

1. Introduction

Advanced technology methods for high-throughput biological studies, such as metabolomics and transcriptomics developed during the last decades, are successfully applied in biomedical research [1], plant studies [2], and microbiology [3]. The wide use of these technologies led to the accumulation of data on biological processes at their multiple levels (metabolic, genetic, enzymatic, physiological, phenotypical, etc.) and called for the development of tools to ease the visualization, analysis, and interpretation of an often complex and multidimensional matrix. Furthermore, the readily available "*omics*" technologies in biological laboratories prompted biologists to enter a field often needing extensive computational knowhow and led to the increased interest in biological interaction networks [4]. Thus, in the recent decades networks describing cellular processes were generated for human [5], yeast [6], and plants [7].

Networks can be presented as graphs, that is, a set of vertices (V) connected by edges (E), and consequently can be analyzed using graph theory, an approach that has been increasingly implemented in biological studies during the last decade. It is commonly accepted that graph theory as a scientific discipline was first used by the Swiss mathematician Leonhard Euler in 1735-1736, tackling the Königsberg bridge problem. Later, in the 19th and 20th centuries, graph theory was formulated and eventually introduced for applied fields, such as physics, computer science, and biology [8]. Today, graph theory consists of many tens of basic definitions and properties [9]. The understanding of the biological networks lies in the nature of the vertices and edges between them; that is, the vertices may represent one of the components of the

three major molecular levels: genes, proteins, or metabolites, while the edges between them represent gene coexpression, protein-protein interactions, or biochemical conversions of metabolites, respectively [10]. However, molecular networks are not delimited to illustrate single-level component interactions. They can also show cross-level interactions. Alternatively, and perhaps a little counterintuitive, a network may incorporate vertices representing a set of metabolic reactions, where the connection between a pair of vertices is established if the reactions share one or multiple metabolites used or produced by these reactions [11, 12]. In other networks, vertices represent a community of molecular components, especially used with very vast data sets (>1000 of components) such as in weighted gene coexpression network analysis (WGCNA). Here, a single vertex delineates a module of genes and edges between vertices represent the correlation between them. This allows reducing the complexity of the network and simultaneously retains most of the information used for the interpretation of the gene coexpression results [13]. In simple words, vertices and edges represent the information as defined by the creator/user of the network.

In the last decade, correlation-based network analysis (CNA) has become a popular data-mining tool for visualizing and analyzing biological relationships within large data sets [13, 14]. In this type of networks, vertices and edges represent molecular elements (e.g., metabolites or genes) and their correlation coefficient (strength and sign), respectively [10, 15, 16]. Edges inferred by correlation analyses reflect a coordinated behavior between vertices across the data set (treatments, genotypes, conditions, and time). The type of correlation has to be selected based on the parametrical distribution of the data. In large population studies, data has to be tested for normality using existing tests, for example, the Shapiro-Wilk test. The Pearson correlation should be applied to normally distributed data, while Spearman's rank correlation should be used for data violating the assumption of normal distribution. CNA was successfully applied to various biological systems; it revealed, for example, metabolic markers related to plant growth and biomass in *Arabidopsis thaliana* recombinant inbred lines (RIL) and introgression lines (IL) [17, 18], the role of gene Col5a2 in myocardial infarction [19], effect of hypoxia on tumor cell biochemistry [20], and recently, identification of genetically based mechanism of the regulation of amino acid metabolism [2].

Graph theory defines a number of network properties that allow successful analysis and interpretation of correlation networks (CN). These properties are a set of measures that describe the graph topology from different vantage points. CNs are undirected graphs, reflecting the coordinated behavior of two or more adjacent vertices (connected vertices) and the biological components they represent and not the effect of one vertex/component onto another, that is, a directed network. Properties that may have biological significance have been reviewed by Toubiana et al. [10]; they include (a) *vertex degree*: the number of edges incident on a given vertex [21], (b) *centrality score*: reflecting the number of shortest paths between a vertex and any other vertex in the network, (c) *network diameter*: the maximal shortest path between any two vertices in the graph, (d) *network density*: the ratio of existing edges to the number of all possible edges of a network, (e) *vertex betweenness centrality*: the relative number of the shortest paths between any two vertices that pass via a specific vertex, and (f) *modules*: subgraphs, within a global network characterized by higher connectivity (biologically interpreted as possible tighter coordination) between their components compared to other regions of the network. The analysis of these modules within the obtained network helped in the prediction of diseases [22, 23]. In this contribution we aim at providing an easy-to-implement pipeline for the generation of CNs for biologists without extensive computational skills. To do so, we are demonstrating the potential use of CNs in cancer studies.

Nowadays, there exist a number of software tools that allow researchers to generate networks, visualize them, and analyze their structure, via the calculation of a number of network properties, based on their own experimental data. Commonly known tools are Cytoscape [24], Gephi [25], and iGraph [26]. Each software has its benefits and disadvantages. For example, while iGraph requires programming skills and knowledge of the R programming language syntax, graphical-user-interface (GUI) based programs, such as Gephi and Cytoscape, do not, simplifying the interaction with the user. On the other hand, while script-based programs allow for the extension of existing functions and integration of compatible libraries, increasing the number of potential properties to be calculated, GUI programs are bound to the functionalities of the version of the software the researcher is using. However, Cytoscape and Gephi both offer a greater and easier-to-use set of visualization tools for networks, whereas the visualization functionalities of iGraph are rather limited and difficult to handle. Cytoscape allows for the integration of externally developed plugins, exerting functionality as desired by its developer. However, this option requires knowledge of the Java programming language and an understanding of how to interface it with the Cytoscape software.

The current proposed stepwise pipeline allows the user to control each step of the network creation, as it provides flexibility in selection of correlation methods and thresholds and describes easy-to-handle options to analyze the network topology. The pipeline works irrespective of the nature of the data set and can be implemented by a combined use of the freely distributed Apache OpenOffice software (http://www.openoffice.org/), built-in packages within the R-environment [27], and Cytoscape [24].

2. Method

The construction of correlation-based networks starts form the calculation of the pairwise correlation coefficients between any two pairs of vectors of a given data set. One of the easiest ways to complete this calculation in big sets of data is to exploit the freely available R-software. There are several packages developed for correlation analysis under the R-environment. It is very important for the output matrix to select the proper type of correlation coefficient (Pearson, Spearman, Kendal, etc., represented as the letter "r") and its corresponding thresholds (r and p). We recommend using the "*psych*" package under the R-environment [27, 28]. This

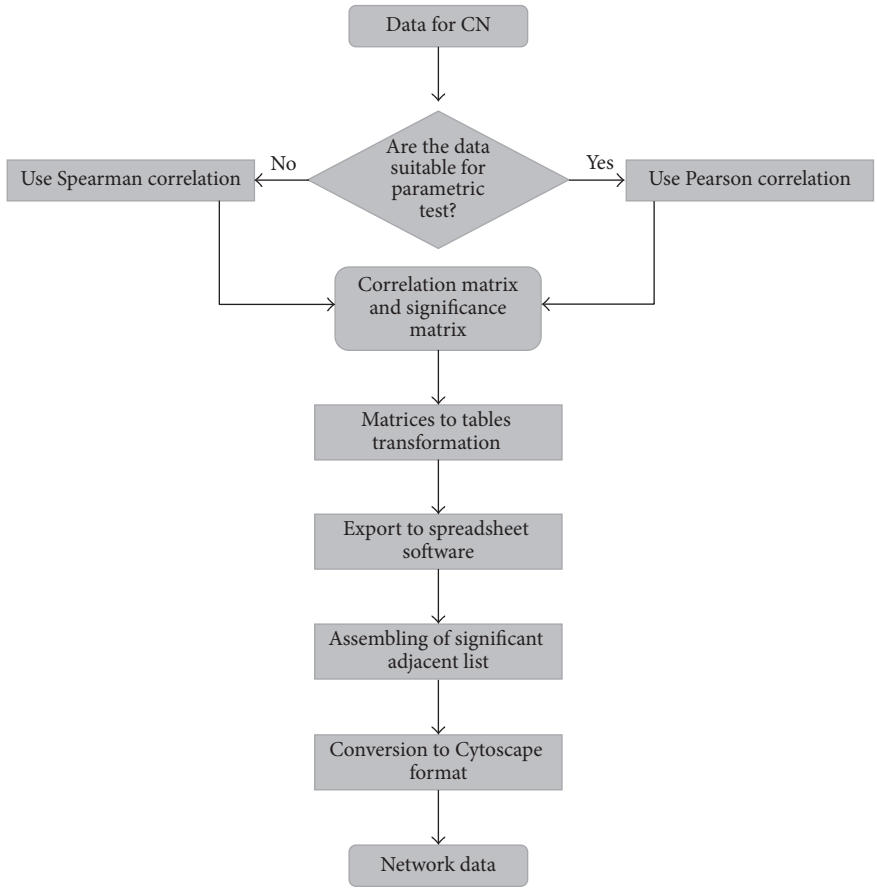


Figure 1: Correlation-based network, pipeline flowchart.

package allows calculation of two diagonal matrices: (1) a symmetric diagonal *r-matrix* and (2) a symmetric diagonal *p-matrix*, where the lower triangle stores the *p-values* and the upper triangle the multiple hypotheses corrected *p-values*, corrected either by the Bonferroni correction or by applying a false discovery rate (FDR) correction. The obtained matrix with both *r*- and raw/adjusted *p-values* can be then transformed to the table view and exported to any spreadsheet software for a supervised selection of significant correlation coefficients. The thresholds of significance should be selected in respect to the nature and size of the data and considering the general suggestions as described in the introduction and elsewhere [29]. The selected significant correlation values can be easily converted to a table, listing in three columns the vertices that are adjacent to each other. This table is subsequently used as a template to illustrate the network using Cytoscape. We have chosen Cytoscape out of the list of network software as it was specifically developed for biological data, because of its intuitively understandable interface, wide range of visualization options, and available additional plugins for calculations of the main network properties. The method's workflow is presented in Figure 1.

2.1. Method Pipeline

2.1.1. Download R-Environment and Required R-Packages. To start the workflow, first download and install the latest version of R-environment from the following website: https://www.r-project.org/. For the processes described here two R-packages will be used: "psych" [28] and "reshape2" [30]. Both packages are freely available for downloading via the R-environment window. As mentioned above, the R-environment is a freely available powerful statistical software often used to analyze biological data. Its benefits stem from the integration of various built-in functions and libraries/packages, supplemented by its ability to complement these by numerous externally developed packages and the freedom to combine them as necessary. Often, different packages offer different functions tackling the same task. For example, to compute correlation coefficients, one may use the core built-in function "cor" or the "rcorr" function of the Hmisc-package [31]. For the current work we have chosen specifically the "psych" package to perform correlation analysis as it conveniently computes the *r* coefficients and its corresponding *p* values and also performs *post hoc* tests to correct for multiple hypothesis testing (MHT). The package "reshape2" allows converting a matrix into a table and was chosen for this work for its easy implementation.

2.1.2. Adjusting the Allocated Memory. Before beginning with the actual analysis, we recommend checking for the size of virtual memory available for R and Cytoscape, considering the potential large size of a data set. To do so for R under

Windows OS type memory.limit() and if the result is smaller than the potential amount of your data set, increase the memory by typing memory.limit(size = 4096). This step allocates 4096 MB, equivalent to 4 GB (maximal number for 32 GB systems) of virtual memory, to the R-software. Unix-based OS's do not offer this function, as their virtual memory management is dynamic, adjusting itself to new and existing processes.

Similarly to the R-software the user may increase the memory allocated to Cytoscape, if, for instance, the size of a network is too large. Cytoscape is a Java-based software, so the first step here will be to access the *Configure Java* option via the Programs list. Next, select the *Java* tab in the displayed window, click on *View* button, and type *-Xms4096m* into the *Runtime parameters* line to allocated 4 GB of memory to the Cytoscape software. The amount of allocated memory is editable.

2.1.3. Producing the Matrices (the R Code Necessary to Complete the Steps Described below Can Be Found in Supplementary Figure 1). After the size of virtual memory is set, the user can start the pipeline according to the protocol presented in Supplementary Figure 1 available online at http://dx.doi.org/10.1155/2016/8313272 . The described protocol represents a set of consequent commands (with an exception to the parallel computation of the *r*- and *p-value* matrices using the "psych" package), where the execution of one step is dependent on the former.

The output of the executed protocol will provide two separate files that can be opened in spreadsheet software. One of the files, "*r*_table.csv," will represent a table view of the correlation matrix, and the second file, "*p*_table," will represent the same table where *r-values* will be replaced by the correspondent *p values*. Probably the single disadvantage of this method is the time of calculation that strongly depends on number of the variables for the analysis and can be problematic for large (more than 500 variables) data sets. Nevertheless, the vast majority of metabolomics data sets does not exceed this amount of variables and usually is much smaller. Thus, the reader should not run into problems when executing the above code.

The obtained files "*r*_table.csv" and "*p*_table.csv" can be opened in any spreadsheet software (in our case OpenOffice). The next step is to remove the first column in each file and copy the rest to a new multisheet file on separate sheets for the *r-values* and the *p-values,* respectively. This step will provide two tables with two identical columns with the names of the variables, for example, metabolites/genes, and different third column with *r*- and *p-values,* respectively. At this stage the correlation threshold has to be selected.

2.1.4. Selection of Significant Interactions and Arrangement of the Data to the Network Format Spreadsheet Software. Correlation coefficients, *r*, are the determining elements in CN construction; the threshold of acceptable *r*-value range and the threshold of its statistical significance will greatly affect the output of the network and its interpretation. The significance of a correlation is a two-factor concept. The first factor, the correlation coefficient (*r*), is expressed as a value ranging from −1 to 1, where positive and negative values represent a relation, alike or inverse, between the changes in the measure of the two variables. The magnitude of the coefficient reveals the strength of this relationship. However, the reliability of the model also depends on a second factor: the probability (*p*) of the detected *r-values*, reflecting a true relation. This value ranges from 0 to 1 and depends to a great extent on the sample size [32] but also on the experimental setup and the biological system of study. The selection of the threshold for both values depends largely on the researcher. It is trivial that $r = 1$ (perfect positive correlation) or $r = -1$ (perfect negative correlation) represent strong coordinated behaviors, while $r = 0$ shows the absence of a relation between the variables. But what can be said about intermediate *r*'s? The "rule of thumb" suggests that there is no absolute *r*-threshold and different scientific disciplines apply different *r-value* thresholds. For example, in biology, thresholds from as low as $|\pm 0.3|$ have been proposed to be relevant, for example, for metabolic data in tomato introgression lines seeds and fruits [33], while in physics, an *r-value* lower than $|\pm 0.9|$ is often considered insignificant. Usually $r \geq |\pm 0.5|$ is considered as "strong" by most of researches in biological systems [34]. The *p-value* that reflects significance of a correlation is usually accepted at three levels: 0.05, 0.01, and 0.001 [32]. However, since correlation analysis is applied on large data sets, *p-values* should usually be corrected by one of the *post hoc* tests for MHT, such as the Bonferroni correction or the false discovery rate (FDR) method, with the aim of avoiding false positives.

After both parameters of significance are decided, create a new sheet and copy the first two columns from any of the sheets (they are identical). In the first cell of the third column input the following formula:

$$= \text{if}\left(\text{and}\left(\text{abs}\left(X\right) > R,\ Y < P\right), 1, 0\right). \quad (1)$$

In this formula X is the value of the 1st cell in the "*r*_values" sheet; R is the selected critical *r*-value; Y is the value of the 1st cell in the "*p*_values" sheet; P is the selected critical *p* value. Expand this formula to the whole table. This will provide an adjacency list that can be easily converted to the network format for Cytoscape software. For this, input in the next column following formula:

$$= \text{if}\left(X = 1, \text{concatenate}\left(Y, \text{“}-\text{”}, Z\right), \text{“”}\right). \quad (2)$$

In this formula X is a number of 1st cell of the obtained column (usually 3rd) on the current sheet; Y and Z are the numbers of 1st cell in the 1st and 2nd column on the current sheet, respectively. At this stage copy 1st, 2nd, and 4th columns as the values to the new sheet, filter out and remove rows with empty last cell, and save the obtained fully filled three-column table in *.txt tab delimited format*.

This file can be imported as a network to the Cytoscape software and analyzed using the built-in NetworkAnalyzer plugin. To import the file run the Cytoscape and select "*import*" from the main menu bar, then locate the previously saved three-column file in *.txt* format, and import it as *"table"* (Supplementary Figure 2). To run the plugin locate *NetworlAnalyzer* in the Tool menu and execute *"Analyze*

Network" (Supplementary Figure 3). The plugin will calculate the degree of vertices, vertex betweenness centrality, vertex clustering coefficient, and edge betweenness. The obtained parameters will be automatically added as attributes of vertices and edges of the network and can be visualized by customizable view options, including color, size, shape, and thickness. Additionally, NetworkAnalyzer can check if a vertex distribution of a network fits the power law, calculates the main properties of the network topology, such as diameter and global transitivity, and shows the average shortest path and other useful parameters.

3. Results and Discussion

Hypoxia is one of the major features of solid tumors affecting their development and treatment selection [35, 36]. The simulation of hypoxia in cancer cells *in vitro* can be used as a model study to understand the alteration of cancer cell metabolism that supports tumor growth under hypoxic conditions, the phenomena known as the "Warburg effect" [20, 37–39].

In short, the experiment included MDA-MB-231 breast adenocarcinoma cells that were incubated in 95% air and 5% CO_2 at 37°C and 95% relative humidity and then were transferred to normoxic (21% oxygen) conditions. After 24 hours cell culture was divided into two groups, and one group was maintained under the normoxic condition, while the second group was transferred to a specific vessel with flow of gas containing 1% O_2 and 5% CO_2 balanced with N_2 (hypoxic conditions). Next, GC-MS metabolic profiling of the two groups was performed [20].

Prior to network construction, we first elaborated the published data, keeping uniquely identified metabolites only. Considering the sample size ($n = 30$) and the relatively large number of missing values in the data set, we decided to use Spearman's rank correlation with the thresholds $r \geq |\pm 0.7|$ and $q_{\text{FDR}} < 0.05$. The applied procedure resulted in two adjacent tables (control (c) and hypoxia (h)) (Supplementary Data 1-2, control/hypoxia_adjacent_table resp. sheet) that were loaded to Cytoscape and visualized as a network (Figures 2(a) and 2(b), Supplementary Data 3). The simple comparison of two graphs revealed that the normal metabolic network was smaller compared to the network under hypoxic conditions. The differences in the number of vertices, v, were not very high ($v_c = 19$ versus $v_h = 23$, control versus hypoxia, resp.), but the number of edges, e, differed significantly ($e_c = 87$ versus $e_h = 144$, control versus hypoxia, resp.).

In order to identify metabolites or metabolic interactions specific to the hypoxic conditions, we used the "*merge*" tool in Cytoscape, selecting the two data networks. The tool gives multiple merging options visualizing either unique or common edges between two (or more) networks (Supplementary Data 3). The resulting merged graph displays common links (the union, for this kind of comparisons graph theory uses set-theory jargon). The comparison of the original graphs with the merged one is done by the same merging tool selecting the "difference" option and eventually it generates a graph (difference graph) for each comparison based on unique edges and vertices of the selected condition. The resulting difference graphs emphasize many condition-specific relations between metabolites existing in the two original networks (Figures 2(c) and 2(d)). The number of vertices changed to $v_c = 16$ and $v_h = 22$ for the control and the hypoxic conditions, respectively, and number of edges changed to $e_c = 27$ and $e_h = 84$, respectively. Thus, the gap (in folds) between the two e values increased from 1.67 to 3.1. The increased number of edges under hypoxia suggests the appearance of alternative metabolic routes to sustain the cell metabolism. Hypoxia treatment is used to mimic the conditions occurring in cancer cells because of high "uncontrolled" growth rate. Here, the unique metabolic relation identified could lead to the isolation of biochemical steps/reactions or common regulatory mechanisms between distantly related metabolites induced by the cancer growth (hypoxia treatment). Eventually the potential to identify markers defined as edges and not as vertices is significantly higher; just consider that the potential number of edges in a correlation (undirected) network with n metabolites can have $n * (n - 1)/2$ interactions.

Table 1: The main properties of the networks with unique interactions under control and hypoxia conditions, respectively.

Network name	Density	Diameter	Transitivity
Unique interactions under normal conditions (Figure 2(c))	0.23	4	0.38
Unique interactions under hypoxia conditions (Figure 2(d))	0.36	3	0.63

We then applied the NetworkAnalyzer plugin to calculate some of the topological properties of the networks such as network density, diameter, and transitivity and vertex degree and betweenness centrality (Tables 1, 2, and 3).

The results of the NetworkAnalyzer analysis of the networks topology suggested a reorganization of the metabolic network under hypoxia. Thus, a smaller (3 versus 4, Table 1) diameter, the longest shortest path between any two vertices in the network, and a larger (0.63 versus 0.38, Table 1) transitivity, the probability to form cliques in the network, suggest that the reorganization of the metabolic network under hypoxic conditions occurs via specific metabolites, namely, Ala, creatinine, 2OG, Tyr, and citrate. They act as hubs as they exhibit the greatest vertex degree and betweenness centrality measures (Table 3). In contrast, the properties of the network under normal conditions showed the topological importance of lactate, Thr, and GABA in Table 2. Surprisingly, lactate, the vertex with the highest betweenness centrality under normal conditions (0.46, Table 2), is absent in the hypoxia network (Figure 2(d)). This can be explained by the fact that nonoxidative metabolism is induced under stronger hypoxic conditions [20]. Alternatively, considering that lactate production is an indicator of inhibited respiratory [40], its absence in the correlation network under hypoxic conditions can suggest a strong specific effect of hypoxia on lactate irrespectively of other related metabolites. The results of GC-MS analysis revealed almost 1.3 times increase of lactate level under hypoxia compared to control conditions

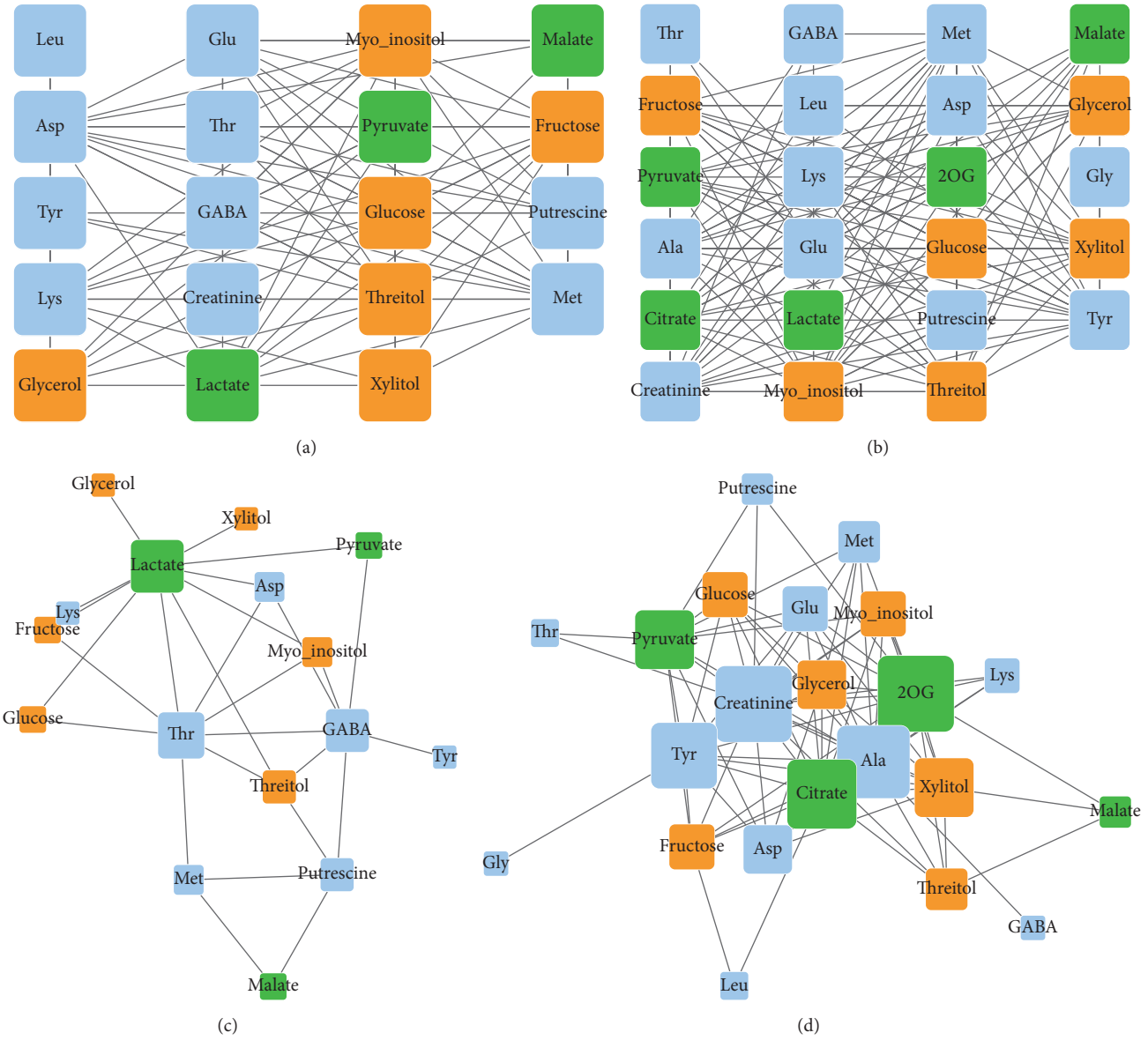


Figure 2: Correlation-based networks of metabolite data sets. (a) Original network under control conditions, (b) original network under hypoxic conditions, (c) network of unique relationships under control conditions compared to hypoxic conditions, and (d) network of unique relationships under hypoxic conditions compared to control conditions. Metabolic profiling of breast cancer cells under control and hypoxia (30 samples each) was used for pairwise correlation analysis between metabolites and network-view production. The data used to generate the network is from Kotze et al., 2013 [20]. Each vertex represents a metabolite; each edge represents a significant correlation between pairs of metabolites across samples. Vertex colors reflect biochemical classes: amino acids and N-compounds (blue), sugars and sugar alcohols (orange), and carboxylic acids (green). Vertex size reflects degree.

and support this suggestion [20]. Furthermore, the oxygen deficient condition leads not only to the increased conversion of glucose to lactate but also to the sharp suppression of citrate production [41]. The results of the GC-MS showed almost a twofold decrease in citrate levels under hypoxia compared to the control conditions [20]. The replacement of lactate to citrate in the metabolic network under hypoxic conditions, the high centrality of citrate in the network according to its vertex degree and betweenness centrality (Table 3), and the appearance of the citrate-2OG edge suggest the shift of citrate production from glucose oxidation to reductive carboxylation of 2OG (Figures 2(c) and 2(d) and Table 3) [41].

Glycolytic activity is high in cancer cells under both normal and hypoxic conditions. In the hypoxia network glycolysis derived pyruvate is strongly correlated with a row of biochemically related amino acids Ala, Asp, and Tyr, while in the network under normal conditions these associations were not detected (Figures 2(c) and 2(d)). Additionally, the unexpected drop of the correlation between pyruvate and GABA under hypoxia and the great centrality of Ala

Table 2: The properties of the network of unique interactions under normal conditions in the descending order according to betweenness centrality. The compound class 1 represents amino acids, 2 sugars and sugar alcohols, and 3 carboxylic acids.

Metabolite	Compound class	Degree	Betweenness centrality
Lactate	3	10	0.460
Thr	1	8	0.255
GABA	1	7	0.224
Threitol	2	4	0.085
Putrescine	1	4	0.084
Met	1	3	0.065
Asp	1	3	0.015
Myo-inositol	2	3	0.015
Pyruvate	3	2	0.015
Fructose	2	2	0.000
Glucose	2	2	0.000
Glycerol	2	1	0.000
Lys	1	1	0.000
Malate	3	2	0.000
Tyr	1	1	0.000
Xylitol	2	1	0.000

Table 3: The properties of the network of unique interactions under hypoxia conditions in the descending order according to betweenness centrality. The compound class 1 represents amino acids, 2 sugars and sugar alcohols, and 3 carboxylic acids.

Metabolite	Compound class	Degree	Betweenness centrality
Ala	1	15	0.147
Creatinine	1	16	0.133
2OG	3	16	0.127
Tyr	1	13	0.113
Citrate	3	14	0.103
Pyruvate	3	11	0.060
Xylitol	2	11	0.037
Fructose	2	7	0.033
Threitol	2	6	0.010
Glycerol	2	8	0.006
Asp	1	8	0.005
Glucose	2	7	0.004
Met	1	6	0.004
Glu	1	7	0.003
Myo-inositol	2	7	0.003
Lys	1	4	0.001
Putrescine	1	3	0.001
GABA	1	1	0.000
Gly	1	1	0.000
Leu	1	2	0.000
Malate	3	3	0.000
Thr	1	2	0.000

in the hypoxia network should be noted. GABA can be used in the transamination of pyruvate to produce alanine and succinic semialdehyde. GABA also accumulates under hypoxia in neurons of rats [42], and the present study shows that the level of GABA increased 1.5 times under hypoxia compared to control conditions [20]. Taken together these results suggest the transamination of pyruvate to Ala, possibly via GABA. Alternatively pyruvate is converted to Ala via alanine transaminase (ALT), involving Glu and 2OG (the latter also exhibiting a high centrality in the network), which act in a concerted action with aspartate transaminase (AST). The AST/ALT ratio in the blood of a human or animal is used in the diagnosis of liver damage or hepatotoxicity. By emphasizing the tight interaction between pyruvate, Ala, and Asp, our results likely show the metabolic reflection of a toxic condition imposed on the cell by hypoxia. Last, Ala is considered a marker of prostate [43] and breast [44] cancers where it significantly accumulates. However, the results by Kotze and coworkers did not reveal this in Ala level under hypoxia. We hypothesize that the changes in content of Ala might not be consistent between systems, while the actual coordinated response of Ala with a few tightly linked metabolites reflected within the network could potentially be a better candidate.

4. Conclusions

The interpretation of the CNs shows the relevance of graph theory in the analysis of biological data in general and specifically in the works dedicated to metabolic and genetic pathways. Implementing a network-based workflow using previously published data, we show how the pipeline can generate and visualize a network and how the network analysis can be used in biological studies. The presented pipeline aims at providing an easy to use but relatively powerful tool for *in silico* analysis of experimental data. The pipeline is not limited to metabolic data and can be effectively applied to gene coexpression network analysis, like the previously identified human disease-associated genes [45], lethal genes combination in yeast, and others [46–48]. This short essay exemplifies that the usage of CNs can lead to biologically sound conclusions on metabolic pathway regulation and original hypothesis generation without the need for complex and capacity consuming approaches. That said, CNs can be used as a part of top-down, complexity reduction approach leading to insights in the search and identification of marker genes or metabolites, respectively. Having said that, we wish to emphasize that the quality of the analysis more often than not depends on the design of the experiment and the sampling strategy.

Disclosure

The current address for Albert Batushansky is University of Missouri-Columbia, Division of Biological Sciences, Laboratory of Seed Metabolism, 1201 Rollins Street, 304 LSC Building, Columbia, MO 65211, USA.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this article.

Acknowledgments

Authors thank Dr. Royston Goodacre and his group from the School of Chemistry, Manchester Institute of Biotechnology, University of Manchester, UK, for permission to use the data obtained during their research which was previously published [20].

References

[1] D. Nagrath, C. Caneba, T. Karedath, and N. Bellance, "Metabolomics for mitochondrial and cancer studies," *Biochimica et Biophysica Acta—Bioenergetics*, vol. 1807, no. 6, pp. 650–663, 2011.

[2] D. Toubiana, A. Batushansky, O. Tzfadia et al., "Combined correlation-based network and mQTL analyses efficiently identified loci for branched-chain amino acid, serine to threonine, and proline metabolism in tomato seeds," *The Plant Journal*, vol. 81, no. 1, pp. 121–133, 2015.

[3] J. Tang, "Microbial metabolomics," *Current Genomics*, vol. 12, no. 6, pp. 391–403, 2011.

[4] X. Ma and L. Gao, "Biological network analysis: Insights into structure and functions," *Briefings in Functional Genomics*, vol. 11, no. 6, pp. 434–442, 2012.

[5] U. Stelzl, U. Worm, M. Lalowski et al., "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[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] 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.

[8] J. L. Gross, J. Yellen, and P. Zhang, *Handbook of Graph Theory*, Chapman & Hall/CRC, 2nd edition, 2013.

[9] R. J. Wilson, *Introduction to Graph Theory*, John Wiley & Sons, New York, NY, USA, 1986.

[10] D. Toubiana, A. R. Fernie, Z. Nikoloski, and A. Fait, "Network analysis: tackling complex data to study plant metabolism," *Trends in Biotechnology*, vol. 31, no. 1, pp. 29–36, 2013.

[11] M. Bekaert, P. P. Edger, J. Chris Pires, and G. C. Conant, "Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints," *The Plant Cell*, vol. 23, no. 5, pp. 1719–1728, 2011.

[12] A. Wagner and D. A. Fell, "The small world inside large metabolic networks," *Proceedings of the Royal Society B: Biological Sciences*, vol. 268, no. 1478, pp. 1803–1810, 2001.

[13] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, article 559, 2008.

[14] F. Kose, W. Weckwerth, T. Linke, and O. Fiehn, "Visualizing plant metabolomic correlation networks using clique-metabolite matrices," *Bioinformatics*, vol. 17, no. 12, pp. 1198–1208, 2001.

[15] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, UK, 1994.

[16] D. Yu, M. Kim, G. Xiao, and T. H. Hwang, "Review of biological network data and its applications," *Genomics & Informatics*, vol. 11, no. 4, pp. 200–210, 2013.

[17] J. Lisec, R. C. Meyer, M. Steinfath et al., "Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations," *Plant Journal*, vol. 53, no. 6, pp. 960–972, 2008.

[18] R. C. Meyer, M. Steinfath, J. Lisec et al., "The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4759–4764, 2007.

[19] F. Azuaje, L. Zhang, C. Jeanty, S.-L. Puhl, S. Rodius, and D. R. Wagner, "Analysis of a gene co-expression network establishes robust association between Col5a2 and ischemic heart disease," *BMC Medical Genomics*, vol. 6, no. 1, article 13, 2013.

[20] H. L. Kotze, E. G. Armitage, K. J. Sharkey et al., "A novel untargeted metabolomics correlation-based network analysis incorporating human metabolic reconstructions," *BMC Systems Biology*, vol. 7, article 107, 2013.

[21] T. F. Cooper, A. P. Morby, A. Gunn, and D. Schneider, "Effect of random and hub gene disruptions on environmental and mutational robustness in *Escherichia coli*," *BMC Genomics*, vol. 7, no. 1, article 237, 2006.

[22] 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.

[23] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," *Nature Communications*, vol. 5, article 3231, 2014.

[24] 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.

[25] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM '09)*, San Jose, Calif, USA, May 2009.

[26] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *Inter Journal Complex Systems*, p. 1695, 2006.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2013.

[28] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Ill, USA, 2015.

[29] D. G. Altman, *Practical Statistics for Medical Research*, Chapman & Hall/CRC, New York, NY, USA, 1990.

[30] H. Wickham, "Reshaping data with the reshape package," *Journal of Statistical Software*, vol. 21, no. 12, pp. 1–20, 2007.

[31] F. E. J. Harrell, *Hmisc Package for R*, 2006.

[32] R. Bloom, *Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient*, 2010.

[33] D. Toubiana, Y. Semel, T. Tohge et al., "Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations," *PLoS Genetics*, vol. 8, no. 3, Article ID e1002612, 2012.

[34] D. J. Rumsey, *Statistics for Dummies*, For Dummies, 2nd edition, 2011.

[35] P. Vaupel and A. Mayer, "Hypoxia in cancer: significance and impact on clinical outcome," *Cancer and Metastasis Reviews*, vol. 26, no. 2, pp. 225–239, 2007.

[36] W. R. Wilson and M. P. Hay, "Targeting hypoxia in cancer therapy," *Nature Reviews Cancer*, vol. 11, no. 6, pp. 393–410, 2011.

[37] J. R. Griffiths, P. M. J. McSheehy, S. P. Robinson et al., "Metabolic changes detected by in vivo magnetic resonance studies of HEPA-1 wild-type tumors and tumors deficient in hypoxia-inducible factor-1β (HIF-1β): evidence of an anabolic role for the HIF-1 pathway," *Cancer Research*, vol. 62, no. 3, pp. 688–695, 2002.

[38] M. G. V. Heiden, L. C. Cantley, and C. B. Thompson, "Understanding the warburg effect: the metabolic requirements of cell proliferation," *Science*, vol. 324, no. 5930, pp. 1029–1033, 2009.

[39] O. Warburg, "On the origin of cancer cells," *Science*, vol. 123, no. 3191, pp. 309–314, 1956.

[40] T. N. Seyfried, R. E. Flores, A. M. Poff, and D. P. D'Agostino, "Cancer as a metabolic disease: implications for novel therapeutics," *Carcinogenesis*, vol. 35, no. 3, pp. 515–527, 2014.

[41] D. R. Wise, P. S. Ward, J. E. S. Shay et al., "Hypoxia promotes isocitrate dehydrogenasedependent carboxylation of α-ketoglutarate to citrate to support cell growth and viability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 49, pp. 19611–19616, 2011.

[42] J. E. Madl and S. M. Royer, "Glutamate dependence of GABA levels in neurons of hypoxic and hypoglycemic rat hippocampal slices," *Neuroscience*, vol. 96, no. 4, pp. 657–664, 2000.

[43] M.-B. Tessem, M. G. Swanson, K. R. Keshari et al., "Evaluation of lactate and alanine as metabolic biomarkers of prostate cancer using ^{1}H HR-MAS spectroscopy of biopsy tissues," *Magnetic Resonance in Medicine*, vol. 60, no. 3, pp. 510–516, 2008.

[44] A. Kubota, M. M. Meguid, and D. C. Hitch, "Amino acid profiles correlate diagnostically with organ site in three kinds of malignant tumors," *Cancer*, vol. 69, no. 9, pp. 2343–2348, 1992.

[45] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets," *Genome Research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[46] X. Pan, D. S. Yuan, D. Xiang et al., "A robust toolkit for functional profiling of the yeast genome," *Molecular Cell*, vol. 16, no. 3, pp. 487–496, 2004.

[47] 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.

[48] X. Zhu, M. Gerstein, and M. Snyder, "Getting connected: analysis and principles of biological networks," *Genes and Development*, vol. 21, no. 9, pp. 1010–1024, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 3594937, 13 pages
http://dx.doi.org/10.1155/2016/3594937

Research Article

Semisupervised Learning Based Disease-Symptom and Symptom-Therapeutic Substance Relation Extraction from Biomedical Literature

Qinlin Feng,[1] Yingyi Gui,[2] Zhihao Yang,[1] Lei Wang,[3] and Yuxia Li[3]

[1]*College of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China*
[2]*School of Optoelectronics, Beijing Institute of Technology, Beijing 100081, China*
[3]*Beijing Institute of Health Administration and Medical Information, Beijing 100850, China*

Correspondence should be addressed to Zhihao Yang; yangzh@dlut.edu.cn and Lei Wang; wangleibihami@gmail.com

Received 24 April 2016; Revised 13 July 2016; Accepted 18 August 2016

Academic Editor: Md. Altaf-Ul-Amin

Copyright © 2016 Qinlin Feng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid growth of biomedical literature, a large amount of knowledge about diseases, symptoms, and therapeutic substances hidden in the literature can be used for drug discovery and disease therapy. In this paper, we present a method of constructing two models for extracting the relations between the disease and symptom and symptom and therapeutic substance from biomedical texts, respectively. The former judges whether a disease causes a certain physiological phenomenon while the latter determines whether a substance relieves or eliminates a certain physiological phenomenon. These two kinds of relations can be further utilized to extract the relations between disease and therapeutic substance. In our method, first two training sets for extracting the relations between the disease-symptom and symptom-therapeutic substance are manually annotated and then two semisupervised learning algorithms, that is, Co-Training and Tri-Training, are applied to utilize the unlabeled data to boost the relation extraction performance. Experimental results show that exploiting the unlabeled data with both Co-Training and Tri-Training algorithms can enhance the performance effectively.

1. Introduction

In recent years, with the rapid growth of biomedical literature, the technology of information extraction (IE) has been extensively applied to relation extraction in this literature, for example, extracting the semantic relations between diseases, drugs, genes, proteins, and so forth [1–3]. The related challenges (e.g., BioCreative II protein-protein interaction (PPI) task [4], DDIExtraction 2011 [5], and DDIExtraction 2013 [6]) have been held successfully.

In our work, we focus on extracting the relations between diseases and their symptoms and symptoms and their therapeutic substances. These relations are defined the same as those in [4–6] and also annotated at the sentence level. The former is the relationship between a disease and its related physiological phenomenon in a sentence. For example, the sentence "many blood- and blood vessel-related characteristics are typical for *Raynaud* patients: *Blood viscosity* and *platelet aggregability* are high" shows that *blood viscosity* and *platelet aggregability* are physiological phenomenon of *Raynaud disease*. The latter is the relationship between a physiological phenomenon and the therapeutic substance that can relieve it in a sentence. For example, the sentence "*fish oil* and *its active ingredient eicosapentaenoic acid (EPA)* lowered *blood viscosity*" shows that *fish oil* and *EPA* can relieve the physiological phenomenon *(blood viscosity)*. These two kinds of relations can be further utilized to extract the relations between disease and therapeutic substance. As shown in the above example, it can be assumed that *fish oil* and *EPA* may relieve or heal *Raynaud disease*. Therefore, such information is important for drug discovery and disease treatment. Currently, a large amount of knowledge on diseases, symptoms, and therapeutic substances remains hidden in the literature and needs to be mined with IE technology.

Generally, the methods of extracting the semantic relation between biomedical entities include cooccurrence-based

methods [7], pattern-based methods [8], and machine learning methods [9]. Cooccurrence-based methods use frequent cooccurrence to extract the relations between entities. This method is simple and shows very low precision for high recall [10]. Yen et al. developed a cooccurrence approach based on an information retrieval principle to extract gene-disease relationships from text [11]. Pattern-based methods define a series of patterns in advance and use pattern matching to extract the relations between entities. Huang et al. used a dynamic programming algorithm to compute distinguishing patterns by aligning relevant sentences and key verbs that describe protein interactions [12]. Since templates are manually defined, its generalization ability is not satisfactory. Machine learning methods, the most popular ones, use classification algorithms to extract the relations between entities from literature, such as support vector machine (SVM) [13], maximum entropy [14], and Naive Bayes [15]. Among others, kernel-based methods are widely used in relation extraction. These methods define different kernel functions to extract the relations between entities, such as graph kernel [16], tree kernel [17], and walk path kernel [18].

The machine learning methods belong to the supervised learning ones which need a large of labeled examples to train the model. However, currently no corpuses for extraction of disease-symptom and symptom-therapeutic substance relations are available. In addition, even if limited labeled data are available, it is still difficult to achieve satisfactory generalization ability for a classifier. To solve the problem, we first manually annotated two training sets for extracting the relations between the disease-symptom and symptom-therapeutic substance and then introduced the semisupervised learning methods to utilize the unlabeled data for training the models.

Semisupervised learning methods attempt to exploit the unlabeled data to help improve the generalization ability of the classifier with limited labeled data. They can be roughly divided into four categories, that is, generative parametric models [19], semisupervised support vector machines (S3VMs) [20], graph-based approaches [21], and Co-Training [22–27]. Co-Training was proposed by Blum and Mitchell [22]. This method requires two sufficient and redundant views which do not exist in most real-world scenarios. In order to relax this constraint, Zhou and Li proposed a Tri-Training algorithm that neither requires the instance space to be described with sufficient and redundant views nor puts any constraints on the supervised learning method [28]. The algorithm uses three classifiers, which can not only tackle the problem of determining how to label the unlabeled data, but also improve generalization ability of a classifier with unlabeled data. Wang et al. made a large number of studies on Co-Training and proved that if two views have large diversity, Co-Training is able to improve the learning performance by exploiting the unlabeled data even with insufficient views [23–25]. Until now, Tri-Training and Co-Training have been widely used in natural language processing. Pierce and Cardie [26] applied Co-Training to noun phrase recognition. They regarded the current word and the k words which appear before the current word in the document as a view and the k words appear after the current word as another view and then trained the classifiers on these two views with Co-Training algorithm. Mavroeidis et al. [29] applied Tri-Training algorithm to spam detection filtering and achieved a satisfactory result.

Meanwhile, the ensemble learning methods have been proposed, which combine the outputs of several base learners to form an integrated output for enhancing the classification performance. There are three popular ensemble methods, that is, Bagging [30], Boosting [31], and Random Subspace [32]. The Bagging method uses random independent bootstrap replicates from a training dataset to construct base learners and calculates the final result by a simple vote [30]. For Boosting method, the base learners are constructed on weighted versions of training set, which are dependent on previous base learners' results and the final result is calculated by a simple vote or a weighted vote [31]. The Random Subspace method uses random subspaces of the feature space to construct the base learners [32].

In our method, we regard three kernels (i.e., the feature kernel, graph kernel, and tree kernel which will be introduced in the following section) as three different views. Co-Training and Tri-Training algorithms are then employed to exploit the unlabeled data with these views and build the disease-symptom model and symptom-therapeutic substance model. Meanwhile, in the Tri-Training process, we adopted the ensemble learning method to integrate three individual kernels and achieved a satisfactory result.

2. Methods

2.1. Feature Kernel. The core work of the feature-based method is feature selection which has a significant impact on the performance. The following features are used in our feature-based kernel.

(1) Word Feature. Word feature uses two disordered sets of words which are between two concept entities (diseases, symptoms, and therapeutic substances) and surrounding two conceptual entities as the eigenvector. The features surrounding two concept entities' names include the left M words of the first concept entity name and the right M words of the second concept entity name (in our experiments, M is set to 4).

(2) N-Gram Word Feature. In our method, we use N-gram (N = 1, 2, and 3 in our experiments) words from the left four words of the first concept entity to the right four words of the second concept as features. N-gram features enrich the word feature and add contextual information, which can effectively express the relation of concept entities.

(3) Position Feature. The relative position information of word feature and N-gram feature for the concept entities has an important influence on relation extraction and, therefore, is introduced into our method. For example, "E1_L_feature" denotes a word feature or N-gram feature appears in the left of first concept entity; "E_B_feature" between two concept entities; "E2_R_feature" in the right of second concept entity.

(4) Interaction Word and Distance Features. Some words such as "induce," "action," and "improve" often imply the

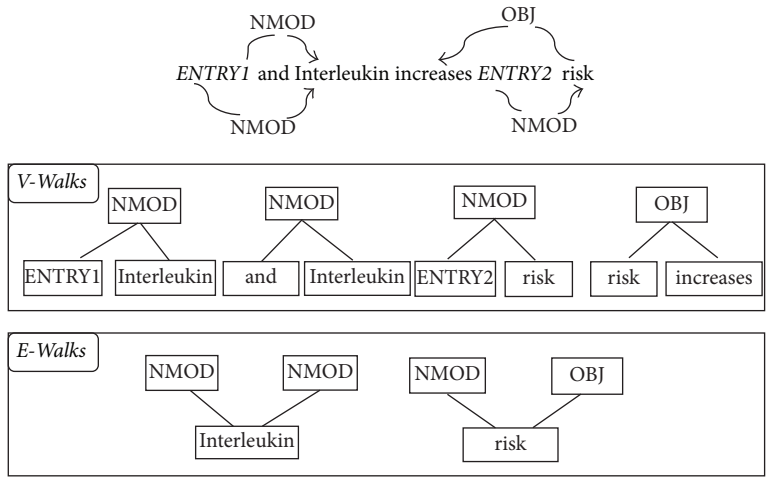


Figure 1: An example of a dependency graph. The candidate interaction pair is marked as "ENTRY1" and "ENTRY2."

existence of relations. Therefore, the existence of these words (we called interaction words) is chosen as a binary feature. In addition, we found that the shorter the distance between two concept entities is, the more likely the two concept entities have an interactive relationship. Therefore, the distance is chosen as a feature. For example, "DISLessThanTree" is a feature value showing that the distance between the two concept entities is less than three.

The initial eigenvector extracted with our feature-based kernel has a high dimension and includes many sparse features. In order to reduce the dimension, we employed the document frequency method [33] to select features. Initially, the feature-based kernel method extracts 248,000 features from the disease-symptom training set and we preserved the features with document frequencies exceeding five (a total of 12,000 features). Similarly, 345,000 features were extracted from the symptom-therapeutic substance training set and 13,700 features were retained.

2.2. Convolution Tree Kernel. In our method, convolution tree kernel $K_c(T_1, T_2)$, a special convolution kernel, is used to obtain useful structural information from substructure. It calculates the syntactic structure similarity between two parse trees by counting the number of common subtrees of the two parse trees rooted by T_1 and T_2:

$$K_c(T_1, T_2) = \sum_{n_1 \in N_1, n_2 \in N_2} \Delta(n_1, n_2), \quad (1)$$

where N_j denotes the set of nodes in the tree T_j and $\Delta(n_1, n_2)$ denotes the number of common subtrees of the two parse trees rooted by n_1 and n_2.

2.2.1. Tree Pruning in Convolution Kernel. In our method, Stanford parser [34] is used to parse the sentences. Before a sentence is parsed, the concept entity pairs in the sentence are replaced with "ENTRY1" and "ENTRY2," and other entities are replaced with "ENTRY." Take gene-gene interaction between C0021764 and interleukin increases C0002395 risk (the sentence is processed with MetaMap, and the two concept entities are represented with their CUIs) for example. It is replaced with "gene-gene interaction between ENTRY1 and interleukin increases ENTRY2 risk." Then, we use Stanford parser to parse the sentence to get a Complete Tree (CT). Since a CT includes too much contextual information which may introduce many noisy features, we used the method described in [35] to obtain the shortest path enclosed tree (SPT),and replace the CT with it. SPT is the smallest common subtree including the two concept entities, which is a part of CT.

2.2.2. Predicate Argument Path. The representation of a predicate argument is a graphic structure, which expresses the deep syntactic and semantic relations between words. In the predicate argument structure, different substructures on the shortest path between the two concept entities have different information. An example of a dependency graph is shown in Figure 1. In our method, v-walk and e-walk features (which are both on the shortest dependency paths) are added into the tree kernel. V-walk contains the syntactic and semantic relations between two words. For example, in Figure 1, the relation between "ENTRY1" and "interleukin" is "NMOD" and the relation between "risk" and "increases" is "OBJ," and so forth. E-walk contains the relations between a word and its two adjacent nodes. Figure 1 shows the relation of "interleukin" with its two adjacent nodes "NMOD" and "NMOD" and the relation of "risk" with its two adjacent nodes "NMOD" and "OBJ."

2.3. Graph Kernel. The graph kernel method uses the syntax tree to express a graph structure of a sentence. The similarity of two graphs is calculated by comparing the relation between two public nodes (vertices). Our method uses the all-paths graph kernel proposed by Airola et al. [16]. The kernel consists of two directed subgraphs, that is, a parse graph and a graph representing the linear order of words. In Figure 2 the upper part is the analysis of the structure subgraph and the lower part is the linear order subgraph. These two subgraphs denote

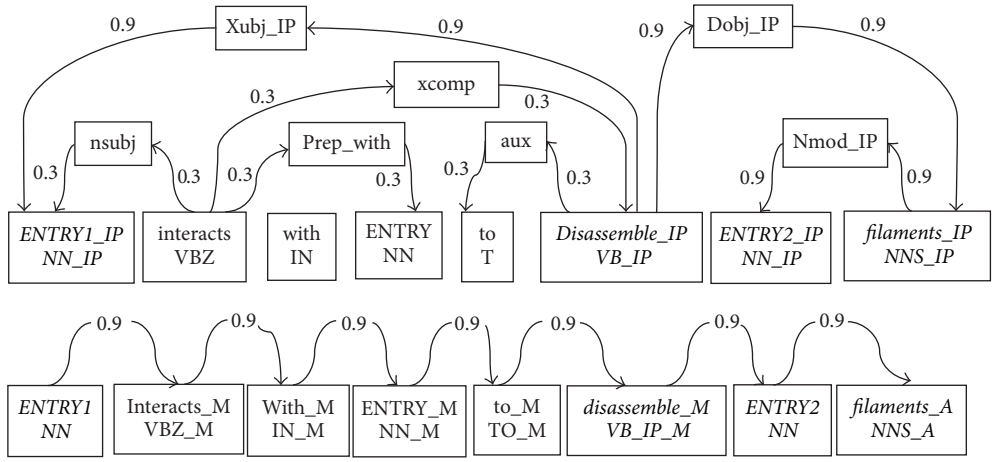


Figure 2: Graph kernel with two directed subgraphs. The candidate interaction pair is marked as "ENTRY1" and "ENTRY2." In the dependency based subgraph all nodes in a shortest path are specialized using a post-tag (IP). In the linear order subgraph possible tags are (B)efore, (M)iddle, and (A)fter.

the dependency structure and linear sequence of a sentence, respectively.

In our method, a simple weight allocation strategy is chosen; that is, the edges of the shortest path are assigned a weight of 0.9; other edges 0.3; all edges in the linear order subgraph 0.9. The representation thus allows us to emphasize the shortest path without completely disregarding potentially relevant words outside of the path. A graph kernel calculates the similarity between two input graphs by comparing the relations between common vertices (nodes). A graph matrix G is calculated as

$$G = L\sum_{n=1}^{\infty} A^n L^T, \tag{2}$$

where A is an edge matrix whose rows and columns are indexed vertices. A_{ij} is a weight if edge V_i is connected to edge V_j. L is the label matrix whose row indicates the label and column indicates the vertex. $L_{ij} = 1$ indicates that vertex V_j contains ith label. The graph kernel $K(G, G')$ is defined by using two input graph matrices G and G' [15].

$$K\left(G, G'\right) = \sum_{i=1}^{L}\sum_{j=1}^{L} G_{ij}G'_{ij}. \tag{3}$$

2.4. Co-Training Algorithm. The initial Co-Training algorithm (or standard Co-Training algorithm) was proposed by Blum and Mitchell [22]. They assumed that the training set has two sufficient and redundant views; namely, the set of attributes meets two conditions. First, each attribute set is sufficient to describe the problem; that is, if the training set is sufficient, each attribute set is able to learn a strong classifier. Second, each attribute set is conditionally independent of the other given the class label. Our Co-Training algorithm is described in Algorithm 1:

Algorithm 1 (Co-Training algorithm).

(1) Input is as follows:

The labeled data L and the unlabeled data U

Initialize training set L_1, L_2 ($L_1 = L_2 = L$)

Sufficient and redundant views: V_1, V_2

Iteration number: N

(2) Process is as follows:

(2.1) Create a pool u of examples by choosing n examples at random from U, $U = U - u$.

(2.2) Use L_1 to train a classifier h_1 in V_1.
Use L_2 to train a classifier h_2 in V_2.

(2.3) Use h_1 and h_2 to label the examples from u.

(2.4) Take m positive examples and m negative examples out, which were consistently labeled by h_1 and h_2. Then take p positive examples out from the m positive examples and add them to L_1 and L_2, respectively. Choose $2m$ examples from U to replenish u, $U = U - 2m$, $N = N - 1$.

(2.5) Repeat the processes (2.2)–(2.4) until the unlabeled corpora U are empty or the number of unlabeled data in u is less than a certain number or $N = 0$.

(3) Outputs are as follows:

The classifiers h_1 and h_2

2.5. Tri-Training Algorithm. The Co-Training algorithm requires two sufficient and redundant views. However, this constraint does not exist in most real-world scenarios. The Tri-Training algorithm neither requires the instance space to be described with sufficient and redundant views and nor puts any constraints on the supervised learning algorithm [28]. In this algorithm, three classifiers are used, which can

Table 1: The details of two corpora.

Corpus	Training set		Test set		Unlabeled data
	Positive	Negative	Positive	Negative	Total
Diseases and symptoms	299	299	249	250	19,298
Symptoms and therapeutic substances	300	300	249	249	19,392

tackle the problem of determining how to label the unlabeled data and produce the final hypothesis. Our Tri-Training algorithm is described in Algorithm 2.

In addition, the different classifiers calculate the similarity with different aspects between the two sentences. Combining the similarities can reduce the danger of missing important features. Therefore, in each Tri-Training round, two different ensemble strategies are used to integrate the three classifiers for further performance improvement. The first strategy integrates the classifiers with a simple voting method. The second strategy assigns each classifier with a different weight. Then the normalized output K of three classifier outputs K_m ($m = 1, 2, 3$) is defined as

$$K = \sum_{m=1}^{M} \sigma_m K_m$$
$$\sum_{m=1}^{M} \sigma_m = 1, \quad \sigma_m \geq 0, \ \forall m, \tag{4}$$

where M represents the number of classifiers ($M = 3$ in our method).

Algorithm 2 (Tri-Training algorithm).

(1) Input is as follows:

The labeled data L and the unlabeled data U

Initializing training set L_1, L_2, L_3 ($L_1 = L_2 = L_3 = L$)

Selecting views: V_1, V_2, and V_3

Iterations number: N

(2) Process is as follows:

(2.1) Create a pool u of examples by choosing n examples at random from U, $U = U - u$.

(2.2) Use L_1 to train a classifier h_1 in V_1.
Use L_2 to train a classifier h_2 in V_2.
Use L_3 to train a classifier h_3 in V_3.

(2.3) Use h_1, h_2, and h_3 to label examples from u.

(2.4) Take m positive examples and m negative examples out, which were consistently labeled by h_1, h_2, and h_3. Then take p_1 positive examples from the m positive examples and add them to L_1, L_2, and L_3, respectively; take p_2 negative examples from the m negative examples and add them to L_1, L_2, and L_3, respectively. Choose $2m$ examples from U to replenish u, $U = U - 2m$, $N = N - 1$.

(2.5) Repeat the processes (2.2)–(2.4) until the unlabeled corpora U are empty or the number of unlabeled data in u is less than a certain number or $N = 0$.

(3) Outputs are as follows:

The classifiers h_1, h_2, and h_3

3. Experiments and Results

3.1. Experimental Datasets. In our experiments, the disease and symptom corpus data was obtained through searching Semantic MEDLINE Database [36] using 200 concepts chosen from MeSH (Medical Subject Headings) with semantic type "*Disease or Syndrome.*" Since these sentences (corpus data) have been processed by SemRep [37], a natural language processing tool based on the rule to identify relationship in the MEDLINE documents, the possibility of the relation between the two concept entities in the sentences is high. To limit the semantic types of two concept entities in a sentence, we only preserved the sentences containing the concepts of the needed semantic types (i.e., *biologic function, cell function, finding, molecular function, organism function, organ or tissue function, pathologic function, phenomenon or process*, and *physiologic function*). Finally, we obtained a total of about 20,400 sentences from which we manually constructed two labeled datasets as the initial training set T_{initial} (598 labeled sentences as shown in Table 1) and test set (499 labeled sentences), respectively.

During the manual annotation, the following criteria are applied: the disease and symptom relationship indicates that the symptom is a physiological phenomenon of the disease. If an instance in a sentence semantically expresses the disease and symptom relationship, it is labeled as a positive example. As in the example provided in Section 1, the sentence "many blood- and blood vessel-related characteristics are typical for *Raynaud* patients: *blood viscosity* and *platelet aggregability* are high" contains two positive examples, that is, *Raynaud* and *blood viscosity* and *Raynaud* and *platelet aggregability*. In addition, some special relationships such as "B in A" and "A can change B" are also classified as the positive examples since they show a physiological phenomenon (B) occurs when someone has the disease (A). However, if a relation in a sentence is only a cooccurrence one, it is labeled as a negative example. For the patterns such as "A is a B" and "A and B" they are labeled as the negative examples since "A is a B" is a "IS A" relation and "A and B" is a coordination relation, which are not the relations we need.

The symptom-therapeutic substance corpus data was obtained as follows. First, some "Alzheimer's disease" related symptom terms were obtained from the Semantic MEDLINE

Database. Then these symptom terms were used to search the database for the sentences which contain the query terms and terms belonging to the semantic types of therapeutic substance (e.g., *pharmacologic substance* and *organic chemical*). We obtained about 20,500 sentences and then manually annotated about 1,100 sentences as the disease-symptom corpora: 600 labeled sentences are used as the initial training set and the remaining 498 labeled sentences as the test set. Similar to the disease and symptom relationship annotation, the following criteria are applied: the symptom-therapeutic substance relationship indicates that a therapeutic substance can relieve a physiological phenomenon. If an instance in a sentence semantically expresses the symptom-therapeutic substance relationship, it is labeled as a positive example. As in the example provided in Section 1, the sentence "*fish oil* and *its active ingredient eicosapentaenoic acid (EPA)* lowered *blood viscosity*" contains two positive examples, that is, *fish oil* and *blood viscosity* and *EPA* and *blood viscosity*.

When the manual annotation process was completed, the level of agreement was estimated. Cohen's kappa scores between each annotator of two corpora are 0.866 and 0.903, respectively, and content analysis researchers generally think of a Cohen's kappa score more than 0.8 as good reliability [38]. In addition, the two corpora are available for academic use (see Supplementary Material available online at http://dx.doi.org/10.1155/2016/3594937).

3.2. Experimental Evaluation. The evaluation metrics used in our experiments are precision (P), recall (R), F-score (F), and Area under Roc Curve (AUC) [39]. They are defined as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F = \frac{2 * P * R}{P + R} \tag{7}$$

$$\mathrm{AUC} = \frac{\sum_{i=1}^{m_+} \sum_{j=1}^{m_-} H\left(x_i - y_j\right)}{m_+ m_-}, \tag{8}$$

where TP denotes true interaction pair; TN denotes true noninteraction pair; FP denotes false interaction pair; and FN denotes false noninteraction pair. F-score is the balanced measure for quantifying the performance of the systems. In addition, the AUC is also used to evaluate the performance of our method. It is not affected by the distribution of data, and it has been advocated to be used for performance evaluation in the machine learning community [40]. In formula (8), m_+ and m_- are the numbers of positive and negative examples, respectively, and $x_1, \ldots, x_{m_+}$ are the outputs of the system for the positive examples, and $y_1, \ldots, y_{m_-}$ are the ones for the negative examples. The function $H(r)$ is defined as follows:

$$H(r) = \begin{cases} 1, & r > 0 \\ 0.5, & r = 0 \\ 0, & r < 0. \end{cases} \tag{9}$$

Table 2: The initial results on the disease-symptom test set. Method 1 integrates three classifiers (the feature kernel, graph kernel, and tree kernel) with the same weight while Method 2 integrates them with a weight ratio of 4 : 4 : 2.

Method	P	R	F-score	AUC
Feature kernel	91.38	62.11	73.95	87.13
Graph kernel	93.87	59.77	73.04	87.21
Tree kernel	69.10	62.89	65.85	73.37
Method 1	92.05	63.28	**75.00**	89.47
Method 2	92.81	60.55	73.29	**89.74**

3.3. The Initial Performance of the Disease-Symptom Model. Table 2 shows the performance of the classifiers on the initial disease-symptom test set. Feature kernel and graph kernel achieve almost the same performance which is better than that of tree kernel. When the three classifiers are integrated with the same weight, the higher F-score (75.00%) is obtained while, when they are integrated with a weight ratio of 4 : 4 : 2, the F-score is a bit lower than that of feature kernel. However, in both cases, the AUC performances are improved, which shows that since different classifiers calculate the similarity with different aspects between two sentences, combining these similarities can boost the performance.

3.3.1. The Performance of Co-Training on the Disease-Symptom Test Set. In our method, the feature set for the disease-symptom model is divided into three views: the feature kernel, graph kernel, and tree kernel. In Co-Training experiments, to compare the results of each combination of two views, the experiments are divided into three groups as shown in Table 3. Each group uses same experimental parameters; that is, u = 4,000, m = 300, and p = 100 (u, m, and p in Algorithm 1). The performance curves of different combinations are shown in Figures 3, 4, and 5, respectively, and their final results with different iteration times (13, 27 and 22, resp.) are shown in Table 3.

From Figures 3, 4, and 5, we can obtain the following observations. (1) With the increase of the iteration time and more unlabeled data added to the training set, the F-score shows a rising trend. The reason is that, as the Co-Training process proceeds, more and more unlabeled data are labelled by one classifier for the other, which improves the performance of both classifiers. However, after a number of iterations, the performance of the classifiers could not be improved any more since too much noise (false positives and false negatives) may be introduced from the unlabeled data. (2) The AUC of classifiers have different trends with different combinations of the views. The AUC of the feature kernel fluctuate around 88% while the ones of the graph kernel fluctuate between 85% and 87%. In contrast, all of the tree kernel's AUC have a rising trend since the performance of the initial tree kernel classifier is relatively low and then improved with the relatively accurate labelled data provided by feature kernel or graph kernel.

In fact, the performance of semisupervised learning algorithms is usually not stable because the unlabeled examples may often be wrongly labeled during the learning

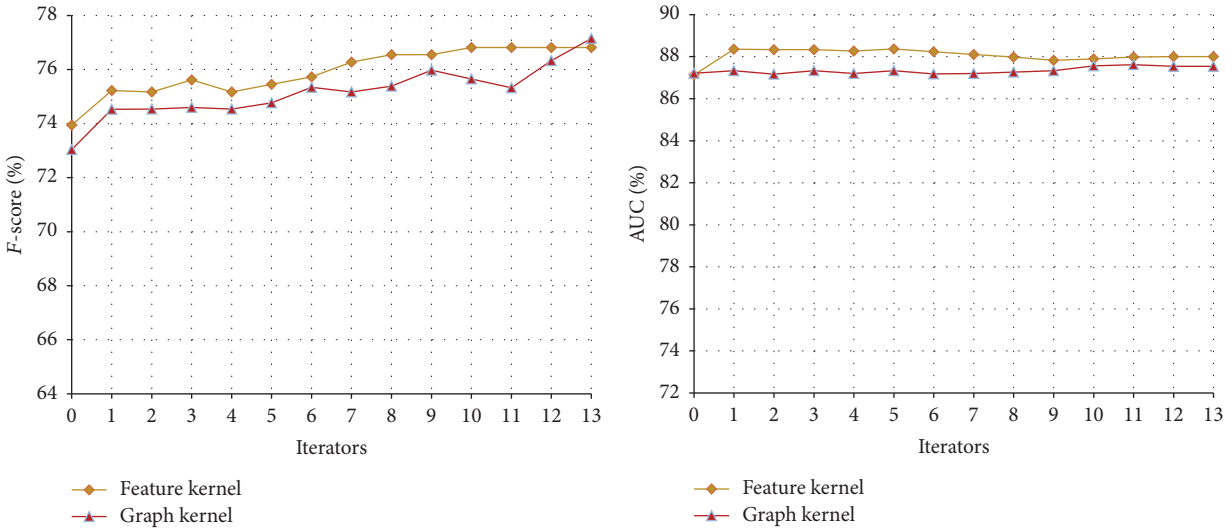


Figure 3: Co-Training performance curve of feature kernel and graph kernel on the disease-symptom test set.

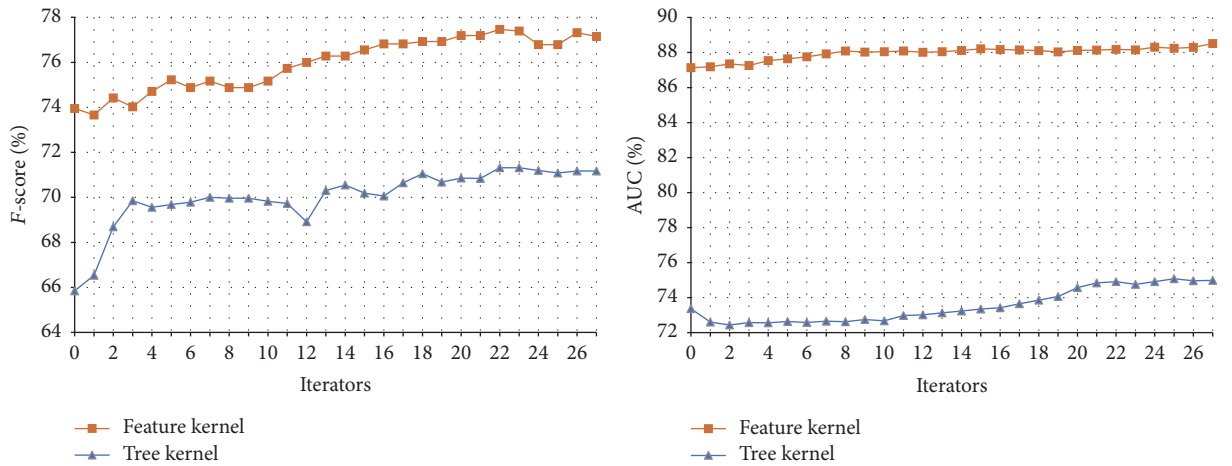


Figure 4: Co-Training performance curve of feature kernel and tree kernel on the disease-symptom test set.

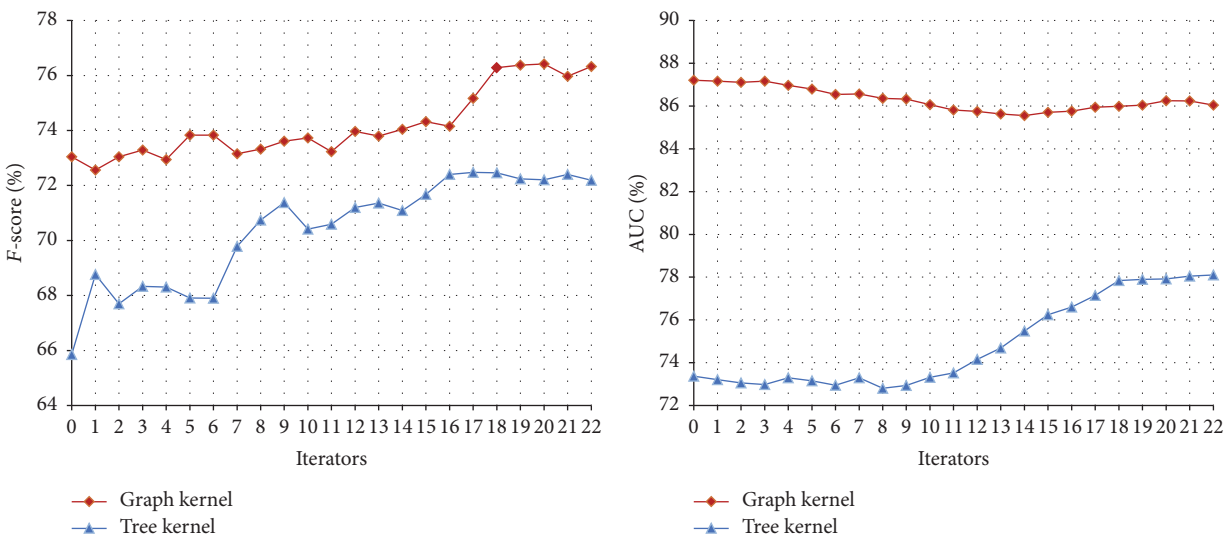


Figure 5: Co-Training performance curve of graph kernel and tree kernel on the disease-symptom test set.

Table 3: The results obtained with Co-Training on the disease-symptom test set. Combination method integrates three classifiers (the feature kernel, graph kernel, and tree kernel) with the same weight.

Combination	View	P	R	F-score	AUC
Feature and graph kernel	Feature kernel	88.32	67.97	76.82	88.01
	Graph kernel	**83.26**	71.88	77.15	87.54
	Combination	74.91	85.16	79.71	**88.66**
Feature and tree kernel	Feature kernel	86.06	69.92	77.15	88.51
	Tree kernel	57.80	92.58	71.17	74.99
	Combination	75.08	87.11	**80.65**	87.18
Graph and tree kernel	Graph kernel	84.04	69.92	76.33	86.04
	Tree kernel	58.10	95.31	72.19	78.10
	Combination	82.43	76.95	79.60	86.84

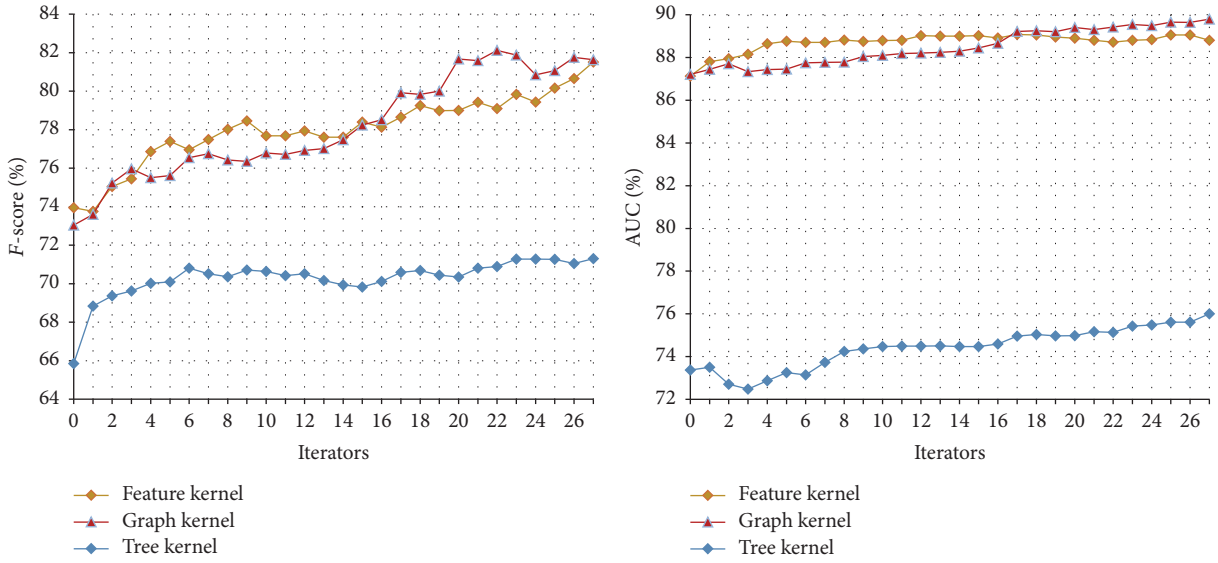


Figure 6: The Tri-Training performance on the disease-symptom test set.

process [28]. At the beginning of the Co-Training, the number of the noises is limited and unlabeled data added to the training set can help the classifiers improve the performance. However, after a number of learning rounds, more and more noises introduced will cause the performance decline.

3.3.2. The Performance of Tri-Training on the Disease-Symptom Test Set. In our method, we select three views to conduct the Tri-Training, that is, the feature kernel, graph kernel, and tree kernel. In each Tri-Training round, SVM is used to train the classifier on each view. The parameters are set as follows: $u = 4{,}000$, $m = 300$, $p_1 = 100$, $p_2 = 0$, and $N = 27$ (u, m, p_1, p_2, and N in Algorithm 2). Here $p_2 = 0$ means that only the positive examples are added into the training set. In this way, the recall of the classifier can be improved (the recall is defined as the number of true positives divided by the total number of examples that actually belong to the positive class and usually more positive examples in the training set will improve the recall) since it is lower compared with the precision (see Table 2). The results are shown in Table 4 and Figure 6.

Table 4: The results obtained with Tri-Training on the disease-symptom test set. Method 1 integrates three classifiers (the feature kernel, graph kernel, and tree kernel) with the same weight while Method 2 integrates them with a weight ratio of 4 : 4 : 2.

Method	P	R	F-score	AUC
Feature kernel	83.00	80.08	81.51	88.80
Graph kernel	77.74	85.94	81.63	89.80
Tree kernel	57.38	94.14	71.30	76.00
Method 1	79.79	87.89	**83.64**	**91.57**
Method 2	79.93	85.55	82.64	90.75

Compared with the performances of the classifiers on the initial disease-symptom test set shown in Table 2, the ones achieved through Tri-Training are significantly improved. This shows that Tri-Training can exploit the unlabeled data and improve the performance more effectively. The reason is that, as mentioned in Section 1, the Tri-Training algorithm can achieve satisfactory results while neither requiring the instance space to be described with sufficient and redundant

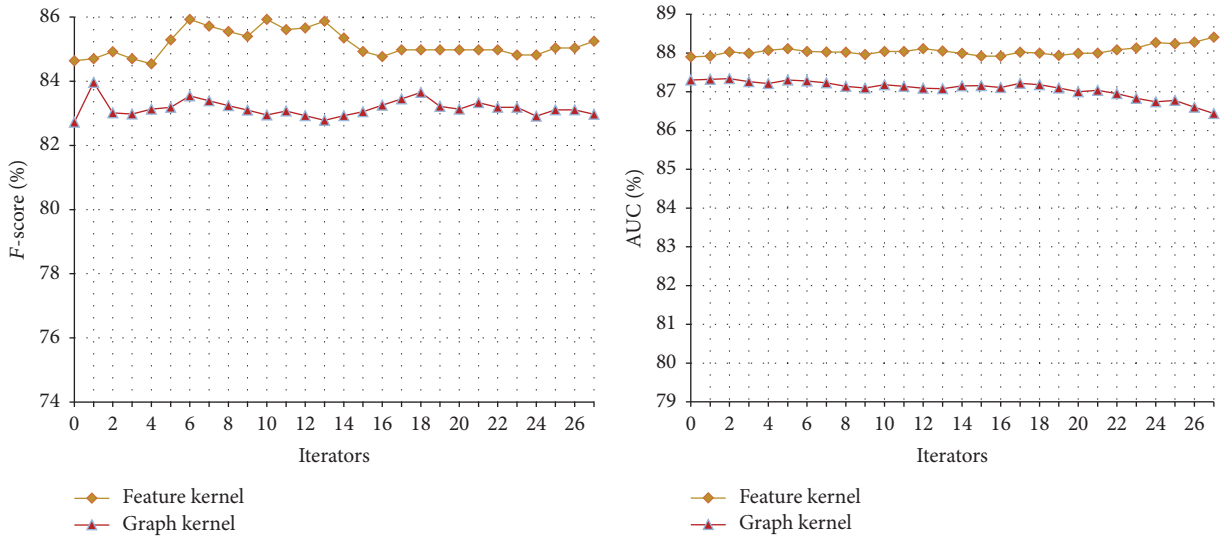


Figure 7: Co-Training performance curve of feature kernel and graph kernel on the symptom-therapeutic substance test set.

views nor putting any constraints on the supervised learning method.

In addition, when three classifiers are integrated either with the same weight or with a weight ratio of 4 : 4 : 2, the higher F-scores and AUCs are obtained. Furthermore, comparing the performance of Co-Training and Tri-Training shown in Tables 3 and 4, we found that, in most cases, Tri-Training outperforms Co-Training. The reason is that, through employing three classifiers, Tri-Training is facilitated with good efficiency and generalization ability because it could gracefully choose examples to label and use multiple classifiers to compose the final hypothesis [28].

3.4. The Performance of the Symptom and Therapeutic Substance Model. Table 5 shows the performances of the classifiers on the initial symptom-therapeutic substance test set. Similar to the results on the initial disease-symptom test set, the feature kernel achieves the best performance while the tree kernel performs the worst. One difference is that when the three classifiers are integrated with a weight ratio of 4 : 4 : 2, the higher F-score and AUC are obtained while, when they are integrated with the same weight, the F-score and AUC are a little lower than those of feature kernel.

3.4.1. The Performance of Co-Training on the Symptom and Therapeutic Substance Test Set. Similar to that in the disease-symptom experiments, the feature set for the symptom-therapeutic substance model is also divided into three views: the feature, graph, and tree kernels. The experiments are divided into three groups. Each group uses the same experimental parameters; that is, u = 4,000, m = 300, and p = 100. The performance curves of different combinations are shown in Figures 7, 8, and 9 and their final results with different iteration times (27, 26, and 9, resp.) are shown in Table 6.

From the figures, we can draw similar conclusions as from the disease-symptom experiments. In most cases, the performance can be improved through the Co-Training process while they are usually not stable since noise will be introduced during the learning process.

Table 5: The initial results on the symptom-therapeutic substance test set.

Method	P	R	F	AUC
Feature kernel	79.30	90.76	84.64	87.90
Graph kernel	76.27	90.36	82.72	87.30
Tree kernel	68.90	82.73	75.18	79.94
Method 1	75.99	92.77	83.54	87.59
Method 2	77.81	94.38	**85.30**	**88.94**

3.4.2. The Performance of Tri-Training on the Symptom and Therapeutic Substance Test Set. In the experiments of Tri-Training on the symptom-therapeutic substance, the parameters are set as follows: u = 4,000, m = 300, p_1 = 100, p_2 = 0, and N = 27 (u, m, p_1, p_2, and N in Algorithm 2). The results are shown in Table 7 and Figure 10.

Compared with the performance of the classifiers on the initial symptom-therapeutic substance test set shown in Table 6, the ones achieved through Tri-Training are also improved as in the disease-symptom experiments. This verifies that the Tri-Training algorithm is effective in utilizing the unlabeled data to boost the relation extraction performance once again. When the three classifiers are integrated with a weight ratio of 4 : 4 : 2, a better AUC is obtained.

Comparing the performance of Co-Training and Tri-Training on the symptom-therapeutic substance test set as shown in Tables 6 and 7, we found that, in most cases, Tri-Training outperforms Co-Training, which is consistent with the results achieved in the disease-symptom experiments. This is due to the better efficiency and generalization ability of Tri-Training over Co-Training.

In addition, the performances of the classifiers on the disease-symptom corpus are improved more than those on

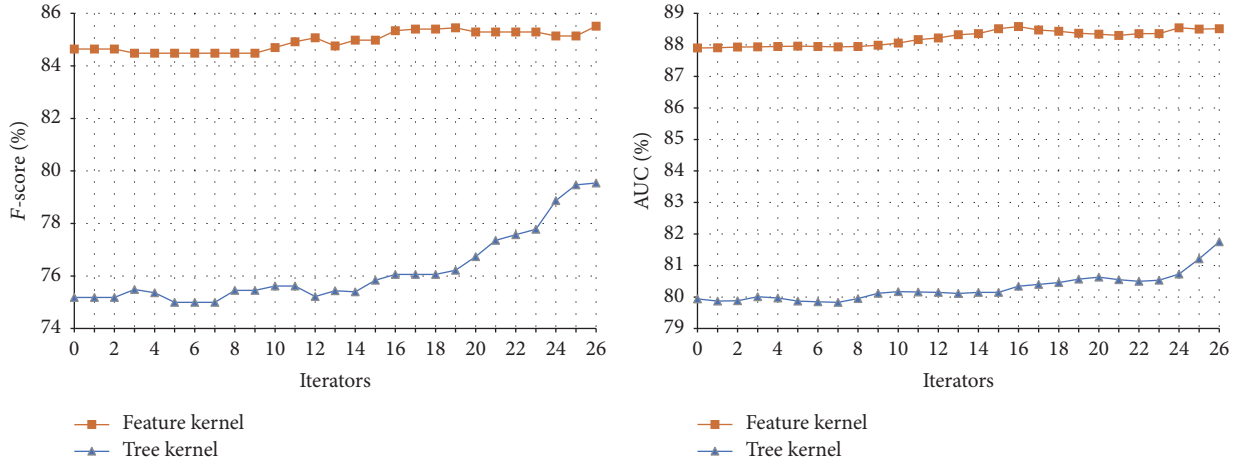


Figure 8: Co-Training performance curve of feature kernel and tree kernel on the symptom-therapeutic substance test set.

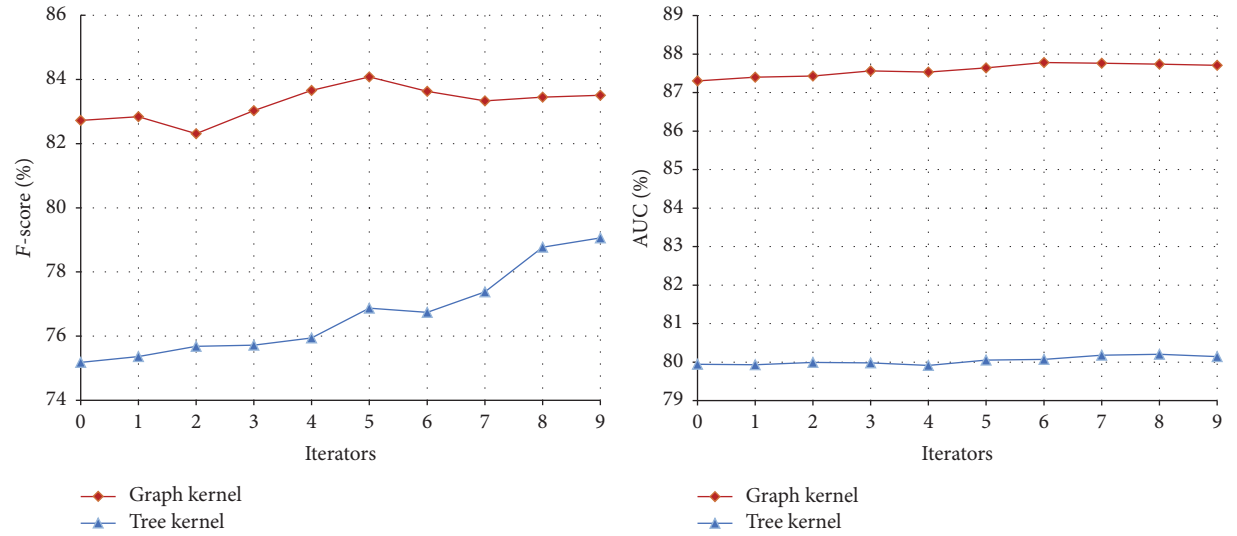


Figure 9: Co-Training performance curve of graph kernel and tree kernel on the symptom-therapeutic substance test set.

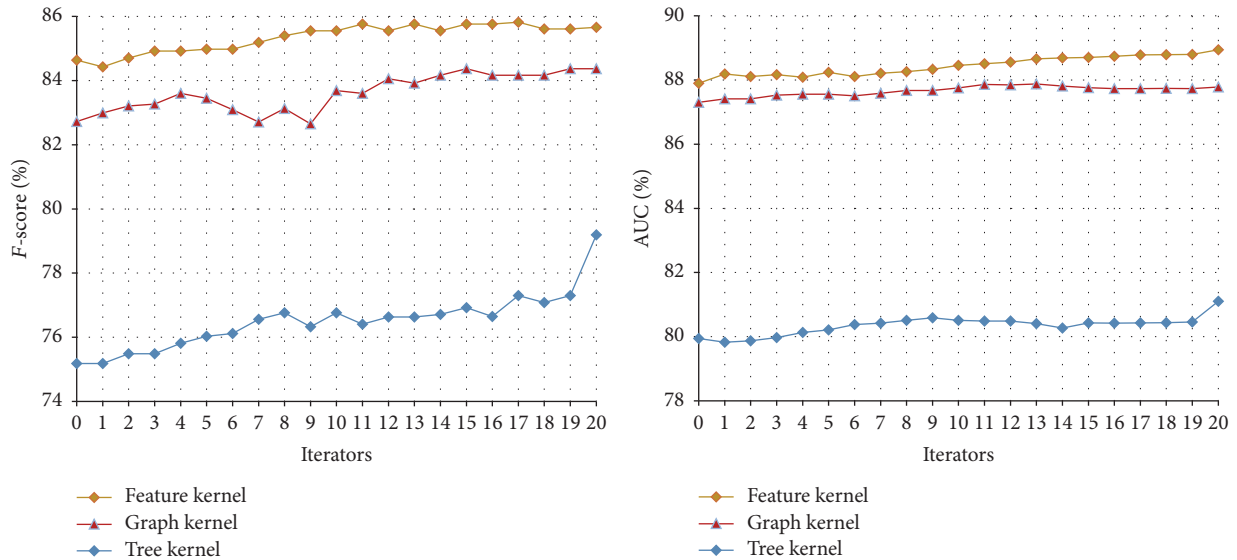


Figure 10: The results of Tri-Training on the symptom-therapeutic substance test set.

Table 6: The results obtained with Co-Training on the symptom-therapeutic substance test set.

Combination	View	P	R	F	AUC
Feature kernel and graph kernel	Feature kernel	78.00	93.98	85.25	88.41
	Graph kernel	71.51	98.80	82.97	86.44
	Combination	77.45	95.18	85.40	**89.10**
Feature kernel and tree kernel	Feature kernel	78.72	93.57	85.51	88.51
	Tree kernel	67.13	97.59	79.54	81.75
	Combination	77.51	96.79	**85.66**	88.61
Graph kernel and tree kernel	Graph kernel	74.14	95.58	83.51	87.71
	Tree kernel	67.82	94.78	79.06	80.14
	Combination	71.05	97.59	82.23	86.24

Table 7: The results of Tri-Training on symptom-therapeutic substance test set.

	P	R	F	AUC
Feature kernel	78.98	93.57	**85.66**	88.94
Graph kernel	74.31	97.59	84.37	87.78
Tree kernel	68.01	94.78	79.19	81.10
Method 1	74.77	98.80	85.12	88.08
Method 2	75.62	98.39	85.51	**89.13**

Table 8: Some disease-symptom relations extracted from biomedical literature.

Disease	Symptom	Sentence
C0020541 (portal hypertension)	C0028778 (block)	*C0020541* as C2825142 of intrahepatic *C0028778* accounted for 83% of the patients (C0023891 65%, meta-C0022346 12%) and C0018920 11%
	C1565860	
	C0035357	
	C0005775	
	C0014867	
	C0232338	

Table 9: Some symptom-therapeutic substance relations extracted from biomedical literature.

Symptom	Therapeutic substance	Sentence
C0028778 (block)	C0017302 (general anesthetic agents)	Use-dependent conduction *C0028778* produced by volatile *C0017302*
	C0006400 (bupivacaine)	Epidural ropivacaine is known to produce less motor *C0028778* compared to *C0006400* at anaesthetic concentrations
	C0053241 (benzoquinone)	In contrast, *C0053241* and hydroquinone led to g2-*C0028778* rather than to a mitotic arrest

the symptom-therapeutic substance corpus. There are two reasons for that. First, on the symptom-therapeutic substance corpus, the classifiers have better performance. Therefore, the Co-training and Tri-training algorithms have less room for the performance improvement. Second, as the Co-training and Tri-training process proceeds, more unlabeled data are added into the training set, which could introduce new information for the classifiers. Therefore, the recalls of the classifiers are improved. Meanwhile, more noise is also introduced causing the precision decline. For the initial classifiers, the higher the precision is, the less the noise is introduced in the iterative process, and the performance of the classifier would be improved. As a summary, if the initial classifiers have big difference, the performance can be improved through two algorithms. In the experiment, when more unlabeled data are added to the training set, the difference between the classifiers becomes smaller. Thus, after a number of iterations, performance could not be improved any more.

3.5. Some Examples for Disease-Symptom and Symptom-Therapeutic Substance Relations Extracted from Biomedical Literatures. Some examples for disease-symptom or symptom-therapeutic substance relations extracted from biomedical literatures are shown in Tables 8 and 9. Table 8 shows some symptoms of disease C0020541 (*portal hypertension*). One sentence containing the relation between *portal hypertension* and its symptom C0028778 (*block*) is provided. Table 9 shows some relations between the symptom C0028778 (*block*) and some therapeutic substances, in which the sentences containing the relations are provided.

4. Conclusions and Future Work

Models for extracting the relations between the disease-symptom and symptom-therapeutic substance are important for further extracting knowledge about diseases and their potential therapeutic substances. However, currently there is no corpus available to train such models. To solve the problem, we first manually annotated two training sets for extracting the relations. Then two semisupervised learning algorithms, that is, Co-Training and Tri-Training, are applied to explore the unlabeled data to boost the performance. Experimental results show that exploiting the unlabeled data with both Co-Training and Tri-Training algorithms can enhance

the performance. In particular, through employing three classifiers, Tri-training is facilitated with good efficiency and generalization ability since it could gracefully choose examples to label and use multiple classifiers to compose the final hypothesis [28]. In addition, its applicability is wide because it neither requires sufficient and redundant views nor puts any constraint on the employed supervised learning algorithm.

In the future work, we will study more effective semisupervised learning methods to exploit the numerous unlabeled data pieces in the biomedical literature. On the other hand, we will apply the disease-symptom and symptom-therapeutic substance models to extract the relations between diseases and therapeutic substances from biomedical literature and predict the potential therapeutic substances for certain diseases [41].

Competing Interests

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

Acknowledgments

This work is supported by the grants from the Natural Science Foundation of China (nos. 61272373, 61070098, 61340020, 61572102, and 61572098), Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China (NCET-13-0084), the Fundamental Research Funds for the Central Universities (nos. DUT13JB09 and DUT14YQ213), and the Major State Research Development Program of China (no. 2016YFC0901902).

References

[1] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, "Using literature-based discovery to identify disease candidate genes," *International Journal of Medical Informatics*, vol. 74, no. 2–4, pp. 289–298, 2005.

[2] M. N. Prichard and C. Shipman Jr., "A three-dimensional model to analyze drug-drug interactions," *Antiviral Research*, vol. 14, no. 4-5, pp. 181–205, 1990.

[3] Q.-C. Bui, S. Katrenko, and P. M. A. Sloot, "A hybrid approach to extract protein-protein interactions," *Bioinformatics*, vol. 27, no. 2, pp. 259–265, 2011.

[4] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, "Overview of the protein-protein interaction annotation extraction task of BioCreative II," *Genome Biology*, vol. 9, supplement 2, article S4, 2008.

[5] I. Segura Bedmar, P. Martinez, and D. Sánchez Cisneros, "The 1st DDIExtraction-2011 challenge task: extraction of Drug-Drug Interactions from biomedical texts," in *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction '11)*, pp. 1–9, Huelva, Spain, September 2011.

[6] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo, *SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)*, Association for Computational Linguistics, 2013.

[7] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology," *Genome Biology*, vol. 6, no. 7, article 224, 2005.

[8] T.-K. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, no. 1, pp. 21–28, 2001.

[9] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB '99)*, pp. 60–67, 1999.

[10] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress," *Briefings in Bioinformatics*, vol. 8, no. 5, pp. 358–375, 2007.

[11] Y. T. Yen, B. Chen, H. W. Chiu, Y. C. Lee, Y. C. Li, and C. Y. Hsu, "Developing an NLP and IR-based algorithm for analyzing gene-disease relationships," *Methods of Information in Medicine*, vol. 45, no. 3, pp. 321–329, 2006.

[12] 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.

[13] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 823–830, Alberta, Canada, July 2004.

[14] J. Xiao, J. Su, G. Zhou et al., "Protein-protein interaction extraction: a supervised learning approach," in *Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine*, pp. 51–59, Hinxton, UK, April 2005.

[15] L. A. Nielsen, "Extracting protein-protein interactions using simple contextual features," in *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pp. 120–121, ACM, June 2006.

[16] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski, "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning," *BMC Bioinformatics*, vol. 9, no. 11, article 52, 2008.

[17] L. Qian and G. Zhou, "Tree kernel-based protein–protein interaction extraction from biomedical literature," *Journal of Biomedical Informatics*, vol. 45, no. 3, pp. 535–543, 2012.

[18] S. Kim, J. Yoon, J. Yang, and S. Park, "Walk-weighted subsequence kernels for protein-protein interaction extraction," *BMC Bioinformatics*, vol. 11, no. 1, article 107, 2010.

[19] D. J. Miller and H. S. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled data," in *Advances in Neural Information Processing Systems*, pp. 571–577, 1997.

[20] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the 16th International Conference on Machine Learning (ICML '99)*, pp. 200–209, 1999.

[21] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, vol. 3, pp. 912–919, Washington, DC, USA, August 2003.

[22] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT '98)*, pp. 92–100, ACM, 1998.

[23] W. Wang and Z. H. Zhou, "Co-training with insufficient views," in *Proceedings of the Asian Conference on Machine Learning*, pp. 467–482, 2013.

[24] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, pp. 1135–1142, June 2010.

[25] W. Wang and H. Zhou Z, "Analyzing co-training style algorithms," in *Machine Learning: ECML 2007*, vol. 4701 of *Lecture Notes in Computer Science*, pp. 454–465, Springer, Berlin, Germany, 2007.

[26] D. Pierce and C. Cardie, "Limitations of co-training for natural language learning from large datasets," in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 1–9, Pittsburgh, Pa, USA, 2001.

[27] S. Kiritchenko and S. Matwin, "Email classification with co-training," in *Proceedings of the Conference of the Center for Advanced Studies on Collaborative Research (CASCON '01)*, pp. 301–312, IBM, Toronto, Canada, November 2011.

[28] Z.-H. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.

[29] D. Mavroeidis, K. Chaidos, S. Pirillos et al., "Using tri-training and support vector machines for addressing the ECML/PKDD 2006 discovery challenge," in *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, pp. 39–47, Berlin, Germany, 2006.

[30] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[31] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[32] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: an experiment on online product reviews," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 46–53, 2010.

[33] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, vol. 97, pp. 412–420, Morgan Kaufmann, San Mateo, Calif, USA, 1997.

[34] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the the 41st Annual Meeting on Association for Computational Linguistics*, vol. 1, pp. 423–430, Association for Computational Linguistics, Sapporo, Japan, July 2003.

[35] M. Zhang, J. Zhang, J. Su et al., "A composite kernel to extract relations between entities with both flat and structured features," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 825–832, Association for Computational Linguistics, 2006.

[36] National Library of Medicine, "Semantic MEDLINE Database," http://skr3.nlm.nih.gov/SemMedDB/.

[37] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 462–477, 2003.

[38] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 248–254, 1996.

[39] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[40] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[41] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, 1986.