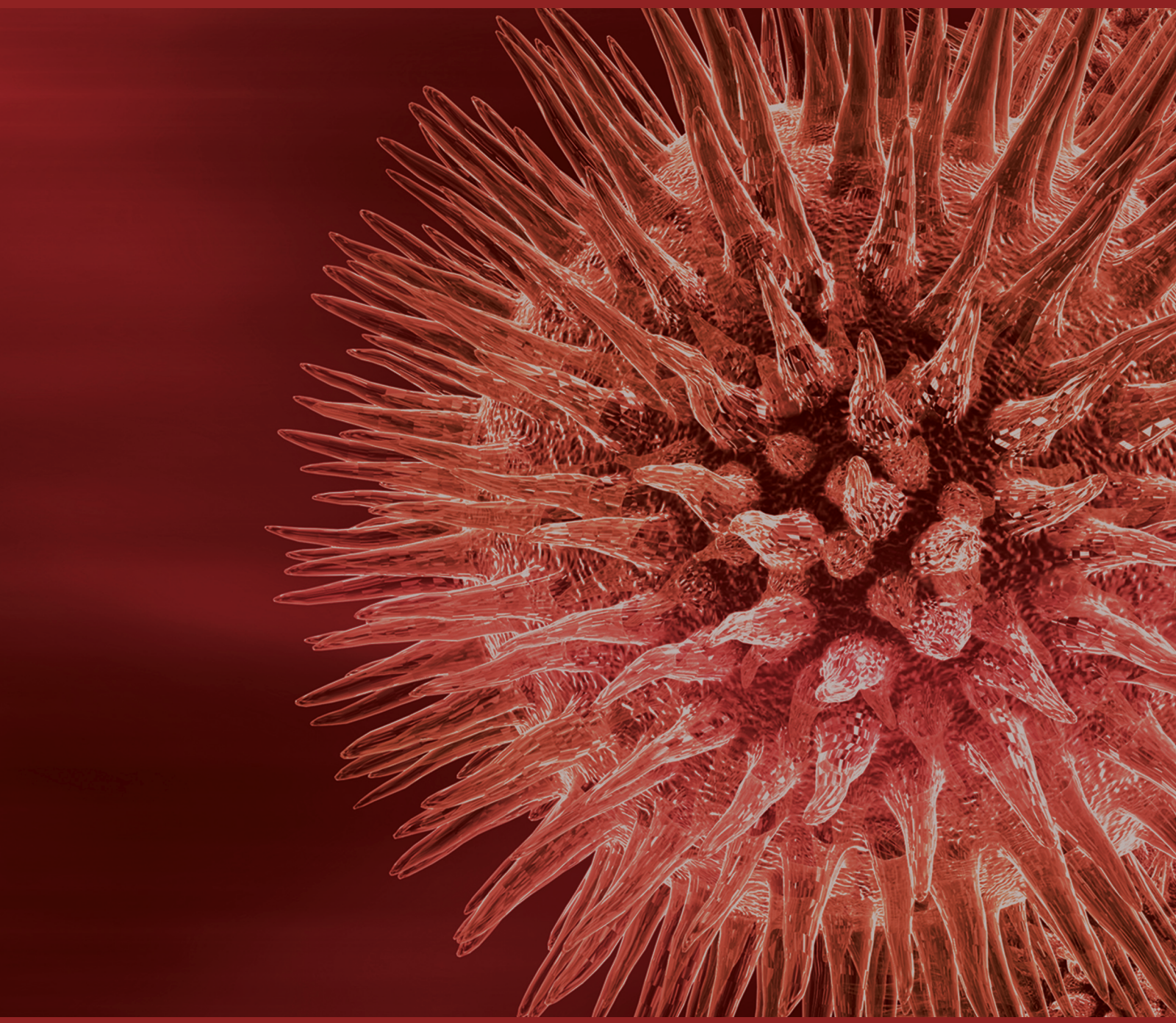
BioMed Research International

Translational Biomedical Informatics and Computational Systems Medicine

Guest Editors: Zhongming Zhao, Bairong Shen, Xinghua Lu, and Wanwipa Vongsangnak



Translational Biomedical Informatics and Computational Systems Medicine

BioMed Research International

Translational Biomedical Informatics and Computational Systems Medicine

Guest Editors: Zhongming Zhao, Bairong Shen, Xinghua Lu, and Wanwipa Vongsangnak

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 237465, 2 pages
http://dx.doi.org/10.1155/2013/237465

Editorial

Translational Biomedical Informatics and Computational Systems Medicine

Zhongming Zhao,[1,2,3,4] Bairong Shen,[5] Xinghua Lu,[6] and Wanwipa Vongsangnak[5]

[1] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA*
[2] *Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[3] *Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37212, USA*
[4] *Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[5] *Center for Systems Biology, Soochow University, Suzhou, Jiangsu 215006, China*
[6] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206, USA*

Correspondence should be addressed to Zhongming Zhao; zhongming.zhao@vanderbilt.edu

Received 7 November 2013; Accepted 7 November 2013

Translational biomedical informatics is rapidly emerging as a new discipline to meet translational medical research demands. This discipline integrates a variety of data from medical research, biological research, and electronic medical records. Computational systems medicine applies computational and systems biology approaches to solve complex problems in medical research; this approach aims for a deeper understanding of disease pathophysiology and a systems level view of disease development. Systems medicine approaches assist investigators with better biomarker discovery and, thus, improve the diagnosis, prognosis, and treatment of complex diseases. Research activities in these areas have rapidly expanded, largely due to the huge volume of data generated from high throughput technologies such as next-generation sequencing (NGS), availability and better management of the massive amount of clinical data, and the demand to effectively link biological and genetic data to clinical records. One example is the B2B program, which includes two iterative components: bench-to-bedside, such that the basic research findings can be translated to clinical practice, and bedside-to-bench, such that the refinements to clinical practice offer new clinical insights and samples for experimental investigation. These complimentary components further enhance translational applications. Among the activities of translational biomedical research and clinical practice, computational approaches, including data curation and management, algorithm and model development, multidimensional data integration, data visualization, and high performance computing, provide fundamental support.

We launched this special issue to address the demand for translational biomedical informatics and discuss the current advances in this field. We are interested in both new theories and tools in this area as well as their applications in translational research. We specifically encouraged the submission of work in areas such as "-omics" data integration and analysis in complex diseases, NGS data analysis and application in medicine, systems biology related research in biomedicine, biomedical data mining and database management, system level modeling and simulation of complex diseases, and visualization of complex data in medicine. Correspondingly, after a rigorous peer review, six papers were selected from the 12 submissions. We briefly describe these papers as follows.

In "*Translational bioinformatics for diagnostic and prognostic prediction of prostate cancer in the next-generation sequencing era,*" J. Chen et al. discussed the current technological advances in molecular biomarker discovery and their translation into the clinical realm for prostate cancer diagnosis and prognosis. The authors reviewed the advances and challenges in the discovery of molecular markers for diagnosis and prognosis of prostate cancer based on high throughput technologies, including microarray and NGS. The authors highlighted 24 prostate cancer NGS studies and

discussed prostate cancer biomarkers at the pathway level. Finally, they provided future direction and perspectives on translational research in prostate cancer.

In "*Exploring the cooccurrence patterns of multiple sets of genomic intervals,*" H. Wu and Z. S. Qin presented a novel statistical method and software tool to characterize the cooccurrence patterns of multiple sets of genomic intervals found in high throughput data such as ChIP-seq. Specifically, they applied a finite mixture model to measure co-occurrence patterns and demonstrated the model's accuracy using simulation and real data. The method is useful to detect co-occurrence patterns in genomic interval-based large datasets.

In "*Diagnosis value of the serum amyloid A test in neonatal sepsis: a meta-analysis,*" H. Yuan et al. performed a meta-analysis of the serum amyloid A (SAA) test as a diagnostic marker for neonatal sepsis. Neonatal sepsis is a common human disorder. It is caused by a bacterial blood stream infection in a newborn baby, which produces a high fever. Through a meta-analysis of studies retrieved from PubMed, EMBASE, the Cochrane Library, and the Google Network between January 1996 and June 2013, the authors found a moderate accuracy of a SAA test in the diagnosis of neonatal sepsis, suggesting that SAA might be promising for the diagnosis of neonatal sepsis.

In "*Characterization of schizophrenia adverse drug interactions through a network approach and drug classification,*" J. Sun et al. first constructed a schizophrenia-specific adverse drug interaction network and then characterized the schizophrenia and adverse drug interactions using the Anatomical Therapeutic Chemical (ATC) classification system. The authors found that schizophrenia drugs tend to have more adverse drug interactions than other drugs. They further revealed the distinct biological features of schizophrenia typical and atypical drugs. This work is the first to characterize the adverse drug interactions in the course of schizophrenia treatment.

In "*CADe system integrated within the electronic health record,*" N. Vállez et al. implemented an electronic health record (EHR) combined with a computer aided detection (CADe) system for breast cancer diagnosis, aiming to provide radiologists with a comprehensive working environment that facilitates the integration, image visualization, and use of aided tools within the EHR. The CADe system allows a user to display, edit, and report results in standardized formats not only for the patient information but also for their medical images. More features will be added in future work.

In "*NCBI2RDF: enabling full RDF-based access to NCBI databases,*" A. Anguita et al. introduced the NCBI2RDF system that provides users with RDF-based access to the complete NCBI data repository. RDF is a standard model for data interchange on the Web and was created by the W3C consortium and accepted as a standard in 2004. The NCBI2RDF, which has two steps (metadata generation and query resolution), enables a user to obtain integrated access to comprehensive data within other existing RDF-based repositories, overcoming current limitations on NCBI data search by implementing its Entrez system.

Acknowledgments

We are grateful to the anonymous reviewers whose critical review helped improve the quality of the papers in this special issue. We would like to acknowledge the organizers and committee members of The First International Conference on Translational Biomedical Informatics (ICTBI 2012, held on December 8–10, 2012) for their efforts to provide a forum to discuss translational biomedical informatics and computational systems medicine, through which this special issue was made possible.

Zhongming Zhao
Bairong Shen
Xinghua Lu
Wanwipa Vongsangnak

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 219407, 14 pages
http://dx.doi.org/10.1155/2013/219407

Research Article

CADe System Integrated within the Electronic Health Record

Noelia Vállez,[1] Gloria Bueno,[1] Óscar Déniz,[1] María del Milagro Fernández,[2] Carlos Pastor,[2] Miguel Ángel Rienda,[2] Pablo Esteve,[2] and María Arias[3]

[1] *VISILAB Group, E.T.S.I. Industriales, University of Castilla-La Mancha, 13071 Ciudad Real, Spain*
[2] *Department of Radiology, University General Hospital of Ciudad Real, 13005 Ciudad Real, Spain*
[3] *Department of Radiology, Hospital of Alcázar de San Juan, 13600 Alcázar de San Juan, Spain*

Correspondence should be addressed to Gloria Bueno; gloria.bueno@uclm.es

Received 26 April 2013; Accepted 10 August 2013

Academic Editor: Xinghua Lu

The latest technological advances and information support systems for clinics and hospitals produce a wide range of possibilities in the storage and retrieval of an ever-growing amount of clinical information as well as in detection and diagnosis. In this work, an Electronic Health Record (EHR) combined with a Computer Aided Detection (CADe) system for breast cancer diagnosis has been implemented. Our objective is to provide to radiologists a comprehensive working environment that facilitates the integration, the image visualization, and the use of aided tools within the EHR. For this reason, a development methodology based on hardware and software system features in addition to system requirements must be present during the whole development process. This will lead to a complete environment for displaying, editing, and reporting results not only for the patient information but also for their medical images in standardised formats such as DICOM and DICOM-SR. As a result, we obtain a CADe system which helps in detecting breast cancer using mammograms and is completely integrated into an EHR.

1. Introduction

Mammograms are difficult to interpret. This fact is especially aggravated in screening campaigns. In these campaigns, radiologists have to examine a large number of mammograms of asymptomatic patients to try to detect breast cancer in its early stages. Therefore, mammograms play a very important role in early detection of this type of cancer. The sensitivity of mammographic studies depends on the quality of the images and the radiologist experience and concentration levels. CADe (Computer-Aided Detection) and CADx (Computer Aided Diagnosis) systems can help radiologists and increase the detection and diagnostic precision by offering a second opinion. This second opinion has lower price than that of a human radiologist. These systems detect suspicious forms in mammograms and mark and classify them according to their malign or benign character. These systems also provide visual results for radiologists [1].

An Electronic Health Record (EHR) is a collection of relevant medical data about the life of patients stored in electronic format [2, 3]. It compounds specific information about patients such as allergies, prescribed medications, immunizations, or major diseases in addition to other information from tests that were performed and the results obtained in them. This information can provide information feedback to health care professionals [4]. Therefore, in an EHR it is possible to find several types of data with different formats, such as, image, video, audio, or text.

Given the possibility that each institution develops its own EHR using different file formats, some standards to facilitate information exchange have been published. The main standard in medicine to store images and results is the Digital Imaging and Communication in Medicine (DICOM) standard [5].

The integration of a display system within the EHR facilitates accessing to patients textual information and displaying medical test images in a complete working environment. This makes all the information concerning a patient accessible from anywhere with appropriate access restrictions.

Most of the current CADe systems are independent of the other applications installed in clinics and hospitals. There is a need to integrate them [6, 7]. Using information from

both the EHR and the CADe system helps to provide a more complete diagnostic [8].

This paper shows the integration of a CADe system within the EHR. Thus, a comprehensive working environment to radiologists which facilitates the integration together with the visualization and aided tools within the EHR has been developed. The aim of the CADe system is to be a tool for displaying medical images and helping in breast cancer detection. Moreover, an EHR in which a CADe system is integrated to complete its functionality has been developed.

Sections 2 and 3 present a study of the current EHRs and CADe systems for mammography in addition to their advantages and disadvantages. Section 4 presents a complete model of the CADe system and the EHR developed. This model is based on the use of a software development methodology, the study of the physical architecture of the system, and compatibility requirements with the most used image format in medicine. Finally, the results and conclusions are outlined in Sections 5 and 6.

2. Electronic Health Record

Nowadays, many health centres offer or are being adapted to offer the possibility of accessing medical information using information systems and advanced technology. Hospitals in Middle East countries [9, 10], most United States clinics [11], and health centres of European countries [12] are some examples.

The use of EHRs has the following advantages [13].

(1) Availability and accessibility of data at any time and anywhere.

(2) Effective patient treatment enhancement by the ability to use saved data for further studies.

(3) Quick and easy access to data.

(4) Reduction of the number of unnecessary tests. Specialists can easily access previous test results that may be still valid.

(5) Reduction in information loss.

(6) Reduction of costs. Using EHRs it is not necessary to print results.

(7) Reduction of data redundancies.

(8) Data access security enhancement.

In contrast, the potential risks that should be considered for the successful design and implementation of EHRs are [14, 15].

(1) Tedious data entry.

(2) Unfriendly user interfaces.

(3) Creating additional and unusual work for physicians.

(4) Interruptions or alterations in the usual workflow.

(5) Lack of network connectivity.

(6) Update failures.

(7) Using nonstandardized terms.

(8) Security failures.

(9) Rejection by system users.

(10) Large response times.

(11) Programming errors.

(12) Limited documentation.

Although there are several studies that have implemented an EHR, we focus our attention on integration of CAD systems within the EHR. As it stands today, there is no one step by step process for developing EHRs. While some authors use Web Services (WS) to access to databases (DB) [2], others use XML files for storing patient data or processing stored data obtained with medical images [16–18].

Other authors use a Web Service Oriented Architecture [19], SOA, that provides interoperability between different systems and programming languages, usability to deploy Web applications, and possibility of reusing implemented functions. SOAs also facilitate application development. In contrast, in some cases, the use of SOA may be a problem because they use a protocol based on plain text in which requests can be larger than in other protocols. Therefore, they can increase the elapsed time and pending client connections.

The EHRs that use XML files to store all the patient information suffer from a large data dispersion along several files. This data dispersion may cause data loss or duplication. Thus, health care record fragmentation must be reduced [20]. Retrieving information about various patients or compiling statistics with a DB offers greater simplicity and speed than XML files.

3. Computer Aided Detection Systems

The computer-aided diagnosis process can be defined as a diagnosis given by radiologists who use an automatic image analysis system to assist them [21]. The system's output acts as a second reader opinion. This second opinion helps radiologists to find previously missed important regions.

The ability to detect lesions by examining a radiological image varies depending on the circumstances of the observer. Fatigue, distraction, or stress are some examples in which a radiologist can ignore early cancer signs. Therefore, a radiologist could detect lesions in some cases and not in others. These facts make CAD systems more suitable for sessions in which a radiologist examines a large number of images with low rate of lesion occurrence.

Most work on CAD systems focuses on breast cancer. The interpretation of mammograms is not easy; even expert radiologists miss early signs of cancer in between 10% and 30% of cases. The use of CAD as a second reading increased cancer detection by 15% [21]. The purpose of computer-aided diagnosis is to provide a second opinion for reducing the number of false negatives (FNs) and false positives (FPs). However, radiologist do not trust CAD results since sometimes CADs can provide a large number of FPs even when the diagnosis was correct [22].

The use of digital image formats to store mammograms, either by direct acquisition of these images or by digitization of printed mammograms, facilitates the work of

radiologists in diagnosing breast cancer. There are two types of CAD systems according to the form by which digital images are obtained: based on digitized screen-film mammograms (SFMs) and based on full-field digital mammograms (FFDMs). SFMs CAD systems have one disadvantage: there is noise in SFMs images produced during the scanning process. By contrast, the use of FFDMs CAD systems offers less noise, larger colour range, and higher contrast than the previous ones [23]. All commercial CAD systems that have obtained FDA approval use FFDMs as input. In general, both types of CAD systems obtain similar accuracy.

In order to make diagnostic decisions, radiologists look for a number of lesion characteristics: contour, density, location, size, and so forth. These characteristics are defined by the ACR (American College of Radiology) [24].

In recent years, many researches have dedicated time, personal and economical efforts to develop this kind of application [25]. In particular, most of the work has been dedicated to the development of the detection and classification functions of CAD systems [26]. Currently, there are at least three mammography CAD systems approved by the FDA: *R2 Technology, Inc.* [27], *Intelligent Systems Software, Inc.* [28], and *CADx Medical Systems.* [29].

However, most of these systems are independent of the other applications that are installed in clinics and hospitals. It is common to find other applications that access images stored in digital format or provide mechanisms to digitize mammographic films in medical centres [6].

4. Materials and Methods

Medical images in DICOM format and test results files in DICOM-SR format (DICOM structured report) are used as input and output files, respectively, by the proposed CAD system. Moreover, it is possible to view images stored in other file formats on the application like *jpg*, *png*, *pgm*, and so on.

In order to design and build both applications, the CAD system for mammography and the EHR, several paradigms have been proposed in the literature. Bernstein et al. [30] describe four different methodologies for developing an EHR: using a semantic model, using the generic model of three layers, using a model based on middleware, and through a model based on communication.

The paradigm proposed in this paper is based on a combination of the generic three-layer model with client-server architecture and a software development methodology such as Unified Process (UP) [31]. DICOM-SR format has been used to store CAD results.

4.1. Images and Reports. Systems based on output files that work in medical institutions do not always satisfy the requirements of some standards that specify the encoding, the storage, and the transfer protocols of medical data. These nonstandardized systems hinder the information interchange between different hospitals.

The DICOM standard was developed to make possible the communication between different clinical and hospital information systems and the appropriate medical image storage and retrieval. In 1993, its creators, ACR (American College of Radiology) and NEMA (National Electrical Manufacturers Association), presented the third version of the standard [5]. The standard describes the file format, the information fields that appear in it, and the image or the images contained in a DICOM file to provide a means to access and interpret all the information of different medical systems. Following the DICOM standard, images are stored together with patient, doctor, or clinic data and the information about results or measures. Therefore, the display system must be able to handle this type of image files.

In addition, DICOM-SR is a part of the DICOM standard that refers to the storage and transmission of clinical reports. The data is stored in a file format similar to the image one.

Another problem is that in some storage systems, results are based on the use of pure text. The pure text is difficult to translate into other languages, may contain ambiguities, and searches using it are not efficient.

Using DICOM-SR format files to store and display the reports in a structured way is a solution to the above problems and adds several advantages [32].

(1) The SR specifies rules that define how structured documents that contain medical results should be created, stored, and transmitted.

(2) Such documents may contain references to other DICOM files such as images and audio files.

(3) SR uses a standard terminology to avoid the ambiguity of natural language, facilitate the automatic interpretation of the content, and improve search.

In our case, it is necessary to store information about images and lesions found in mammograms. The main lesions in breast images are masses and microcalcifications [33]. Figure 1 shows the hierarchy data results that must be saved in case of working with a regular mammographic study which contains four images (a craniocaudal projection (CC) and a mediolateral oblique (MLO) projection by each breast).

4.2. Application Development. For developing the complete application we focus our attention, first, on the physical structure. Currently, the ability to access applications from anywhere and any time is desirable. There is an increasing necessity to use physical architectures that facilitate data access from different locations and devices. To deal with this situation, the client-server architecture is a great solution. In this architecture, each system that operates as a client creates an information demand to the server system. The server systems provide the information required responding to the demand made by the client.

Applications based on this type of hardware architecture do not need to change the whole system with every update. It is sufficient to make small changes in the client side or in the server side independently. It is also possible to move, change, or update some computers on the network without any repercussions on the rest of devices. Sometimes it is possible that changes in server applications affect, minimally, client applications or vice versa.

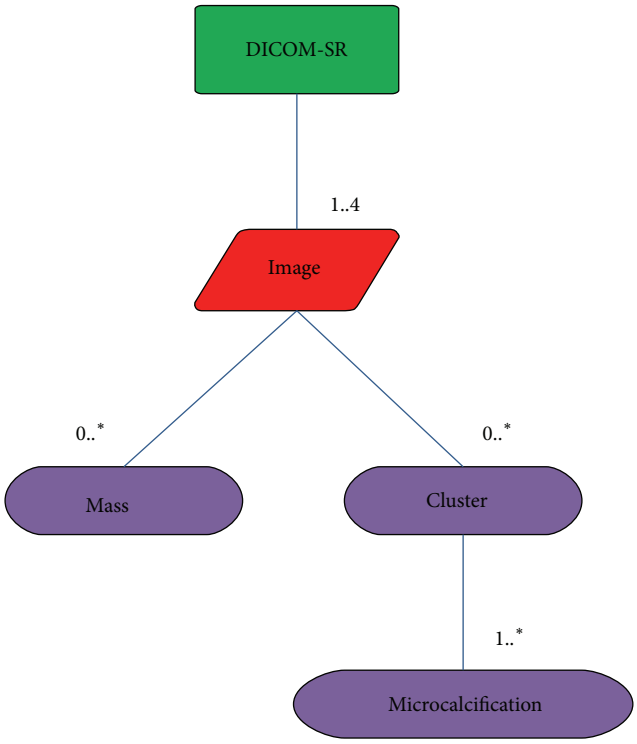


Figure 1: Mammographic structured report hierarchy.

Figure 2: Extract of the implemented EHR principal view.

The EHR proposed in this paper bases its operation on the client-server architecture. This is an application that can be installed on an application server and can be accessed through a web browser from any computer with network and user privilege access. The application server is responsible for carrying out the appropriate operations and facilitating the communication with the DB to provide the necessary information to the clients. The Web browser is not usually connected directly to the application server. In most cases, the connection is done through a web server that acts as an intermediary. The web server, the application server, and the data server can be found in the same physical machine or separated in different computers.

Traditionally, this type of applications has been developed using Common Gateway Interface (CGI). A few years ago Java Servlets appeared for programming server side applications [34]. Servlets are small programs that run on the server after a request made by the client. Java Server Pages (JSP) is an improvement of Servlets and facilitates the work making programming easier and reducing the time used for this task. For coding on the server side Microsoft created Active Server Pages (ASP.NET). We decided to use ASP.NET for developing the EHR as in [35].

The EHR has been developed based on radiologic templates and the ACR standard [24]. These templates are created from the requirements obtained in different meetings and questionnaires with some specialists like in [36]. Moreover, it is important to take into account some EHR specific requirements [37]. When the user starts the application, the EHR shows only the basic fields. If the user needs other data or information, it is possible to display more fields by pressing the corresponding buttons. This makes managing patient information easier. Figures 2 and 3 show the appearance of the mammographic EHR.

In order to develop and integrate the CAD system within the EHR, web oriented programming languages should be used. Java and JavaScript are examples of this type of web programming languages. Specifically, Java Applets are used to develop the CAD system. The use of Java Applets offers the following advantages.

(1) Applets can run on any operating system which has previously installed a JVM (Java Virtual Machine).

(2) New versions of the JVM are able to run applets created for previous versions.

(3) Applets are supported by most Web browsers.

(4) You can have full access to the machine on which the applet is running if the user permits it.

(5) The work can be moved from servers to clients. Sometimes, all the work of several clients should be centralized on the server to prevent overloading and reduce queues.

Although Java Applets have some security restrictions, such as the impossibility to access client resources, it is possible to deal with them packaging the Applet in a JAR-file, that enables to bundle multiple files into a single archive file and sign it. Signed applets do not have the security restrictions that are imposed on unsigned applets and can run outside the security sandbox. Users who verify the signature can grant the JAR-bundled software security privileges.

The EHR and the CAD system have a bidirectional communication. The EHR is responsible for providing the necessary parameters for the CAD initialization by using the ⟨applet⟩ label existing in HTML or the ⟨object⟩ label in HTML5. Once the CAD is initialized, the communication between the EHR and the CAD is made by JavaScript. The JSObject class allows Java and Java Applets to manipulate

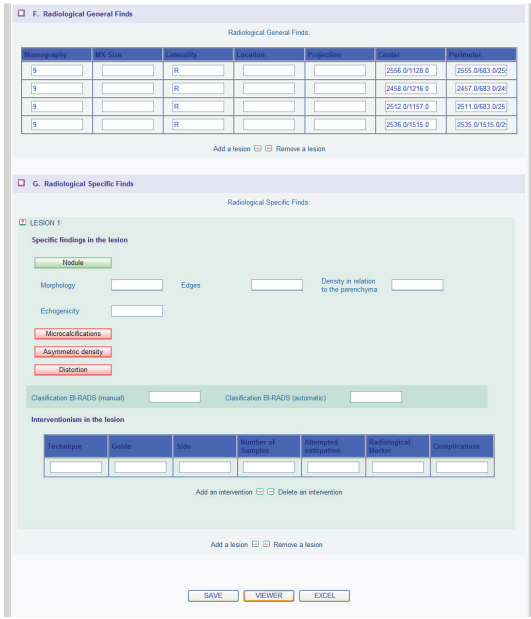


Figure 3: Extract of the implemented EHR information about mammographic findings and CAD link with the button "VIEWER."

objects that are defined in JavaScript. The EHR also has two combo-boxes to select the images that are loaded in the CAD panels. If a different value is selected in any of the combo-boxes, a call to a JavaScript function is made. This function sends the new parameters to a function of the CAD system that handles the change of the images loaded. Moreover, the CAD system is able to communicate results to the EHR by means of calling JavaScript functions residing in the EHR. After receiving the results, the EHR is the responsible for their storing, retrieval, and use.

SR files are used to store CAD results. Once a radiologist select, the final results of the CAD system it is possible to store them in a SR file and in the EHR database. In this step, it is important to control that there are no inconsistencies between data stored in SR files and data stored in the EHR database. For creating SR files the *dcm4che* library is used [33]. To store the results in the EHR database, the CAD system calls to a JavaScript function of the EHR.

For the code structure and the software development process of the application we propose to use the three-layer architecture and the Unified Process (UP). The three-layer model performs a code division according to the responsibilities that has every part of the application code. The classical division of the code divides it into a presentation layer, a domain layer, and a data layer. In our application the use of this architecture

(1) facilitates reuse and migration,

(2) improves human resources distribution,

(3) makes maintenance easier,

(4) limits application changes propagation between layers.

Finally, the software development process is selected according to the software structure that has to be developed. The developed software quality often depends on the methodology used for its construction. Currently, one of the more used methods to develop a complete system is the UP [31]. This development methodology is perfectly compatible with the three-layer architecture. The design is made from the use cases and the system architecture of the application. It is iterative, incremental and tries to predict the risks from the start of the project. The main reasons why we select this development methodology are as follows.

(1) As the system is developed according to the use case diagrams, in which each use case specifies the system functionality, it contains all the functions that should be fulfilled (Figure 4).

(2) The use of UP minimizes the effects of risks by identifying them from the start.

4.3. Image Preprocessing. The idea of using computers in radiographic image analysis is not new. In 1964 an automatic system that determines the proportion of the heart in the thoracic chest X-ray was proposed [38]. In 1967 a system for automatic analysis of mammograms based on bilateral comparison was developed [39]. In 1975 an algorithm for detection of microcalcifications in mammograms based on identifying peaks of grey values was described [21]. After these developments, and due to poor image quality, there was no significant progress in this issue until the late 80s.

Over recent years, various techniques have been developed to detect and automatically characterize masses, microcalcifications, asymmetries, and other lesions in mammography. In addition, many studies on the development of CAD systems have been made [26]. In recent decades there have been important advances in signal processing. The rapidly increasing processing power is responsible for these advances. This increment represents the viability of the use and the study of new algorithms with higher computational costs than before.

Although it is very important to detect and classify masses and microcalcifications, the correct image preparation using preprocessing and editing tools is a crucial step [40, 41]. When digital images and a display system are used, the system must be able to simulate the actual tools that are used by specialists in printed images analysis (Figure 5). Changing the size of the image that is being displayed or performing actions on it such as contrast enhancement or brightness adjustment helps to perform tasks faster, easier, and in a less costly way than if they were carried out manually.

There are several preprocessing techniques that improve image perception highlighting some areas, illuminating the image, or removing the noise [40, 41]. Contrast enhancement is one of the most widely used techniques for improving image visualization and has great application oriented to medical images. Improvements that are mainly achieved are intermediate zones of detail enhancement, edge definition, and overall image contrast [42]. In this case, options for image preprocessing such as changing the brightness, the contrast, and obtaining the negative were added to the visualization

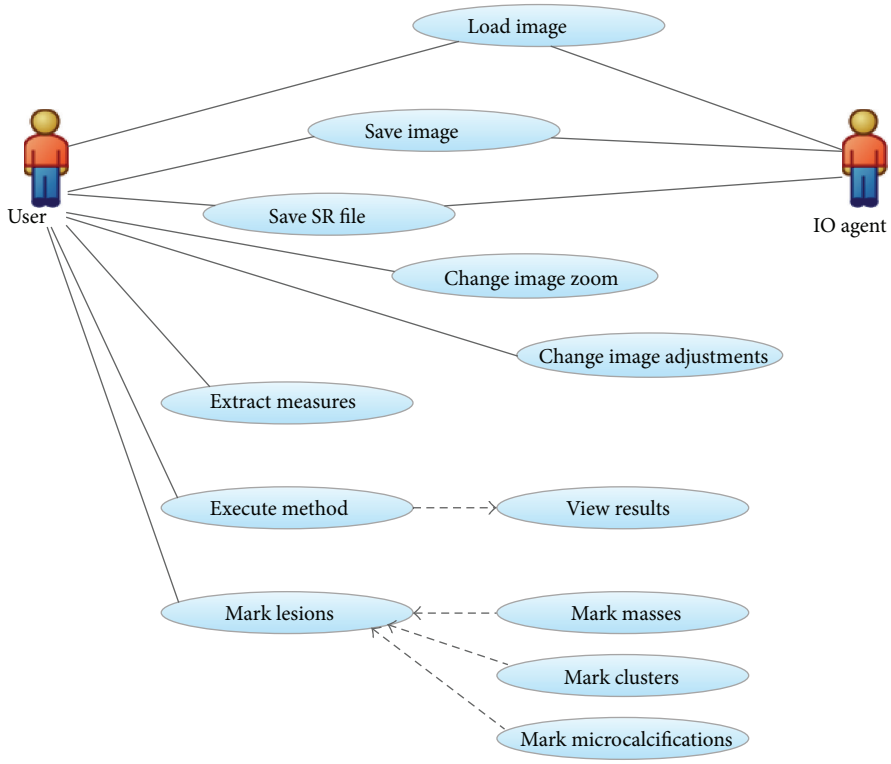


Figure 4: CAD application use case diagram.

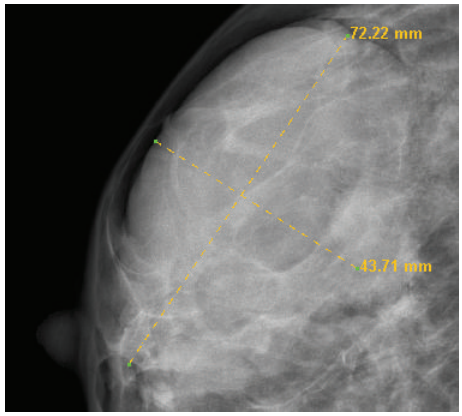


Figure 5: Lesion measures. The values represent the real distance in millimetres.

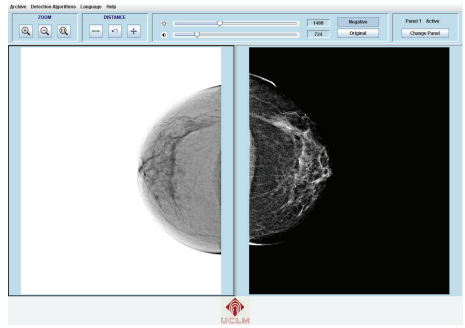


Figure 6: Negative, brightness, and contrast effects on images.

part of the CAD application (Figure 6). In some cases it is easier to identify lesions in the image obtained by inverting the original colour.

4.4. Image Processing. In addition to the preprocessing tools the CAD system should have tools for the detection of two main breast lesions: masses and microcalcifications.

An automatic filter to highlight the visibility of microcalcifications on breast tissue is usually applied. Sometimes it is sufficient to detect clusters of these lesions and not each microcalcification individually [21].

A common practice in mass detection is the bilateral comparison of asymmetric densities. The first stage is characterized by the correct alignment between left and right mammograms. Once the images are well aligned, a subtraction between grey values of the two images is performed, obtaining in this step the density difference [21]. However, this technique has problems because there are natural density differences between the two breasts and this could cause a significant number of false positives.

When processing mammograms it is possible to take two different approaches: make algorithms for contrast enhancement that allow distinguishing lesions from background or performing algorithms to detect and mark them. In [33] several algorithms for helping in detecting masses, microcalcifications, and distortions were developed. These algorithms were integrated into the CAD tool.

(i) *Fuzzy Logic Clustering.* Fuzzy logic clustering helps in mass contour outline. Fuzzy sets are a generalization of

conventional set theory introduced by Zadeh in 1965 as a mathematical way to represent uncertainties [43]. Fuzzy set theory applied to image segmentation is a fuzzy partition of the image data, $I(x, y) = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, into c fuzzy subsets or c specified classes. That is, $\{\mathbf{x_m}(i, j) :\rightarrow [0, 1] : m = 1, \ldots, c\}$, which replaces a crisp membership function that divides the image into c regions by means of a nonfuzzy segmentation process. Thus, image data is clustered into the c different classes by using an unsupervised fuzzy C-mean (FCM) clustering algorithm. This is achieved by computing a measure of membership, called fuzzy membership, at each pixel [44]. The fuzzy membership function, constrained to be between 0 and 1, reflects the degree of similarity between the data value at that location and the centroid of its class. Thus, a high membership value near 1 means that the data value (pixel intensity) at that location is close to the centroid for that particular class. The FCM algorithm is then formulated as the minimization of the squared error with respect to the membership functions, U, and the set of centroids, $\{V\} = \{v_1, v_2, \ldots, v_n\}$:

$$J(U, V : X) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \left\| \mathbf{x_k} - v_i \right\|, \tag{1}$$

where $u_{ik} = u_i(\mathbf{x_k})$ is the membership of $\mathbf{x_k}$ in class i and $m \geq 1$ is a weighting exponent of each fuzzy membership.

The FCM objective function, (1), is minimized when high membership values are assigned to pixels whose intensities are close to the centroid for its particular class and low membership values are assigned when the pixel intensity is far from the centroid. Taking the first derivatives of (1), with respect to u_{ik} and v_k, and setting those equations to zero yield necessary conditions for (1) to be minimized. Then u_{ik} and v_i are defined as follows:

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{\left\| \mathbf{x_k} - v_i \right\|}{\left\| \mathbf{x_k} - v_j \right\|} \right)^{(2/(m-1))} \right]^{-1},$$
$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \quad \forall i, k. \tag{2}$$

Iterating through these conditions leads to a grouped coordinate descent scheme for minimizing the objective function. The stop criterion is determined for each iteration as $E_i < \varepsilon$ where

$$E_i = \sum_{i=1}^{c} \left\| v_{i,t+1} - v_{i,t} \right\| \quad \forall t. \tag{3}$$

The resulting fuzzy segmentation can be converted to a hard or crisp segmentation by assigning each pixel solely to the class that has the highest membership value for that pixel. This is known as a *maximum membership* segmentation. Once the method has converged, a matrix U with the membership or degree to which every pixel is similar to all of the c different classes is obtained. Every pixel is assigned the class for which the maximum membership is found. That is, if $\max(u_{ik}) = u_{ik}$, then $\mathbf{x_k}$ is assigned the label associated with class i. Further modifications of the FCM algorithm may be done by introducing more terms to the objective function (1), for example, in order to cope with noisy images [44].

The result of the FCM algorithm may be rather variable according to the number of selected clusters, c, and the position of the centroids, $\{V\}$. In order to apply the FCM algorithm to the problem of interest a proper configuration of both c and $\{V\}$ has been found from analysis of image histograms [45].

(ii) *Contrast Enhancement Using β-Spline*. It tries to get a picture with the effect of raised areas from the original image using the 3rd order β-Spline transform with an expansion degree of 1:

$$B_1^3(z)^{-1} = \frac{6}{z + 4 + z^{-1}}. \tag{4}$$

The B-spline filter, $(B_1^3)^{-1}$, has been implemented recursively by using a causal ($f_c(z)$) and anticausal ($f_{\text{ac}}(z)$) filter:

$$f_c(z) = \frac{6}{1 - z_1 z^{-1}}$$
$$f_{\text{ac}}(z) = \frac{-z_1}{1 - z_1 z} \tag{5}$$

with $z_1 = -0.268$.

Rewriting the above equation with f_c and f_{ac} filters, the equations will be as follows:

$$B_1^3(z)^{-1} = \frac{-36 z_1}{(1 - z_1 z^{-1})(1 - z_1 z)}$$
$$B_2^3(z) = \frac{z^3 + z^{-3} + 110\left(z^2 + z^{-2}\right)}{2304} + \frac{1087\left(z + z^{-1}\right) + 2212}{2304}. \tag{6}$$

Once the images have been processed with the cubic spline model the first derivative is applied both in X- and Y-axes direction obtaining new coefficients that are rescaled from 0 to 255 for visualization purposes. The results simulate raised areas in the image. This is due to the intensity changes produced on the image when converting from discrete to continuous coefficients with the previous transform.

A useful characteristic highlighted by the clinicians is that the B-spline transform keeps the original size of the calcium nodes.

Finally, the modulus of the B-spline filtering is calculated for exact detection. Areas with higher intensity values and greater variation of these values will have greater prominence. A breast lesion over an enhanced background is obtained with this technique. This method enhances both masses and microcalcifications.

(iii) *Adaptive Filtering*. It is based on the fact that the outline of objects within an image is characterized by discontinuities. It tries to find and mark such discontinuities by means of least squares, in a neighbourhood defined by a mask

known as the 2D lattice structure. This algorithm is useful for microcalcification detection.

The lattice structure is composed of multi-inputs and multioutputs defined by a coefficient of variation. The outputs are backward and forward prediction errors, $e_M^b(n)$ and $e_M^f(n)$, calculated simultaneously. Thus, the algorithm is based on a 2D predictor-estimator within a region, R, of size $(m+1, n+1)$. The output of a signal at point $x(n_i, n_j)$, with i, j belonging to R, is the estimation of the following element, calculated as a linear combination (l.c.) of the $(m+1, n+1)$ elements of the lattice structure; that is,

$$\begin{aligned} \overline{x}(n_1, n_2) &= \sum_{i,j,R} a_{i,j} x(n_1 - i,\, n_2 - j) \\ &= x^f(n_1, n_2) \\ \overline{x}(n_3 - m,\, n_4 - n) &= \sum_{i,j,R} a_{i,j} x(n_3 - i + 1,\, n_4 - j + 1) \\ &= x^b(n_3, n_4). \end{aligned} \tag{7}$$

Once the prediction error is obtained for each pixel, a threshold, γ, is defined to establish if the process is stationary or not. From a geometrical point of view, the prediction error is the projection of the input signal over the l.c. calculated with the elements of the 2D lattice structure. Thus, an angular coordinate, θ, is defined as the angle from the signal to the prediction space; those values of θ close to 0 mean a good approximation. The threshold, γ, is therefore defined by the following equation:

$$\gamma_R(n) = \cos^2(\theta) \tag{8}$$

being close to 1 when the prediction error is small and varying with discontinuities.

As shown in (8), γ also varies with the prediction process; thus, the threshold must be based on the gradient image. The FP fraction detection is dependant on a good selection of the threshold. Finally, those pixels detected by the adaptative filter are grouped together by a region growing algorithm. In order to avoid FP the region growing algorithm discharges those ROI smaller than a certain size (of about 5 pixels).

The algorithm works well for fatty tissue and small lesions, being satisfactory to detect suspicious and isolated lesions. However, it may fail with many FP detections if a proper configuration of γ is not selected. The number of FP is higher on dense and fatty-glandular tissues due to their low contrast; it has been shown how the B-spline algorithm works better for this type of tissue.

5. Results and Discussion

An EHR and a CAD system have been developed and integrated. Figures 2 and 3 show the appearance of the mammographic EHR. It is possible to add, edit, and save data, test results, generate reports, and show images associated with a radiological study. These images are displayed on the CADe application where it is possible to use editing, preprocessing, and processing tools to find and measure lesions (see Figures 5 and 6). Moreover, the CAD system provides tools to mark lesions once they have been located (Figure 7). After marking lesions, results can be stored persistently through the database and the DICOM-SR output files.

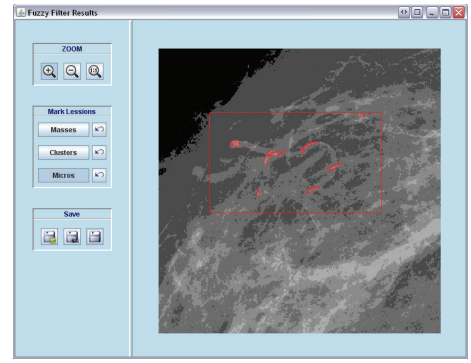


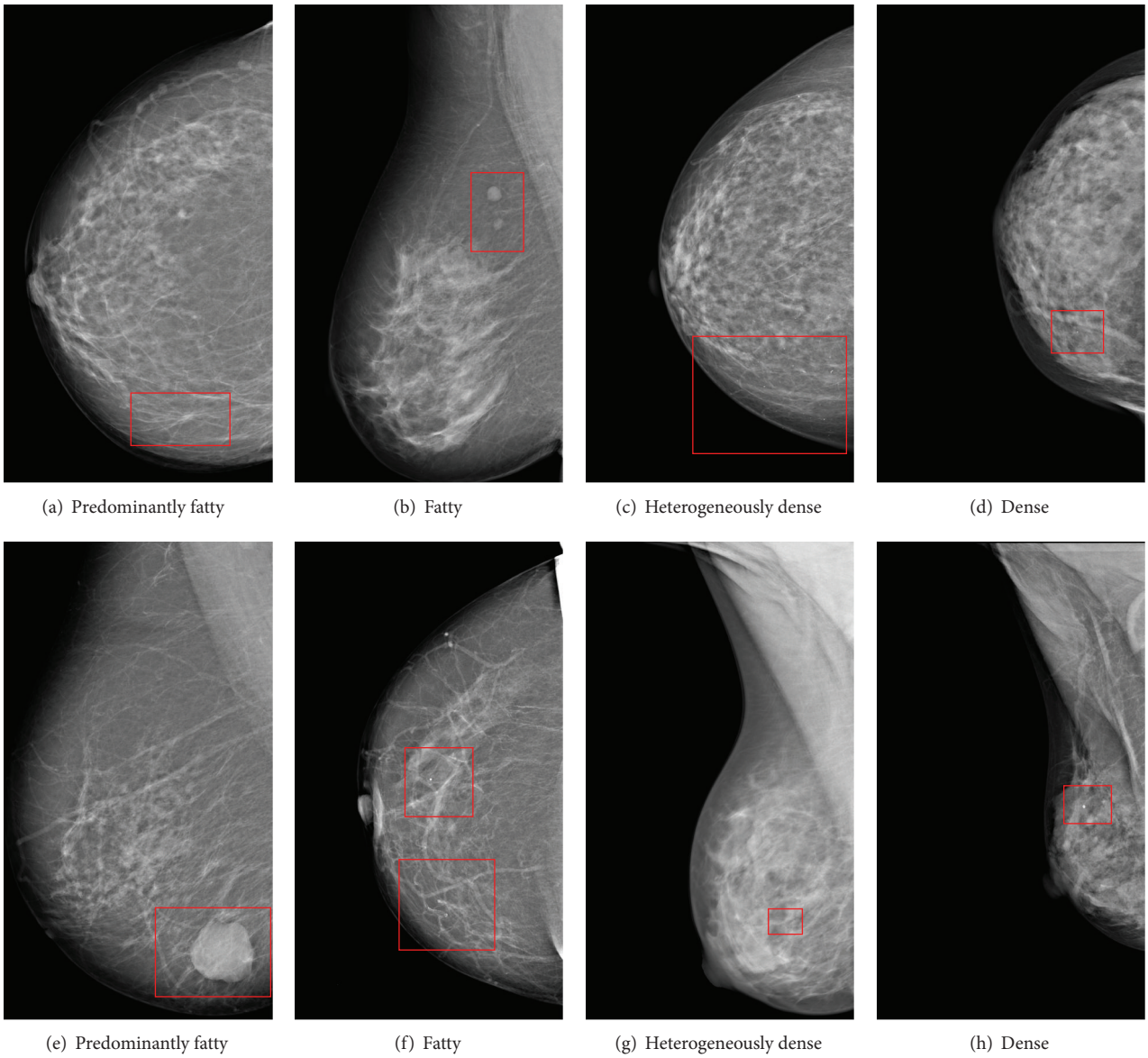
Figure 7: A marked cluster of microcalcifications.

5.1. Results of the CAD Lesion Detection Algorithms. The processing tools implemented into the CADe system have been tested and qualitatively validated by expert clinicians at Hospital General Universitario of Ciudad Real. The results obtained with the CAD segmentation algorithms on eight of the images are illustrated in Figures 8, 9, and 10. These images represent the four tissue types specified by the ACR in the BI-RADS Atlas [24], predominantly fatty, fatty, heterogeneously dense, and dense [24]. In addition, each image belongs to different patients and projections (CC and MLO).

Figure 8(a) shows a spiculated mass at the bottom of the breast tissue and no microcalcifications were present (Figures 9(a)–9(d)). It has a predominantly fatty breast tissue. Making the contrast enhancement using fuzzy clustering we can obtain an output that adequately defines the contours of the lesion. This new image provides uniformity to the background and highlights the lesions assigning similar values of neighbouring pixels. The result from the adaptive filter shows a small mark in the spiculated lesion. In case of using β-Spline it is possible to get an image that enhances the lesion clearly distinguished from the background.

In Figure 8(b) it is possible to see two well-defined masses on top of the mammographic tissue (Figures 9(e)–9(h)). The breast tissue in this case is fatty. As in the previous case this image does not have microcalcifications and making the contrast enhancement using fuzzy clustering an output that adequately defines the contours of the lesion is obtained. The adaptive filter marks some FP points in a dense area of the parenchyma. Using β-Splines it is possible to get an image that highlights the lesions through a raised area.

Figure 8(c) contains a heterogeneous tissue. There are microcalcifications at the bottom of the breast tissue in this case (Figures 9(i)–9(l)). With fuzzy logic clustering an image that highlights the microcalcifications and their contour is achieved. In this case, the adaptive filter does not produce marks. Using β-Splines the image obtained as output presents a raised effect in the lesions and makes these quickly and clearly visible.



(a) Predominantly fatty (b) Fatty (c) Heterogeneously dense (d) Dense

(e) Predominantly fatty (f) Fatty (g) Heterogeneously dense (h) Dense

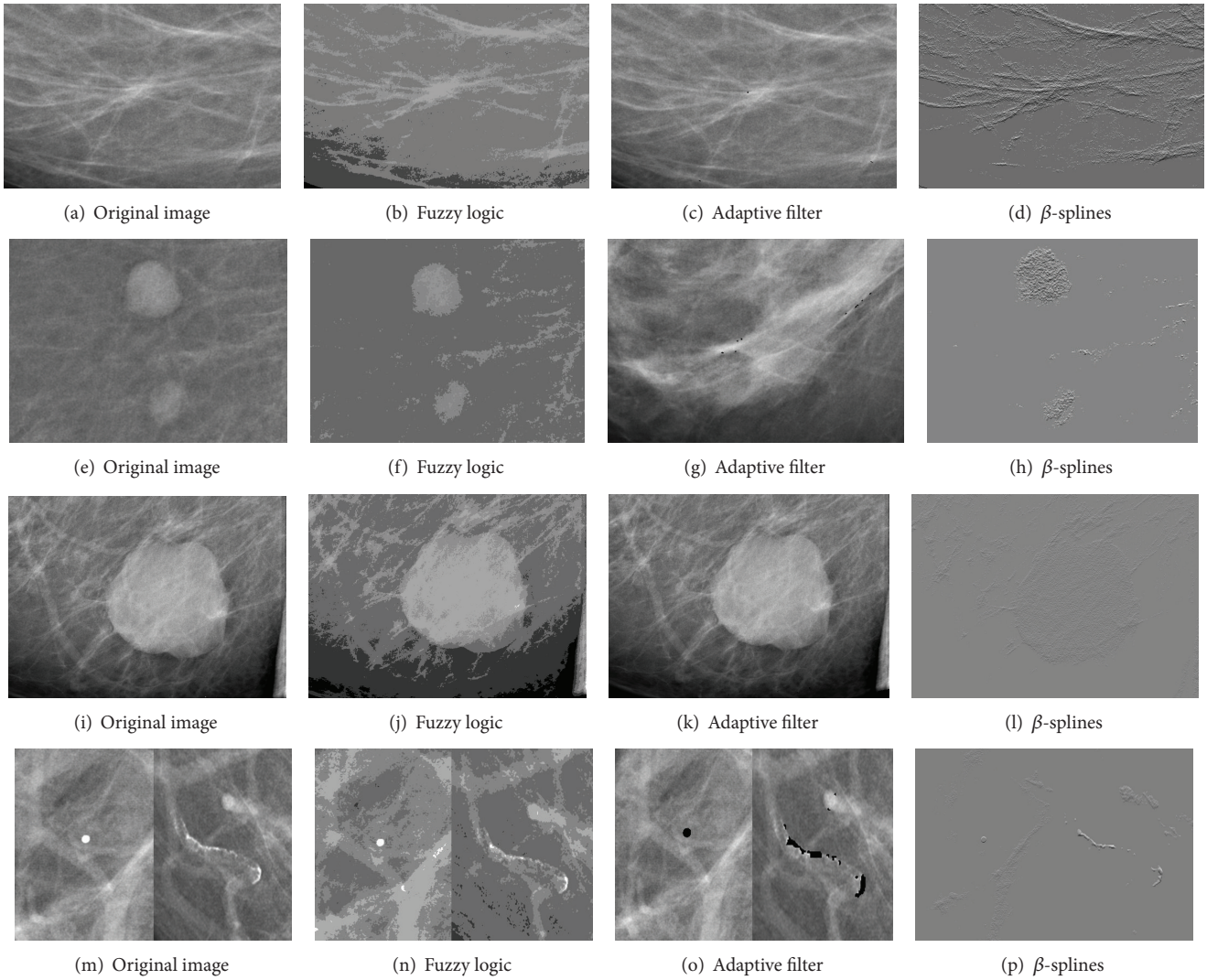
Figure 8: Some test images with (a) spiculated mass, (b) two well-defined masses, (c) region microcalcifications, (d) microcalcifications, (e) a well-defined mass, (f) a rounded microcalcification and some vascular microcalcifications, (g) a small microcalcification, and (h) a microcalcification.

Figure 8(d) contains an extremely dense mammographic tissue. At the bottom of the mammographic tissue appears an isolated microcalcification (Figures 9(m)–9(p)). With fuzzy logic clustering it is possible to obtain an image that facilitates lesion recognition and contour definition. The adaptive filter marks correctly the microcalcification but, in this tissue type, some false-positives appear caused by the high intensity of the breast tissue. By means of β-Splines an image with an homogeneous background and a small raised area in the lesion is obtained.

In Figure 8(e) it is possible to see a large well surrounded mass at the bottom of the breast tissue (Figures 10(a)–10(d)). The breast tissue does not present microcalcifications in this case and follows a predominantly fatty pattern. Enhancing contrast by using fuzzy clustering allows to define the contours of the lesion correctly. In case of using β-Spline present a raised area from the background.

Figure 8(f) has an isolated microcalcification in the center of the image and some vascular microcalcifications at the bottom (Figures 10(e)–10(h)). The breast tissue in this case is fatty. With the contrast enhancement by using fuzzy clustering an output that adequately defines the contours of the lesions is obtained. The adaptive filter marks correctly the microcalcification in both cases. In case of using β-Spline, the image obtained presents a raised area distinguishing the lesion from the background.

Figure 8(g) contains a heterogeneously tissue. There is only one small microcalcification in the center of the breast



(a) Original image (b) Fuzzy logic (c) Adaptive filter (d) β-splines

(e) Original image (f) Fuzzy logic (g) Adaptive filter (h) β-splines

(i) Original image (j) Fuzzy logic (k) Adaptive filter (l) β-splines

(m) Original image (n) Fuzzy logic (o) Adaptive filter (p) β-splines

Figure 9: Results of the processing tools on predominately fatty (a, e) and fatty images (b, f) from Figure 8.

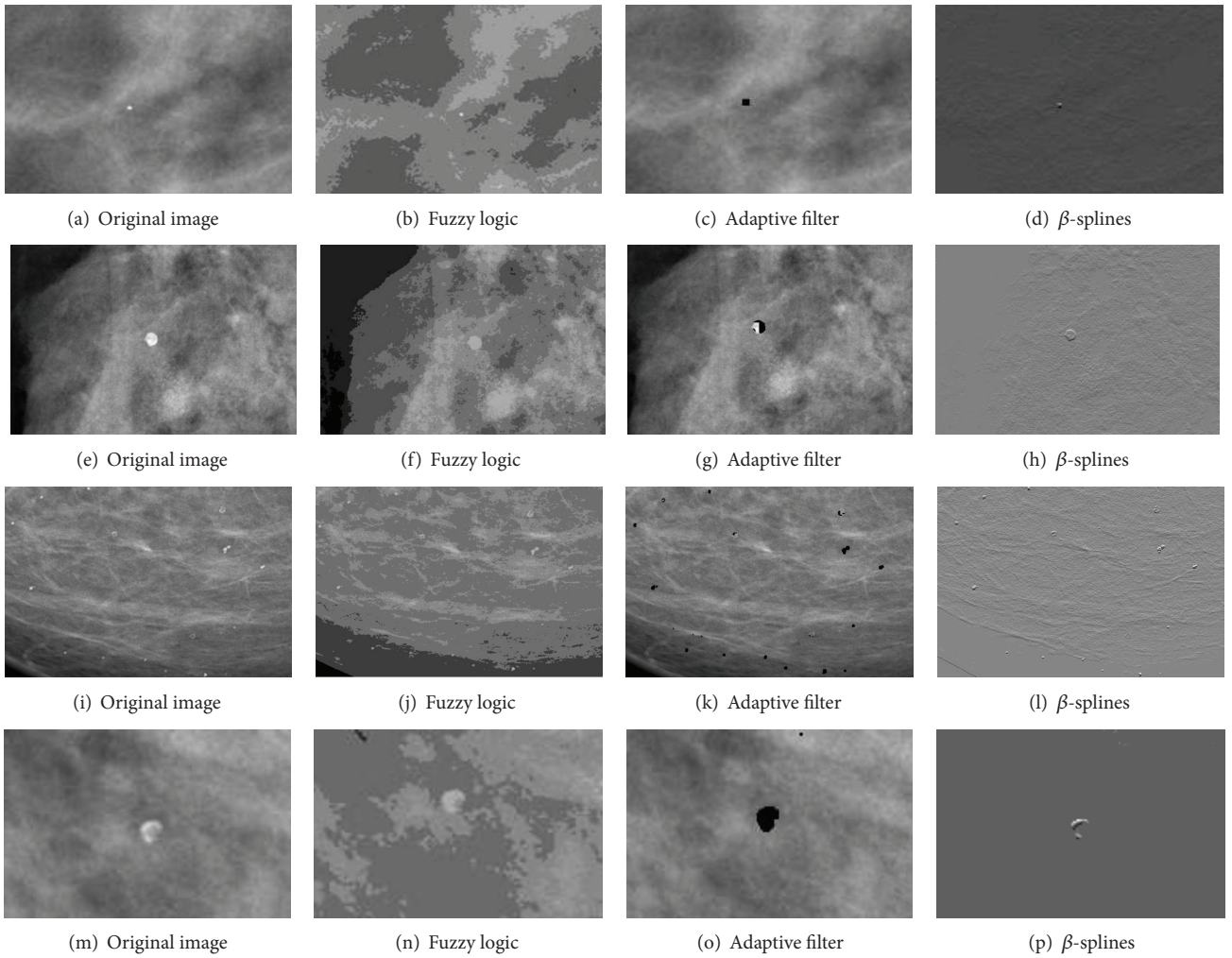
tissue (Figures 10(i)–10(l)). With fuzzy logic clustering an image that highlights the microcalcifications and their contour is achieved. Applying the adaptive filter the microcalcification is detected and marked correctly. Using β-Splines, the image obtained as output presents a raised area that makes the lesion clearly visible.

Finally, Figure 8(h) contains an extremely dense mammographic tissue. This image contains also a microcalcification (Figures 10(m)–10(p)). With fuzzy logic clustering it is possible to obtain an image that defines correctly the contour. The adaptive filter marks correctly the microcalcification. By means of β-Splines an image with a homogeneous background and a raised area in the lesion is obtained.

The use of a specific algorithm or another depends on the lesion that we are looking for. For masses, the algorithm which provides better results is the fuzzy logic clustering method. This algorithm helps to define lesion contours and offers computational time significantly lower than the rest of the methods. In case of detecting microcalcifications we can arrive at different conclusions depending on the tissue type from the mammogram. On mammograms whose tissue type is extremely dense the algorithm that gives better results is the algorithm based on β-Spline contrast enhancement. With this tissue type, using the fuzzy logic clustering algorithm does not adequately distinguish microcalcifications. For the other tissues, both, the adaptive filter or the β-Spline based-algorithm, may be used.

The filtering presented has shown to be successful in highlighting breast lesions on different types of tissues. It is worth mentioning the comments made by the clinicians. The tools improve the resolution, in terms of the detectability of lesions, and additionally, they are able to distinguish their degrees of attenuation.

5.2. Effectiveness of the Whole System. In order to test the effectiveness of the implemented system a study composed of 28 cases belonging to 26 patients, that is, a total of 194 images from the parenchyma area, has been carried out. The 194 images are divided into 82 images without lesions, 89 with benign lesions, and 23 with malign lesions.



(a) Original image (b) Fuzzy logic (c) Adaptive filter (d) β-splines

(e) Original image (f) Fuzzy logic (g) Adaptive filter (h) β-splines

(i) Original image (j) Fuzzy logic (k) Adaptive filter (l) β-splines

(m) Original image (n) Fuzzy logic (o) Adaptive filter (p) β-splines

Figure 10: Results of the processing tools on heterogeneously dense (c, g) and dense images (d, h) from Figure 8.

The study was carried out by 4 radiologists from local hospitals. The EHR was populated with general information of the patient and specific information which may influence breast cancer diagnosis as well as the image density and therefore the amount of false positive detections. Table 1 shows the results of the complete system, comparing the performance of the CADe with and without using the specific information provided by the EHR. The algorithms incorporated into the CAD do not distinguish between benign and malign lesions. However, when the specific EHR information was used and this information was hinted at a malign lesion, then if the CAD detected a lesion this was highlighted in red, in other cases the adaptive filter was run again with different parameters. The adaptive filter was also run twice with different parameters in the case where a lesion was detected and the data from the EHR was not sensitive of being cancer.

The results were validated by the radiologists, comparing the automatic results given by the CAD with those given by the radiologist. Thus, the integration of the EHR information was useful to increase the number of true positive detections (TPDs) and reduce the number of false positives (FPDs).

Table 1: Performance of the whole system (% TPD and % FPD).

% Detections	CADe	CADe + EHR
TPD	75%	85%
FPD	25%	12%

Moreover, the evaluation of the usability of the system made by the clinicians through some questionnaires, like in [46], found the whole system very useful because (a) it allows to do prospective studies and review previous detected lesions, (b) it has all the useful information integrated in a single application, and (c) the EHR may help classifying the lesions by means of rule based classification methods and developing dedicated ontologies.

5.3. Computational Time of the CAD Tools. In any system it is important to take into account the computational time required to obtain the results. The computational times of the preprocessing and the image processing algorithms have been obtained during the system tests. All times were obtained with an Intel Xeon CPU E5440 2.83 GHz with

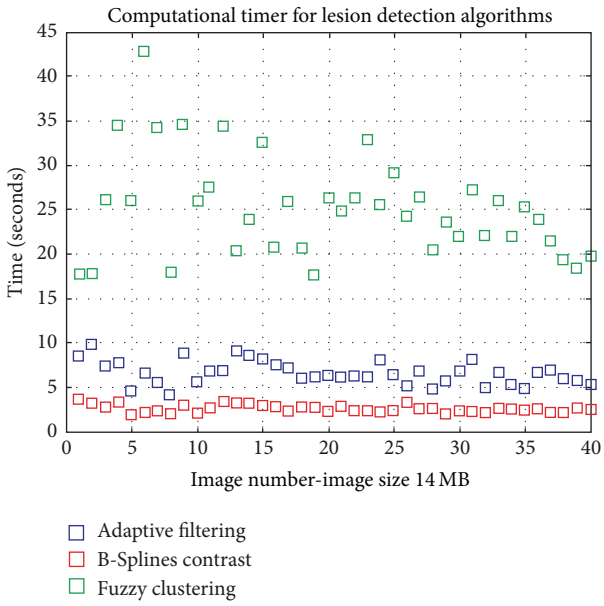


Figure 11: Computational times of image processing algorithms.

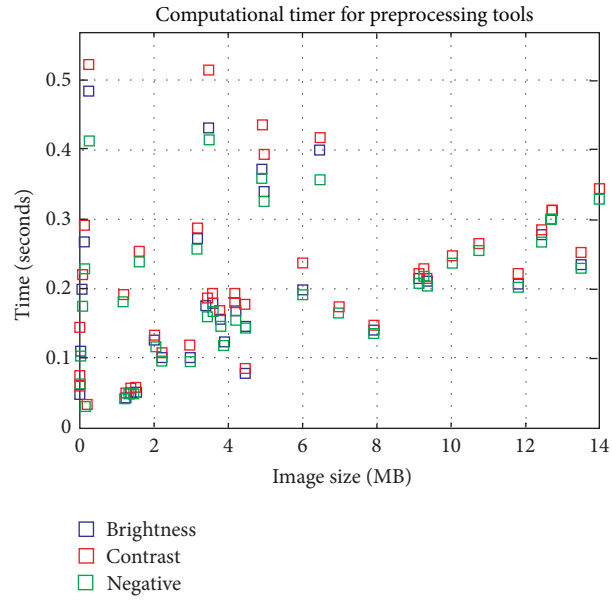


Figure 12: Computational times of preprocessing tools.

3 GB of RAM and with a precision of milliseconds. Figure 11 shows the computational times for a dataset with a subset of 40 mammograms in DICOM format with 3328 × 4084 pixels. Figure 12 shows the computational times of the preprocessing tools using different file formats.

The results show computational times lower than 1 minute for all detection algorithms, being under 10 seconds for the adaptive filtering and β-Splines. Moreover, the computational time for the preprocessing tools is also very low, under 0.5 seconds. These computational times make it possible to integrate these tools in the EHR and their use in the daily work.

6. Conclusions

In some medical centres it is still common that medical tests results are printed on paper and the patient has the responsibility of carrying these results to the hospital. This caused delays and, sometimes, the results could be damaged. Moreover, in case of using radiological tests, an X-ray view box and a magnifying glass are necessary to view and amplify the results.

Thanks to the EHRs, as the one developed in this work, it is possible to access and process clinical stored data about patient information and clinical tests and then solve the above-mentioned problems.

Furthermore, the use of digital imaging in breast cancer detection makes it possible to develop tools to assist radiologists in the diagnosis of this disease. In this work we have implemented a CADe system that integrates some tools to offer adjustments of the image zoom, mechanisms to change image brightness and contrast, the ability to see more clearly some areas, and the possibility to take measurements with a few mouse clicks. Such tools are now, according to expert radiologists in breast cancer, widely used in their daily work.

We also include in the CAD system more complex algorithms to help radiologists detect the two most common types of mammographic tissue lesions that are sometimes associated with breast cancer.

The fact that the CAD application has been developed to be integrated within the EHR makes all the applications that a specialist need options within a main application and promotes the usability of the entire application. Moreover, using information from both the EHR and the CADe system helps to provide a more complete diagnostic and therefore to reduce the number of FP detections and increase the TP.

Finally, using the DICOM standard to develop this application (DICOM for image management and storage and DICOM-SR for reports) facilitates the compatibility with other programs. The use of DICOM-SR for reports is also a novelty of the CADe implemented in this work.

Further ongoing work is devoted to provide some tools to classify the type of lesions into benign and malign, using also the specific information of the EHR.

Acknowledgments

The authors acknowledge partial financial support from the Spanish Research Ministry and Junta de Comunidades de Castilla-La Mancha through projects RETIC COMBIOMED and PI-2006/01.1. The authors want to thank their collaborator, technician at VISILAB, Felipe Terriza.

References

[1] R. M. Rangayyan, F. J. Ayres, and J. E. Leo Desautels, "A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs," *Journal of the Franklin Institute*, vol. 344, no. 3-4, pp. 312–348, 2007.

[2] D. G. Katehakis, S. G. Sfakianakis, G. Kavlentakis, D. N. Anthoulakis, and M. Tsiknakis, "Delivering a lifelong integrated electronic health record based on a service oriented architecture," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 6, pp. 639–650, 2007.

[3] H. F. Marin, "Nursing informatics: advances and trends to improve health care quality," *International Journal of Medical Informatics*, vol. 76, supplement 2, pp. S267–S269, 2007.

[4] S. N. van der Veer, N. F. de Keizer, A. C. J. Ravelli, S. Tenkink, and K. J. Jager, "Improving quality of care. A systematic review on how medical registries provide information feedback to health care providers," *International Journal of Medical Informatics*, vol. 79, no. 5, pp. 305–323, 2010.

[5] DICOM Standard, http://medical.nema.org.

[6] S. M. Astley and F. J. Gilbert, "Computer-aided detection in mammography," *Clinical Radiology*, vol. 59, no. 5, pp. 390–399, 2004.

[7] D. Protti, I. Johansen, and F. Perez-Torres, "Comparing the application of Health Information Technology in primary care in Denmark and Andalucía, Spain," *International Journal of Medical Informatics*, vol. 78, no. 4, pp. 270–283, 2009.

[8] N.-H. T. Trinh, S. J. Youn, J. Sousa et al., "Using electronic medical records to determine the diagnosis of clinical depression," *International Journal of Medical Informatics*, vol. 80, no. 7, pp. 533–540, 2011.

[9] M. A. Farsi and D. J. West Jr., "Use of electronic medical records in Oman and physician satisfaction," *Journal of Medical Systems*, vol. 30, no. 1, pp. 17–22, 2006.

[10] M. Končar and D. Gvozdanović, "Primary healthcare information system—the cornerstone for the next generation healthcare sector in Republic of Croatia," *International Journal of Medical Informatics*, vol. 75, no. 3-4, pp. 306–314, 2006.

[11] A. K. Jha, T. G. Ferris, K. Donelan et al., "How common are electronic health records in the United States? A summary of the evidence," *Health Affairs*, vol. 25, no. 6, pp. w496–w507, 2006.

[12] L. L. Fragidis and P. D. Chatzoglou, "The use of electronic health record in greece: current status," in *Proceedings of the 11th IEEE International Conference on Computer and Information Technology (CIT '11)*, pp. 475–480, September 2011.

[13] S. Abraham, "Technological trends in health care: electronic health record," *Health Care Manager*, vol. 29, no. 4, pp. 318–323, 2010.

[14] J. F. Wilson, "Making electronic health records meaningful," *Annals of Internal Medicine*, vol. 151, no. 4, pp. 293–296, 2009.

[15] D. A. Handel and J. L. Hackman, "Implementing electronic health records in the emergency department," *Journal of Emergency Medicine*, vol. 38, no. 2, pp. 257–263, 2010.

[16] B. Jung, "DICOM-X—seamless integration of medical images into the EHR," in *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pp. 203–207, June 2005.

[17] P. Marcheschi, A. Mazzarisi, S. Dalmiani, and A. Benassi, "HL7 clinical document architecture to share cardiological images and structured data in next generation infrastructure," in *Proceedings of the Computers in Cardiology*, pp. 617–620, September 2004.

[18] S. Cohen, F. Gilboa, and U. Shani, "PACS and electronic health records," in *Medical Imaging*, Proceedings of SPIE, pp. 288–298, February 2002.

[19] J. Mykkänen, A. Riekkinen, M. Sormunen, H. Karhunen, and P. Laitinen, "Designing web services in health information systems: from process to application level," *International Journal of Medical Informatics*, vol. 76, no. 2-3, pp. 89–95, 2007.

[20] V. Lapsia, K. Lamb, and W. A. Yasnoff, "Where should electronic records for patients be stored?" *International Journal of Medical Informatics*, vol. 81, no. 12, pp. 821–827, 2012.

[21] U. Bick and K. Doi, "Tutorial on Computer-Aided Diagnosis," in *Proceedings of the CARS*, 2008.

[22] S. J. Kim, W. K. Moon, M. H. Seong, N. Cho, and J. M. Chang, "Computer-aided detection in digital mammography: false-positive marks and their reproducibility in negative mammograms," *Acta Radiologica*, vol. 50, no. 9, pp. 999–1004, 2009.

[23] J. Wei, B. Sahiner, L. M. Hadjiiski et al., "Computer-aided detection of breast masses on full field digital mammograms," *Medical Physics*, vol. 32, no. 9, pp. 2827–2838, 2005.

[24] *Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)*, American College of Radiology, Reston, Va, USA, 2003.

[25] K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," *British Journal of Radiology*, vol. 78, pp. S3–S19, 2005.

[26] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 236–251, 2009.

[27] United States Food and Drug Administration P970058, *Summary of Safety and Effectiveness Data: R2 Technologies*, 1998.

[28] United States Food and Drug Administration P970038, *Summary of Safety and Effectiveness Data: ISSI*, 2002.

[29] United States Food and Drug Administration P970034, *Summary of Safety and Effectiveness Data: CADx Medical Systems*, 2002.

[30] K. Bernstein, M. Bruun-Rasmussen, S. Vingtoft, S. K. Andersen, and C. Nøhr, "Modelling and implementing electronic health records in Denmark," *International Journal of Medical Informatics*, vol. 74, no. 2–4, pp. 213–220, 2005.

[31] I. Jacobson, G. Booch, and J. Rumbaugh, *The Unified Software Development Process*, Addison-Wesley Professional, 1999.

[32] R. Hussein, U. Engelmann, A. Schroeter, and H.-P. Meinzer, "DICOM structured reporting—part 1. Overview and characteristics," *Radiographics*, vol. 24, no. 3, pp. 891–896, 2004.

[33] G. Bueno, N. Vállez, O. Déniz et al., "Automatic breast parenchymal density classification integrated into a CADe system," *International Journal of Computer Assisted Radiology and Surgery*, vol. 6, no. 3, pp. 309–318, 2011.

[34] A. Saimi, T. Syomura, H. Suganuma, and I. Ishida, "Presentation layer framework of web application systems with server-side Java technology," in *Proceedings of the IEEE 24th Annual International Computer Software and Applications Conference (COMPSAC '00)*, pp. 473–478, October 2000.

[35] Y. Zhang, P. Yu, and J. Shen, "The benefits of introducing electronic health records in residential aged care facilities: a multiple case study," *International Journal of Medical Informatics*, vol. 81, no. 10, pp. 690–704, 2012.

[36] Y. Sumita, M. Takata, K. Ishitsuka, Y. Tominaga, and K. Ohe, "Building a reference functional model for EHR systems," *International Journal of Medical Informatics*, vol. 76, no. 9, pp. 688–700, 2007.

[37] S. Yoo, S. Kim, S. Lee et al., "A study of user requests regarding the fully electronic health record system at seoul national university bundang hospital: challenges for future electronic health record systems," *International Journal of Medical Informatics*, 2012.

[38] P. H. Meyers, C. M. Nice, H. C. Becker, N. J. Nettleton, J. W. Sweeney, and G. R. Meckstroth, "Automated computer analysis of radiographic images," *Radiology*, vol. 83, pp. 1029–1034, 1964.

[39] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis," *Radiology*, vol. 89, pp. 211–215, 1967.

[40] M. F. Angelo, A. C. Patrocinio, H. Schiabel, R. B. Medeiros, and S. R. Pires, "Comparing mammographic images," *IEEE Engineering in Medicine and Biology Magazine*, vol. 27, no. 3, pp. 74–81, 2008.

[41] T.-L. Ji, M. K. Sundareshan, and H. Roehrig, "Adaptive image contrast enhancement based on human visual properties," *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 573–586, 1994.

[42] I. N. Bankman, Ed., *Handbook of Medical Imaging, Processing and Analysis*, Academic Press, 2000.

[43] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[44] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data," *IEEE Transactions on Medical Imaging*, vol. 21, no. 3, pp. 193–199, 2002.

[45] G. Bueno, "Fuzzy systems and deformable models," in *Intelligent and Adaptive Systems in Medicine*, pp. 305–330, Taylor & Francis, 2008.

[46] F. Fritz, S. Balhorn, M. Riek, B. Breil, and M. Dugas, "Qualitative and quantitative evaluation of EHR-integrated mobile patient questionnaires regarding usability and cost-efficiency," *International Journal of Medical Informatics*, vol. 81, no. 5, pp. 303–313, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 458989, 10 pages
http://dx.doi.org/10.1155/2013/458989

Research Article

Characterization of Schizophrenia Adverse Drug Interactions through a Network Approach and Drug Classification

Jingchun Sun,[1,2] Min Zhao,[1] Ayman H. Fanous,[3,4] and Zhongming Zhao[1,2,5,6]

[1] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA*
[2] *Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[3] *Mental Health Service Line, Washington VA Medical Center, 50 Irving St. NW, Washington, DC 20422, USA*
[4] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA*
[5] *Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37212, USA*
[6] *Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*

Correspondence should be addressed to Zhongming Zhao; zhongming.zhao@vanderbilt.edu

Received 5 July 2013; Accepted 8 August 2013

Academic Editor: Bairong Shen

Antipsychotic drugs are medications commonly for schizophrenia (SCZ) treatment, which include two groups: typical and atypical. SCZ patients have multiple comorbidities, and the coadministration of drugs is quite common. This may result in adverse drug-drug interactions, which are events that occur when the effect of a drug is altered by the coadministration of another drug. Therefore, it is important to provide a comprehensive view of these interactions for further coadministration improvement. Here, we extracted SCZ drugs and their adverse drug interactions from the DrugBank and compiled a SCZ-specific adverse drug interaction network. This network included 28 SCZ drugs, 241 non-SCZs, and 991 interactions. By integrating the Anatomical Therapeutic Chemical (ATC) classification with the network analysis, we characterized those interactions. Our results indicated that SCZ drugs tended to have more adverse drug interactions than other drugs. Furthermore, SCZ typical drugs had significant interactions with drugs of the "alimentary tract and metabolism" category while SCZ atypical drugs had significant interactions with drugs of the categories "nervous system" and "antiinfectives for systemic uses." This study is the first to characterize the adverse drug interactions in the course of SCZ treatment and might provide useful information for the future SCZ treatment.

1. Introduction

Schizophrenia (SCZ) is a common, complex mental disorder with a worldwide prevalence of approximately 1%, creating a substantial healthcare challenge in the world. Over the past several decades, antipsychotic drugs have been the commonly used medications to treat psychiatric disorders such as SCZ and bipolar disorder [1]. These drugs are classified as two types: typical and atypical antipsychotics. Typical antipsychotics are known as first generation antipsychotics, while atypical antipsychotics are known as second generation antipsychotics. Their significant difference lies in their different ability to produce extrapyramidal side effects (EPS), block dopamine type 2 receptors, improve negative symptoms, and others [2]. Antipsychotics can be effective in the treatment of SCZ but vary in efficacy and side effects. As the representative drugs of typical and atypical antipsychotics, respectively, haloperidol and clozapine are often utilized to characterize the effects of each of the two drug categories [2, 3]. In addition to the positive, negative, and cognitive symptoms, which are the hallmarks of psychotic illness, patients with schizophrenia are highly likely to have comorbid medical and other conditions such as anxiety disorders, depression, cardiovascular disease, and diabetes [4, 5]. Therefore, prescribed coadministration of antipsychotics and nonantipsychotics drugs has been increasing for the treatment of SCZ patients along with various aspects of their illness [6].

In healthcare, serious adverse effects have been reported in the coadministration of multiple drugs in many diseases such as heart disease [7], glaucoma [8], and cancer [9, 10]. It is estimated that approximately 20–30% of all adverse reactions are caused by interactions between drugs [11]. The adverse

drug-drug interaction is defined as the phenomenon that occurs when the effects of a drug are altered by prior administration or coadministration of a second drug. It could happen during the drug absorption, the metabolism process, and the binding process of drug targets [12]. As the coadministration of multiple drugs to treat the psychiatric and nonpsychiatric comorbid medical conditions of schizophrenia increases, the potential of adverse drug-drug interactions (DDIs) is becoming an important consideration in the treatment of SCZ [6].

Numerous studies have focused on adverse drug-drug interactions associated with antipsychotics [13, 14]. However, few of them have comprehensively characterized the adverse interactions of these antipsychotics either among themselves or with non-SCZ drugs. Most of these studies focused on the collection of adverse interactions and their clinical characteristics [13, 14]. Additionally, for typical and atypical antipsychotics, various studies have been conducted to illustrate both adverse effects pinpointed in clinical trials and underlying molecular interactions between these two types of drugs and their targets [15–17]. However, the differences between their adverse interaction characteristics have never been investigated. These limitations were mainly attributed to an absence of comprehensive clinical and molecular data. These limitations have recently been largely eased thanks to a comprehensive and publicly available database DrugBank [18–20], a unique resource that combines detailed drug data. Thus, the ability to characterize the drug interactions of these antipsychotics is emerging; this in turn will provide a wide-ranging view of the drug-drug interactions of SCZ drugs and potential information for SCZ drug coadministration and prediction of adverse drug-drug interactions.

In this study, we extracted 32 SCZ drugs based on the records in multiple fields from the DrugBank database and the suggestions from one of us (AHF), a practicing psychiatrist treating mostly psychotic disorders. We then collected the adverse drug-drug interactions of these SCZ drugs from the DrugBank to build an adverse drug-drug interaction network. Combing the network analysis with the drug classification from the Anatomical Therapeutic Chemical (ATC) systems [21], we characterized the adverse drug interactions of SCZ drugs as compared to other types of drugs. Additionally, we compared the properties of adverse interactions of SCZ typical and atypical drugs. This study represents the first to investigate the adverse drug interactions of SCZ drugs. The results may assist researchers to develop better diagnostic tests, effective medications, and coadministration strategies.

2. Materials and Methods

2.1. Collection of Drugs to Treat Schizophrenia. To collect a comprehensive list of the drugs that are routinely used to treat schizophrenia patients, we first utilized the schizophrenia related keywords, schizophrenia, schizophrenias, schizophrenic, schizophrenics, schizotypy, and schizotypal, to search the data from the DrugBank (version 3.0) [18]. The DrugBank contained 6796 drugs including 1571 approved drugs. Our search resulted in 46 drugs. Then, we checked the "indication" field, which describes common names of diseases that a drug is used to treat. We obtained 38 drugs that had the SCZ related keywords in the "indication" field and eight drugs that did not. Next, we accessed the DrugBank website and manually checked if each of these drugs has been used to treat schizophrenia. We found that five of these 38 drugs were not related to SCZ treatment. Thus, we obtained 32 SCZ drugs. Among them, 28 drugs have adverse drug-drug interactions according to the DrugBank. To obtain information regarding typical or atypical classification for these 28 drugs, we manually checked data from multiple resources: DrugBank, PubMed Health (http://www.ncbi.nlm.nih.gov/pubmedhealth/), Wikipedia (http://www.wikipedia.org/), and several textbooks.

2.2. Collection of Adverse Drug Interactions and Construction of Adverse Drug-Drug Interaction Networks. The DrugBank consists of adverse drug-drug interactions and represents the most complete, publicly accessible collection of its kind [19]. These adverse drug-drug interactions are the events that occur when the effects of a drug are altered by prior administration or coadministration of another drug. In DrugBank, for a given drug, the "interaction" field includes drugs and their corresponding adverse descriptions with the drug. These descriptions were compiled from a variety of web and textbook resources and verified by accredited pharmacists. We extracted all adverse drug interactions from the DrugBank data. Considering that the descriptions of adverse drug-drug interactions are very complicated, in this study, we formed a pair of drugs if both are involved in one adverse drug-drug interaction, but the direction of the interaction in each pair was ignored. Based on these drug pairs, we constructed an adverse drug interaction network as the human adverse DDI network, in which a node denotes a drug and an edge indicates that the two drugs would have some adverse events if they were coadministrated. From all adverse interactions, we extracted the adverse drug interactions of SCZ drugs including both interactions among SCZ drugs and interactions between SCZ drugs and non-SCZ drugs. Based on these SCZ drug interactions, we constructed a SCZ-specific DDI network, in which nodes are SCZ drugs and non-SCZ drugs and edges are the adverse interactions among SCZ drugs or the adverse interactions between SCZ drugs and non-SCZ drugs.

2.3. ATC Drug Classification. The World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology developed and maintains the Anatomical Therapeutic Chemical (ATC) classification database (http://www.whocc.no/). The classification system curates drugs into different groups according to their therapeutic, chemical, and pharmacological properties. It has five levels that represent progressively finer classifications. For example, the letter "N" represents the top level of the classification "nervous system." The N class is further divided into, for example, N05 (psycholeptics) on the second level, N05A (antipsychotics) on the third level, N05AH (diazepines,

oxazepines, thiazepines, and oxepines) on the fourth level, and N05AH02 (clozapine) on the fifth level.

In this study, we applied the first-level classification to examine the general network properties of SCZ drugs. The first-level classification indicates the anatomical main group and consists of 14 main groups represented by 14 letters (codes). We further employed the third-level classification to examine the difference between classification of SCZ typical and atypical drugs. The third level classifies drugs based on mixed criteria involving therapeutic or pharmacological properties. We obtained drug ATC codes from DrugBank and the Kyoto Encyclopedia of Genes and Genomes (KEGG DRUG) database [22].

2.4. Network Topological Analysis and Visualization. In this study, we applied the node connectivity to examine the network topological property difference of SCZ drugs. For a given node in a network, node connectivity (degree) is the number of edges linked to the node, which is the network's most elementary characteristics [23]. Considering that SCZ drugs belong to the "nervous system" category, we compiled another two drug sets: other nervous system drugs (other N-drugs) and drugs excluding nervous systems drugs (non-N-drugs); we then compared their degree distributions in the context of all human adverse DDIs. We utilized the software Cytoscape for network visualization [24].

2.5. Statistical Tests. We employed the Wilcoxon rank-sum test to compare degree distribution. To compare adverse interaction tendencies of SCZ typical and atypical drugs, we divided the SCZ drug interactions into 14 categories according to their linked non-SCZ drugs' ATC first-level classification, and then we performed Fisher's exact test for each category. For a given category, we calculated a 2×2 contingency table, which includes four counts: n, $N-n$, r, and $R-r$, where n is the count of the links that SCZ typical drugs have in the category, N is the count of total links that SCZ typical drugs have in all 14 categories, r is the count of the links that SCZ atypical drugs have in the category, and R is the count of total links that SCZ atypical drugs have in all 14 categories. We utilized the R package (http://www.r-project.org/) to calculate Pvalues followed by multiple testing correction using the Bonferroni method [25].

3. Results

In this study, we collected 32 drugs that are mainly used to treat the schizophrenia patients, denoted as SCZ drugs. Among them, 28 drugs had adverse drug-drug interactions according to DrugBank data. Within this list, 18 SCZ drugs belonged to typical antipsychotics while the other 10 belonged to atypical antipsychotics category (Table 1). To explore the characteristics of SCZ adverse drug interactions, we first collected all the adverse DDIs from DrugBank and constructed a human DDI network, which included 10,931 pairs involving 1087 drugs. Among these 1087 drugs, 1005 had at least one ATC annotation, and they were classified into 14 drug sets based on their ATC first-level classification.

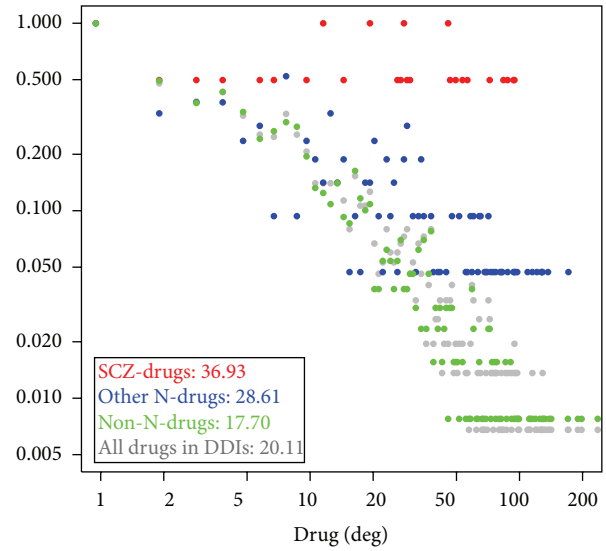


Figure 1: Degree distributions and average degrees (vertical lines) of four drug sets. Y-axis represents the proportion of drugs that have a degree while the X-axis is the drug degree. "SCZ-drugs" (red) denotes the 28 schizophrenia (SCZ) drugs. "other N-drugs" (blue) denotes the drugs belonging to the "nervous systems" category after exclusion of SCZ drugs. "non-N-drugs" (green) denotes the drugs not belonging to "nervous systems." "all drugs in DDIs" (grey) denotes all drugs in the human DDIs. The inserted table summarizes the average degree for each drug set.

3.1. SCZ Drugs Had a Significant Higher Degree of Adverse Interactions Than Other Drugs. For comparison, we compiled three drug sets based on all drugs in human DDIs: SCZ drugs, other N-drugs, and non-N-drugs. Then, we calculated degree distributions and average degrees for three drug sets and all drugs in DDIs. Figure 1 displays their degree distributions and average degrees. The average degree of SCZ drugs was 36.93, which was significantly higher than that of other N-drugs (28.61, Wilcoxon's test, P value = 0.0209) or that of non-N-drugs (17.70, P value = 6.45×10^{-6}). This observation indicated that, compared to non-SCZ drugs, SCZ drugs tended to have more adverse drug interactions with other drugs. Moreover, other N-drugs had significantly more adverse drug interactions than that of the non-N-drugs (P value = 5.98×10^{-7}). This observation indicated that drugs belonging to the "nervous systems" category tended to have more adverse drug interactions with other drugs.

To explore this tendency in detail, we examined if the N-drug set is different from the other 13 drug categories in all DDIs based on the ATC first-level annotation. According to the ATC first-level annotation, 1005 drugs in DDIs with at least one ATC drug annotation could be grouped into 14 groups. Among the 1005 drugs, those with multiple ATC codes were assigned to multiple groups. We performed Wilcoxon's rank sum test to examine if the degree distribution of the N-drugs is different from that of each of the other 13 groups of drugs. We found that N-drugs had significantly more adverse drug interactions than the other eight groups (P value < 0.05) (Table 2). These other eight

Table 1: Antipsychotics used to treat schizophrenia patients.

DrugBank ID	Drug name	Number of adverse drug interactions	Typical/atypical[a]
DB01063	Acetophenazine	15	Typical
DB06288	Amisulpride	48	Atypical
DB01238	Aripiprazole	12	Atypical
DB00477	Chlorpromazine	51	Typical
DB01239	Chlorprothixene	4	Typical
DB00363	Clozapine	55	Atypical
DB00875	Flupenthixol	20	Typical
DB00623	Fluphenazine	28	Typical
DB00502	Haloperidol	58	Typical
DB00408	Loxapine	20	Typical
DB00933	Mesoridazine	74	Typical
DB01403	Methotrimeprazine	47	Typical
DB01618	Molindone	7	Atypical
DB00334	Olanzapine	10	Atypical
DB01267	Paliperidone	12	Atypical
DB00850	Perphenazine	31	Typical
DB01100	Pimozide	47	Typical
DB01621	Pipotiazine	6	Typical
DB00433	Prochlorperazine	29	Typical
DB01224	Quetiapine	29	Atypical
DB00734	Risperidone	27	Atypical
DB06144	Sertindole	2	Atypical
DB01622	Thioproperazine	3	Typical
DB00679	Thioridazine	86	Typical
DB01623	Thiothixene	96	Typical
DB00831	Trifluoperazine	30	Typical
DB00246	Ziprasidone	90	Atypical
DB01624	Zuclopenthixol	97	Typical

[a]Antipsychotic drugs are classified as typical and atypical mainly based on their different ability to cause extrapyramidal side effects (EPS).

Table 2: Comparison of drugs belonging to the "nervous Systems" category with drugs from the other categories based on their degrees in human adverse drug-drug interaction network.

ATC first-level classification (ATC code)	Number of drugs	Average degree	Wilcoxon test P-value	Bonferroni adjusted P-value
Antineoplastic and immunomodulating agents (L)	118	15.25	3.01×10^{-8}	4.22×10^{-7}
Various (V)	22	7.14	5.03×10^{-6}	7.04×10^{-5}
Alimentary tract and metabolism (A)	119	15.52	1.49×10^{-5}	2.09×10^{-4}
Blood and blood forming organs (B)	45	21.00	3.02×10^{-4}	0.0042
Sensory organs (S)	85	19.66	0.0089	0.1242
Respiratory system (R)	75	20.27	0.0148	0.2068
Antiinfectives for systemic use (J)	143	25.19	0.0228	0.3186
Musculoskeletal system (M)	56	14.39	0.0287	0.4021
Cardiovascular system (C)	153	22.75	0.7008	1.000
Dermatologicals (D)	56	25.27	0.4775	1.000
Genitourinary system and sex hormones (G)	61	21.38	0.1570	1.000
Systemic hormonal preparations, excluding sex hormones and insulins (H)	21	20.76	0.5294	1.000
Antiparasitic products, insecticides, and repellents (P)	20	19.95	0.2517	1.000
Nervous system (N)	219	29.68	—	—

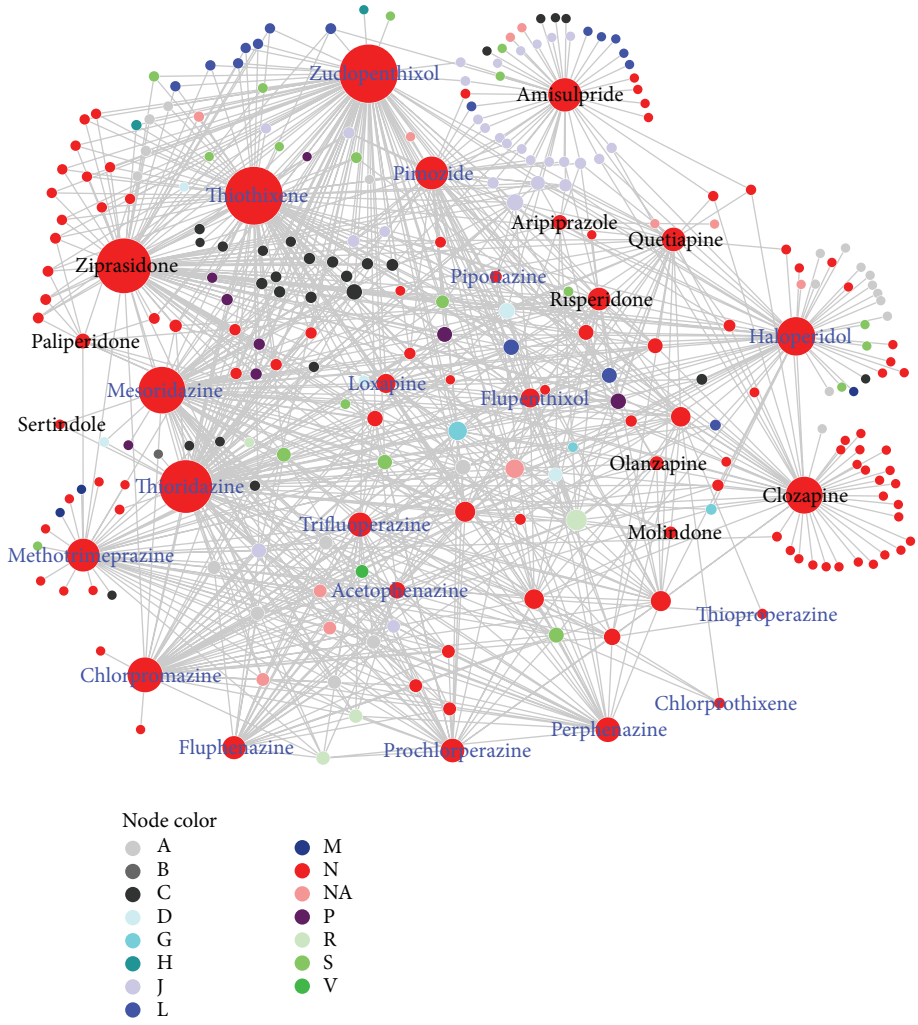


Figure 2: Adverse drug-drug interaction network for schizophrenia (SCZ) drugs. Node color corresponds to ATC first-level classification code. With the exception of "NA" for the drugs without ATC classification, the representations of these letters are detailed in Table 2. Nodes with blue labels are SCZ typical drugs, and nodes with black labels are SCZ atypical drugs. Node size corresponds to the number of the adverse interactions that the drug had in the network.

groups were "antineoplastic and immunomodulating agents" (L), "various" (V), "alimentary tract and metabolism" (A), "blood and blood forming organs" (B), "sensory organs" (S), "respiratory system" (R), "anti-infectives for systemic use" (J), and "musculoskeletal system" (M). Notably, among the eight groups, four passed the stringent Bonferroni multiple testing correction (Bonferroni adjusted P-value < 0.05). These four were "antineoplastic and immunomodulating agents" (L), "various" (V), "alimentary tract and metabolism" (A), and "blood and blood forming organs" (B). These results indicated that the N-drugs had different adverse interaction tendencies compared to other groups of drugs; specifically their tendencies were different from the drugs in groups L, V, A, and B.

3.2. SCZ Adverse DDI Network. From the human DDIs compiled in this study, 28 SCZ drugs had 991 interactions in total, in which 43 interactions occurred among SCZ drugs and 948 interactions occurred between SCZ drugs and non-SCZ drugs (Figure 2). The average degree of the 28 drugs was 36.93. Among these SCZ drugs, 11 drugs had more than forty interactions. These 11 drugs were zuclopenthixol (degree: 97), thiothixene (96), ziprasidone (90), thioridazine (86), mesoridazine (74), haloperidol (58), clozapine (55), chlorpromazine (51), amisulpride (48), methotrimeprazine (47), and pimozide (47). Among the 11 drugs, three were atypical drugs: clozapine, ziprasidone, and amisulpride. The average degree of the 241 non-SCZ drugs was 3.92. Among these 241 drugs, 22 drugs had more than 10 interactions with SCZ drugs. They were triprolidine (23), tacrine (22), tetrabenazine (22), galantamine (21), donepezil (21), trospium (19), trimethobenzamide (19), voriconazole (15), rivastigmine (15), tacrolimus (13), bromocriptine (13), vorinostat (13), lumefantrine (13), nilvadipine (12), toremifene (12), cisapride (12), trimipramine (12), guanethidine (12), artemether (12), desvenlafaxine (12), levofloxacin (11), and sparfloxacin (11).

Table 3: Comparison of interaction categories between typical and atypical SCZ drugs using ATC first-level classification.

ATC first-level classification (ATC code)	Number of typical drug interactions (%[a])	Number of atypical drug interactions (%[b])	Fisher's exact test P-value
Alimentary tract and metabolism (A)	74 (11.60)	5 (1.94)	3.94×10^{-7}
Nervous system (N)	216 (33.86)	119 (46.12)	0.0008
Antiinfectives for systemic use (J)	66 (10.34)	46 (17.83)	0.0035
Sensory organs (S)	51 (7.99)	12 (4.65)	0.0839
Various (V)	8 (1.25)	0 (0)	0.1136
Respiratory system (R)	35 (5.49)	8 (3.10)	0.1668
Antineoplastic and immunomodulating agents (L)	37 (5.80)	21 (8.14)	0.2297
Antiparasitic products, insecticides, and repellents (P)	36 (5.64)	10 (3.88)	0.3191
Cardiovascular system (C)	68 (10.66)	23 (8.91)	0.4662
Musculoskeletal system (M)	3 (0.47)	0 (0)	0.5612
Genitourinary system and sex hormones (G)	19 (2.98)	6 (2.33)	0.6618
Dermatologicals (D)	21 (3.29)	7 (2.71)	0.8325
Systemic hormonal preparations, excluding sex hormones and insulins (H)	3 (0.47)	1 (0.39)	1.0000
Blood and blood forming organs (B)	1 (0.16)	0 (0)	1.0000

[a]The percentage was calculated by the number of interactions with SCZ atypical in each category divided by all the numbers of interactions with SCZ atypical drugs.
[b]The percentage was calculated by the number of interactions with SCZ typical in each category divided by all the numbers of interactions with SCZ typical drugs.

Among 991 interactions, 43 interactions occurred among 16 SCZ drugs, while 948 occurred between 28 SCZ drugs and 241 non-SCZ drugs. According to the description of the 43 adverse drug interactions, 41 were related to increased risk of cardiotoxicity and arrhythmias, which indicated that the coadministration of SCZ drugs might increase the risk of heart disease, especially for prolonging QT intervals [26].

Since all of the 28 SCZ drugs involved in the 991 interactions belonged to the "nervous systems" category, we grouped the 991 interactions into 14 categories based on their interacting drugs' ATC first-level annotation. Among the 241 drugs that had interactions with 28 SCZ drugs, 229 drugs had an ATC annotation. Based on these drug ATC first-level codes, we categorized the interactions into 14 categories (Additional file, Table S1 available online at http://dx.doi.org/10.1155/2013/458989). Among them, the top 3 types of interactions are "nervous system" (N), "anti-infectives for systemic use" (J), and "cardiovascular system" (C). This observation indicated that SCZ drugs tended to have adverse interactions with these three types of non-SCZ drugs.

3.3. Comparison of SCZ Typical and Atypical Drugs. Among the 28 SCZ drugs with adverse drug-drug interactions, 18 were typical antipsychotic drugs and 10 were atypical antipsychotic drugs. The average degree of typical drugs in the SCZ DDI network was 41.22, while that of atypical drugs was 29.20. The former was much higher than the later, indicating that typical drugs had more adverse interactions than atypical drugs.

Among the 948 interactions between 28 SCZ drugs and 241 non-SCZ drugs, 680 of the interactions occurred between 18 typical drugs and 188 non-SCZ drugs while the other 268 occurred between 10 atypical drugs and 164 non-SCZ drugs. To examine the difference in adverse drug interactions between typical and atypical drugs, we divided these interactions into 14 groups according to their interaction with non-SCZ drugs' ATC first-level classifications. Among the 241 non-SCZ drugs, 230 drugs had at least one ATC annotation. For the 680 typical SCZ drug interactions, 638 could be grouped into 14 groups, and for the 268 atypical SCZ drug interactions, 258 could be grouped into 11 groups. We performed Fisher's exact test for each group (Table 3). These results showed that SCZ typical and atypical drugs had significant differences at three ATC first-level categories: "alimentary tract and metabolism" (A), "nervous system" (N), and "anti-infectives for systemic use" (J). More specifically, for the category "alimentary tract and metabolism" (A), the interaction number and percentage of typical SCZ drugs were higher than those of SCZ atypical drugs. In contrast, for the categories "nervous system" (N) and "anti-infectives for systemic use" (J), the interaction percentage of SCZ atypical drugs was higher than that of SCZ typical drugs. These observations revealed that SCZ typical drugs had significant interactions with drugs belonging to the category "alimentary tract and metabolism" (A), while SCZ atypical drugs tended to have more interactions with drugs belonging to the categories "nervous system" (N) and "anti-infectives for systemic use" (J).

To further interrogate the detailed difference between the two SCZ drug categories, we performed the comparison of adverse interactions between SCZ typical and atypical drugs using the ATC third-level classification. According to the third-level classification, among the 680 adverse interactions

Table 4: Comparison of interaction categories between SCZ typical and atypical drugs using ATC third-level classification.

ATC third-level classification (ATC code)	Number of typical drug interactions (%[a])	Number of atypical drug interactions (%[b])	Fisher's exact test P-value
Antiepileptics (N03A)	3 (0.47)	21 (3.30)	1.95×10^{-8}
Anxiolytics (N05B)	1 (0.16)	11 (1.73)	2.91×10^{-5}
Hypnotics and sedatives (N05C)	3 (0.47)	10 (1.57)	1.13×10^{-3}
Psychostimulants, agents used for ADHD and nootropics (N06B)	28 (4.40)	2 (0.31)	0.0012
Direct acting antivirals (J05A)	13 (2.04)	17 (2.67)	0.0051
Quinolone antibacterials (J01M)	18 (2.83)	1 (0.16)	0.0107
Drugs for treatment of tuberculosis (J04A)	2 (0.31)	6 (0.94)	0.0161

[a]The percentage was calculated by the number of interactions with SCZ atypical in each category divided by all the numbers of interactions with SCZ atypical drugs.
[b]The percentage was calculated by the number of interactions with SCZ typical in each category divided by all the numbers of interactions with SCZ typical drugs.

of 18 SCZ typical drugs, 637 could be categorized into 57 third-level categories while the remaining 43 interactions could not be similarly classified due to a lack of ATC annotation for interacting non-SCZ drugs. Similarly, among 268 adverse interactions of 10 SCZ atypical drugs, 251 adverse interactions could be sorted into 44 third-level categories while the other 17 interactions could not. Among these 57 and 44 third-level categories, there were 39 common categories in both SCZ typical and atypical drug adverse interactions sets, 18 categories were specific for SCZ typical drugs, and 5 were specific to SCZ atypical drugs (Additional file, Table S2). Among these categories that are specific for typical and atypical drugs, most of them had only a few interactions, except for categories "antiobesity preparations, excluding diet products" (A08A), and "antiglaucoma preparations and miotics" (S01E). SCZ typical drugs had 45 interactions (7.06%) in A08A and 13 interactions (2.04%) in S01E while SCZ atypical drugs had none in either category. The category A08A included antiobesity drugs excluding diet products while S01E included the drugs used to treat glaucoma and related diseases. The observation may indicate that, compared to atypical drugs, typical drugs tended to have adverse drug-drug interactions with antiobesity drugs and antiglaucoma drugs.

Among the 39 common categories of SCZ drug adverse interactions, seven had significant difference between typical and atypical drugs (Fisher's exact test, P value < 0.05, Table 4). Among them, four belonged to the category "nervous systems" (N) and the other three belonged to the category "anti-infectives for systemic use" (J). This is consistent with the above observations. For the categories "psychostimulants, agents used for ADHD and nootropics" (N06B), and "quinolone antibacterials" (J01M), the percentage of SCZ typical drug adverse interactions was significantly higher than that of SCZ atypical drug adverse interactions, indicating that SCZ typical drugs tended to have adverse interactions with these drugs belonging to the N06B and J01M categories. On the contrary, the percentage of SCZ atypical drug adverse interactions was significantly higher than those of SCZ typical drug adverse interactions in the categories "antiepileptics" (N03A), "anxiolytics" (N05B), "hypnotics and sedatives" (N05C), and "drugs for treatment of tuberculosis" (J04A). This result suggested that SCZ atypical drugs might have adverse interactions with these drugs belonging to the categories N03A, N05B, N05C, and J04A.

3.4. Clozapine Tended to Have More Adverse Interactions with Non-N-Drugs but Not with SCZ Drug. Haloperidol and clozapine are the representative drugs for SCZ typical and atypical drugs, respectively. Haloperidol had adverse interactions with 58 drugs while clozapine had adverse interactions with 55 drugs. Figure 3 showed the merged subnetwork for haloperidol and clozapine adverse drug-drug interactions. Among the 58 drugs having adverse interactions with haloperidol, 27 (46.55%) belonged to the "nervous system" category, while among the 55 drugs that have adverse interactions with clozapine, 40 (72.73%) belonged to the same category. The results showed that haloperidol might tend to have adverse drug interactions with non-N-drugs while clozapine might tend to have adverse drug interactions with N-drugs. However, in contrast to clozapine, haloperidol had adverse interactions with five other SCZ drugs (mesoridazine, thioridazine, thiothixene, ziprasidone, and zuclopenthixol) while clozapine had only adverse interactions with haloperidol (Figure 3 and Additional file, Figure S1). This observation revealed that, for SCZ drugs, haloperidol had adverse interactions with other SCZ drugs but clozapine did not.

4. Discussion

In this study, we began with a comprehensive compilation of schizophrenia (SCZ) drugs and their adverse drug-drug interactions, and then we performed comprehensive comparisons between SCZ drugs and other types of drugs as well as SCZ typical and atypical antipsychotics. The results in this study had shown that SCZ drugs had different adverse interaction tendencies, and these differences extended to SCZ typical and atypical drugs, as well. This study provides

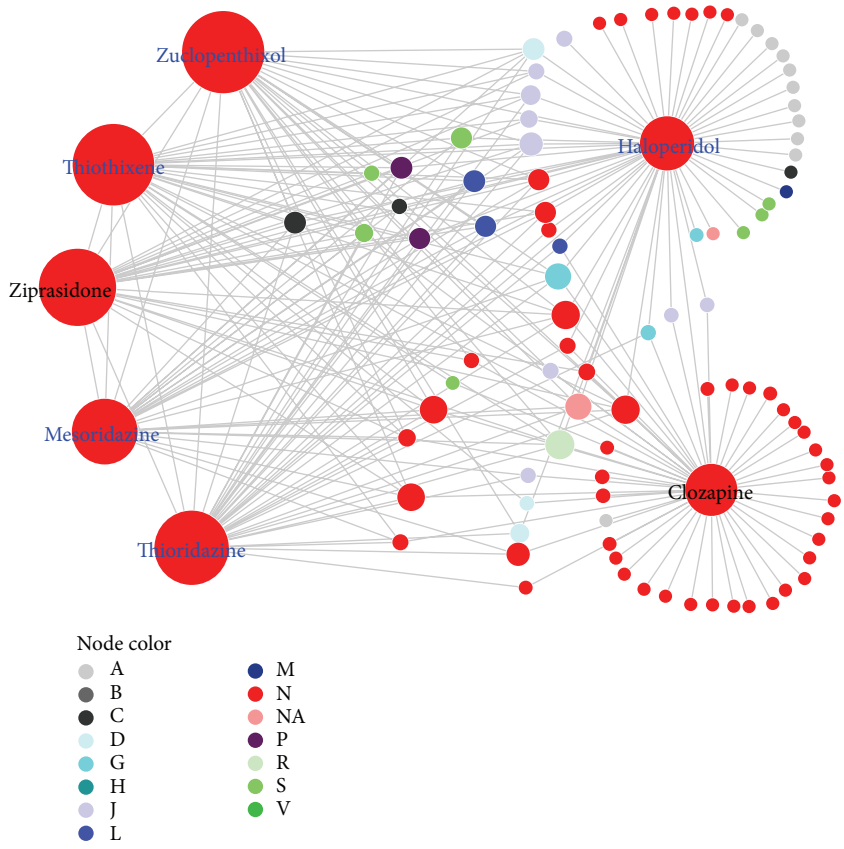


Figure 3: Haloperidol and clozapine adverse drug-drug interaction subnetwork extracted from the schizophrenia (SCZ) adverse drug-drug interaction network. Node color corresponds to ATC first-level classification code. With the exception of "NA" for the drugs without ATC classification, the representations of these letters are detailed in Table 2. Nodes with blue labels are SCZ typical drugs, and nodes with black labels are SCZ atypical drugs. Node size corresponds to the number of adverse interactions that the drug had in the SCZ drug-drug interaction network.

the global view of the characteristics of antipsychotics adverse interactions either among themselves or with non-SCZ drugs. Additionally, this study might provide potential adverse drug-drug interactions between SCZ drugs and non-SCZ drugs, which have never been used to treat SCZ or have no records of adverse drug-drug interactions. If these non-SCZ drugs have ATC classification annotation, physicians may consider whether the characteristics of non-SCZ drug categories have a higher probability to cause adverse drug-drug interactions with SCZ drugs. Thus, before prescribing these drugs belonging to these categories to SCZ patients, physicians would be more cautious.

One important output of this study was that, compared to other types of drugs, the drugs belonging to the "nervous system" (N) tended to have more adverse drug interactions, especially SCZ drugs. This was not surprising since it is well known that SCZ drugs affect neurotransmitter systems that are common to psychotropic medications used to treat other disorders. For example, while atypical antipsychotics block the dopamine and serotonin receptors, most antidepressants increase serotonergic levels while others increase levels of both serotonin and dopamine. Furthermore, many antidepressants as well as antipsychotics have known anticholinergic properties.

During the past several decades, numerous studies have revealed the difference in side effects between typical and atypical antipsychotics. However, the mechanisms underlying this difference are unclear. To characterize the effects of each of the two drug categories, researchers often utilize two drugs, haloperidol and clozapine, as the representatives of typical and atypical antipsychotics. In this study, we observed that haloperidol tended to have adverse interactions with other SCZ drugs but this is not the case for clozapine. This observation might suggest that their functional pathways in their drug action be different. Therefore, understanding the molecular mechanisms underlying those drug actions is critical for developing effective diagnostic tests and medications. Neurotransmitter receptors are primary targets of antipsychotics, and their interactions are important for drug efficacy. However, additional therapeutic properties may not directly relate to the receptor mechanism but to the intracellular signaling cascades. Through these cascades, the chemical signals

of drug receptor interactions transfer to alter gene expression, further affecting the formation of phenotypes. Therefore, future research endeavors may ensure investigating the post-receptor mechanisms of antipsychotics.

We mainly extracted the adverse drug-drug interaction data from DrugBank for this study. Though the study provides a review of the adverse drug-drug interactions of SCZ drugs, this investigation still needs to be improved since the current data utilized here is neither complete nor bias-free. Thus, future research in this area should include more adverse drug interaction data from multiple databases and other text resources such as Drug Interaction Facts [12], Drug Interaction Analysis and Management [27], Micromedex Drug-REAX [28], and the KEGG DRUG database [22]. Additionally, as a large volume of electronic medical records (EMRs) and FDA drug labels become available and effective text mining approaches were developed [29], the development of the novel methods to integrate multidimensional data sources and build a comprehensive resource for adverse drug interactions will be possible and practical.

Additionally, DDIs can occur through several biological processes of drug disposition, which might affect the same targets, the same metabolic pathways, or the same signaling pathways [30]. Considering the limitation of current molecular data, in this study, we did not explore the molecular mechanisms underlying the adverse drug interactions of SCZ drugs. However, a large volume of genome-wide molecular neuropharmacology data, such as microarray gene expression [31] and genome-wide association studies [32], is available, and more large-scale data will be available in the near future due to the rapid advances in genome-wide technologies and strong support from pharmacology communities. Therefore, it is possible and necessary to develop novel detection methods for investigation of adverse DDIs based on the molecular data. This will not only provide the valuable information for physicians, but also create a deeper understanding of the molecular mechanisms underlying adverse drug-drug interaction side effects, thereby furthering the ability to detect potential drug interactions.

5. Conclusions

In this study, we presented a comprehensive investigation of adverse drug-drug interactions of the antipsychotics used to treat schizophrenia. We integrated the network analysis with ATC drug classifications, which provided the adverse drug interaction characteristics of SCZ drugs as well as typical and atypical drugs. However, much more work is needed to collect more adverse drug interaction information and develop advanced pharmacogenomics network approaches. Potential findings could be used to predict adverse drug-drug interactions and improve the coadministration of multiple drugs, which in turn may lead to the avoidance of the drug-drug interaction adverse effects.

Conflict of Interests

All authors declare that there is no conflict of interests.

Acknowledgments

The authors would like to thank Dr. Qi Liu for the discussion regarding the statistical methods used in this study. This work was supported by a 2010 NARSAD Young Investigator Award to Jingchun Sun.

References

[1] B. A. Ellenbroek, "Psychopharmacological treatment of schizophrenia: what do we have, and what could we get?" *Neuropharmacology*, vol. 62, no. 3, pp. 1371–1380, 2012.

[2] H. Y. Meltzer, "Update on typical and atypical antipsychotic drugs," *Annual Review of Medicine*, vol. 64, pp. 393–406, 2013.

[3] K. S. Nandra and M. Agius, "The differences between typical and atypical antipsychotics: the effects on neurogenesis," *Psychiatria Danubina*, vol. 24, supplement 1, pp. S95–S99, 2012.

[4] J. Sokal, E. Messias, F. B. Dickerson et al., "Comorbidity of medical illnesses among adults with serious mental illness who are receiving community psychiatric services," *Journal of Nervous and Mental Disease*, vol. 192, no. 6, pp. 421–427, 2004.

[5] D. C. Goff, C. Cather, A. E. Evins et al., "Medical morbidity and mortality in schizophrenia: guidelines for psychiatrists," *Journal of Clinical Psychiatry*, vol. 66, no. 2, pp. 183–273, 2005.

[6] M. Zink, S. Englisch, and A. Meyer-Lindenberg, "Polypharmacy in schizophrenia," *Current Opinion in Psychiatry*, vol. 23, no. 2, pp. 103–111, 2010.

[7] S. P. Dunn, D. R. Holmes Jr., and D. J. Moliterno, "Drug-drug interactions in cardiovascular catheterizations and interventions," *JACC: Cardiovascular Interventions*, vol. 5, pp. 1195–1208, 2012.

[8] M. Huber, M. Kolzsch, R. Stahlmann et al., "Ophthalmic drugs as part of polypharmacy in nursing home residents with glaucoma," *Drugs & Aging*, vol. 30, pp. 31–38, 2013.

[9] J. Lees and A. Chan, "Polypharmacy in elderly patients with cancer: clinical implications and management," *The Lancet Oncology*, vol. 12, no. 13, pp. 1249–1257, 2011.

[10] C. D. Scripture and W. D. Figg, "Drug interactions in cancer therapy," *Nature Reviews Cancer*, vol. 6, no. 7, pp. 546–558, 2006.

[11] J. Kuhlmann and W. Mück, "Clinical-pharmacological strategies to assess drug interaction potential during drug development," *Drug Safety*, vol. 24, no. 10, pp. 715–725, 2001.

[12] D. S. Tatro, *Drug Interaction Facts*, Facts & Comparisons, 2012.

[13] R. R. Conley and D. L. Kelly, "Drug-drug interactions associated with second-generation antipsychotics: considerations for clinicians and patients," *Psychopharmacology Bulletin*, vol. 40, no. 1, pp. 77–97, 2007.

[14] N. B. Sandson, S. C. Armstrong, and K. L. Cozza, "An overview of psychotropic drug-drug interactions," *Psychosomatics*, vol. 46, no. 5, pp. 464–494, 2005.

[15] M. A. Raggi, R. Mandrioli, C. Sabbioni, and V. Pucci, "Atypical antipsychotics: pharmacokinetics, therapeutic drug monitoring and pharmacological interactions," *Current Medicinal Chemistry*, vol. 11, no. 3, pp. 279–296, 2004.

[16] M. Bubser and A. Y. Deutch, "Differential effects of typical and atypical antipsychotic drugs on striosome and matrix compartments of the striatum," *European Journal of Neuroscience*, vol. 15, no. 4, pp. 713–720, 2002.

[17] J. Sun, H. Xu, and Z. Zhao, "Network-assisted investigation of antipsychotic drugs and their targets," *Chemistry & Biodiversity*, vol. 9, pp. 900–910, 2012.

[18] C. Knox, V. Law, T. Jewison et al., "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1035–D1041, 2011.

[19] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. 1, pp. D901–D906, 2008.

[20] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, pp. D668–D672, 2006.

[21] World Health Organization (WHO), *The Anatomical Therapeutic Chemical*, Classification System, 2003.

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

[23] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[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] J. M. Bland and D. G. Altman, "Multiple significance tests: the Bonferroni method," *British Medical Journal*, vol. 310, no. 6973, article 170, 1995.

[26] D. D. Short, J. M. Hawley, and M. F. McCarthy, "Management of schizophrenia with medical disorders: cardiovascular, pulmonary, and gastrointestinal," *Psychiatric Clinics of North America*, vol. 32, no. 4, pp. 759–773, 2009.

[27] J. R. Horn and P. D. H. Hansten, "Prophylactic dose adjustments in management of drug interactions," *Pharmacy Times*, vol. 77, no. 11, 2011.

[28] Truven Health Analytics Inc, "Micromedex Healthcare Series," 2013.

[29] D. M. Roden, H. Xu, J. C. Denny, and R. A. Wilke, "Electronic medical records as a tool in clinical pharmacology: opportunities and challenges," *Clinical Pharmacology & Therapeutics*, vol. 91, pp. 1083–1086, 2012.

[30] J. F. Carlquist and J. L. Anderson, "Pharmacogenetic mechanisms underlying unanticipated drug responses," *Discovery Medicine*, vol. 11, pp. 469–478, 2011.

[31] 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.

[32] S. F. Kingsmore, I. E. Lindquist, J. Mudge, D. D. Gessler, and W. D. Beavis, "Genome-wide association studies: progress and potential for drug discovery and development," *Nature Reviews Drug Discovery*, vol. 7, no. 3, pp. 221–230, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 520294, 9 pages
http://dx.doi.org/10.1155/2013/520294

Review Article

Diagnosis Value of the Serum Amyloid A Test in Neonatal Sepsis: A Meta-Analysis

Haining Yuan,[1] Jie Huang,[2] Bokun Lv,[1] Wenying Yan,[1,3] Guang Hu,[1] Jian Wang,[2] and Bairong Shen[1,3]

[1] *Center for Systems Biology, Soochow University, Suzhou 215006, China*
[2] *Affiliated Children's Hospital of Soochow University, Suzhou 225121, China*
[3] *Suzhou Zhengxing Translational Biomedical Informatics Ltd., Taicang 215400, China*

Correspondence should be addressed to Guang Hu; huguang@suda.edu.cn

Received 29 April 2013; Accepted 4 July 2013

Academic Editor: Zhongming Zhao

Neonatal sepsis (NS), a common disorder for humans, is recognized as a leading global public health challenge. This meta-analysis was performed to assess the accuracy of the serum amyloid A (SAA) test for diagnosing NS. The studies that evaluated the SAA test as a diagnotic marker were searched in Pubmed, EMBASE, the Cochrane Library, and Google Network between January 1996 and June 2013. A total of nine studies including 823 neonates were included in our meta-analysis. Quality of each study was evaluated by the quality assessment of diagnostic accuracy studies tool (QUADAS). The SAA test showed moderate accuracy in the diagnosis of NS both at the first suspicion of sepsis and 8–96 h after the sepsis onset, both with $Q^* = 0.91$, which is similar to the PCT and CRP tests for the diagnosis of NS in the same period. Heterogeneity between studies was also explained by cut-off point, SAA assay, and age of included neonates. On the basis of our meta-analysis, therefore, SAA could be promising and meaningful in the diagnosis of NS.

1. Introduction

Neonatal sepsis (NS) is recognized as a leading global public health challenge because of the important contribution to neonatal morbidity and mortality in both low and high income countries [1–3]. It is mainly responsible for most of neonatal deaths, and the incidence of NS is about 3–40 per 1000 live births, the mortality rate of whom varies from 9% to 20% of affected neonates [4, 5].

Neonatal sepsis may be defined, both clinically and/or microbiologically, by positive blood and/or cerebrospinal fluid cultures. Early gradual signs and symptoms of NS are often indefinite and subtle, especially at the onset, which is easily confused with other common noninfectious causes [6]. Clinicians often treat with a broad-spectrum antibiotic and prolong treatment with empirical antibiotics, which are associated with adverse outcomes. Furthermore, physical signs are used to identify neonates at risk of sepsis with limited specificity, generally resulting in large number of unnecessary referrals, and antibiotics treatment in the setting of negative cultures may not be benign [7, 8]. Early diagnosis and treatment are vital to improve gradual outcomes; if diagnosed early and with aggressive supportive care, it is possible to save most cases of NS [9]. To this aim, clinicians have sought reliable markers to detect early NS for a long time and to exclude diseases of noninfectious origin [10, 11]. Till now, a large number of markers have been proposed for early diagnosis of sepsis [12], especially C-reactive protein (CRP) and procalcitonin (PCT). CRP is an acute-phase protein found in the blood that is produced by the liver because of infection or tissue injury, while PCT is a 116-amino acid peptide involved as a precursor in calcium homeostasis, and both of them have been widely used as useful markers for the diagnosis of neonatal sepsis [13–17]. However, the specificity and the value of CRP and PCT still have challenges; thus, there is a continuous need for searching for better biomarkers of sepsis.

Serum amyloid A (SAA), the precursor protein in inflammation-associated reactive amyloidosis, whose level in the

blood increases up to 1000 fold in response to information, is synthesized in the liver. SAA is also an acute phase reactant like PCT and CRP, which has been proven to be a prognostic marker in late-onset sepsis in preterm infants [18–20]. Arnon et al. [21] reported that SAA had an overall better diagnostic accuracy than CRP for predicting early onset sepsis. Also, they showed that SAA was a useful inflammatory marker during late-onset sepsis in preterm infants [22]. However, some studies showed an opposite opinion [23, 24]. In view of this contradiction, a more comprehensive study is needed to discuss the accuracy of the SAA test for the diagnosis of NS.

Our primary objective is to systematically and quantitatively evaluate all published studies about the diagnostic use of SAA for NS.

2. Methods

2.1. Retrieval and Selection of Studies. The common approach of computer-aided literature search is used to search PUBMED, EMBASE (http://www.embase.com/), the Cochrane Library (http://www.thecochranelibrary.com/view/0/index.html), and Google Network for relevant citations from January 1996 to June 2013. Our search terms included "serum amyloid A," "SAA," "sepsis," "septicemia," "neonate," "newborn," "infant," and mutual combinations. We have examined the references of known articles to fully retrieve.

The following criteria were applied to identify studies for inclusion in our meta-analysis: (1) studies that assessed the diagnostic accuracy of the SAA test on NS; (2) studies providing both sensitivity and specificity or sufficient information to construct the 2×2 contingency tables; and (3) studies with the sample containing only neonates. Articles including studies that evaluated SAA levels as diagnostic markers for NS are appropriate. The gold standard for the diagnosis of neonatal sepsis is microbial culture blood or other sterile body fluids in these included studies. Furthermore, the change of the SAA in the research sample is the index test for the diagnosis of neonatal sepsis. Selection of articles was performed by two investigators independently to ensure the high accuracy.

2.2. Data Extraction. We extracted data from selected articles, which include first authors, years of publication, study population, region, methods of SAA assay, diagnostic cut-off point and time, and methods quality. Accurate data was extracted to construct 2 × 2 table at a specific time.

2.3. Quality Assessment. Quality assessment of studies was performed based on the quality assessment of diagnostic accuracy studies (QUADAS) tool [25]. This tool consists of 11 key items. Each item was assessed by scoring it as "low," "high," "unclear," which are phrased to the answer, such as "yes" indicates low risk of bias.

2.4. Statistical Analysis. The statistical analysis was performed using Review Manager 5.0 and Meta-DiSc 1.4 software. Studies included in the meta-analysis were divided into two groups according to the time of SAA test for diagnosis of NS. The neonates at the onset of sepsis were set as a group, and those 8–96 h after onset were set as another. The sensitivity, specificity, and diagnostic odds ratio (OR) with corresponding 95% confidence intervals (CI) were calculated for each study. Meanwhile, the pooled sensitivity, specificity and diagnosis OR were also calculated for each group. The diagnostic OR expresses how much greater are the odds of having sepsis for neonates with a positive test result than for neonates with a negative test result [26]. Heterogeneity among included studies is assessed by using the Cochrane Q statistic and quantified with the I^2 lying between 0% and 100% [27]. In general, I^2 (>50%) shows that heterogeneity among studies produce some impact, whereas I^2 (<50%) shows that homogeneity is good for the reliability of meta-analysis. We further performed sensitivity analysis to explore the reasons of heterogeneity and examined characteristics of included studies. To summarize these results, we constructed a summary receiver operator characteristic (SROC) curve, which shows the relationship between sensitivity and the proportion of false positives (1-specificity). Q^* values, defined by the point where sensitivity equals specificity, were calculated from the SROC curves. Meanwhile, the area under SROC curve (AUC) was also calculated to show the probability of the correctly ranked diagnostic test values for a random pair of diseased and nondiseased subjects [28].

3. Results

3.1. Characteristics and Quality of the Included Studies. The literature search was performed as described, and 57 potentially relevant articles were identified. Only 9 articles met our inclusion criteria. Figure 1 shows the process of selecting studies. Detailed characteristics and data of each included study are presented in Tables 1 and 2, respectively.

All the conditions and methods of the included studies, as shown in Figure 2, were used for different quality assessment of diagnostic accuracy. Each included study was strictly judged on the basis of 11 QUADAS tool criteria [25]: withdrawals explained; uninterpretable results reported; relevant clinical information; index test results blinded; reference standard results blinded; incorporation avoided; differential verification avoided; partial verification avoided; Acceptable delay between tests; acceptable reference standard; and representative spectrum. We also appraise them quantitatively according to 11 QUADAS tool criteria, as shown in Table 2.

3.2. Accuracy of the SAA Test in the Diagnosis of NS. Nine articles, which meet inclusion criteria, estimate the use of the SAA test in the diagnosis of NS. We set studies at the first suspicion of NS as a group and those at 8–96 h after sepsis onset as another group on the basis of time point for the SAA test.

Nine studies from included papers evaluated the use of the SAA test at the first suspicion of sepsis. The sensitivity ranged from 23% to 100% (pooled sensitivity: 0.84, 95% CI 80%–87%), whereas specificity ranged from 44% to 100% (pooled sensitivity: 0.89, 95% CI 86%–92%). The detailed descriptions are shown in Figure 3. We found significant heterogeneity among studies (sensitivity, I^2 = 92.7%; specificity,

Table 1: Characteristics of studies included in the meta-analysis of the diagnosis of neonatal sepsis using a SAA test.

Study and year	Study population	Patients (*n*)	Region	Assay method	Time of SAA test: cutoff (mg/L)
Enguix et al. 2001 [31]	Cases: NICU neonates with sepsis Control: neonates without sepsis	46	Spain	AMLN	Onset: 41.3
Arnon et al. 2007 [21]	Cases: full-term neonates with sepsis Control: neonates without sepsis	104	Israel	ALPIA	Onset: 8 24 h after the onset: 10
Çetinkaya et al. 2009 [29]	Cases: NICU neonates with probable sepsis Control: neonates without sepsis	163	Turkey	INMM	Onset: 68 48 h after the onset: 68
Yildiz et al. 2008 [23]	Cases: NICU newborns with suspected sepsis control: noninfected newborns	72	Turkey	ELISA	Onset: 5.5 96 h after the onset: 5.7
Arnon et al. 2005 [22]	Cases: neonates with proven or clinical sepsis Control: noninfected newborns	116	Israel	ELISA	Onset: 10 8 h after the onset: 10
Arnon et al. 2002 [44]	Cases: preterm infants with sepsis or suspected sepsis Control: healthy preterm infants	94	Israel	ELISA	Onset: 10
Edgar et al. 2010 [24]	Cases: term/preterm neonates with infection Control: term/preterm neonates with infection	68	England	ELISA	Onset: 1
Mostafa et al. 2011 [45]	Cases: infants with sepsis or suspected sepsis Control: healthy neonates	100	Egypt	ELISA	Onset: >10
Mohsen et al. 2012 [30]	Cases: term/preterm neonate with sepsis Control: healthy neonates	60	Egypt	ELISA	Onset: cases, 40.16 ± 35.17 control, 6.45 ± 2.42

AMLN: automatic laser nephelometry; ALPIA: automated latex photometric immunoassay; INMM: immunonephelometric method; ELISA: enzyme-linked immunoassay; HSAIA: highly sensitive automated immunoassays, and NICU: neonatal intensive care unit.

Table 2: True positive, Fp, Fn, Tn, Se, Sp, time, and QUADAS of included studies for the diagnosis of NS.

Study and year	Tp	Fp	Fn	Tn	Se	Sp	Time	QUADAS
Enguix et al. 2001 [31]	19	2	1	24	0.95	0.92	Onset	9
Arnon et al. 2007 [21]	22	4	1	77	0.96	0.95	Onset	7
	22	2	1	79	0.96	0.98	24 h after the onset	
Çetinkaya et al. 2009 [29]	94	0	29	40	0.76	1	Onset	6
	80	0	41	42	0.66	1	48 h after the onset	
Yildiz et al. 2008 [23]	27	20	9	16	0.75	0.44	Onset	5
	31	24	5	12	0.86	0.33	96 h after the onset	
Arnon et al. 2005 [22]	40	5	2	68	0.95	0.93	Onset	8
	37	12	0	67	1	0.85	8 h after the onset	
Arnon et al. 2002 [44]	42	4	0	48	1	0.92	Onset	6
Edgar et al. 2010 [24]	6	3	20	39	0.23	0.93	Onset	8
Mostafa et al. 2011 [45]	80	2	2	16	0.98	0.89	Onset	6
Mohsen et al. 2012 [30]	26	2	4	28	0.87	0.93	Onset	6

Tp: true positive; Fp: false positive; Fn: false negative; Tn: true negative; Se: sensitivity; and Sp: specificity.

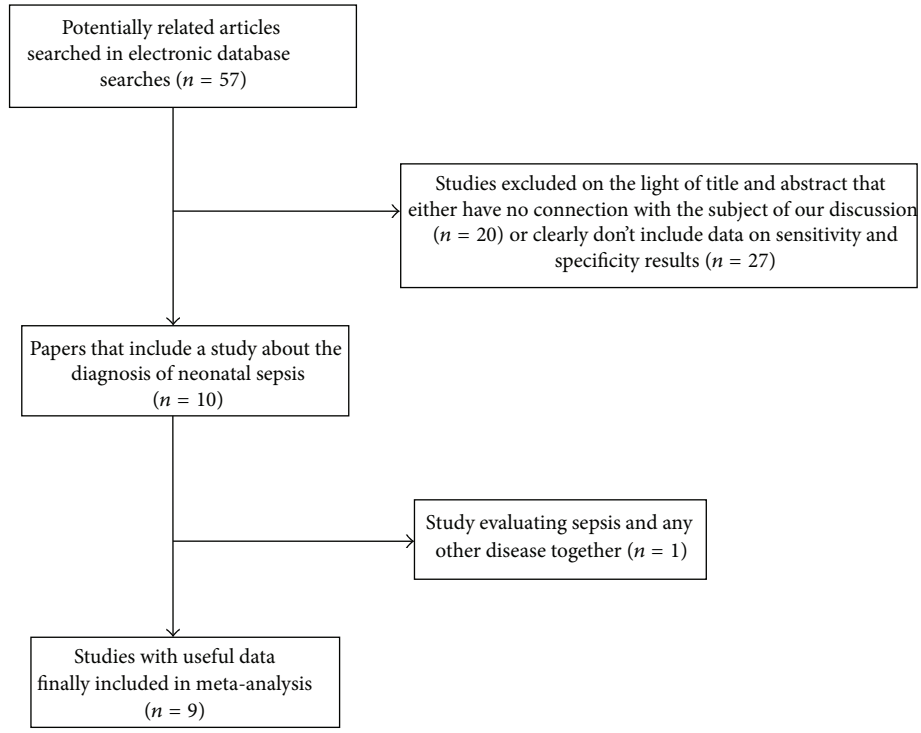


Figure 1: Flow chart of study evaluation and inclusion in the meta-analysis of studies involving diagnosis of neonatal sepsis using a SAA test.

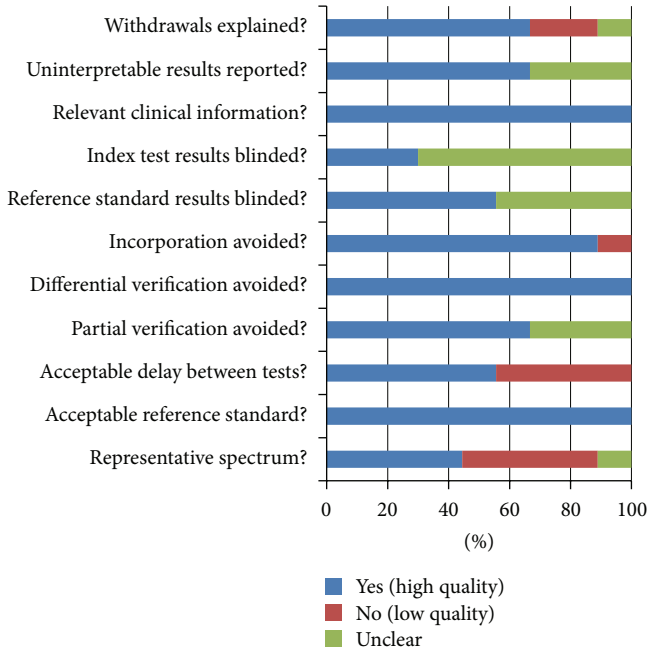


Figure 2: Summary of the methodological quality assessment of the included studies according to 11 QUADAS tool criteria, which are presented as percentages.

I^2 = 86.5%), which indicated that patient selection or other covariates might be responsible for heterogeneity.

The value of DOR of SAA was 91.84 (95% CI, 16.78–502.80), as shown in the forest plot of Figure 4. Among these studies, we also detected significant heterogeneity (I^2 = 86.8%). The corresponding SROC curve was plotted in Figure 5, which shows the AUC was 0.96 with standard error of 0.02, and the pooled diagnostic accuracy (Q^*) was 0.90 with standard error of 0.03, showing a high overall accuracy of SAA for NS.

Four studies from included papers evaluated the use of the SAA test at 8–96 h after the first suspicion of sepsis. Accordingly, Figure 6 shows the pooled sensitivity and specificity, and the diagnostic OR and SROC curve were shown in Figures 7 and 5, respectively.

3.3. Comparison of the Diagnostic Accuracy of Markers for NS. CRP and PCT have been proved to be useful biomarkers for the diagnosis of neonatal sepsis [15–17]. To show the value of diagnosis of the SAA test for NS, we compared SAA with CRP and PCT. Six trials evaluated the diagnosis of both SAA [21–24] and CRP [29, 30]. Compared with 0.67 (95% CI 0.62–0.73) for the CRP test, the pooled sensitivity for the SAA test was better (0.78 (95% CI 0.73–0.83)).

Pooled specificity for the SAA test was slightly lower than for the CRP test, which was 0.89 (95% CI 0.84–0.92) versus 0.92 (95% CI 0.89–0.95). Their difference was not statistically significant ($P < 0.05$). However, the pooled diagnostic OR for the SAA test was smaller than that for the CRP test: 54.95 (95% CI 6.25–483.10) versus 77.16 (95% CI 9.79–248.21). Although this difference was statistically significant ($P > 0.05$), the Q^* value for the SAA test was almost the same as that for the CRP test (0.89 VS 0.84), when the SROC curves for SAA and CRP tests were plotted, respectively. Meanwhile, the AUC for the detection of neonatal sepsis in the SAA

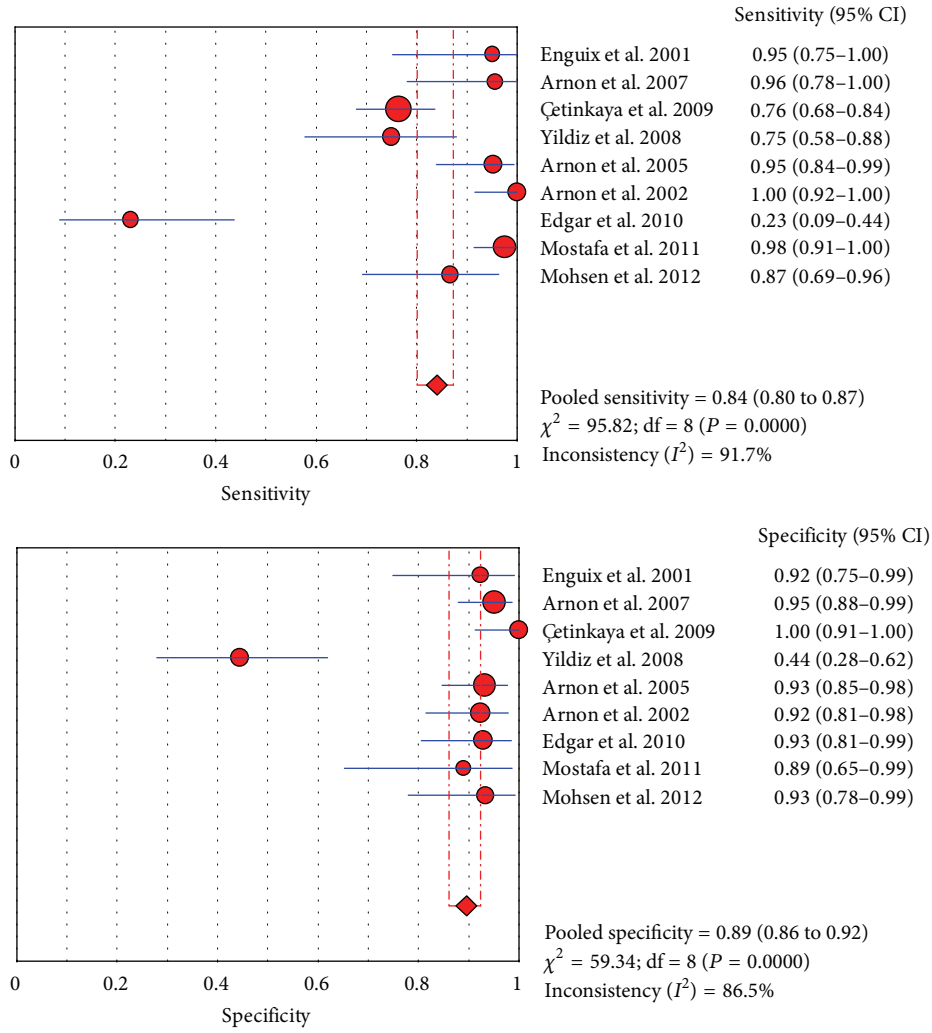


Figure 3: Forest plot [46] for sensitivity and specificity of the SAA test to diagnose neonatal sepsis at the first suspicion of sepsis.

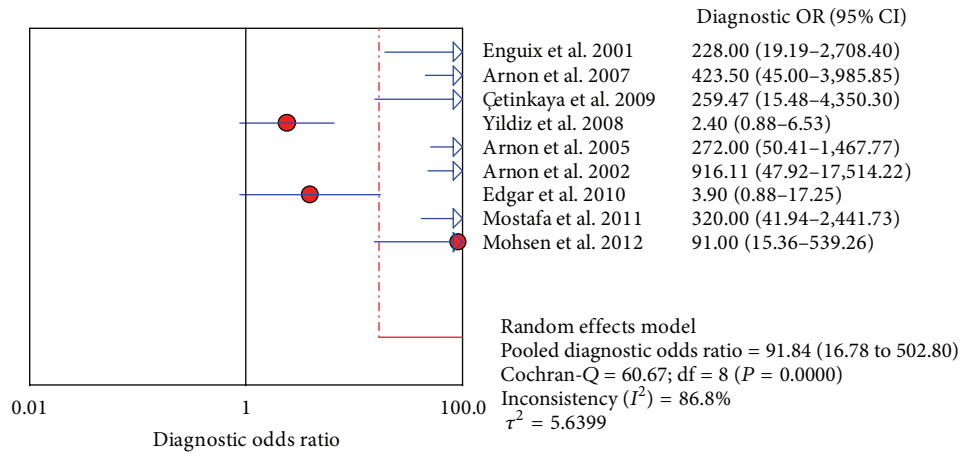


Figure 4: Forest plot for diagnostic OR of the SAA test to diagnose neonatal sepsis at the first suspicion of sepsis.

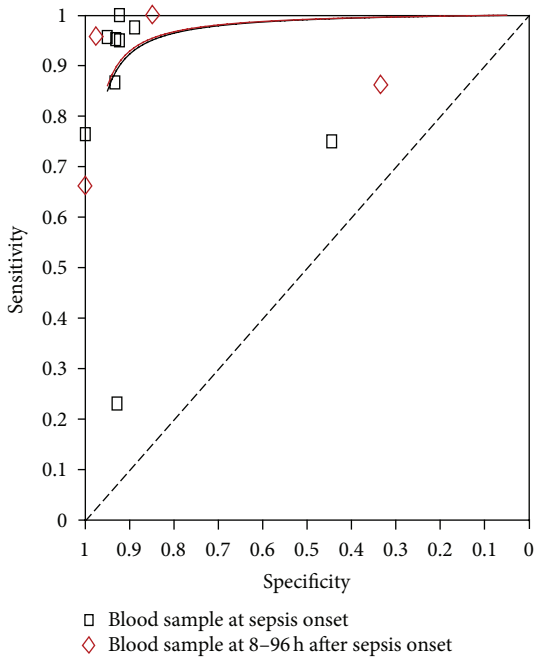


Figure 5: Summary receiver operating characteristic (SROC) curve of the SAA test for the diagnosis of neonatal sepsis.

test is larger than that in the CRP test (0.95 VS 0.91), which means that the SAA test is slightly better than CRP test for the diagnosis of NS by judging from the whole accuracy.

Two studies from included papers were selected to evaluate the use of the SAA and PCT for NS [23, 31]. The pooled sensitivity for the SAA test is slightly lower than for the PCT (0.78 (95% CI 0.73–0.82) versus 0.89 (95% CI 0.78–0.96)), but value of their specificity are similar to each other (0.87 versus 0.87). The pooled diagnosis OR for the SAA test was also almost similar to that for the PCT test (59.74 (95% CI 6.08–586.65) versus 60.17 (95% CI 8.09–447.31)). Since current trials are not sufficient to evaluate the use of SAA and PCT, the SROC curve for the PCT test cannot be plotted, and the Q^* value and the AUC failed to be calculated.

3.4. Analysis of Heterogeneity. In a meta-analysis of diagnostic test, heterogeneity is an important issue to understand the possible factors that influence accuracy diagnosis and estimate the appropriateness of statistical pooling of results from various studies. Variations are brought by several factors, such as the cutoff, assay method, and the age of patients.

In fact, heterogeneity may not be entirely avoided in meta-analysis, so it is necessary for us to explore the reason and extent of heterogeneity. Generally, one of the most important sources of heterogeneity is the threshold effect in a diagnostic study. So, we firstly explored the threshold effect, which was evaluated with the Spearman correlation coefficient with Moses' model weighted by inverse variance. We found that there was no statistically significant difference (Spearman's correlation coefficient = 0.133, P-value =0.731 > 0.5). The source of heterogeneity was explored by meta-regression analysis. The results show that cutoff (≥10 mg/L) is the largest factor (RDOR (relative diagnostic odds ratio) = 14.47, P-value =0.0963), while the effects of the age (RDOR = 0.35, P-value =0.591), assay method (RDOR = 0.08, P-value =0.2405), region (RDOR = 0.5, P-value =0.4997), and QUADAS (RDOR = 1.54, P-value =0.8076) are small. The sensitivity analysis was also performed to identify further analysis of these sources. Homogeneity (I^2 = 0%) was showed among studies when we removed studies of Yildiz et al., 2008 [23], and Edgar, 2010 [24], including relatively small cutoff. If we only collect those studies in the diagnosis of late-onset neonates sepsis (postnatal age > 72 h), heterogeneity (I^2 = 89.9%) was found. When the study of Yildiz et al., 2008 [23] was removed, the remaining studies also showed homogeneity (I^2 = 0%). As the amount of the included studies is too small, we cannot do further analysis of other factors, such as subgroup analysis. We except that our results will become more convincing if more and more studies about the SAA test of the diagnosis of neonatal sepsis are published in the future.

3.5. Publication Bias. The Deeks test [32] was performed to detect publication bias by using the Stata 11 software. As shown in Figure 8, this was not statistically significant for the studies of the SAA test in the diagnosis of neonatal sepsis (P = 0.406 > 0.05). So, the result indicates no potential for publication bias. However, our included studies are so few that this result may be biased.

4. Discussion

Neonatal sepsis is the most common cause of neonatal deaths with high mortality; thus, its identification is vital to improve bad results. Clinical signs are subtle and nonspecific, and laboratory tests including blood culture are not always reliable [33]. So far, many markers for NS have been suggested, such as CRP, PCT, IL-6, and TNF-α. However, a single biomarker is not sufficiently reliable for identification of NS at present. More researchers focus on the combination of different biomarkers in different clinical settings and hope to achieve clearer conclusions [34, 35]. Therefore, it is necessary for us to clearly study important markers for future research of NS. Here, we demonstrated that the SAA test shows appropriate accuracy for the diagnosis of NS.

An ideal biomarker should have high sensitivity and specificity, that is, have a high diagnostic accuracy [36]. In this meta-analysis, pooled sensitivity of the SAA test was 0.84 for the diagnosis of NS, the pooled specificity was 0.89, and Q^* value was 0.91 at the first suspicion of sepsis; then, 8–96 h after the first suspicion of sepsis, a pooled sensitivity of the SAA test was 0.78 for the diagnosis of NS, the pooled specificity was 0.84, and Q^* value was 0.91. These results show that the SAA test is a valuable biomarker and also has a long diagnostic cycle for the diagnosis of NS.

Certainly, the limitations of our work should be considered. This meta-analysis only contained nine studies, though we tried our best to retrieve appropriate papers. In addition, we divided the sample into onset and 8–96 h after onset entirely based on characteristics of the sample, which accords with clinical facts, while the sample was divided into early-onset and late-onset neonatal sepsis generally. Different

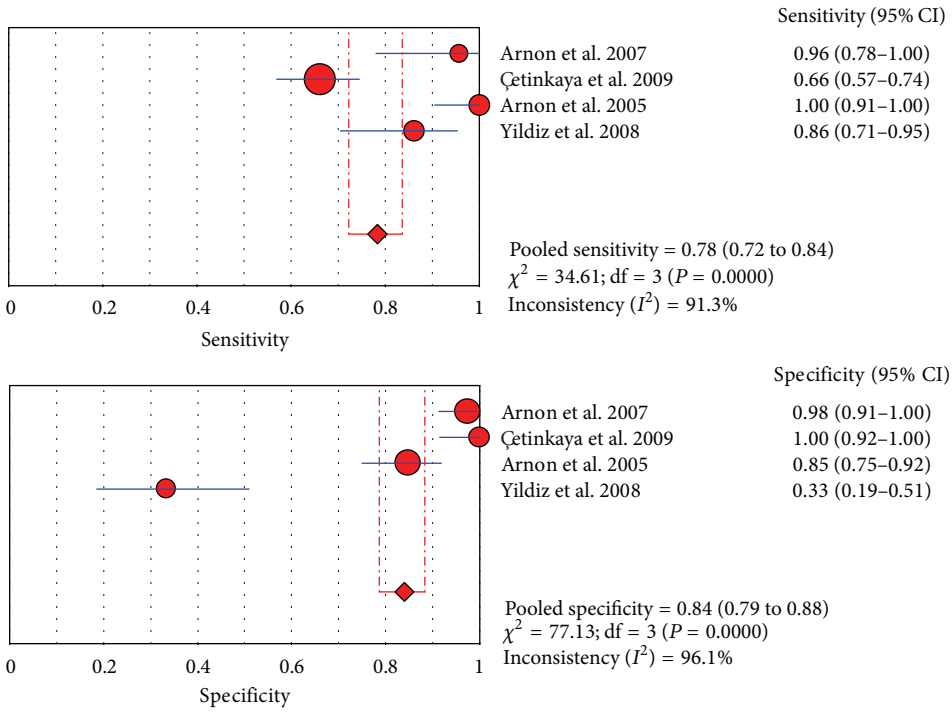


Figure 6: Forest plots for sensitivity and specificity of the SAA test to diagnose neonatal sepsis at 8–96 h after the first suspicionof sepsis.

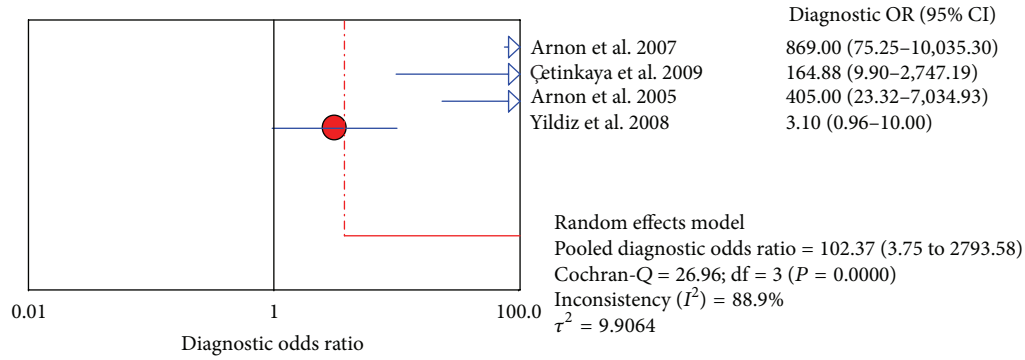


Figure 7: Forest plot for diagnostic OR of the SAA test to diagnose neonatal sepsis at 8–96 h after the first suspicion of sepsis.

cut-off values, ranging from 1 to 75.17 mg/L, exist in the included studies. Publication bias might be generally difficult to be avoided in the meta-analysis, so we tried to include more studies and even adverse studies about the accuracy of the diagnosis of SAA to reduce the bias. To explore the reasons for heterogeneity, we considered the diagnostic cut-off point, which may partially explain this heterogeneity because of differences between studies. Therefore, we suspected that heterogeneity will be reduced by considering more studies for the SAA test, which might reduce the effect of different sampling, clinical settings, and other heterogeneity-caused factors.

CRP, a traditional useful marker, has been applied in clinic [37]. It is triggered by cytokines IL-6, TNF-α, and so forth and is a late marker of neonatal sepsis which increases evidently at 24 h after sepsis onset [22, 38]. Its sensitivity was 30%–97%, and its specificity ranged from 75% to 100% [39]. In included studies of the meta-analysis, the sensitivity of SAA varied from 26% to 100%, and specificity varied from 44% to 100%. To be truth, it is difficult to evaluate which marker is good or not. Lannergård et al. [40] also thought that there are positive correlations between SAA and CRP in infectious diseases. However, SAA rises earlier and more sharply than CRP, especially during the first 24 h after sepsis onset [41, 42], and it has showed a moderate accuracy at 8–96 h after the first suspicion of sepsis in our meta-analysis, which means that SAA is useful in the diagnosis of NS.

In addition, PCT is also a useful marker which shows better accuracy than CRP for the diagnotic of NS in some aspects [16]. In our meta-analysis, the pooled diagnotic OR for the SAA test was also almost similar to that for the PCT test (59.74 versus 60.17). So, we believe that SAA is

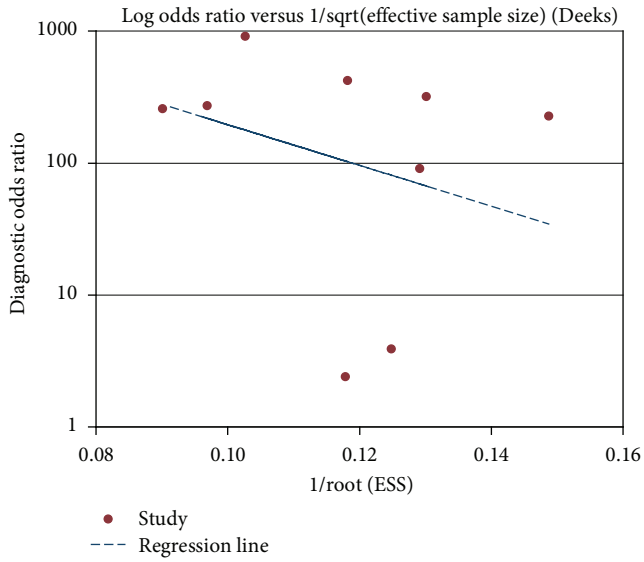


Figure 8: Test for the assessment of potential publication bias in the SAA test for the diagnosis of neonatal sepsis.

also a useful marker for the diagnosis of NS like CRP, PCT. Certainly, SAA cannot be replaced easily because it has a prognostic value as early as eight hours after the onset and before clinical signs. For example, des-arginine variants of SAA were identified as the most promising biomarkers, which can make neonatologists withhold treatment in 45% of nonsepsis neonates [43]. Moreover, SAA can be combined with sICAM-1, CRP, and sE-selectin to improve the diagnosis results, and it can be also added to PCT and CRP, which may increase the rate of sepsis diagnosis by about 10% [24, 29]. We believe that its usefulness will be showed if we evaluate it in combination with other markers or perform a clearer research on it alone in future studies.

In summary, SAA showed moderate accuracy and a longer diagnostic cycle in the diagnosis of neonatal sepsis. Furthermore, the SAA test has showed better accuracy than the CRP test for the diagnosis of neonatal sepsis in the first suspicion of sepsis. It not only has higher accuracy at the first suspicion of sepsis, but also keeps this usefulness at 8–96 after the first suspicion of sepsis. Accordingly, we believe that the combination of SAA with CRP and PCT will improve the diagnosis. To further analyze the diagnostic accuracy of the SAA test for the diagnosis of NS and correlation with other biomarkers in depth, follow-up clinical validation is needed.

Authors' Contribution

Haining Yuan and Jie Huang have equally contributed to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Grants (81272143, 21203131), the Natural Science Foundation of Jiangsu Province (K200509), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (12KJB180014), Jiangsu Innovation Team Grant LJ201141, and Program of Innovative and Entrepreneurial Talent.

References

[1] S. A. Qazi and B. J. Stoll, "Neonatal sepsis: a major global public health challenge," *Pediatric Infectious Disease Journal*, vol. 28, no. 1, supplement, pp. S1–S2, 2009.

[2] Y.-J. Shin, M. Ki, and B. Foxman, "Epidemiology of neonatal sepsis in South Korea," *Pediatrics International*, vol. 51, no. 2, pp. 225–232, 2009.

[3] S. Vergnano, M. Sharland, P. Kazembe, C. Mwansambo, and P. T. Heath, "Neonatal sepsis: an international perspective," *Archives of Disease in Childhood: Fetal and Neonatal Edition*, vol. 90, no. 3, pp. F220–F224, 2005.

[4] E. Persson, B. Trollfors, L. L. Brandberg, and I. Tessin, "Septicaemia and meningitis in neonates and during early infancy in the Göteborg area of Sweden," *Acta Paediatrica*, vol. 91, no. 10, pp. 1087–1092, 2002.

[5] V. Sundaram, P. Kumar, S. Dutta et al., "Blood culture confirmed bacterial sepsis in neonates in a north Indian tertiary care center: changes over the last decade," *Japanese Journal of Infectious Diseases*, vol. 62, no. 1, pp. 46–50, 2009.

[6] P. C. Ng, "Clinical trials for evaluating diagnostic markers of infection in neonates," *Biology of the Neonate*, vol. 87, no. 2, pp. 111–112, 2005.

[7] M. W. Weber, J. B. Carlin, S. Gatchalian et al., "Predictors of neonatal sepsis in developing countries," *The Pediatric Infectious Disease Journal*, vol. 22, no. 8, pp. 711–717, 2003.

[8] N. Tripathi, C. M. Cotten, and P. B. Smith, "Antibiotic use and misuse in the neonatal intensive care unit," *Clinics in Perinatology*, vol. 39, no. 1, pp. 61–68, 2012.

[9] I. M. Stefanovic, "Neonatal sepsis," *Biochemia Medica (Zagreb)*, vol. 21, no. 3, pp. 276–281, 2011.

[10] U. K. Mishra, S. E. Jacobs, L. W. Doyle, and S. M. Garland, "Newer approaches to the diagnosis of early onset neonatal sepsis," *Archives of Disease in Childhood: Fetal and Neonatal Edition*, vol. 91, no. 3, pp. F208–F212, 2006.

[11] P. C. Ng and H. S. Lam, "Diagnostic markers for neonatal sepsis," *Current Opinion in Pediatrics*, vol. 18, no. 2, pp. 125–131, 2006.

[12] C. Pierrakos and J.-L. Vincent, "Sepsis biomarkers: a review," *Critical Care*, vol. 14, no. 1, article R15, 2010.

[13] E. Bilavsky, H. Yarden-Bilavsky, S. Ashkenazi, and J. Amir, "C-reactive protein as a marker of serious bacterial infections in hospitalized febrile infants," *Acta Paediatrica, International Journal of Paediatrics*, vol. 98, no. 11, pp. 1776–1780, 2009.

[14] A. Kordek, M. Hałasa, and W. Podraza, "Early detection of an early onset infection in the neonate based on measurements of procalcitonin and C-reactive protein concentrations in cord blood," *Clinical Chemistry and Laboratory Medicine*, vol. 46, no. 8, pp. 1143–1148, 2008.

[15] J. P. S. Caldas, S. T. M. Marba, M. H. S. L. Blotta, R. Calil, S. S. Morais, and R. T. D. Oliveira, "Accuracy of white blood cell count, C-reactive protein, interleukin-6 and tumor necrosis factor alpha for diagnosing late neonatal sepsis," *Jornal de Pediatria*, vol. 84, no. 6, pp. 536–542, 2008.

[16] Z. Yu, J. Liu, Q. Sun, Y. Qiu, S. Han, and X. Guo, "The accuracy of the procalcitonin test for the diagnosis of neonatal sepsis:

a meta-analysis," *Scandinavian Journal of Infectious Diseases*, vol. 42, no. 10, pp. 723–733, 2010.

[17] B. S. Naher, M. A. Mannan, K. Noor, and M. Shahiddullah, "Role of serum procalcitonin and C-reactive protein in the diagnosis of neonatal sepsis," *Bangladesh Medical Research Council Bulletin*, vol. 37, no. 2, pp. 40–46, 2011.

[18] S. Urieli-Shoval, R. P. Linke, and Y. Matzner, "Expression and function of serum amyloid A, a major acute-phase protein, in normal and disease states," *Current Opinion in Hematology*, vol. 7, no. 1, pp. 64–69, 2000.

[19] D. B. Jovanovic, "Clinical importance of determination of serum amyloid A," *Srpski Arhiv Za Celokupno Lekarstvo*, vol. 132, no. 7-8, pp. 267–271, 2004.

[20] S. Arnon, I. Litmanovitz, R. Regev, M. Lis, R. Shainkin-Kestenbaum, and T. Dolfin, "The prognostic virtue of inflammatory markers during late-onset sepsis in preterm infants," *Journal of Perinatal Medicine*, vol. 32, no. 2, pp. 176–180, 2004.

[21] S. Arnon, I. Litmanovitz, R. H. Regev, S. Bauer, R. Shainkin-Kestenbaum, and T. Dolfin, "Serum amyloid A: an early and accurate marker of neonatal early-onset sepsis," *Journal of Perinatology*, vol. 27, no. 5, pp. 297–302, 2007.

[22] S. Arnon, I. Litmanovitz, R. Regev et al., "Serum amyloid A protein is a useful inflammatory marker during late-onset sepsis in preterm infants," *Biology of the Neonate*, vol. 87, no. 2, pp. 105–110, 2005.

[23] B. Yildiz, B. Ucar, M. A. Aksit et al., "Serum amyloid A, procalcitonin, tumor necrosis factor-α, and interleukin-1β levels in neonatal late-onset sepsis," *Mediators of Inflammation*, vol. 2008, Article ID 737141, 7 pages, 2008.

[24] J. D. M. Edgar, V. Gabriel, J. R. Gallimore, S. A. McMillan, and J. Grant, "A prospective study of the sensitivity, specificity and diagnostic performance of soluble intercellular adhesion molecule 1, highly sensitive C-reactive protein, soluble E-selectin and serum amyloid A in the diagnosis of neonatal infection," *BMC Pediatrics*, vol. 10, article 22, 2010.

[25] P. Whiting, A. W. S. Rutjes, J. B. Reitsma, P. M. M. Bossuyt, and J. Kleijnen, "The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews," *BMC Medical Research Methodology*, vol. 3, article 1, 2003.

[26] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.

[27] J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *British Medical Journal*, vol. 327, no. 7414, pp. 557–560, 2003.

[28] S. D. Walter, "Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data," *Statistics in Medicine*, vol. 21, no. 9, pp. 1237–1256, 2002.

[29] M. Çetinkaya, H. Özkan, N. Köksal, S. Çelebi, and M. Hacimustafaoğlu, "Comparison of serum amyloid A concentrations with those of C-reactive protein and procalcitonin in diagnosis and follow-up of neonatal sepsis in premature infants," *Journal of Perinatology*, vol. 29, no. 3, pp. 225–231, 2009.

[30] L. M. Mohsen et al., "Study on diagnostic value of serum amyloid A protein during late- onset sepsis in preterm and full term neonates," *Australian Journal of Basic and Applied Sciences*, vol. 6, no. 12, pp. 530–536, 2012.

[31] A. Enguix, C. Rey, A. Concha, A. Medina, D. Coto, and M. A. Diéguez, "Comparison of procalcitonin with C-reactive protein and serum amyloid for the early diagnosis of bacterial sepsis in critically ill neonates and children," *Intensive Care Medicine*, vol. 27, no. 1, pp. 211–215, 2001.

[32] J. J. Deeks, P. Macaskill, and L. Irwig, "The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed," *Journal of Clinical Epidemiology*, vol. 58, no. 9, pp. 882–893, 2005.

[33] J. S. Gerdes, "Diagnosis and management of bacterial infections in the neonate," *Pediatric Clinics of North America*, vol. 51, no. 4, pp. 939–959, 2004.

[34] S. Arnon and I. Litmanovitz, "Diagnostic tests in neonatal sepsis," *Current Opinion in Infectious Diseases*, vol. 21, no. 3, pp. 223–227, 2008.

[35] Y. Fan and J. L. Yu, "Umbilical blood biomarkers for predicting early-onset neonatal sepsis," *World Journal of Pediatrics*, vol. 8, no. 2, pp. 101–108, 2012.

[36] J. C. Marshall and K. Reinhart, "Biomarkers of sepsis," *Critical Care Medicine*, vol. 37, no. 7, pp. 2290–2298, 2009.

[37] M. M. Levy, M. P. Fink, J. C. Marshall et al., "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference," *Intensive Care Medicine*, vol. 29, no. 4, pp. 530–538, 2003.

[38] B. Janković, D. Veljković, S. Pasić, Z. Rakonjac, D. Jevtić, and J. Martić, "C-reactive protein and cytokines in the diagnosis of neonatal sepsis," *Medicinski pregled*, vol. 59, no. 11-12, pp. 545–549, 2006.

[39] T. Chan and F. Gu, "Early diagnosis of sepsis using serum biomarkers," *Expert Review of Molecular Diagnostics*, vol. 11, no. 5, pp. 487–496, 2011.

[40] A. Lannergård, A. Larsson, P. Kragsbjerg, and G. Friman, "Correlations between serum amyloid A protein and C-reactive protein in infectious diseases," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 63, no. 4, pp. 267–272, 2003.

[41] C. Pizzini, M. Mussap, M. Plebani, and V. Fanos, "C-reactive protein and serum amyloid A protein in neonatal infections," *Scandinavian Journal of Infectious Diseases*, vol. 32, no. 3, pp. 229–235, 2000.

[42] M. Mussap, A. Noto, F. Cibecchini, and V. Fanos, "The importance of biomarkers in neonatology," *Seminars in Fetal & Neonatal Medicine*, vol. 18, no. 1, pp. 56–64, 2013.

[43] P. C. Ng, I. L. Ang, R. W. K. Chiu et al., "Host-response biomarkers for diagnosis of late-onset septicemia and necrotizing enterocolitis in preterm infants," *Journal of Clinical Investigation*, vol. 120, no. 8, pp. 2989–3000, 2010.

[44] S. Arnon, I. Litmanovitz, R. Regev, M. Lis, R. Shainkin-Kestenbaum, and T. Dolfin, "Serum amyloid A protein in the early detection of late-onset bacterial sepsis in preterm infants," *Journal of Perinatal Medicine*, vol. 30, no. 4, pp. 329–332, 2002.

[45] M. S. Mostafa et al., "Serum amyloid A an early diagnostic marker for neonatal sepsis," *Life Science Journal*, vol. 8, no. 3, pp. 271–277, 2011.

[46] S. Lewis and M. Clarke, "Forest plots: trying to see the wood and the trees," *British Medical Journal*, vol. 322, no. 7300, pp. 1479–1480, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 983805, 9 pages
http://dx.doi.org/10.1155/2013/983805

Research Article

NCBI2RDF: Enabling Full RDF-Based Access to NCBI Databases

Alberto Anguita, Miguel García-Remesal, Diana de la Iglesia, and Victor Maojo

Biomedical Informatics Group, Artificial Intelligence Laboratory, School of Computer Science, Universidad Politécnica de Madrid, Campus de Montegancedo S/N, Boadilla del Monte, 28660 Madrid, Spain

Correspondence should be addressed to Alberto Anguita; aanguita@infomed.dia.fi.upm.es

Received 3 May 2013; Accepted 30 June 2013

Academic Editor: Wanwipa Vongsangnak

RDF has become the standard technology for enabling interoperability among heterogeneous biomedical databases. The NCBI provides access to a large set of life sciences databases through a common interface called Entrez. However, the latter does not provide RDF-based access to such databases, and, therefore, they cannot be integrated with other RDF-compliant databases and accessed via SPARQL query interfaces. This paper presents the NCBI2RDF system, aimed at providing RDF-based access to the complete NCBI data repository. This API creates a virtual endpoint for servicing SPARQL queries over different NCBI repositories and presenting to users the query results in SPARQL results format, thus enabling this data to be integrated and/or stored with other RDF-compliant repositories. SPARQL queries are dynamically resolved, decomposed, and forwarded to the NCBI-provided E-utilities programmatic interface to access the NCBI data. Furthermore, we show how our approach increases the expressiveness of the native NCBI querying system, allowing several databases to be accessed simultaneously. This feature significantly boosts productivity when working with complex queries and saves time and effort to biomedical researchers. Our approach has been validated with a large number of SPARQL queries, thus proving its reliability and enhanced capabilities in biomedical environments.

1. Introduction

Over the last decade, there has been a paradigm shift regarding how biomedical data is used for biomedical research, moving from a single database based approach towards an integrative one based on the seamless access and analysis of data from multiple heterogeneous sources. Different technological advances have contributed to this shift in focus, from which two of them stand out over the rest. On one hand, the breakthrough in high-throughput techniques for "omics" data—for example, genomic, proteomic, transcriptomic, epigenomic, cytomic, and so forth—generation has led to the development of novel databases providing a myriad of original data ready to be exploited [1, 2]. On the other hand, advances in telecommunications, improvement of transfer bandwidths, and the increasing ability to access remotely located databases over the Internet have hugely facilitated the access and sharing of biomedical repositories. This way, researchers have gained access to vast amounts of data, enabling them to undertake new lines of research [3–7].

The benefits of the integrative approach are manifold, mainly in the area of enhanced diagnosis and treatment of diverse diseases [8]. For example, Petrik et al. identified biomarkers for brain tumors by jointly analyzing "omics" data from brain tumor tissue [9]. Zirn et al. employed integrated clinical and genomic data to obtain genetic biomarkers that allow creating a personalized treatment for each patient [10]. More recent examples are described by Ferrara et al., by integrating metabolomic and transcriptional profiling to "construct causal networks for control of specific metabolic processes in liver" [11]. Connor et al. performed integration of metabolomics and transcriptomics data to discover biomarkers related to type 2 diabetes [12]. Elkan-Miller et al. identified several miRNAs functionally important in cases of deafness in mammals after integrating transcriptomics, proteomics, and microRNA analyses [13]. Yi et al. carried out integration of genomic and epigenomic data to "identify key genes and pathways altered in colorectal cancers (CRC)," leading to a prognostic signature in colon cancer [14].

However, the integrative data access approach involves a more complex data handling process. Therefore, researchers and/or database curators will often need carrying out a homogenization and integration step prior to analyzing the data. Heterogeneities among disparate data sources greatly

hamper this task. Biomedical researchers demand new data access techniques that relief them from the hassle of handling multiple heterogeneous data sources at a time and that allow them to automatically perform the data homogenization process. To overcome this problem, there have been numerous efforts in the bioinformatics community towards providing methods, tools, and standards aimed at facilitating the integrated access to heterogeneous data sources. One of the most important achievements has been the development of RDF (http://www.w3.org/RDF/), a framework for describing generic resources. RDF was created by the W3C consortium and accepted as a standard in 2004. The development of RDF has facilitated the creation of numerous resources for describing specific knowledge areas. These resources are used in the database integration field as shared vocabularies, providing unified frameworks—that is, shared conceptualizations—that simplify the homogenization process of disparate data [15]. In this sense, ontologies have been established as shared vocabularies [16], generally using RDF itself as the representation language, or any of its extensions (OWL (http://www.w3.org/2004/OWL/), OWL2 (http://www.w3.org/TR/owl2-primer/)). The development of ontologies over the last few years has provided an extensive collection of formal domain descriptions, especially in the area of biomedicine. Some of the most well-known contributions are, for instance, the Gene Ontology (GO), offering a representation of gene and gene product attributes across species [17], the Foundational Model of Anatomy (FMA), built as a symbolic representation of the phenotypic structure of the human body [18], the ACGT Master Ontology (ACGTMO), a thorough description of the area of modern clinical trials on cancer [19], and the Protein Ontology (PRO), a "formal, logically-based classification of specific protein classes including structured representations of protein isoforms, variants, and modified forms" [20]. On top of these developments, there also exist several initiatives targeted at gathering collections of relevant ontologies and providing them publicly. One of these initiatives is the OBO Foundry, a consortium dedicated to the establishment of good practices in ontology development [21]. Another relevant initiative is BioPortal, which "provides access via Web Services and Web browsers to ontologies developed in OWL, RDF, OBO format and Protégé frames" [22].

Regarding state-of-the-art formalisms for querying RDF-based data sources, SPARQL (http://www.w3.org/TR/rdf-sparql-query/) is the most widespread query language at the time of writing this paper. It was developed by the W3C consortium and became a standard in 2008. Nowadays, RDF and SPARQL have been established as the *de facto* data model and query language for representing and accessing biomedical information, respectively. Most biomedical research institutions provide RDF-based access to their repositories, and there are several initiatives targeted at automatically providing RDF views of existing data—for example, Bio2RDF [23]—and SPARQL endpoints to access them. There are also approaches that propose adopting RDF as a solution to increasingly overwhelming sizes of biomedical databases [24].

Among the biomedical data sources that researchers often access and use in their work, a very important one is the public set of databases hosted by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) (NCBI). The NCBI was founded more than two decades ago with the mission of providing researchers with access to the most relevant biomedical databases. However, and despite its importance, the NCBI data access service lacks the ability to access the data in an RDF-compliant form. In this paper, we present NCBI2RDF, a system designed to provide RDF-based access to the NCBI databases through SPARQL query endpoint. Furthermore, NCBI2RDF offers increased expressivity and enhanced functionalities embedded in its query endpoint, compared to those offered by the native interface of NCBI.

Next section describes in detail the NCBI database system and its native querying interface. Section 3 explains our approach for translating and decomposing SPARQL queries into simple queries supported by the NCBI databases. In Section 4, we describe how our system would service a sample query. Section 5 discusses the benefits of our approach and compares it with other related initiatives. Finally, Section 6 provides a summary and the conclusions.

2. Background

The NCBI was established in 1988 with the goal of offering computerized access to a wide set of biomedical data repositories. Its infrastructure has been continuously growing by adding new databases, services, and tools. In 2009, PubMed itself was accessed almost 100 million times each month (http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmed.html). This has pushed the NCBI data sources as one of the most important biomedical data resources for biomedical researchers (http://1degreebio.org/blog/?bid=146/).

The NCBI manages a large set of biomedical databases storing different types of data, all free to access. These include, as of May 2013, over 50 databases ranging from citations and abstracts (PubMed) to genetic repositories (Gene, GenBank, etc.). The background of users accessing these resources includes physicians, biologists, and medical informaticians, bioinformaticians, medical students. In order to provide these users with Internet-based access to the data, the NCBI offers a web-based interface for accessing and displaying the data of their hosted databases. This interface, called Entrez (http://www.ncbi.nlm.nih.gov/sites/gquery/) [25], provides a simple entry point for searching biomedical information based on keyword-based searches—including terms from selected lexical resources such as the MeSH thesaurus—and links among related data.

The Entrez system is focused on ease of use. Its interface features a simple HTML form for specifying search filters over either one of the databases or the entire repository set. This provides users lacking technical background an adequate access point to the stored data. Searches through the Entrez system produce web pages containing a list of UIDs, the global identifier used in NCBI to refer to every entry, independent of the database. Each UID provides a link

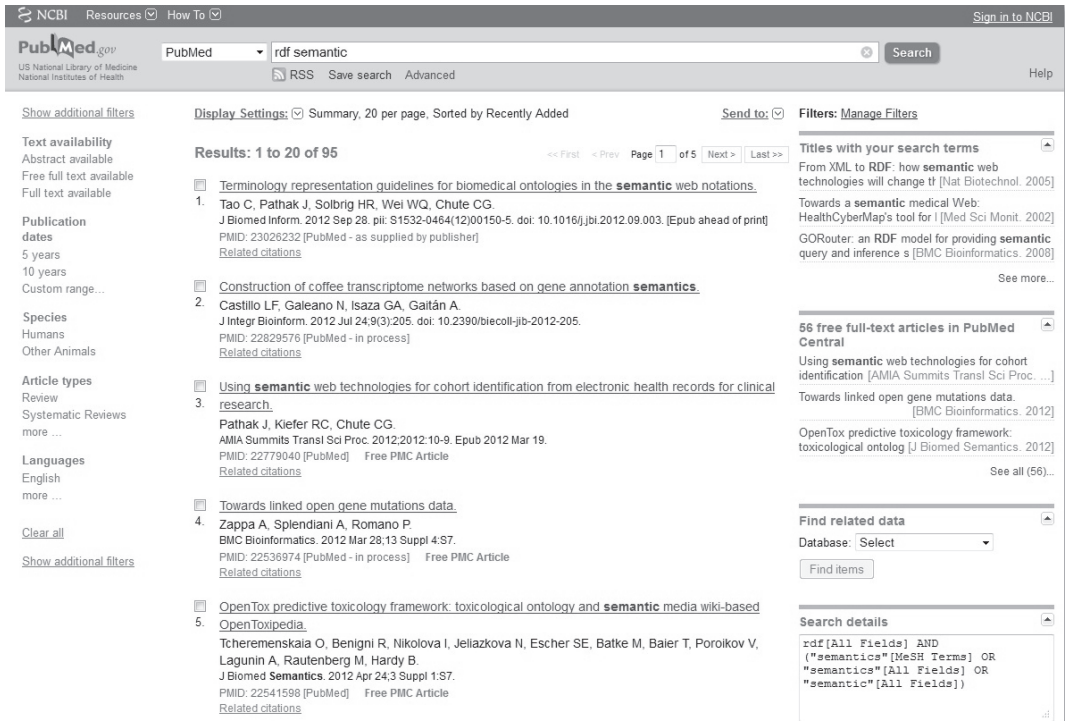


Figure 1: Screen capture of the Entrez system. This screen shows the UIDs of the results for the search of the term "*rdf semantic*" over the PubMed database. For example, the first result shows the UID 23026232.

to an HTML page displaying detailed information about the selected entry—a scientific publication in the case of the PubMed database or a gene description in the case of the Gene database. The inner details of these HTML pages depend on the nature of the queried database. For instance, results from PubMed include details about the retrieved scientific publication—title, journal, authors, abstract, and so forth. These individual result pages may also cross-reference other UIDs which are related to the currently selected result. Figure 1 depicts a screenshot of the Entrez interface showing the results of the sample query "rdf semantic."

The Entrez navigation-based system allows nontechnical users to easily access the data stored at the NCBI repositories. The interface allows either performing a general term search in all NCBI databases or defining a more complex query for one specific database. However, the system lacks the ability to enable users to customize the structure of the results—that is, choosing the data fields to be retrieved for each item on the result set—or even to display compound results created by integrating records from different databases. These constraints greatly limit the expressivity of allowed queries in Entrez. Therefore, we believe that Entrez is a suitable interface for performing simple searches, but impractical for more complex situations involving accessing large amounts of interrelated data.

The NCBI offers a second approach for accessing their data: the Entrez Programming Utilities (E-utilities) (http://www.ncbi.nlm.nih.gov/books/NBK25500/). The E-utilities are a set of web-based services where queries are submitted as URLs and results are provided in simple HTML pages or XML documents. The URLs contain all the information needed by the server to resolve the query—database, filters, number of desired results, and so forth— and can be constructed using a simple set of rules. This interface is targeted at developers who intend to build applications that access the NCBI repositories. The E-utilities also implement a powerful feature for reusing results from queries: the history server. The history server maintains recently retrieved lists of UIDs and provides keys for accessing them. The goal is twofold: on one hand, client applications are relieved from repeating queries for accessing frequent data, and on the other hand, the NCBI servers receive a lesser amount of requests. The adoption of the E-utilities by other applications is rather simple. URL codification and result parsing is straightforward. Nevertheless, this is a proprietary format that does not follow any existing standard. In addition, limitations regarding the type of queries allowed with respect to Entrez persist, since queries must be targeted to the whole repository set or to a specific database, excluding any sort of complex query that uses *join* statements.

3. Methods

3.1. Overview. NCBI2RDF is a Java API for setting up SPARQL-query endpoints over the NCBI databases. It provides a channel for performing SPARQL queries over the NCBI repositories and retrieving data in SPARQL results format. The goal is twofold. On one hand, the API provides RDF-based access to the entire NCBI data, facilitating its

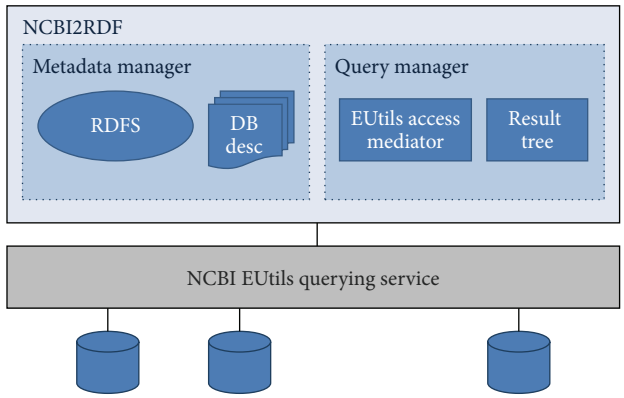


Figure 2: NCBI2RDF system architecture.

```
<eInfoResult>
   <DbList>
      <DbName>pubmed</DbName>
      <DbName>protein</DbName>
      <DbName>nuccore</DbName>
      <DbName>nucleotide</DbName>
      <DbName>nucgss</DbName>
      ...
   </DbList>
</eInfoResult>
```

Figure 3: The master XML file listing all available databases.

integration with other biomedical resources. On the other hand, more powerful queries can be performed compared to the native NCBI querying system, enabling users to launch complex queries involving multiple NCBI databases.

NCBI2RDF adopts a dynamic query translation approach for resolving queries. For each SPARQL query posed against the system, NCBI2RDF sets up a workflow of requests to NCBI that allows fetching the data requested by users. This approach stands upon two main processes: metadata generation and query resolution. The metadata generation process consists of building a set of formal descriptions of the data sources available at the NCBI and how to access them. The query resolution process is the one in charge of solving SPARQL queries posed by users. Figure 2 depicts the system architecture.

The next subsections describe these two processes in detail.

3.2. Metadata Generation. The first step in the development of NCI2RDF was obtaining formal descriptions of the databases hosted by the NCBI. NCBI2RDF requires these descriptions to build the RDF schema for the NCBI databases and to be able to access them. The data required by NCBI2RDF includes a list of identifiers of databases stored by the NCBI, a list of fields that each of those databases includes, distinguishing between retrievable fields, that is, those that can be shown within the results, and filterable fields, that is, those that can be used to constrain the searches, and the identifiers of the existing relations between databases. All these descriptions can be obtained through specific web services provided by the NCBI.

In a first attempt, we decided to generate the needed metadata manually. However, the rather frequent changes that these descriptions undergo forced us to develop an automated process capable of retrieving the repository descriptions and generating the metadata files automatically. This process, called metadata generation, must be triggered regularly to keep the API up to date with the changes in the NCBI databases structure.

NCBI provides formal descriptions of all its databases through XML files. Through the web service available at http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi/, a master file listing all available databases can be retrieved. Figure 3 shows a code snippet of this master file.

Using the available NCBI database names, it is also possible to retrieve XML files describing each individual database. These XMLs provide general information about the database—that is, a natural language description of the database—and details on its structure, including a thorough description of which fields can be filtered and what relations to other databases are supported. Figure 4 shows a snippet of the XML file related to the PubMed database, as extracted from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=pubmed.

```
<eInfoResult>
  <DbInfo>
    <DbName>pubmed</DbName>
    <MenuName>PubMed</MenuName>
    <Description>PubMed bibliographic record</Description>
    <Count>22595116</Count>
    <LastUpdate>2013/03/19 03:50</LastUpdate>
    <FieldList>
      <Field>
        <Name>ALL</Name>
        <FullName>All Fields</FullName>
        <Description>All terms from all searchable fields</Description>
        <TermCount>139835112</TermCount>
        <IsDate>N</IsDate>
        <IsNumerical>N</IsNumerical>
        <SingleToken>N</SingleToken>
        <Hierarchy>N</Hierarchy>
        <IsHidden>N</IsHidden>
      </Field>
      <Field>
      ...
</eInfoResult>
```

Figure 4: The XML file describing the structure of the PubMed database.

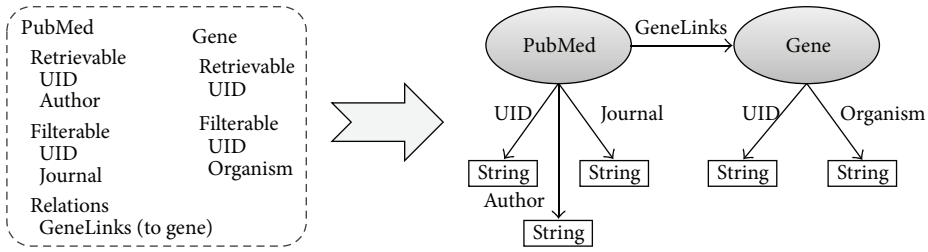


Figure 5: The databases PubMed and Gene in the NCBI repository set produce two related classes in the corresponding RDF model.

Although these files provide the necessary information about filterable fields and relations between related databases, the list of retrievable cross-linked fields is still missing. The NCBI database management system handles filterable and retrievable fields in a different way, meaning that there exist fields that can be filtered but not retrieved—for example "PDAT"—and vice versa— for example, "LANG."

The list of retrievable fields is created as follows. For the databases that support the fetch operation—such as PubMed—a related Document Type Definition (DTD) that specifies the retrievable fields can be obtained through a dedicated web service. For any other database which does not support the fetch operation, a valid result document must be manually analyzed for obtaining the final list of retrievable fields. These results are obtained by performing a simple test query.

Once the list of retrievable and filterable fields and the properties relating different databases has been generated, NCBI2RDF is able to automatically generate the RDF schema that will allow users building correct SPARQL queries. For each NCBI database, an RDF class named after the database is automatically created. Each filterable and/or retrievable field is translated into a *String* datatype in this class. Finally, the generated classes are linked by object properties according to the previously discovered relations between the NCBI databases. Figure 5 shows a piece of the RDF schema generated from part of the information of the PubMed and the Gene databases.

It must be noted that the asymmetry between retrievable fields and filterable fields is lost when constructing the RDF schema given the inability of this paradigm to model this situation. While our API is able to handle this difference, users will not be able to obtain this information from the RDF schema alone. For this reason, the RDF schema is accompanied with side documentation describing this situation. Users are expected to generate SPARQL queries compliant both with the RDF schema and with the side documentation that forbids that some specific fields are retrieved and other fields are filtered.

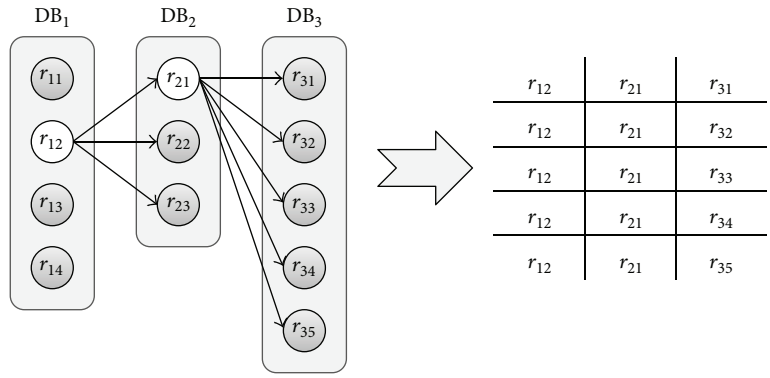


Figure 6: This diagram explains how the tree of results translates into actual results of the queries that the system receives. Each leaf of the tree produces one result, composed of the leaf itself and all the super nodes of that leaf node.

3.3. Query Resolution. The query resolution process is in charge of accepting SPARQL queries in terms of the previously generated RDF schema and translating them into an equivalent set of requests to the E-utilities services that effectively allow solving the user query. Obviously, SPARQL queries might contain joins between databases—expressed as relations between classes of the RDF schema. As it was described in the background section, NCBI does not allow this type of queries. To overcome this issue, we were forced to develop a solution that would allow us to raise the expressivity of the queries to the desired level, just by using the functions available at the E-utilities services. The solution relies on the simulation of this behavior by means of workflows of simple requests to NCBI that fulfill the original query.

The workflows of requests to NCBI are built in a dynamic fashion. The results of each request help determining the subsequent requests; therefore, it is not possible to assess the complete sequence of requests at once. In essence, each database is queried separately with its own parameters—retrieved variables and filters. Each result produced with this access allows fetching more results from a related database in the query. This process produces what we call "tree of results." The branches of this tree are built sequentially in a depth-first mode, as each level corresponds to one database in the query. Whenever a branch reaches the last level, results to the original query are gathered from all the different branches that reach from the root node to one leaf. When the results of a branch are exhausted, the algorithm moves backwards in the tree and explores another branch. Figure 6 depicts this process.

The retrieval of each database's results implies several requests to the E-utilities services. As it was described before, these web services provide means to query the NCBI data by external applications. However, complex queries involving more than one database are, as well as with the Entrez interface, not permitted. The operations offered by the E-utilities services are described in the following—only the operations relevant to our query execution process are shown.

(i) *eSearch* allows retrieving a list of entries of a single database matching some specific criteria. This is the equivalent of the SELECT operation in SQL. The actual result of this operation is not the entries themselves, but the entry identifiers. It is not possible to retrieve entries from more than one database in a single *eSearch* operation—that is, *joins* are not permitted.

(ii) *eFetch/eSummary* allows accessing the information of a single entry from a single database, once the entry identifier has been retrieved using an eSearch operation. The *eFetch* operation retrieves all the fields available for the entry, while the *eSummary* operation offers a subset of this information. Not all NCBI databases support the eFetch operation.

(iii) *eLink* retrieves the list of entry identifiers of a database that are related to a specific entry in another database. This operation resembles the join in SQL. Only one relation can be specified in an *eLink* operation.

Using these operations, a dynamic query resolution workflow is generated. The following algorithm describes this process.

(1) The first database in the user query is accessed according to the specified retrievable fields and filters, and its results are stored in the first level of the tree (eSearch + eFetch/eSummary). This is now the current level.

(2) If the current level does not correspond to the last database in the sequence, go to 3. Else, the paths from the root of the tree to all the generated leaves form results to the original query. Gather these results and move up one level.

(3) Go through the nodes of the current level and, for each node *i*, do:

  (a) Retrieve the entry identifiers from the next database in the sequence that are related to *i* (*eLink*). These identifiers are stored in a new level hanging from the node *i*.

  (b) Retrieve the required information for the new nodes (eFetch/eSummary).

  (c) Make the new level the current one and go to 2

```
PREFIX base: <http://RDFEutilsWrapper#>
SELECT ?pubmedUID ?geneUID ?pubmedUID2
WHERE {
    ?pubmed base:pubmed_UID ?pubmedUID.
    ?pubmed base:pubmed_TITL ?title.
    ?pubmed base:pubmed_gene ?gene.
    ?gene base:gene_UID ?geneUID.
    ?gene base:gene_pubmed ?pubmed2.
    ?pubmed2 base:pubmed_UID ?pubmedUID2.
    FILTER (?title = "wilms tumor").
}
```

Figure 7: The tested SPARQL query that retrieves articles from PubMed, then related records from Gene, and finally articles that talk about those records from PubMed again.

(4) Once all nodes were used, prune this branch. If there are levels above the current level, go up one level. Else, finish the data retrieval process.

Once the results are retrieved, they are automatically translated into SPARQL results format and given back to the user. The result returning is performed in a progressive manner, due to the time it requires to completely solve some queries—for example, in queries involving the gene database, there can be millions of results. With our API, clients must request results one by one. To do this, the API offers the typical "iteration" programming schema, with the *hasNext* and *next* methods for exploring the result set.

The constructed workflows are always designed to produce a minimum possible count of requests to the NCBI resources. This helps avoiding the performance penalties imposed for launching multiple requests in a short time lapse. In addition, all accessions to the NCBI databases by means of the E-utilities services are performed using the E-utilities own history service. This allows reducing overload on the NCBI data management system and optimizing performance of our API.

4. Experiments

This section describes how our system deals with an incoming query, providing details of what NCBI-compliant queries are performed to obtain the requested results. The data we intend to retrieve involves three different NCBI databases. We will retrieve papers indexed by PubMed which contain the string "Wilms tumor" in their title. For each of these results, we will retrieve related genes—to Wilms tumor—from the Gene database. Finally, for each of these genes, we will gather again the PubMed database searching for papers that mention that specific gene. Using the NCBI2RDF system, this query can be performed in a single step. The resulting query in SPARQL is depicted in Figure 7.

To process this query, NCBI2RDF automatically generates a dynamic workflow, which is executed as follows. First, a single *eSearch* query targeted at PubMed and including the filter *TITLE="wilms tumor"* is generated, producing 3026 unique results which are stored in the NCBI history server. Then, the first result of this set is retrieved with an *eFetch* query and stored locally—by parsing the XML file representing this result. A third *eLink* query is performed using the recently obtained UID and the *pubmed_gene* relation as arguments, which generates one single result. This result is again retrieved with an *eFetch* query, and a new *eLink* query is realized with the relation *gene_pubmed*. This query returns 557 results from the PubMed database, all of which are fetched simultaneously. The concatenation of the first UID from PubMed, the subsequent UID from Gene, and each of these UIDs composes the first set of retrieved results. To complete the query, the algorithm backtracks twice—since Gene only produces a single result—and selects the second result from the initial PubMed query. This process is repeated until all branches have been fully explored.

Processing a similar query with the NCBI native web interface would involve manually visiting hundreds of thousands of pages. For each of the 3026 initial results in PubMed, the user would have to visit each result individually, and after that, navigate the set of related results in the Gene database. In addition, each Gene result would have to be opened separately, together with its related PubMed records. In the sample query previously shown, our system performed all these steps automatically, producing an average of 50 results per second—note however that this value depends on the complexity of the query.

5. Discussion

The example presented in the previous section shows the advantages that our system presents when compared to the native NCBI interface. The sample query in Section 4 belongs to a test set that we created to validate the system, which included over 200 queries. By raising the expressiveness of the NCBI query processing system, our API enables users to launch complex queries while relieving them from navigating through the NCBI pages that display each result.

To our knowledge, only the Bio2RDF [23] system is targeted at offering RDF-based access to the NCBI repositories. This system was designed to provide data from multiple

Table 1: Comparison of features provided by Entrez, Bio2RDF, and NCBI2RDF for accessing the NCBI databases.

	Query language	Support for complex queries	Support for full NCBI database structure	Access to up-to-date data	Adaptability to changes in NCBI databases
Entrez	HTML form	No	Yes	Yes	Yes
Bio2RDF	HTML form	No	No	No	No
NCBI2RDF	SPARQL	Yes	Yes	Yes	Yes

biomedical databases in an RDF-compliant form—data can be retrieved in different formats, such as RDF, N3, or plain HTML. However, they base their approach on the manual analysis of HTML documents representing results of queries in order to map its contents to a prefixed RDF structure. This RDF structure was manually created (the Bio2RDF ontology) and only covers some general concepts which are common to all the covered databases. Therefore, Bio2RDF is only capable of providing a few fields for each database. Furthermore, the Bio2RDF interface does not support SPARQL queries and instead resorts to a simple HTML form that allows specifying either single results—through a result ID—or a general search term. This approach allows a quick adoption of new data sources—the system is prepared to be adapted for new relational, HTML, XML, or unstructured databases, but lacks the ability to cover the complexity of each integrated database. Conversely, our approach focuses on maintaining all the information contained in all the NCBI databases and providing SPARQL querying capability over these databases. Moreover, the automated metadata generation procedure permits our system to seamlessly adapt to future changes in the structure of the NCBI databases, or even cover new databases not currently included in the NCBI repositories. Bio2RDF, however, bases its approach on Java Server Pages manually codified for each database, which must be reencoded with each database structure modification. Table 1 compares the features provided by NCBI2RDF with those of Bio2RDF and the Entrez interface. Conversely, our approach focuses on automatically building a new RDF schema that reflects all databases, relations among these databases, variables, and filterable fields of the NCBI repositories. It is possible, in addition, to define a mapping of this schema to any existing domain ontology—for example, GO, FMA, and PRO—thus enabling the integration of the NCBI data in terms of these vocabularies. The NCBI2RDF approach solves all syntactic heterogeneities between NCBI and the rest of RDF-compliant biomedical data sources.

This research has been carried out in the context of a large-scale European-funded research project, p-Medicine [26]. This project is aimed at creating a technological infrastructure with data integration capabilities for advanced knowledge discovery in clinical trials in cancer. By integrating the genomic information stored in the NCBI databases within the RDF-enabled p-medicine data infrastructure—which already includes RDF-based access to other relevant sources such as ArrayExpress [27]—we hope to further enhance clinical and genomic information to foster the development of novel personalized drugs for cancer patients based on their genomic profile.

6. Conclusions

In this paper, we present NCBI2RDF, an API for providing SPARQL-based access to the NCBI databases. This is achieved by dynamically building native NCBI query workflows. Results from different databases are merged to service complex SPARQL queries involving multiple repositories. The API has been thoroughly tested with a wide range of queries, one of which was presented in Section 4. The presented system effectively provides RDF-based access to all the databases managed by the NCBI.

Our approach is based on two steps: metadata generation and query resolution. The metadata generation stage gathers information about the structure of the NCBI databases in order to build the subsequent query workflows. This stage is mostly automatic, although some human intervention is still required. However, metadata generation is seldom needed, as the structure of the NCBI databases does not undergo frequent changes. Conversely, the query resolution aims at dynamically constructing the query workflows that will effectively allow servicing SPARQL queries. These workflows emulate the manual work that a researcher would have to carry out in order to retrieve the information distributed across several databases—as explained in Section 4. The workflows are designed to minimize the interaction with the NCBI querying system, in order to save resources.

Our system has two advantages compared to other existing systems—including the NCBI query services themselves. First, the query expressiveness level is raised, since multiple databases can be specified in a single query, saving both time and resources for researchers who wish to perform complex queries against the NCBI system. Second, it enables the semantic integration of the data hosted by the NCBI with other RDF-compliant biomedical resources.

Current biomedical research is highly dependent on the ability of researchers to uniformly access different data sources—both private and public. However, this capability is mainly hampered by heterogeneities in the data structure, formats, and interfaces. By providing RDF-compliance to the NCBI databases, these heterogeneities are automatically solved, enabling the integrated access of these data with other existing RDF-based repositories.

7. Availability

The software can be freely downloaded as a Java library. Detailed instructions of use are included, as well as complete Javadocs. The project homepage is located at http://www.bio-informatics.org/ncbi2rdf/.

Acknowledgments

This work was funded by the European Commission, through the p-medicine project (FP7-ICT-2009-270089), and INTEGRATE (FP7-ICT-2009-270253). The authors would also like to thank Esther Peinado for developing the software that implements the algorithm and for providing useful suggestions.

References

[1] F. L. Kiechle, X. Zhang, and C. A. Holland-Staley, "The -omics era and its impact," *Archives of Pathology and Laboratory Medicine*, vol. 128, no. 12, pp. 1337–1345, 2004.

[2] A. R. Joyce and B. Ø. Palsson, "The model organism as a system: integrating "omics" data sets," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 3, pp. 198–210, 2006.

[3] V. B. Bajic, V. Brusic, J. Li, S. K. Ng, and L. Wong, "From informatics to bioinformatics," in *Proceedings of the 1st Asia-Pacific Bioinformatics Conference on Bioinformatics*, vol. 19, pp. 3–12, 2003.

[4] "Making data dreams come true," *Nature*, vol. 428, no. 6980, article 239, 2004.

[5] J. A. Sagotsky, L. Zhang, Z. Wang, S. Martin, and T. S. Deisboeck, "Life sciences and the web: a new era for collaboration," *Molecular Systems Biology*, vol. 4, article 201, 2008.

[6] L. Martin, A. Anguita, V. Maojo, and J. Crespo, "Integration of omics data for cancer research," in *An Omics Perspective on Cancer Research*, W. C. S. Cho, Ed., pp. 249–266, Springer, Dordrecht, Netherlands, 2010.

[7] K. P. H. Pritzker and L. B. Pritzker, "Bioinformatics advances for clinical biomarker development," *Expert Opinion on Medical Diagnostics*, vol. 6, no. 1, pp. 39–48, 2012.

[8] V. Maojo and M. Tsiknakis, "Biomedical informatics and HealthGRIDs: a European perspective—past and current efforts and projects in the synergy of bionformatics and medical informatics," *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 3, pp. 34–41, 2007.

[9] V. Petrik, A. Loosemore, F. A. Howe, B. A. Bell, and M. C. Papadopoulos, "OMICS and brain tumour biomarkers," *British Journal of Neurosurgery*, vol. 20, no. 5, pp. 275–280, 2006.

[10] B. Zirn, O. Hartmann, B. Samans et al., "Expression profiling of Wilms tumors reveals new candidate genes for different clinical parameters," *International Journal of Cancer*, vol. 118, no. 8, pp. 1954–1962, 2006.

[11] C. T. Ferrara, P. Wang, E. C. Neto et al., "Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling," *PLoS Genetics*, vol. 4, no. 3, Article ID e1000034, 2008.

[12] S. C. Connor, M. K. Hansen, A. Corner, R. F. Smith, and T. E. Ryan, "Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes," *Molecular BioSystems*, vol. 6, no. 5, pp. 909–921, 2010.

[13] T. Elkan-Miller, I. Ulitsky, R. Hertzano et al., "Integration of transcriptomics, proteomics, and microRNA analyses reveals novel microrna regulation of targets in the mammalian inner ear," *PLoS ONE*, vol. 6, no. 4, Article ID e18195, 2011.

[14] J. M. Yi, M. Dhir, L. van Neste et al., "Genomic and epigenomic integration identifies a prognostic signature in colon cancer," *Clinical Cancer Research*, vol. 17, no. 6, pp. 1535–1545, 2011.

[15] A. Anguita, L. Martín, D. Pérez-Rey, and V. Maojo, "A review of methods and tools for database integration in biomedicine," *Current Bioinformatics*, vol. 5, no. 4, pp. 253–269, 2010.

[16] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International Journal of Human*, vol. 43, no. 5-6, pp. 907–928, 1995.

[17] G. O. Consortium, "The Gene Ontology (GO) project in 2006," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D322–D326, 2006.

[18] C. Rosse and J. L. V. Mejino Jr., "A reference ontology for biomedical informatics: the Foundational Model of Anatomy," *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 478–500, 2003.

[19] M. Brochhausen, A. D. Spear, C. Cocos et al., "The ACGT Master Ontology and its applications—towards an ontology-driven cancer research and management system," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 8–25, 2011.

[20] D. A. Natale, C. N. Arighi, W. C. Barker et al., "The Protein Ontology: a structured representation of protein forms and complexes," *Nucleic Acids Research*, vol. 39, no. 1, pp. D539–D545, 2011.

[21] B. Smith, M. Ashburner, C. Rosse et al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology*, vol. 25, no. 11, pp. 1251–1255, 2007.

[22] N. F. Noy, N. H. Shah, P. L. Whetzel et al., "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, vol. 37, no. 2, pp. W170–W173, 2009.

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

[24] B. Mons and J. Velterop, "Nano-Publication in the e-science era," SURF Fundation, 2009.

[25] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, "Entrez: molecular biology database and retrieval system," *Methods in Enzymology*, vol. 266, pp. 141–161, 1996.

[26] S. Rossi, M. L. Christ-Neumann, S. Rüping et al., "p-Medicine: from data sharing and integration via VPH models to personalized medicine," *Ecancermedicalscience*, vol. 5, no. 1, article 218, 2011.

[27] A. Anguita, L. Martin, M. Garcia-Remesal, and V. Maojo, "RDF-Builder: a tool to automatically build RDF-based interfaces for MAGE-OM microarray data sources," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 220–227, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 901578, 13 pages
http://dx.doi.org/10.1155/2013/901578

Review Article

Translational Bioinformatics for Diagnostic and Prognostic Prediction of Prostate Cancer in the Next-Generation Sequencing Era

Jiajia Chen,[1,2] Daqing Zhang,[1] Wenying Yan,[1] Dongrong Yang,[3] and Bairong Shen[1]

[1] *Center for Systems Biology, Soochow University, Suzhou 215006, China*
[2] *School of Chemistry, Biology and Material Engineering, Suzhou University of Science and Technology, Suzhou 215011, China*
[3] *Department of Urology, The Second Affiliated Hospital of Soochow University, Suzhou 215004, China*

Correspondence should be addressed to Bairong Shen; bairong.shen@suda.edu.cn

Received 1 May 2013; Accepted 22 June 2013

Academic Editor: Xinghua Lu

The discovery of prostate cancer biomarkers has been boosted by the advent of next-generation sequencing (NGS) technologies. Nevertheless, many challenges still exist in exploiting the flood of sequence data and translating them into routine diagnostics and prognosis of prostate cancer. Here we review the recent developments in prostate cancer biomarkers by high throughput sequencing technologies. We highlight some fundamental issues of translational bioinformatics and the potential use of cloud computing in NGS data processing for the improvement of prostate cancer treatment.

1. Introduction

Prostate cancer (PCa) is the most common cancer and the second leading cause of cancer deaths among males in western societies [1]. It is estimated that 241740 new PCa cases were diagnosed and that 28170 men died from it in the United States in 2012. Since its discovery over 20 years ago, Prostate Specific Antigen (PSA) has been the mainstay for diagnosis and prognosis of prostate cancer. However, the routine use of PSA screening remains controversial, owing to its limited specificity. PSA fails to differentiate PCa from common prostate disorders; moreover, it cannot discriminate between aggressive tumors and low-risk ones that may otherwise never have been diagnosed without screening [2]. As such, overdetection and overtreatment represent critical consequences of PSA-based screening [3]. The ongoing debate highlights the need for more sensitive and specific tools to enable more accurate diagnosis and prognosis.

During the last decade, the ability to interrogate prostate cancer genomes has rapidly advanced. The resolution for genomic mutation discovery was improved first with array-based methods and now with next-generation sequencing (NGS) technologies. These high throughput technologies open up the possibility to individualize the diagnosis and treatment of cancer. However significant challenges, particularly with respect to integration, storage, and computation of large-scale sequencing data, will have to be overcome to translate NGS achievements into the bedside of the cancer patient. Translational informatics evolves as a promising methodology that can provide a foundation for crossing such "translational barriers" [4, 5].

Here we overview the NGS-based strategies in prostate cancer research, with focus on upcoming biomarker candidates that show promise for the diagnosis and prognosis of prostate cancer. We also outline future perspectives for translational informatics and cloud computation to improve prostate cancer management.

2. Microarray Based Diagnosis and Prognosis of PCa

In the past two decades, high-throughput microarray profiling has been utilized to track complex molecular aberrations during PCa carcinogenesis. We performed a comprehensive search in the Gene Expression Omnibus (GEO) for the

Table 1: Number of PCa-associated GEO series generated by microarray and NGS.

Methodology	Gene expression profiling	Noncoding RNA profiling	Genome binding/occupancy profiling	Genome methylation profiling	Genome variation profiling
Microarray	266	34	17	21	35
NGS	11	1	18	2	2

array-based profiles in human PCa. The retrieved GEO series generally fall into 5 categories: gene expression profiling, noncoding RNA profiling, genome binding/occupancy profiling, genome methylation profiling, and genome variation profiling. The number of GEO series for each category is summarized in Table 1.

Together these array-based technologies have shed light on the genetic alterations in the PCa genome. Among the abnormalities affecting prostate tumors, the copy number alteration is the most common one [6].

Numerous early studies have used comparative genome hybridization (CGH) or single nucleotide polymorphism (SNP) arrays to assess copy number changes in tumor DNA. As a result, multiple genomic regions that displayed frequent gain or loss in the PCa genome [6–11] have been revealed. Chromosome 3p14, 8p22, 10q23, 13q13, and 13q14 are found to display broad copy number deletion. Key genes mapping within these deleted regions include NKX3.1, PTEN [12], BRCA2, C13ORF15, SIAH3 [11], RB1, HSD17B2 [9], FOXP1, RYBP, and SHQ1 [6]. High-level copy number gains are detected at 5p13, 14q21, 7q22, Xq12, and 8q13 [9]. Key amplified genes mapping within these regions include SKP2, FOXA1, AR [11], and HSD17B3 [9].

Using microarray, substantial efforts have also been made to characterize prostate cancer gene expression profiles. Differentially expressed genes identified in these studies point to a plethora of candidate biomarkers with diagnostic or prognostic value.

A diagnostic marker is able to differentiate prostate cancer with other prostatic abnormalities. There are many emerging markers that show promise for PCa diagnosis, such as alpha-methylacyl-CoA racemase (AMACR) [13], prostate cancer gene 3 (PCA3) [14], early prostate cancer antigen (EPCA)-2 [15], Hepsin [16], kallikrein-related peptidase 2 (KLK2) [17], and polycomb group protein enhancer of zeste homolog 2 (EZH2) [18]. The most prominent of these is PCA3, which was found to exhibit higher sensitivity and specificity for PCa detection than PSA. PCA3 thus provides a potential complement to PSA for the early diagnosis of PCa.

Prognostic biomarkers for prediction of prostate cancer patient outcome have also been identified. Increased serum levels of IL-6 and its receptor (IL-6R) are associated with metastatic and hormone refractory disease [19]. Elevated serum chromogranin A levels are indicative of poor prognosis and decreased survival [20]. Other differentially expressed molecules with prognostic potential include the urokinase plasminogen activation (uPA) [21], TGF-β1 [22, 23], MUC1 [24], CD24 [25], hCAP-D3 [26], vesicular monoamine transporter 2 (SLC18A2) [27], TEA domain family member 1 (TEAD1), c-Cbl [28], SOX7 and SOX9 [29], nuclear receptor binding protein 1 (NRBP1) [30], CD147 [31], and Wnt5a [32]. Each of these markers will require proper validation to ensure their clinical utility.

While microarray technology represents a wonderful opportunity for the detection of genomic alterations, there are significant issues that must be considered. For example, array-based methods are impossible to detect variations at a low frequency (many well below 1%) in the samples. In addition, microarrays can only provide information about the genes that are already included on the array. The emerging next-generation sequencing technology, also called massively parallel sequencing, however, helps to overcome the challenges by generating actual sequence reads [33].

3. NGS Based Diagnosis and Prognosis of PCa

Key feature of the next-generation sequencing technology is the massive parallelization of the sequencing process. By virtue of the massively parallel process, NGS generates hundreds of millions of short DNA reads (100–250 nucleotides) simultaneously, which are then assembled and aligned to reference genomes.

A number of NGS systems are available commercially, including Genome Analyzer/HiSeq 2000/MiSeq from Illumina, SOLiD/PGM/Proton from Life Sciences, GS-FLX (454)/GS Junior from Roche, as well as novel single molecule sequencers, for example, Heliscope from Helicos Biosciences and SMRT offered by Pacific Biosciences. All these technologies provide digital information on DNA sequences and make it feasible to discover genetic mutations at unprecedented resolution and lower cost.

It is generally accepted that cancers are caused by the accumulation of genomic alterations. NGS methods can well squeeze all the alteration information that remains hidden within the genome, including point mutations, small insertions and deletions (InDels), copy number alterations (CNV), chromosomal rearrangements, and epigenetic alterations. For these reasons NGS has become the method of choice for large-scale detection of somatic cancer genome alterations and is changing the way how cancer genome is analyzed. An NGS-based research pipeline for PCa biomarkers is given in Figure 1.

Currently, next-generation sequencing is being applied to cancer genome study in various ways:

(1) genome-based sequencing (DNA-Seq), yielding information on sequence variation, InDels, chromosomal rearrangements, and copy number variations,

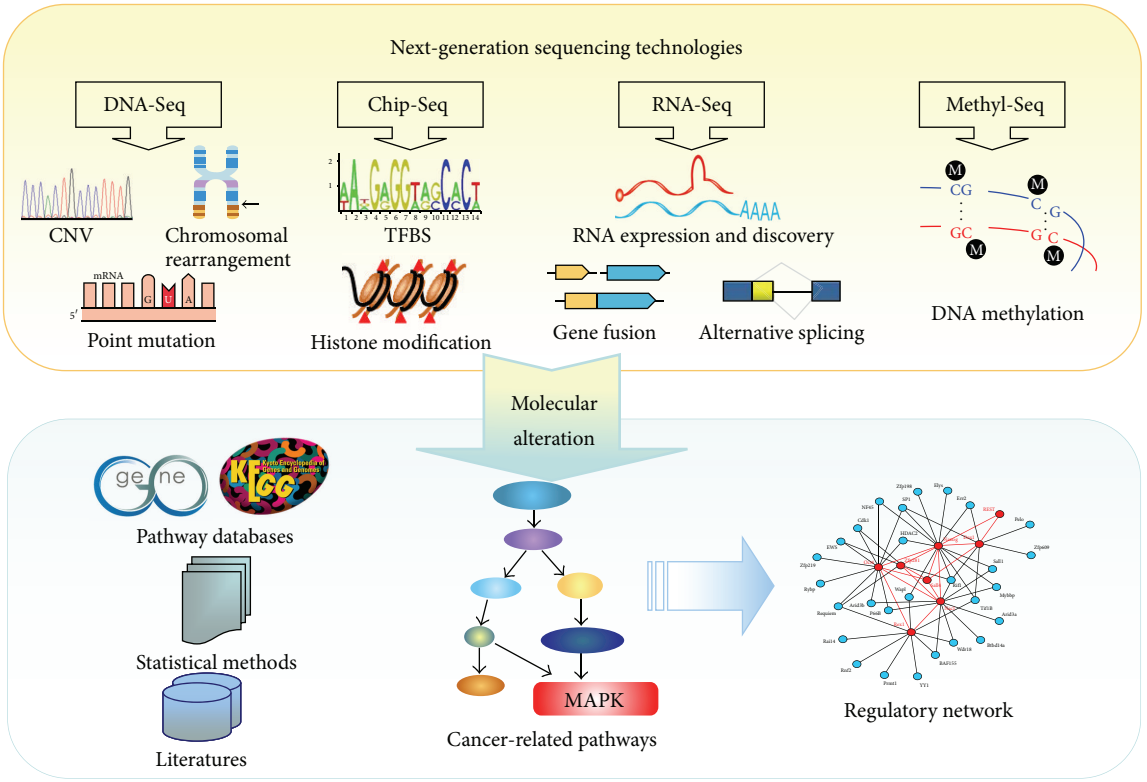


Figure 1: NGS-based pipeline for cancer marker discovery.

(2) transcriptome-based sequencing (RNA-Seq), yielding quantitative information on transcribed regions (total RNA, mRNA, or noncoding RNAs),

(3) interactome-based sequencing (ChIP-Seq), yielding information on protein binding sequences and histone modification,

(4) methylome-based sequencing (Methl-seq), yielding quantitative information on DNA methylation and chromatin conformation.

In the following part, we will introduce the main applications to next-generation sequencing of prostate cancer, using examples from the recent scientific literature (summarized in Table 2).

3.1. DNA-Seq. According to the proportion of the genome targeted, DNA-Seq is categorized into whole-genome sequencing and exome sequencing. The goal of whole-genome sequencing is to sequence the entire genome, not just coding genes, at a single-base resolution. By whole-genome sequencing, recent studies have provided detailed landscape of genomic alterations in localized prostate cancers [6, 11, 46, 68, 69]. The full range of genomic alterations that drive prostate cancer development and progression, including copy number gains and losses, single nucleotide substitutions, and chromosomal rearrangements, are readily identified.

3.1.1. Copy Number Alteration. Most prostate cancers exhibit somatic copy number alterations, with genomic deletions outnumbering amplifications [6]. Early methods for copy number analysis involve fluorescence in situ hybridizations and array-based methods (CGH arrays and SNP arrays). More recently, NGS technologies have been utilized and offer substantial benefits for copy number analysis.

NGS used changes in sequencing depth (relative to a normal control) to identify copy number changes. The digital nature of NGS therefore allows accurate estimation of copy number levels at higher resolution. In addition, NGS can provide novel gene copy information such as homozygous and heterozygous deletions and gene amplifications, whereas traditional sequencing approaches cannot. For example, by next-generation sequencing of castrate-resistant prostate cancer (CRPC), Collins et al. [34] identified a homozygous 9p21 deletion spanning the MTAP, CDKN2, and ARF genes and deficiency of MTAP was suggested as an exploitable tumor target.

3.1.2. Somatic Nucleotide Substitutions. While whole-genome sequencing provides the most comprehensive characterization of the cancer genome, it is the most costly. Alternatively, targeted sequencing approaches, such as exome sequencing, assemble multiple cancer genomes for analysis in a cost-effective manner. Whole-exome sequencing captures the coding exons of genes that contain the vast majority of disease causing mutations. Relative to structural alterations, point mutations are less common in prostate cancer [6, 70] and the average mutation rate was estimated at 1.4 Mb^{-1} in localized PCa [35] and 2.0 Mb^{-1} in CRPC [36].

Table 2: Summary of NGS-based studies on prostate cancer.

Discoveries	Method	References
Copy number loss of MTAP, CDKN2, and ARF genes	DNA-Seq	[34]
Somatic mutations in MTOR, BRCA2, ARHGEF12, and CHD5 genes		[11]
NCOA2, p300, the AR corepressor NRIP1/RIP140, and NCOR2/SMRT		[6]
Somatic mutations in SPOP, FOXA1, and MED12		[35]
Somatic mutations in MLL2 and FOXA1		[36]
Somatic mutations in TP53, DLK2, GPC6, and SDF4.		[37]
TMPRSS2:ERG, TMPRSS2:ETV1	RNA-Seq	[38]
TMPRSS2:ETV4		[39]
TMPRSS2:ETV5, SLC45A3:ETV5		[40]
TMPRSS2:ELK4		[41]
SLC45A3:ETV1, HERV-K_22q11.23:ETV1, HNRPA2B1:ETV1, and C15ORF21:ETV1		[42]
KLK2:ETV4 and CANT1:ETV4		[43]
SLC45A3:BRAF or ESRP1:RAF1		[44]
C15orf21:Myc		[45]
EPB41:BRAF		[46]
TMEM79:SMG5		[47]
Differential expression of PCAT-1		[48]
Differential expression of miR-16, miR-34a, miR-126*, miR-145, and miR-205		[49]
HDACs and EZH2 work as ERG corepressors	Chip-Seq	[50]
AP4 as a novel co-TF of AR		[51]
POU2F1 and NKX3-1		[52]
Runx2a regulates secretion invasiveness and membrane secretion		[53]
A novel transcriptional regulatory network between NKX3-1, AR, and the RAB GTPase signaling pathway		[54]
Distinct patterns of promoter methylation around transcription start sites	Methyl-Seq	[55]

Capillary-based exome sequencing has extensively been performed in localized PCa and CRPC, and a handful of oncogenic point mutations have been defined. Remarkably, Taylor et al. [6] performed focused exon resequencing in 218 prostate cancer tumors and identified multiple somatic alterations in the androgen receptor (AR) gene as well as its upstream regulators and downstream targets. For example, the AR coactivator NCOA2 and p300, the AR corepressor NRIP1/RIP140 and NCOR2/SMRT were found to harbor somatic mutations. Other genes including KLF6, TP53, AR, EPHB2, CHEK2, and ATBF1 [6, 71–74] have also been reported to harbor somatic mutations in localized prostate cancer.

Recently NGS is becoming increasingly routine for exome sequencing analysis. Whole-exome sequencing using next-generation sequencing (NGS) technologies compares all exon sequences between tumors and matched normal samples. Multiple reads that show nonreference sequence are detected as point mutations. In this way a number of driver mutations in prostate cancer have been uncovered. Robbins et al. [11] used NGS-based exome sequencing in 8 metastatic prostate tumors and revealed novel somatic point mutations in genes including MTOR, BRCA2, ARHGEF12, and CHD5. Kumar et al. [37] performed whole-exome sequencing of lethal metastatic tumors and high-grade primary carcinomas. They also observed somatic mutations in TP53, DLK2, GPC6, and SDF4. More recently Barbieri et al. [35] and Grasso et al. [36] systematically analyzed somatic mutations in large cohorts of prostate tumors. Barbieri et al. [35] investigated 112 primary tumor-normal pairs and revealed novel recurrent mutations in SPOP, FOXA1, and MED12. Grasso et al. [36] sequenced the exomes of 11 treatment-naive and 50 lethal CRPC and identified recurrent mutations in multiple chromatin- and histone-modifying genes, including MLL2 and FOXA1. These two studies also reported mutated genes (SPOP [35] and CHD1 [36]) that may define prostate cancer subtypes which are ETS gene family fusion negative.

Together these findings present a comprehensive list of specific genes that might be involved in prostate cancer and prioritize candidates for future study.

3.2. RNA-Seq. In addition to genome applications, NGS will also dramatically enhance our ability to analyze transcriptomes. Before NGS, microarrays have been the dominant technology for transcriptome analysis. Microarray technologies rely on sequence-specific probe hybridization, and fluorescence detection to measure gene expression levels. It is subject to high noise levels, cross-hybridization and limited dynamic range. Compared to microarrays, the emerging RNA-Seq provides digital gene expression measurements that offer significant advantages in resolution, dynamic range, and reproducibility. The goal of transcriptome sequencing is to sequence all transcribed genes, including both coding and noncoding RNAs. It is independent of prior knowledge and offers capacity to identify novel transcripts and mutations

that microarrays could not achieve, such as fusion genes, noncoding RNAs, and splice variants.

3.2.1. Gene Fusions. Recurrent gene fusion is a prevalent type of mutation resulting from the chromosomal rearrangements, which can generate novel functional transcripts that serve as therapeutic targets. Early studies relied on cytogenetic methods to detect chromosomal rearrangements. However this method is only applicable in cases of simple genomes and is vulnerable in complex genomes of epithelial cancers such as PCa.

Complete sequencing of prostate cancer genomes has provided further insight into chromosomal rearrangements in prostate cancer. NGS technologies, for example, paired-end sequencing approaches are sufficiently sensitive to detect break point crossing reads and are extremely powerful for the discovery of fusion transcripts and potential break points.

The first major recurrent fusion to be identified in prostate cancer was discovered by Tomlins et al. using Cancer Outlier Profile Analysis (COPA) algorithm [38]. The fusion discovered places two oncogenic transcription factors from the ETS family (ETV1 and ERG) under control of the prostate-specific gene TMPRSS2.

While the TMPRSS2:ETV1 fusion is rare and occurs in 1–10% of prostate cancers [75], the TMPRSS2:ERG fusion is present in roughly half of prostate cancers and is the most common genetic aberration so far described in prostate cancer. Furthermore TMPRSS2:ERG is unique only to prostatic nonbenign cancers [76]. Given this high specificity, its clinical application as ancillary diagnostic test or prognostic biomarker is promising. The expression of TMPRSS2:ERG fusion gene has been proposed as a diagnostic tool, alone or in combination with PCA3 [77]. In addition, many studies have suggested that TMRSS2:ERG could be a prognostic biomarker for aggressive prostate cancer.

Following Tomlins' pioneering discovery, subsequent research has identified a host of similar ETS family gene fusions. Other oncogenic ETS transcription factors, for example, ETV4 [39], ETV5 [40], and ELK4 [41], have been identified as additional fusion partners for TMPRSS2. Other unique $5'$ fusion partner genes to ETS family members have also been identified, such as SLC45A3, HERV-K_22q11.23, HNRPA2B1, and C15ORF21 in fusion with ETV1 [42], KLK2 and CANT1 in fusion with ETV4 [43], and SLC45A3:ETV5 [40].

For ETS fusion-negative prostate cancers subtypes, novel gene fusions have also been identified, including SLC45A3:BRAF, ESRP1:RAF1, SLC45A3:BRAF, ESRP1:RAF1 [44], C15orf21:Myc [45], EPB41:BRAF [46], and TMEM79:SMG5 [47].

3.2.2. Noncoding RNAs. In addition to gene fusions, RNA-Seq also enables discovery of new noncoding RNAs (ncRNAs) with the potential to serve as cancer markers. Transcriptome sequencing of a prostate cancer cohort has identified an unannotated ncRNA PCAT-1 as a transcriptional repressor linked to PCa progression [48]. RNA- Seq was also applied to identify differentially expressed microRNAs (e.g., miR-16, miR-34a, miR-126*, miR-145, and miR-205) associated with metastatic prostate cancer [49]. These findings establish the utility of RNA-Seq to identify disease-associated ncRNAs that could provide potential biomarkers or therapeutic targets.

3.3. ChIP-Seq. Another application of NGS capitalizes on the ability to analyze protein-DNA interactions, as for ChIP-Seq. ChIP-Seq provides clear indications of transcription factor binding sites (TFBSs) at high resolution. It is also well suited for detecting patterns of modified histones in a genome-wide manner [78].

Much of gene regulation occurs at the level of transcriptional control. In addition, aberrant histone modifications (methylation or acetylation) are also associated with cancer. Therefore experimental identification of TFBSs or histone modifications has been an area of high interest. In traditional ChIP-chip approaches, DNA associated with a transcription factor or histone modification of interest is first selectively enriched by chromatin immunoprecipitation, followed by probing on DNA microarrays. In contrast to ChIP-chip, ChIP-Seq uses NGS instead of custom-designed arrays to identify precipitated DNA fragments, thus yielding more unbiased and sensitive information about target regions.

Many efforts have employed ChIP-Seq approaches to characterize transcriptional occupancy of AR, and many novel translational partners of AR have been identified. Using ChIP-Seq, Chng et al. [50] performed global analysis of AR and ERG binding sites. They revealed that ERG promotes prostate cancer progression by working together with transcriptional corepressors including HDACs and EZH2. Zhang et al. [51] developed a comotif scanning program called CENTDIST and applied it on an AR ChIP-Seq dataset from a prostate cancer cell line. They correctly predicted all known co-TFs of AR as well as discovered AP4 as a novel AR co-TF. Little et al. [53] used genome-wide ChIP-Seq to study Runx2 occupancy in prostate cancer cells. They suggested novel role of Runx2a in regulating secretion invasiveness and membrane secretion. Tan et al. [54] showed that NKX3-1 colocalizes with AR and proposed a critical transcriptional regulatory network between NKX3-1, AR, and the RAB GTPase signaling pathway in prostate cancer. Urbanucci et al. [79] identified AR-binding sites and demonstrated that the overexpression of AR enhances the receptor binding to chromatin in CRPC. By comparing nucleosome occupancy maps using nucleosome-resolution H3K4me2 ChIP-Seq, He et al. [52] found that nucleosome occupancy changes can predict transcription factor cistromes. This approach also correctly predicted the binding of two factors, POU2F1 and NKX3-1. By high-resolution mapping of intra- and interchromosome interactions, Rickman et al. [80] demonstrated that ERG binding is enriched in hotspots of differential chromatin interaction. Their result indicated that ERG overexpression is capable of inducing changes in chromatin structures.

Taken together, these studies have provided a complete regulatory landscape in prostate cancer.

3.4. Methyl-Seq. Another fertile area for NGS involves assessment of the genome-wide methylation status of DNA. Methylation of cytosine residues in DNA is known to silence parts of the genome by inducing chromatin condensation.

DNA hypermethylation probably remains most stable and abundant epigenetic marker.

Several DNA methylation markers have been identified in prostate cancer. The most extensively studied one is CpG island hypermethylation of glutathione-S-transferase P (GSTP1) promoter DNA, resulting in the loss of GSTP1 expression [81]. Today GSTP1 hypermethylation is most frequently evaluated as diagnostic biomarker for prostate cancer. It is also an adverse prognostic marker that predicts relapse of patients following radical prostatectomy [82].

With the aid of NGS technologies, genome-wide mapping of methylated cytosine patterns in cancer cells become feasible. Based on established epigenetics methods, there are emerging next-generation sequencing applications for the interrogation of methylation patterns, including methylation-dependent immunoprecipitation sequencing (MeDIP-Seq) [83], cytosine methylome sequencing (MethylC-Seq) [84], reduced representation bisulfite sequencing (RRBS-Seq) [85], methyl-binding protein sequencing (MBP-Seq) [86], and methylation sequencing (Methyl-Seq) [87]. A number of aberrant methylation profiles have been developed so far and are being evaluated as potential markers for early diagnosis and risk assessment. As an example, using MethylPlex-Seq, Kim et al. [55] mapped the global DNA methylation patterns in prostate tissues and revealed distinct patterns of promoter methylation around transcription start sites. The comprehensive methylome map will further our understanding of epigenetic regulation in prostate cancer progression.

4. PCa Biomarkers in Combinations

Diagnostic and prognostic markers with relevance to PCa are routinely identified. Nevertheless, we should stay aware that candidate markers obtained are often irreproducible from experiment to experiment and very few molecules will make it to the routine clinical practice. Cancer is a nonlinear dynamic system that involves the interaction of many biological components and is not driven by individual causative mutations. In most cases, no single biomarker is likely to dictate diagnosis or prognosis success. Consequently, the future of cancer diagnosis and prognosis might rely on the combination of a panel of markers. Kattan and associates [88] have established a prognostic model that incorporates serum TGF-β1 and IL-6R for prediction of recurrence following radical prostatectomy. This combination shows increased predictive accuracy from 75 to 84%. Furthermore, a predictive model incorporating GSTP1, retinoic acid receptor β2 (RAR β2), and adenomatous polyposis coli (APC) has been assessed. But no increased diagnostic accuracy was shown compared with PSA level alone [89]. More recently a panel of markers, PCA3, serum PSA level, and % free PSA show improved predictive value compared with PSA level and % free PSA. Again the clinical utility of these combinations needs to be evaluated in large-scale studies.

5. PCa Biomarkers at Pathway Level

Recently, an importance of pathway analysis has been emphasized in the study of cancer biomarkers. Pathway-based approach allows biologists to detect modest expression changes of functionally important genes that would be missed in expression-alone analysis. In addition, this approach enables the incorporation of previously acquired biological knowledge and makes a more biology-driven analysis of genomics data. Pathway analysis typically correlates a given set of molecular changes (e.g., differential expression, mutation, and copy number variation data) by projecting them onto well-characterized biological pathways. A number of curated databases are available for canonical signaling and metabolic pathways, such as Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/), Molecular Signatures Database (MSigDB), IngenuityPathway Analysis (IPA, http://www.ingenuity.com/), GeneGO by MetaCore (http://www.genego.com/), and Gene Set Enrichment Analysis (GSEA, http://www.broadinstitute.org/gsea/). Enrichment of pathways can be evaluated by overrepresentation statistics. The overall flowchart of the proposed pathway-based biomarker approach is illustrated in Figure 1. Using this pipeline, the pathways enriched with aberrations are identified and then proposed to be the potential candidate markers.

Some impressive progress has been made to identify pathways with relevance to the pathophysiology of prostate cancer. Rhodes et al. [90] were of the first to perform pathway analysis of the microarray expression datasets. By meta-analysis of 4 independent microarray datasets, they generated a cohort of genes that were commonly deregulated in PCa. The authors then mapped the identified deregulated genes to functional annotations and pinpointed polyamine and purine biosynthesis as critical pathway altered in PCa. Activation of Wnt signaling pathway was reported to be key pathways defining the poor PCa outcome group [91]. A comparison of castration-resistant and castration-sensitive pairs of tumor lines highlighted the Wnt pathway as potentially contributing to castration resistance [37]. Using a similar pathway-based approach, Wang et al. found Endothelin-1/EDNRA transactivation of the EGFR a putative novel PCa related pathway [92]. More recently, an integrative analysis of genomic changes revealed the role of the PI3K, RAS/RAF, and AR pathways in metastatic prostate cancers [6]. The above insights provide a blueprint for the design of novel pathway inhibitors in targeted therapies for prostate cancer.

Thus far the pathway-based approach holds great promise for cancer prediction. However, the known pathways correspond merely to a small fraction of somatic alterations. The alterations that have not been assigned to a definitive pathway undermine the basis for a strictly pathway-centric marker discovery. In addition, the cross-talk between different signaling pathways further complicated the pathway-based analysis. Thus, biomarker discovery has to shift toward an integrative network-based approach that accounts more extensive genomic alteration.

6. PCa-Specific Databases

High throughput research in PCa has led to vast amounts of comprehensive datasets. There has been a growing desire to integrate specific data types into a centralized database and make them publicly available. Considerable efforts were undertaken and to date various national, multicenter, and

institutional databases in the context of prostate cancer research are available. Prostate gene database (PGDB, http://www.urogene.org/pgdb/) is a curated database on genes or genomic loci related to human prostate and prostatic diseases [93]. Another database, prostate expression database (PEDB, http://www.pedb.org/) is a curated database that contains tools for analyzing prostate gene expression in both cancerous and normal conditions [94]. Dragon database of genes associated with prostate cancer (DDPC, http://cbrc.kaust.edu.sa/ddpc/) [95] is an integrated knowledge database that provides a multitude of information related to PCa and PCa-related genes. ChromSorter [96] collects PCa chromosomal regions associated with human prostate cancer. PCaMDB is a genotype-phenotype database that collects prostate cancer related gene and protein variants from published literatures. These specific databases tend to include large numbers of patients from different geographic regions. Their generalizability and statistical power offer researchers a unique opportunity to conduct prostate cancer research in various areas.

7. Translational Bioinformatics in PCa: A Future Direction

Recent advances in NGS technologies have resulted in huge sequence datasets. This poses a tremendous challenge for the emerging field of translational bioinformatics. Translational bioinformatics, by definition, is the development of storage, analytic, and interpretive methods to optimize the translation from bench (laboratory-based genomic discoveries) to bedside (evidence-based clinical applications). The aim of translational bioinformatics is to combine the innovations and resources across the entire spectrum of translational medicine towards the betterment of human health. To achieve this goal, the fundamental aspects of bioinformatics (e.g., bioinformatics, imaging informatics, clinical informatics, and public health informatics) need to be integrated (Figure 2).

As the first aspect, bioinformatics is concerned with applying computational approaches to comprehend the intricate biological details elucidating molecular and cellular processes of cancer. Imaging informatics is focused on what happens at the level of tissues and organs, and informatics techniques are used for image interpretation. Clinical bioinformatics focuses on data from individual patients. It is oriented to provide the technical infrastructure to understand clinical risk factors and differential response to treatment at the individual levels. As for public health informatics, the stratified population of patients is at the center of interest. Informatics solutions are required to study shared genetic, environmental, and life style risk factors on a population level. In order to achieve the technical and semantic interoperability of multidimensional data, fundamental issues in information exchange and repository are to be addressed.

8. Future Perspectives

The prospects for NGS-based biomarkers are excellent. However, compared with array-based studies, real in-depth NGS is still costly and also the analysis pipeline is less established, thus challenging the use of NGS. A survey in the GEO indicated that NGS is currently finding modest application in the identification of PCa markers. Table 1 listed the number of published GEO series on human PCa using NGS versus microarrays. Therefore, further work is still required before NGS can be routinely used in the clinic. Issues regarding data management, data integration, and biological variation will have to be tackled.

8.1. NGS in the Cloud. The dramatic increase in sequencer output has outpaced the improvements in computational infrastructure necessary to process the huge volumes of data. Fortunately, an alternative computing architecture, cloud computing, has recently emerged, which provides fast and cost-effective solutions to the analysis of large-scale sequence data. In cloud computing, high parallel tasks are run on a computer cluster through a virtual operating system (or "cloud"). Underlying the clouds are compact and virtualized virtual machines (VMs) hosting computation-intensive applications from distributed users. Cloud computing allows users to "rent" processing power and storage virtually on their demand and pay for what they use. There has been considerable enthusiasm in the bioinformatics community for use of cloud computing services. Figure 3 shows a schematic drawing of the cloud-based NGS analysis.

Recently exploratory efforts have been made in cloud-based DNA sequence storage. O'Connor et al. [97] created SeqWare Query Engine using cloud computing technologies to support databasing and query of information from thousands of genomes. BaseSpace is a scalable cloud-computing platform for all of Illumina's sequencing systems. After a sequencing run is completed, data from sequencing instruments is automatically uploaded to BaseSpace for analysis and storage. DNAnexus, a company specialized in Cloud-based DNA data analysis, has leveraged the storage capacity of Google Cloud to provide high-performance storage for NGS data.

Some initiatives have utilized preconfigured software on such cloud systems to process and analyze NGS data. Table 3 summarizes some tools that are currently available for sequence alignment, short read mapping, single nucleotide polymorphism (SNP) identification, and RNA expression analysis, amongst others.

Although cloud computing seems quite attractive, there are also issues that are yet to be resolved. The most significant concerns pertain to information security and bandwidth limitation. Transferring massive amounts of data (on the order of petabytes) to the cloud may be time consuming and prohibitively expensive. For most sequencing centers that require substantial data movement on a regular basis, cloud computing currently does not make economic sense. While it is clear that cloud computing has great potential for research purposes, for small labs and clinical applications using benchtop genome sequencers of limited throughput, like MiSeq, PGM, and Proton, cloud computing does have some practical utility.

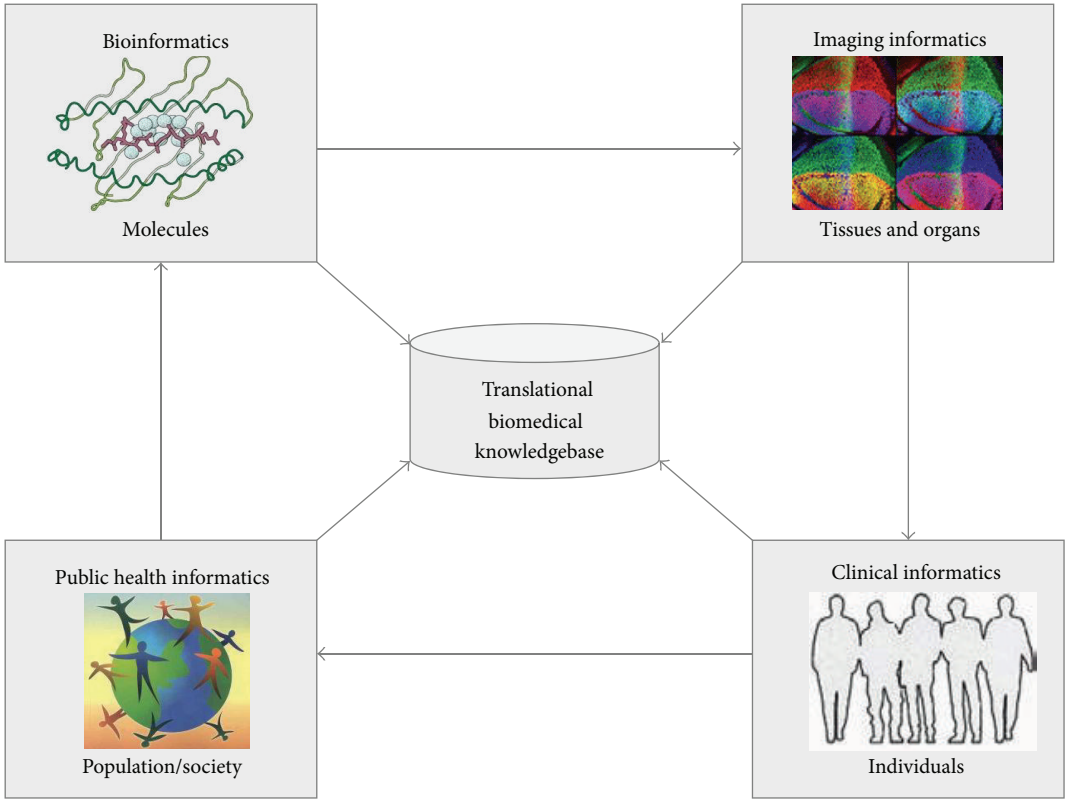


Figure 2: Translational bioinformatics bridges knowledge from molecules to populations. Four subdisciplines of translational bioinformatics and their respective focus areas are depicted in boxes. The success of translational bioinformatics will enable a complete information logistics chain from single molecules to the entire human population and thus link innovations from bench to bedside.

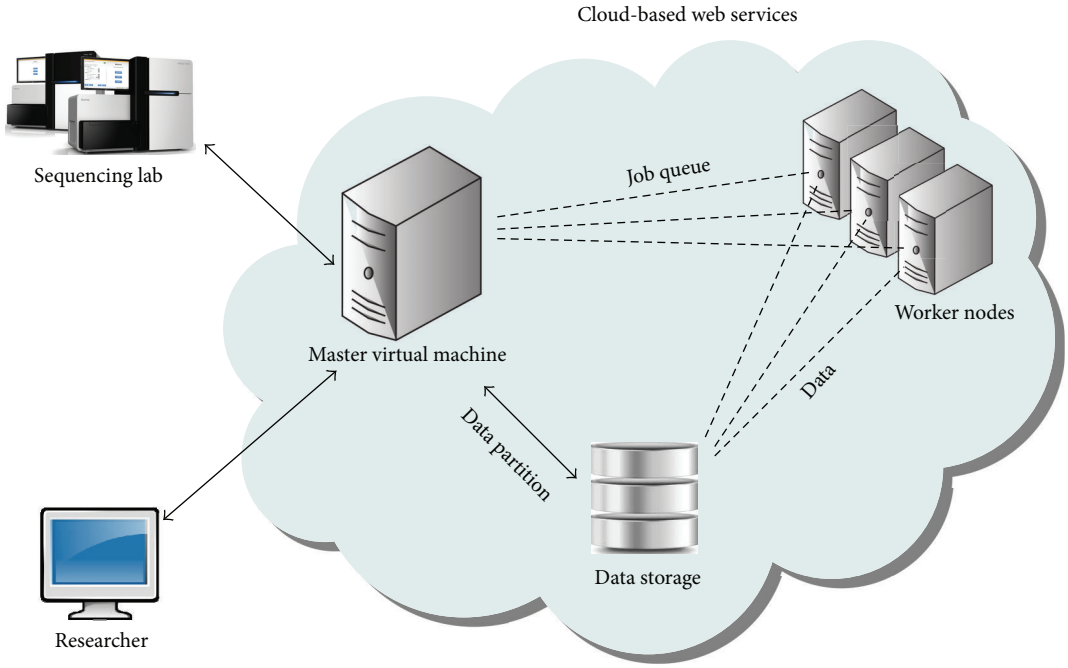


Figure 3: Schematic of the cloud-based NGS analysis. Local computers allocate the cloud-based web services over the internet. Web services comprised a cluster of virtual machines (one master node and a chosen number of worker nodes). Input data are transferred to the cloud storage and the program code driving the computation is uploaded to master nodes, by which worker nodes are provisioned. Each worker node downloads reads from the storage and run computation independently. The final result is stored and meanwhile transferred to the local client computer and the job completes.

Table 3: The cloud computing software for NGS data analysis.

Software	Website	Description	References
Crossbow	http://bowtie-bio.sourceforge.net/crossbow/	Read mapping and SNP calling	[56]
CloudBurst	http://cloudburst-bio.sourceforge.net/	Reference-based read mapping	[57]
Contrail	http://contrail-bio.sourceforge.net/	De novo read assembly	[58]
Cloud-MAQ	http://sourceforge.net/projects/cloud-maq/	Read mapping and assembly	[59]
Bioscope	http://www.lifescopecloud.com/	Reference-based read mapping	[60]
GeneSifter	http://www.geospiza.com/Products/AnalysisEdition.shtml	Customer oriented NGS data analysis services	[61]
CloudAligner	http://sourceforge.net/projects/cloudaligner/	Read mapping	[62]
Roundup	http://rodeo.med.harvard.edu/tools/roundup	Optimized computation for comparative genomics	[63]
PeakRanger	http://www.modencode.org/software/ranger/	Peak caller for ChIP-Seq data	[64]
Myrna	http://bowtie-bio.sf.net/myrna/	Differential expression analysis for RNA-Seq data	[65]
ArrayExpressHTS	http://www.ebi.ac.uk/Tools/rwiki/	RNA-Seq data processing and quality assessment	[66]
SeqMapreduce	Not available	Read mapping	[67]
BaseSpace	https://basespace.illumina.com/home/index		

8.2. Biomarker Discovery Using Systems Biology Approach. NGS makes it possible to generate multiple types of genomic alterations, including mutations, gene fusions, copy number alterations, and epigenetic changes simultaneously in a single test. Integration of these genomic, transcriptomic, interactomic, and epigenomic pieces of information is essential to infer the underlying mechanisms in prostate cancer development. The challenge ahead will be developing a comprehensive approach that could be analysed across these complementary data, looking for an ideal combination of biomarker signatures.

Consequently, the future of biomarker discovery will rely on a systems biology approach. One of the most fascinating fields in this regard is the network-based approaches to biomarker discovery, which integrate a large and heterogeneous dataset into interactive networks. In such networks, molecular components and interactions between them are represented as nodes and edges, respectively. The knowledge extracted from different types of networks can assist discovery of novel biomarkers, for example, functional pathways, processes,or subnetworks, for improved diagnostic, prognostic, and drug response prediction. Network-based discovery framework has already been reported in several types of cancers [98–102] including PCa [103, 104]. In a pioneering study, Jin et al. [103] built up a prostate-cancer-related network (PCRN) by searching the interactions among identified molecules related to prostate cancer. The network biomarkers derived from the network display high-performances in PCa patient classification.

As the field high-throughput technologies continues to develop, we will expect enhancing cooperation among different disciplines in translational bioinformatics, such as bioinformatics, imaging informatics, clinical informatics, and public health informatics. Notably, the heterogeneous data types coming from various informatics platforms are pushing for developing standards for data exchange across specialized domains.

8.3. Personalized Biomarkers. The recent breakthroughs in NGS also promise to facilitate the area of "personalised" biomarkers. Prostate cancer is highly heterogeneous among individuals. Current evidence indicates that the inter-individual heterogeneity arises from genetical environmental and lifestyle factors. By deciphering the genetic make-up of prostate tumors, NGS may facilitate patient stratification for targeted therapies and therefore assist tailoring the best treatment to the right patient. It is envisioned that personalized therapy will become part of clinical practice for prostate cancer in the near future.

9. Conclusions

Technological advances in NGS have increased our knowledge in molecular basis of PCa. However the translation of multiple molecular markers into the clinical realm is in its early stages. The full application of translational bioinformatics in PCa diagnosis and prognosis requires collaborative efforts between multiple disciplines. We can envision that the cloud-supported translational bioinformatics endeavours will promote faster breakthroughs in the diagnosis and prognosis of prostate cancer.

Conflict of Interests

The authors declare that there is no conflict of interests.

Authors' Contribution

Jiajia Chen, Daqing Zhang, and Wenying Yan contributed equally to this work.

Acknowledgments

The authors gratefully acknowledge financial support from the National Natural Science Foundation of China Grants (91230117, 31170795), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113201110015), International S&T Cooperation Program of Suzhou (SH201120), the National High Technology Research and Development Program of China (863 program, Grant no. 2012AA02A601), and Natural Science Foundation for Colleges and Universities in Jiangsu Province (13KJB180021).

References

[1] S. Carlsson, A. J. Vickers, M. Roobol et al., "Prostate cancer screening: facts, statistics, and interpretation in response to the us preventive services task force review," *Journal of Clinical Oncology*, vol. 30, no. 21, pp. 2581–2584, 2012.

[2] L. D. F. Venderbos and M. J. Roobol, "PSA-based prostate cancer screening: the role of active surveillance and informed and shared decision making," *Asian Journal of Andrology*, vol. 13, no. 2, pp. 219–224, 2011.

[3] O. Sartor, "Randomized studies of PSA screening: an opinion," *Asian Journal of Andrology*, vol. 13, no. 3, pp. 364–365, 2011.

[4] I. N. Sarkar, "Biomedical informatics and translational medicine," *Journal of Translational Medicine*, vol. 8, article 22, 2010.

[5] C. A. Kulikowski and C. W. Kulikowski, "Biomedical and health informatics in translational medicine," *Methods of Information in Medicine*, vol. 48, no. 1, pp. 4–10, 2009.

[6] B. S. Taylor, N. Schultz, H. Hieronymus et al., "Integrative genomic profiling of human prostate cancer," *Cancer Cell*, vol. 18, no. 1, pp. 11–22, 2010.

[7] W. Liu, S. Laitinen, S. Khan et al., "Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer," *Nature Medicine*, vol. 15, no. 5, pp. 559–565, 2009.

[8] X. Mao, Y. Yu, L. K. Boyd et al., "Distinct genomic alterations in prostate cancers in Chinese and Western populations suggest alternative pathways of prostate carcinogenesis," *Cancer Research*, vol. 70, no. 13, pp. 5207–5212, 2010.

[9] T. W. Friedlander, R. Roy, S. A. Tomlins et al., "Common structural and epigenetic changes in the genome of castration-resistant prostate cancer," *Cancer Research*, vol. 72, no. 3, pp. 616–625, 2012.

[10] J. Sun, W. Liu, T. S. Adams et al., "DNA copy number alterations in prostate cancers: a combined analysis of published CGH studies," *Prostate*, vol. 67, no. 7, pp. 692–700, 2007.

[11] C. M. Robbins, W. A. Tembe, A. Baker et al., "Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors," *Genome Research*, vol. 21, no. 1, pp. 47–55, 2011.

[12] M. J. Kim, R. D. Cardiff, N. Desai et al., "Cooperativity of Nkx3.1 and Pten loss of function in a mouse model of prostate carcinogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 5, pp. 2884–2889, 2002.

[13] J. Luo, S. Zha, W. R. Gage et al., "α-methylacyl-CoA racemase: a new molecular marker for prostate cancer," *Cancer Research*, vol. 62, no. 8, pp. 2220–2226, 2002.

[14] M. Tinzl, M. Marberger, S. Horvath, and C. Chypre, "DD3PCA3 RNA analysis in urine—a new perspective for detecting prostate cancer," *European Urology*, vol. 46, no. 2, pp. 182–187, 2004.

[15] E. S. Leman, G. W. Cannon, B. J. Trock et al., "EPCA-2: a highly specific serum marker for prostate cancer," *Urology*, vol. 69, no. 4, pp. 714–720, 2007.

[16] J. Luo, D. J. Duggan, Y. Chen et al., "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling," *Cancer Research*, vol. 61, no. 12, pp. 4683–4688, 2001.

[17] M. F. Darson, A. Pacelli, P. Roche et al., "Human glandular kallikrein 2 (hK2) expression in prostatic intraepithelial neoplasia and adenocarcinoma: a novel prostate cancer marker," *Urology*, vol. 49, no. 6, pp. 857–862, 1997.

[18] S. Varambally, S. M. Dhanasekaran, M. Zhou et al., "The polycomb group protein EZH2 is involved in progression of prostate cancer," *Nature*, vol. 419, no. 6907, pp. 624–629, 2002.

[19] S. F. Shariat, B. Andrews, M. W. Kattan, J. Kim, T. M. Wheeler, and K. M. Slawin, "Plasma levels of interleukin-6 and its soluble receptor are associated with prostate cancer progression and metastasis," *Urology*, vol. 58, no. 6, pp. 1008–1015, 2001.

[20] A. Berruti, A. Mosca, M. Tucci et al., "Independent prognostic role of circulating chromogranin A in prostate cancer patients with hormone-refractory disease," *Endocrine-Related Cancer*, vol. 12, no. 1, pp. 109–117, 2005.

[21] S. F. Shariat, C. G. Roehrborn, J. D. McConnell et al., "Association of the circulating levels of the urokinase system of plasminogen activation with the presence of prostate cancer and invasion, progression, and metastasis," *Journal of Clinical Oncology*, vol. 25, no. 4, pp. 349–355, 2007.

[22] S. F. Shariat, M. W. Kattan, E. Traxel et al., "Association of pre- and postoperative plasma levels of transforming growth factor β1 and interleukin 6 and its soluble receptor with prostate cancer progression," *Clinical Cancer Research*, vol. 10, no. 6, pp. 1992–1999, 2004.

[23] S. F. Shariat, J. H. Kim, B. Andrews et al., "Preoperative plasma levels of transforming growth factor beta(1) strongly predict clinical outcome in patients with bladder carcinoma," *Cancer*, vol. 92, no. 12, pp. 2985–2992, 2001.

[24] J. Lapointe, C. Li, J. P. Higgins et al., "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 3, pp. 811–816, 2004.

[25] G. Kristiansen, C. Pilarsky, C. Wissmann et al., "Expression profiling of microdissected matched prostate cancer samples reveals CD166/MEMD and CD24 as new prognostic markers for patient survival," *Journal of Pathology*, vol. 205, no. 3, pp. 359–376, 2005.

[26] J. Lapointe, S. Malhotra, J. P. Higgins et al., "hCAP-D3 expression marks a prostate cancer subtype with favorable clinical behavior and androgen signaling signature," *American Journal of Surgical Pathology*, vol. 32, no. 2, pp. 205–209, 2008.

[27] K. D. Sørensen, P. J. Wild, A. Mortezavi et al., "Genetic and epigenetic SLC18A2 silencing in prostate cancer is an independent adverse predictor of biochemical recurrence after radical prostatectomy," *Clinical Cancer Research*, vol. 15, no. 4, pp. 1400–1410, 2009.

[28] J. F. Knight, C. J. Shepherd, S. Rizzo et al., "TEAD1 and c-Cbl are novel prostate basal cell markers that correlate with poor clinical outcome in prostate cancer," *British Journal of Cancer*, vol. 99, no. 11, pp. 1849–1858, 2008.

[29] W. D. Zhong, G. Q. Qin, Q. S. Dai et al., "SOXs in human prostate cancer: implication as progression and prognosis factors," *BMC Cancer*, vol. 12, no. 1, p. 248, 2012.

[30] C. Ruiz, M. Oeggerli, M. Germann et al., "High NRBP1 expression in prostate cancer is linked with poor clinical outcomes and increased cancer cell growth," *Prostate*, 2012.

[31] N. Pértega-Gomes, J. R. Vizcaíno, V. Miranda-Gonçalves et al., "Monocarboxylate transporter 4 (MCT4) and CD147 overexpression is associated with poor prognosis in prostate cancer," *BMC Cancer*, vol. 11, p. 312, 2011.

[32] A. S. Syed Khaja, L. Helczynski, A. Edsjö et al., "Elevated level of Wnt5a protein in localized prostate cancer tissue is associated with better outcome," *PloS ONE*, vol. 6, no. 10, Article ID e26539, 2011.

[33] M. Q. Yang, B. D. Athey, H. R. Arabnia et al., "High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics," *BMC Genomics*, vol. 10, no. 1, article I1, 2009.

[34] C. C. Collins, S. V. Volik, A. V. Lapuk et al., "Next generation sequencing of prostate cancer from a patient identifies a deficiency of methylthioadenosine phosphorylase, an exploitable tumor target," *Molecular Cancer Therapeutics*, vol. 11, no. 3, pp. 775–783, 2012.

[35] C. E. Barbieri, S. C. Baca, M. S. Lawrence et al., "Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer," *Nature Genetics*, vol. 44, no. 6, pp. 685–689, 2012.

[36] C. S. Grasso, Y. M. Wu, D. R. Robinson et al., "The mutational landscape of lethal castration-resistant prostate cancer," *Nature*, vol. 487, no. 7406, pp. 239–243, 2012.

[37] A. Kumar, T. A. White, A. P. MacKenzie et al., "Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 41, pp. 17087–17092, 2011.

[38] S. A. Tomlins, D. R. Rhodes, S. Perner et al., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, no. 5748, pp. 644–648, 2005.

[39] S. A. Tomlins, R. Mehra, D. R. Rhodes et al., "TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer," *Cancer Research*, vol. 66, no. 7, pp. 3396–3400, 2006.

[40] B. E. Helgeson, S. A. Tomlins, N. Shah et al., "Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer," *Cancer Research*, vol. 68, no. 1, pp. 73–80, 2008.

[41] Z. Shaikhibrahim, M. Braun, P. Nikolov et al., "Rearrangement of the ETS genes ETV-1, ETV-4, ETV-5, and ELK-4 is a clonal event during prostate cancer progression," *Human Pathology*, vol. 43, no. 11, pp. 1910–1916, 2012.

[42] S. A. Tomlins, B. Laxman, S. M. Dhanasekaran et al., "Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer," *Nature*, vol. 448, no. 7153, pp. 595–599, 2007.

[43] K. G. Hermans, A. A. Bressers, H. A. Van Der Korput, N. F. Dits, G. Jenster, and J. Trapman, "Two unique novel prostate-specific and androgen-regulated fusion partners of ETV4 in prostate cancer," *Cancer Research*, vol. 68, no. 9, pp. 3094–3098, 2008.

[44] N. Palanisamy, B. Ateeq, S. Kalyana-Sundaram et al., "Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma," *Nature Medicine*, vol. 16, no. 7, pp. 793–798, 2010.

[45] C. Wu, A. W. Wyatt, A. V. Lapuk et al., "Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer," *Journal of Pathology*, vol. 227, no. 1, pp. 53–61, 2012.

[46] H. Beltran, R. Yelensky, G. M. Frampton et al., "Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity," *European Urology*, vol. 63, no. 5, pp. 920–926, 2013.

[47] K. Kannan, L. Wang, J. Wang, M. M. Ittmann, W. Li, and L. Yen, "Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 22, pp. 9172–9177, 2011.

[48] J. R. Prensner, M. K. Iyer, O. A. Balbin et al., "Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression," *Nature Biotechnology*, vol. 29, no. 8, pp. 742–749, 2011.

[49] A. Watahiki, Y. Wang, J. Morris et al., "MicroRNAs associated with metastatic prostate cancer," *PLoS ONE*, vol. 6, no. 9, Article ID e24950, 2011.

[50] K. R. Chng, C. W. Chang, S. K. Tan et al., "A transcriptional repressor co-regulatory network governing androgen response in prostate cancers," *EMBO Journal*, vol. 31, pp. 2810–2823, 2012.

[51] Z. Zhang, C. W. Chang, W. L. Goh, W.-K. Sung, and E. Cheung, "CENTDIST: discovery of co-associated factors by motif distribution," *Nucleic Acids Research*, vol. 39, no. 2, pp. W391–W399, 2011.

[52] H. H. He, C. A. Meyer, M. W. Chen, V. C. Jordan, M. Brown, and X. S. Liu, "Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics," *Genome Research*, vol. 22, no. 6, pp. 1015–1025, 2012.

[53] G. H. Little, H. Noushmehr, S. K. Baniwal, B. P. Berman, G. A. Coetzee, and B. Frenkel, "Genome-wide Runx2 occupancy in prostate cancer cells suggests a role in regulating secretion," *Nucleic Acids Research*, vol. 40, no. 8, pp. 3538–3547, 2012.

[54] P. Y. Tan, C. W. Chang, K. R. Chng, K. D. Senali Abayratna Wansa, W.-K. Sung, and E. Cheung, "Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival," *Molecular and Cellular Biology*, vol. 32, no. 2, pp. 399–414, 2012.

[55] J. H. Kim, S. M. Dhanasekaran, J. R. Prensner et al., "Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer," *Genome Research*, vol. 21, no. 7, pp. 1028–1041, 2011.

[56] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biology*, vol. 10, no. 11, article R134, 2009.

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

[58] M. C. Schatz, D. D. Sommer, D. R. Kelley, and M. Pop, "De Novo assembly of large genomes using cloud computing," in *Proceedings of the Cold Spring Harbor Biology of Genomes Conference*, New York, NY, USA, May 2010.

[59] A. K. Talukder, S. Gandham, H. A. Prahalad, and N. P. Bhattacharyya, "Cloud-MAQ: the cloud-enabled scalable whole genome reference assembly application," in *Proceedings of the 7th IEEE and IFIP International Conference on Wireless and Optical Communications Networks (WOCN '10)*, Colombo, Sri Lanka, September 2010.

[60] M. H. Rahimi, *Bioscope: Actuated Sensor Network for Biological Science*, University of Southern California, Los Angeles, Calif, USA, 2005.

[61] C. Sansom, "Up in a cloud?" *Nature Biotechnology*, vol. 28, no. 1, pp. 13–15, 2010.

[62] 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 171, 2011.

[63] P. Kudtarkar, T. F. DeLuca, V. A. Fusaro, P. J. Tonellato, and D. P. Wall, "Cost-effective cloud computing: a case study using the comparative genomics tool, roundup," *Evolutionary Bioinformatics*, vol. 2010, no. 6, pp. 197–203, 2010.

[64] X. Feng, R. Grossman, and L. Stein, "PeakRanger: a cloud-enabled peak caller for ChIP-seq data," *BMC Bioinformatics*, vol. 12, article 139, 2011.

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

[66] A. Goncalves, A. Tikhonov, A. Brazma, and M. Kapushesky, "A pipeline for RNA-seq data processing and quality assessment," *Bioinformatics*, vol. 27, no. 6, pp. 867–869, 2011.

[67] Y. Li and S. Zhong, "SeqMapReduce: software and web service for accelerating sequence mapping," *Critical Assessment of Massive Data Anaysis (CAMDA)*, vol. 2009, 2009.

[68] M. F. Berger, M. S. Lawrence, F. Demichelis et al., "The genomic complexity of primary human prostate cancer," *Nature*, vol. 470, no. 7333, pp. 214–220, 2011.

[69] I. N. Holcomb, D. I. Grove, M. Kinnunen et al., "Genomic alterations indicate tumor origin and varied metastatic potential of disseminated cells from prostate cancer patients," *Cancer Research*, vol. 68, no. 14, pp. 5599–5608, 2008.

[70] Z. Kan, B. S. Jaiswal, J. Stinson et al., "Diverse somatic mutation patterns and pathway alterations in human cancers," *Nature*, vol. 466, no. 7308, pp. 869–873, 2010.

[71] L. Agell, S. Hernández, S. De Muga et al., "KLF6 and TP53 mutations are a rare event in prostate cancer: distinguishing between Taq polymerase artifacts and true mutations," *Modern Pathology*, vol. 21, no. 12, pp. 1470–1478, 2008.

[72] J.-T. Dong, "Prevalent mutations in prostate cancer," *Journal of Cellular Biochemistry*, vol. 97, no. 3, pp. 433–447, 2006.

[73] X. Sun, H. F. Frierson, C. Chen et al., "Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer," *Nature Genetics*, vol. 37, no. 6, pp. 407–412, 2005.

[74] L. Zheng, F. Wang, C. Qian et al., "Unique substitution of CHEK2 and TP53 mutations implicated in primary prostate tumors and cancer cell lines," *Human Mutation*, vol. 27, no. 10, pp. 1062–1063, 2006.

[75] C. Kumar-Sinha, S. A. Tomlins, and A. M. Chinnaiyan, "Recurrent gene fusions in prostate cancer," *Nature Reviews Cancer*, vol. 8, no. 7, pp. 497–511, 2008.

[76] S. Perner, J.-M. Mosquera, F. Demichelis et al., "TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion," *American Journal of Surgical Pathology*, vol. 31, no. 6, pp. 882–888, 2007.

[77] S. A. Tomlins, S. M. J. Aubin, J. Siddiqui et al., "Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA," *Science Translational Medicine*, vol. 3, no. 94, Article ID 94ra72, 2011.

[78] D. B. Seligson, S. Horvath, T. Shi et al., "Global histone modification patterns predict risk of prostate cancer recurrence," *Nature*, vol. 435, no. 7046, pp. 1262–1266, 2005.

[79] A. Urbanucci, B. Sahu, J. Seppälä et al., "Overexpression of androgen receptor enhances the binding of the receptor to the chromatin in prostate cancer," *Oncogene*, vol. 31, no. 17, pp. 2153–2163, 2012.

[80] D. S. Rickman, T. D. Soong, B. Moss et al., "Oncogene-mediated alterations in chromatin conformation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 23, pp. 9083–9088, 2012.

[81] X. Lin, M. Tascilar, W.-H. Lee et al., "GSTP1 CpG island hypermethylation is responsible for the absence of GSTP1 expression in human prostate cancer cells," *American Journal of Pathology*, vol. 159, no. 5, pp. 1815–1826, 2001.

[82] E. Rosenbaum, M. O. Hoque, Y. Cohen et al., "Promoter hypermethylation as an independent prognostic factor for relapse in patients with prostate cancer following radical prostatectomy," *Clinical Cancer Research*, vol. 11, no. 23, pp. 8321–8325, 2005.

[83] O. Taiwo, G. A. Wilson, T. Morris et al., "Methylome analysis using MeDIP-seq with low DNA concentrations," *Nature Protocols*, vol. 7, no. 4, pp. 617–636, 2012.

[84] R. Lister, M. Pelizzola, R. H. Dowen et al., "Human DNA methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[85] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5868–5877, 2005.

[86] C. Gebhard, L. Schwarzfischer, T.-H. Pham et al., "Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia," *Cancer Research*, vol. 66, no. 12, pp. 6118–6128, 2006.

[87] A. L. Brunner, D. S. Johnson, W. K. Si et al., "Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver," *Genome Research*, vol. 19, no. 6, pp. 1044–1056, 2009.

[88] M. W. Kattan, S. F. Shariat, B. Andrews et al., "The addition of interleukin-6 soluble receptor and transforming growth factor beta1 improves a preoperative nomogram for predicting biochemical progression in patients with clinically localized prostate cancer," *Journal of Clinical Oncology*, vol. 21, no. 19, pp. 3573–3579, 2003.

[89] J. Baden, G. Green, J. Painter et al., "Multicenter evaluation of an investigational prostate cancer methylation assay," *Journal of Urology*, vol. 182, no. 3, pp. 1186–1193, 2009.

[90] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Research*, vol. 62, no. 15, pp. 4427–4433, 2002.

[91] G. V. Glinsky, A. B. Glinskii, A. J. Stephenson, R. M. Hoffman, and W. L. Gerald, "Gene expression profiling predicts clinical outcome of prostate cancer," *Journal of Clinical Investigation*, vol. 113, no. 6, pp. 913–923, 2004.

[92] Y. Wang, J. Chen, Q. Li et al., "Identifying novel prostate cancer associated pathways based on integrative microarray data analysis," *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 151–158, 2011.

[93] L.-C. Li, H. Zhao, H. Shiina, C. J. Kane, and R. Dahiya, "PGDB: a curated and integrated database of genes related to the prostate," *Nucleic Acids Research*, vol. 31, no. 1, pp. 291–293, 2003.

[94] V. Hawkins, D. Doll, R. Bumgarner et al., "PEDB: the prostate expression database," *Nucleic Acids Research*, vol. 27, no. 1, pp. 204–208, 1999.

[95] M. Maqungo, M. Kaur, S. K. Kwofie et al., "DDPC: dragon database of genes associated with prostate cancer," *Nucleic Acids Research*, vol. 39, no. 1, pp. D980–D985, 2011.

[96] A. Etim, G. Zhou, X. Wen et al., "ChromSorter PC: a database of chromosomal regions associated with human prostate cancer," *BMC Genomics*, vol. 5, article 27, 2004.

[97] B. D. O'Connor, B. Merriman, and S. F. Nelson, "SeqWare Query Engine: storing and searching sequence data in the cloud," *BMC Bioinformatics*, vol. 11, no. 12, article S2, 2010.

[98] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[99] I. W. Taylor, R. Linding, D. Warde-Farley et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature Biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[100] M. Xu, M.-C. J. Kao, J. Nunez-Iglesias, J. R. Nevins, M. West, and X. J. Jasmine, "An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer," *BMC Genomics*, vol. 9, no. 1, article S12, 2008.

[101] Y.-C. Wang and B.-S. Chen, "A network-based biomarker approach for molecular investigation and diagnosis of lung cancer," *BMC Medical Genomics*, vol. 4, article 2, 2011.

[102] Y. Zhang, S. Wang, D. Li et al., "A systems biology-based classifier for hepatocellular carcinoma diagnosis," *PLoS ONE*, vol. 6, no. 7, Article ID e22426, 2011.

[103] G. Jin, X. Zhou, K. Cui, X.-S. Zhang, L. Chen, and S. T. C. Wong, "Cross-platform method for identifying candidate network biomarkers for prostate cancer," *IET Systems Biology*, vol. 3, no. 6, pp. 505–512, 2009.

[104] R. Ummanni, F. Mundt, H. Pospisil et al., "Identification of clinically relevant protein targets in prostate cancer with 2D-DIGE coupled mass spectrometry and systems biology network platform," *PLoS ONE*, vol. 6, no. 2, Article ID e16833, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 617545, 7 pages
http://dx.doi.org/10.1155/2013/617545

Research Article

Exploring the Cooccurrence Patterns of Multiple Sets of Genomic Intervals

Hao Wu[1] and Zhaohui S. Qin[1,2,3]

[1] *Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA*
[2] *Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA*
[3] *Center for Comprehensive Informative, Emory University, Atlanta, GA, USA*

Correspondence should be addressed to Hao Wu; hao.wu@emory.edu and Zhaohui S. Qin; zhaohui.qin@emory.edu

Received 27 March 2013; Accepted 4 May 2013

Academic Editor: Zhongming Zhao

Background. Exploring the spatial relationship of different genomic features has been of great interest since the early days of genomic research. The relationship sometimes provides useful information for understanding certain biological processes. Recent advances in high-throughput technologies such as ChIP-seq produce large amount of data in the form of genomic intervals. Most of the existing methods for assessing spatial relationships among the intervals are designed for pairwise comparison and cannot be easily scaled up. *Results*. We present a statistical method and software tool to characterize the cooccurrence patterns of multiple sets of genomic intervals. The occurrences of genomic intervals are described by a simple finite mixture model, where each component represents a distinct cooccurrence pattern. The model parameters are estimated via an EM algorithm and can be viewed as sufficient statistics of the cooccurrence patterns. Simulation and real data results show that the model can accurately capture the patterns and provide biologically meaningful results. The method is implemented in a freely available R package `giClust`. *Conclusions*. The method and the software provide a convenient way for biologists to explore the cooccurrence patterns among a relatively large number of sets of genomic intervals.

1. Introduction

Exploring the spatial relationships of different genomic features has been of great interest since the early days of genomic research. The relationships often provide important information for certain biological processes. One famous example is that people detected CpG islands (CGI) from the DNA sequence as short, CG rich genomic regions and then found that they significantly overlap gene promoters [1]. The CGI/promoter overlaps shed light on the function of DNA methylation on gene expressions.

In modern functional genomics research, one major goal is to understand the regulatory mechanism of gene expression. The transcriptional process involves the combinatory effects of different DNA-binding proteins and histone modifications. To decipher the complex process, an important first step is to detect the protein binding or histone modification sites and then explore the spatial relationships among them. The spatial relationships provide evidence for interactions among various regulatory elements. For example, if the binding sites of two proteins significantly overlap, it is likely that they interact. Recent advances in high-throughput technologies such as ChIP-seq [2] make the genome-wide profiling of proteins binding or histone modification an easy task. It is now common for a biologist to map the binding sites for a few proteins and then compare them with each other or with some public data. Since the protein binding or histone modification sites are represented as genomic intervals, such task requires the comparison of multiple sets of genomic intervals. Hereafter a set of genomic intervals will be referred to as a "track."

Although good tools for comparing tracks are immediately needed, there are only a few existing methods, and most of them are designed for pairwise comparison. The easiest way to compare two tracks is to compute their overlaps and then represent them by a Venn diagram [3]. This method,

however, does not evaluate the statistical significance of the overlaps. Several more statistically rigorous methods are recently proposed to evaluate the overlap or "closeness" of two tracks. Favorov et al. developed an R package called "GenometriCorr" to evaluate the spatial correlation between two tracks [4]. They implemented several test procedures to measure the "closeness" of the two tracks and report P values. Chikina and Troyanskaya proposed a similar distance based method for comparing two tracks [5]. One advantage of their method is that the relationship can be evaluated within a user-defined genomic regions (termed as "domain set" in the paper). All the methods discussed above are designed for comparing two tracks. Because these methods are based on pairwise overlaps or distances, they cannot be easily scaled up for comparing multiple tracks. To that end, a generalized linear model based approach was proposed to explore the dependence of one track on several others [6]. It first converts each track into a binary vector based on genome wide occurrence; then the occurrence of one track is modeled as a function of the occurrence of other tracks through a log-linear model. One drawback of such an approach is that it only measures the marginal dependence. For example, if track A significantly overlaps track B only in a small proportion of the genome, the marginal dependence will be small and likely to be overlooked. A newly developed method "ChromHMM" was proposed to discover the genome-wide chromatin state [7]. It first splits the genome into equal sized bins and then applies a hidden Markov model (HMM) with certain number of states to segment the genome according to the combinations of chromatin modifications. In this approach, the emission probabilities of the HMM are based on overlapping patterns of multiple chromatin modifications. However the method itself is not specifically designed for evaluating relationships of multiple sets of genomic regions. A newly developed method jMOSAiCS [8] implements a joint model for multiple sets of ChIP-seq data. Its main goal is to improve peak calling and cannot be directly applied to assess the spatial correlations among peak lists from different ChIP-seq data.

In this work, we aim to develop an intuitive tool for comparing multiple tracks. There are different types of spatial relationships. Consider the pairwise comparison, the most straightforward relationship is overlapping. In addition, two tracks could be close to each other but do not overlap. A biological example for the "closeness" relationship is differentially methylated regions (DMRs) and CGI. It was reported that DMRs are at the CGI "shores" [9], meaning that the two tracks are close but not overlapping. Moreover, two tracks could "exclude" each other. In this work, we focus only on the overlapping relationship and present a statistical method to characterize the cooccurrence patterns of multiple tracks. The patterns are described by a finite mixture model. An EM algorithm [10] is devised to infer the model parameters. Given multiple tracks and a set of user provided genomic regions, the method clusters the regions into certain number of groups according to the cooccurrence patterns among the input tracks. Moreover, the model parameters "sufficiently" summarize the cooccurrence patterns among input tracks. The sufficiency means that any joint or conditional occurrence probabilities in a subset of the inputs can be calculated from the parameters alone. For example, questions like "what percent of the sites bound by both protein A and B are also bound by protein C" can be readily answered without the raw data.

2. Methods

Assume that there are D input tracks and one wants to explore their cooccurrence patterns over a set of N genomic intervals (referred to as "regions of interest" hereafter). The regions of interest are specified by the user. An example is the promoter regions of all known genes. The first step in analysis is to count the overlaps of input tracks and the regions of interest. Let Y_{id} indicate that the ith region of interest overlaps some interval from dth track (=1) or not (=0), $i = 1, \ldots, N$, $d = 1, \ldots, D$. The Y_{id} matrix is the input data for the following procedures.

2.1. A Finite Mixture Model for the Cooccurrence Patterns. The cooccurrence pattern of the input tracks is described by a finite mixture model. Each mixture component is represented by a product of Bernoulli distributions. Assume that the regions of interest are from a mixture of K clusters. In each cluster, the input tracks exhibit distinct cooccurrence patterns. Let Z_i denote the cluster indicator for the ith region of interest. Assume that the prior probability for a region of interest being in the kth cluster is π_k; for example, $\Pr(Z_i = k) = \pi_k$, $\sum_{k=1}^{K} \pi_k = 1$. For a region of interest in cluster k, let the probability of intervals from track d occurring in this regions be q_{kd}; in other words, $P(Y_{id} = 1 \mid Z_i = k) = q_{kd}$. Define $\mathbf{Q} \equiv \{q_{kd}; k = 1, \ldots, K; d = 1, \ldots, D\}$, a $K \times D$ matrix. This matrix characterizes the cooccurrence patterns of input tracks over the regions of interest. We further assume that within a cluster, the occurrences of input tracks are independent; for example, $P(Y_{il} = 1, Y_{il'} = 1 \mid Z_i = k) = P(Y_{il} = 1 \mid Z_i = k) P(Y_{il'} = 1 \mid Z_i = k)$ for all l, l'.

2.2. Parameter Estimation. Under the proposed model, Y_{id}s are observed data, Z_is are missing indicator variables, and q_{kd} and π_ks are model parameters. K represent the dimension of the model, which can be either specified by the user or estimated from data. Given K, the model parameters can be estimated by the following EM algorithm.

First let $\boldsymbol{\Psi} = \{\pi_k, q_{kd}; k = 1, \ldots, K; d = 1, \ldots, D\}$ denote all model parameters. The complete data log-likelihood of the parameters can be derived as follows (details are provided in the Supplementary Material available online at http://dx.doi.org/10.1155/2013/617545):

$$L(\boldsymbol{\Psi}) = \sum_i \sum_k \delta(Z_i = k) \left\{ \log \pi_k + \sum_d \left[Y_{id} \log q_{kd} + (1 - Y_{id}) \times \log(1 - q_{kd}) \right] \right\}. \quad (1)$$

Here $\delta(\cdot)$ is an indicator function. Let $\mathbf{Y}$ denote all observed data, and define $\mu_{ik} = \delta(Z_i = k)$. The E-step calculates the expected value of μ_{ik} given the observed data and the parameter values at the current step, denoted by $\boldsymbol{\Psi}^{(t)}$:

$$\begin{aligned}\mu_{ik}^{(t)} &\equiv E\left[\delta\left(Z_i = k\right) \mid \mathbf{Y}, \boldsymbol{\Psi}^{(t)}\right] = \Pr\left(Z_i = k \mid \mathbf{Y}, \boldsymbol{\Psi}^{(t)}\right) \\ &= \frac{\pi_k^{(t)} P\left(\mathbf{Y}_i \mid Z_i = k, \boldsymbol{\Psi}^{(t)}\right)}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} P\left(\mathbf{Y}_i \mid Z_i = k', \boldsymbol{\Psi}^{(t)}\right)}.\end{aligned} \tag{2}$$

Plugging the expected values into (1), one obtains the Q function as follows:

$$\begin{aligned}&Q\left(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(t)}\right) \\ &\quad = E\left[l\left(\boldsymbol{\Psi}\right) \mid \mathbf{Y}, \boldsymbol{\Psi}^{(t)}\right] \\ &\quad = \sum_i \sum_k \mu_{ik}^{(t)} \left\{ \log \pi_k + \sum_d \left[Y_{ij} \log q_{kj} + \left(1 - Y_{ij}\right) \right.\right. \\ &\qquad \left.\left. \times \log\left(1 - q_{kj}\right)\right]\right\}.\end{aligned} \tag{3}$$

The M-step maximizes the Q function with respect to parameters. By solving $\partial Q/\partial \pi_k = 0$ and $\partial Q/\partial q_{kd} = 0$, we obtain the update for π_k and q_{kd} as follows:

$$\pi_k^{(t+1)} = \frac{\sum_i \mu_{ik}^{(t)}}{I}, \qquad q_{kd}^{(t+1)} = \frac{\sum_i \mu_{ik}^{(t)} Y_{ij}}{\sum_i \mu_{ik}^{(t)}}. \tag{4}$$

Because the EM algorithm can sometimes converge to a local maxima, using good starting values is very important. In practice, we choose $\boldsymbol{\Psi}^0$ based on K-means clustering results. To be specific, we first run K-means clustering on $\mathbf{Y}$ for 10 times and then take the one with the smallest total within cluster distances. The cluster centers and cluster sizes are used as starting values for $\boldsymbol{\Psi}$. The EM algorithm then iterates between (2) and (4) until convergence.

2.3. Choosing the Number of Clusters. The above EM algorithm is derived with the number of clusters K given. Choosing K is a model selection problem. A widely used method for obtaining the optimal K is the Bayesian Information Criterion (BIC) [11], which is defined as $\mathrm{BIC}_K = -2 \log L_K + C_K * \log T$. Here L_K is the likelihood from the model with K clusters, computed based on (1). C_K is the number of parameters, which equals $K * (D + 1) - 1$ in a model with K clusters. T is the total number of data points, which is $N * D$. The BIC is computed for different values of K, and then the K associated with the smallest BIC is deemed the to be optimal solution.

The BIC works well in simulation settings. However in practice, we found that the BIC criteria often favor bigger models. This is a fairly common problem in model selection for genomic data. Because the sample sizes (in this problem, N) are often huge, the penalty in BIC is not strong enough to offset the gain in likelihood from bigger models even when effect sizes are small. As a result, BIC often selects a bigger model. In real data analysis, a smaller model is more desirable for interpretability. There are some methods proposed to generate smaller model for genomic data analysis, for example, based on model stability [12] or pruned the larger model down [13]. In this work, we take an easy approach and adopt the recommendation in [14] for choosing number of components in K-means clustering. We plot the log-likelihood versus K and choose the K at the "elbow" point of the curve as the optimal solution. We will show that this *ad hoc* method provides good results in practice.

2.4. Interpreting the Model Parameters. The model parameters are directly interpretable: π_k represents cluster sizes and q_{kd} represents the probability of occurrence of track d in regions of interest from cluster k. Under the model assumptions, the parameters sufficiently describe the cooccurrence relationships among input tracks. Any joint or conditional occurrence probabilities can be directly computed from the model parameters. The sufficiency can be shown in the following simple example. The joint probability $\Pr(Y_{ia} = 1, Y_{ib} = 0)$ for all $1 \le a, b \le D$ can be computed as follows:

$$\begin{aligned}&\Pr\left(Y_{ia} = 1,\, Y_{ib} = 0\right) \\ &\quad = \sum_{k=1}^{K} \Pr\left(Y_{ia} = 1,\, Y_{ib} = 0 \mid Z_i = k\right) \Pr\left(Z_i = k\right) \\ &\quad = \sum_{k=1}^{K} \Pr\left(Y_{ia} = 1 \mid Z_i = k\right) \Pr\left(Y_{ib} = 0 \mid Z_i = k\right) \pi_k \\ &\quad = \sum_{k=1}^{K} q_{ka}\left(1 - q_{kb}\right) \pi_k.\end{aligned} \tag{5}$$

Other joint probabilities can be derived in a similar way. The conditional probabilities can be computed from the ratios of proper joint probabilities. These joint/conditional probabilities answer questions like "what is the probability of cobindings of protein A and B?" or "what is the probability of protein A binding at a region, given that protein B binds at that region?" The answers to these questions can be derived from the model parameters without the raw data (e.g., the $\mathbf{Y}$ matrix). This demonstrates an added advantage of the method: the parameters work like sufficient statistics for cooccurrence patterns among input tracks.

2.5. Implementation. The proposed method has been implemented in an R package `giClust`, which is freely available at http://www.sph.emory.edu/~hwu/giClust.html. The package takes multiple lists of genomic intervals in BED format as inputs. With two lines of R code the package can generate the results, which include the estimated parameters and the best group assignment for each region of interest.

3. Results

We conducted simulations and real data studies to illustrate the usefulness of `giClust`. Here we only present the results from two real data tests. The results of two simulation studies can be found in the Supplementary Material.

3.1. The Cobinding Patterns of 15 Proteins in Mouse ES Data. In a seminal paper, Chen et al. mapped the binding sites of 15 different proteins (13 transcription factors and 2 transcription regulators) to study the transcriptional network in mouse embryonic stem (ES) cells [3]. In this example, `giClust` was applied to discover the cobinding pattern of these proteins. The data were obtained from Gene Expression Omnibus (GEO) database [15] under accession number GSE11431. We first took a union of all called peaks and then used the union as the regions of interest to study the cobinding patterns of all proteins. The union contains 168,439 intervals, with mean length of 542 base pairs. We counted the overlaps between binding sites of all proteins and the unions to construct the $\mathbf{Y}$ matrix then applied `giClust` on $\mathbf{Y}$, allowing the number of clusters (K) to vary from 1 to 15.

Figures 1(a)-1(b) plot the BIC and log-likelihood versus K. It shows that the BIC kept decreasing with K so one cannot obtain an optimal K from that. However, looking at the log-likelihood, the amount of increments with K is initially large but becomes smaller with larger K. The "elbow" point of the curve in Figure 1(b) is at near $K = 8$, so we decided to use $K = 8$ as the optimal number of clusters. The estimated model parameters are summarized in Table 1. The $\widehat{\mathbf{Q}}$ matrix is also represented as a heatmap in Figure 1(c).

It shows that a large proportion of the regions of interest are dominated by a single protein; for example, clusters 1, 2, 3, 5, and 6 are almost exclusively bound by CTCF, Esrrb, Tcfcp2l1, E2f1, and Nanog, respectively. Cluster 4 has sparse binding for a number of proteins. Clusters 7 and 8 are the most interesting groups, which are termed as Multiple Transcription Factor-Binding Loci (MTL) in [3]. Cluster 7 (occupies 7% of the regions of interest) is the Myc specific cluster, which shows strong binding tendency from c-Myc, n-Myc, along with several other proteins including Klf4, Esrrb, Tcfcp2l1, Zfx, and E2f1. Cluster 8 (occupies 3% of the regions of interest) is the Nanog-Oct4-Sox2 specific cluster that is defined as "ES-Cell Enhanceosomes" in [3]. This cluster includes strong binding from Klf4, Esrrb, Tcfcp2l1, and E2f1. Moreover 84% of all P300 binding sites belong to this cluster.

We further looked at the locational distributions of regions of interest in different clusters. Figure 2 shows the percentage of regions in each cluster overlapping some predefined genomic regions, including transcriptional starting site (TSS), transcriptional end site (TES), and exonic/intronic/intergenic regions. The percentages are compared with the expected values (marked as "Random" in Figure 2). The expected values are computed based on total lengths of the predefined genomic features. For example, the total length of TSS regions (defined as ±500 bps around the transcriptional starting location) equals 1% of the total genome. So the expected percentage of genomic regions overlapping TSS is roughly 1%. Note that the percentages for each cluster should sum up to 1. One striking result is that over 60% of intervals in cluster 7 (the Myc specific cluster) overlap TSS, suggesting the regulatory role of Myc proteins in mouse ES cells.

All the findings are consistent with the results reported in [3]. These demonstrate that `giClust` can discover the cobinding patterns of relatively large number of proteins in a quick analysis and provide biologically meaningful results.

3.2. The Histone Modification Pattern over the Gene Promoters in K562 Cell Line. It was well known that the combinatory patterns of histone modifications correlate with gene expressions. In this example, we used `giClust` to analyze a number of histone modification datasets. Nine ChIP-seq datasets for profiling different histone modifications in K562 cell lines were obtained from ENCODE. The histones include H3k4me1, H3k4me2, H3k4me3, H3k9me1, H4k20me1, H3k27me3, H3k36me3, H3k9ac, and H3k27ac. In this example, we focused the analysis on the genes and used the promoter regions of Refseq genes as regions of interest.

Similar to the previous example, we first obtained the overlapping matrix $\mathbf{Y}$ and then ran `giClust` for different number of clusters ($K = 1, \ldots, 10$). Plots of the BIC and log-likelihood versus number of clusters are shown in Supplemental Figure S3(a)-(b). The appearance of the curves is similar to that in the mouse ES data. Using a similar method, we picked $K = 5$ as the optimal number of clusters to perform further analyses. The estimated model parameters are listed in Table 2. The heatmap representation of $\widehat{\mathbf{Q}}$ is shown in Supplemental Figure S3(c).

We further compared the expressions of the genes in different clusters. We obtained the RNA-seq data from ENCODE and computed RPKM (reads per kilo-bp per million reads) to represent the gene expressions. Figure 3 shows the boxplot of expressions, represented as square root of RPKM, for genes in different clusters. It shows that genes in the first cluster (44% of all genes) have the highest average expressions. These genes have modifications on almost all histone marks at promoters except the repressive mark H3K27me3. The second cluster contains 21% genes and shows almost no histone modifications. The expressions for these genes are very low. The third cluster is very interesting. Genes in this cluster have strong modifications on both repressive mark (H3K27me3) and marks associated with activation (H3K9me1 and H4K20me1), yet the gene expressions are low for this cluster. The promoters for these genes probably correspond to the "poised" state as defined in [13]. The fourth cluster (11% of all genes) shows strong enrichment of activation marks, so the gene expressions are relatively high. The difference between the first and fourth clusters is that the genes in the fourth cluster lack the H3k9me1 and H4k20me1 modifications. Although these genes are mostly active, their expressions are on average lower than those in cluster 1. This is possibly due to the additive effects of different activating marks. The fifth cluster (8% of all genes) is enriched with both elongation (H3K36me3) and active marks (H3k9me1) but depleted of the canonical activating marks (H3k4me, H3k9ac, and H3k27ac). The expressions for these genes are relatively

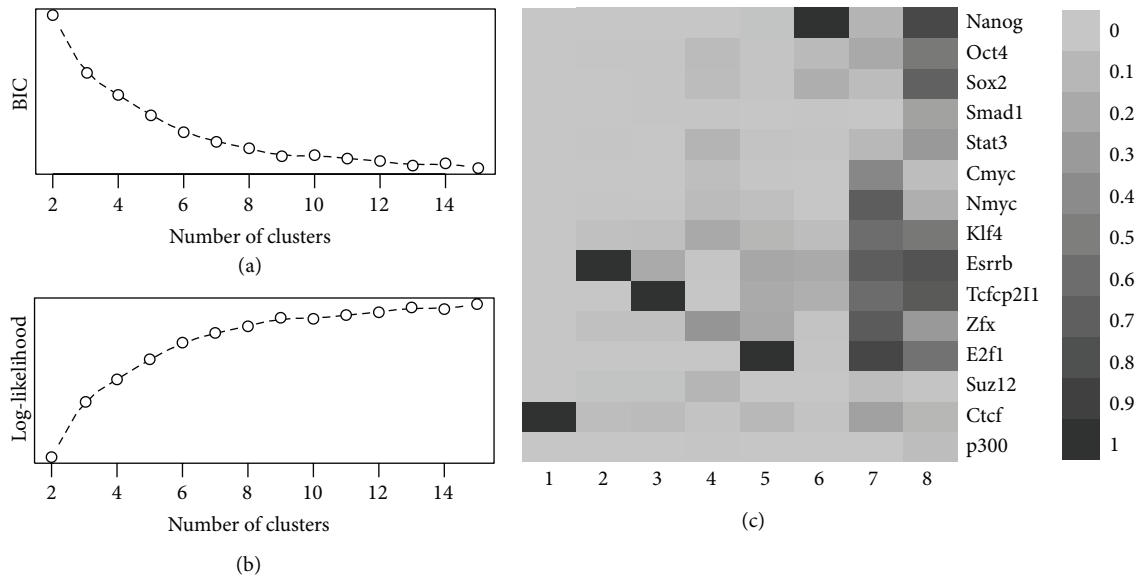


Figure 1: Model fitting results of mouse ES data. (a) BIC versus number of clusters. (b) Log-likelihood versus number of clusters. (c) Estimated $\widehat{\mathbf{Q}}$ represented as heatmap.

Table 1: The estimated model parameters from mouse ES data with 8 clusters. The $\widehat{\pi}$ row shows the estimated cluster sizes. The rest of the table shows $\widehat{\mathbf{Q}}$, the probability of occurrence of protein binding in different clusters.

Cluster	1	2	3	4	5	6	7	8
$\widehat{\pi}$	0.23	0.23	0.18	0.12	0.08	0.07	0.07	0.03
Nanog	0.00	0.00	0.00	0.00	0.03	1.00	0.13	0.85
Oct4	0.00	0.01	0.01	0.07	0.02	0.08	0.20	0.53
Sox2	0.00	0.01	0.01	0.07	0.02	0.17	0.07	0.69
Smad1	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.25
Stat3	0.00	0.01	0.01	0.12	0.02	0.02	0.10	0.31
Cmyc	0.00	0.00	0.00	0.05	0.01	0.00	0.43	0.06
Nmyc	0.01	0.01	0.01	0.08	0.05	0.00	0.71	0.16
Klf4	0.01	0.04	0.05	0.20	0.11	0.06	0.60	0.53
Esrrb	0.00	1.00	0.19	0.01	0.21	0.20	0.71	0.78
Tcfcp2I1	0.00	0.00	1.00	0.00	0.19	0.16	0.60	0.73
Zfx	0.01	0.04	0.04	0.34	0.21	0.03	0.73	0.31
E2f1	0.00	0.00	0.00	0.00	1.00	0.02	0.86	0.56
Suz12	0.00	0.03	0.03	0.12	0.01	0.01	0.06	0.02
Ctcf	1.00	0.06	0.08	0.02	0.10	0.03	0.25	0.11
p300	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.06

high. This potentially suggests alternative mechanisms for gene activation.

In general, the results from this example agree with existing knowledge of the effects of histone modification on gene expression. Yet, they provide some new biological findings worth further exploring. This example further illustrates that a quick analysis by `giClust` can provide biologically meaningful results.

4. Conclusions

With the increasing popularity of ChIP-seq technology, a large amount of data in the form of genomic intervals is generated to represent protein binding or histone modification regions under different biological contexts. Up to date, there has not been an intuitive method available to assess the spatial relationships in these data. To meet this need, we have proposed a method and developed a software tool called "`giClust`" to characterize the cooccurrence pattern of multiple sets of genomic intervals. The method is based on a finite mixture model. It clusters the regions of interest into several groups according to the cooccurrence patterns of inputs. The model parameters sufficiently capture the overlapping structures in input data. Simulation results showed the method accurately estimates the patterns. Two real datasets showed that the proposed method provides

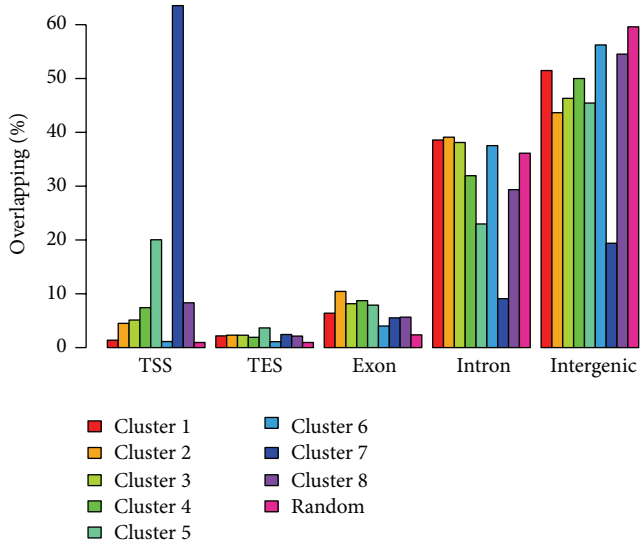


Figure 2: Percentage of genomic regions in each cluster overlapping certain genomic tracks. For example, 63% of the regions in cluster 7 overlap TSS, whereas the expected percentage is only 1%.

Table 2: The estimated model parameters from K562 histone data with 5 clusters.

Cluster	1	2	3	4	5
$\widehat{\pi}$	0.44	0.21	0.16	0.11	0.08
H3k27ac	0.95	0.00	0.01	0.74	0.09
H3k27me3	0.33	0.33	1.00	0.33	0.27
H3k36me3	0.88	0.05	0.02	0.45	0.85
H3k4me1	0.93	0.03	0.20	0.63	0.42
H3k4me2	1.00	0.03	0.18	0.99	0.07
H3k4me3	0.99	0.00	0.12	0.97	0.02
H3k9ac	1.00	0.00	0.02	0.89	0.13
H3k9me1	0.98	0.03	0.79	0.20	0.96
H4k20me1	0.87	0.02	0.87	0.17	0.74

biologically meaningful results. Because there are no existing method or software serving the same purpose, comparisons cannot be performed.

It is important to distinguish `giClust` from ChromHMM [7, 13], although they share some similarities. The goal of ChromHMM is to segment the genome according to the combinatory patterns of multiple tracks (in their case, histone modifications), whereas `giClust` aims to explore the correlation of different tracks. ChromHMM, based on a hidden Markov model (HMM), considers the transitions between consecutive genomic windows, whereas `giClust` assumes that the regions of interest are independent. The only similarity is the way to model the joint likelihood of multiple tracks of each genomic region, where both methods use product of Bernoulli distributions. Overall `giClust` and ChromHMM solve different problems and have no dependence on each other.

The proposed method is designed to work for genomic intervals instead of raw sequencing counts. As discussed in [13], converting the raw count data to presence/absence of peaks before clustering reduces the parameter space and produces more stable and interpretable results. It is also much more computationally efficient since the peak calling results (in the form of genomic intervals) are several orders of magnitude smaller than the raw ChIP-seq sequencing files. Moreover, because different ChIP-seq experiments have distinct technical specifications, clustering directly on raw counts tends to be influenced more by the ones with higher signal to noise ratio. Nevertheless, the method can be extended to include the distribution of raw read counts data in a hierarchical model similar to that one in [16]. This is future research topic worth exploring.

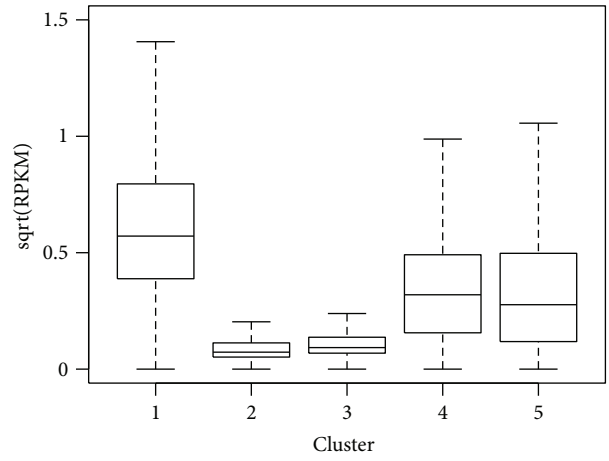


Figure 3: Expressions of genes in different clusters for K562 cell line.

Choosing the number of clusters K is an important step in the analysis. Traditional model selection methods such as BIC tend to favor larger models, which is often difficult to interpret. A software like ChromHMM requires that the user specify number of states. Similarly, we recommend an *ad hoc* procedure to choose K from the log-likelihood versus number of cluster curves. We have shown in simulation studies that such a method works reasonably well and is able to capture the major cluster component. However it is still subjective and relies on user input. Automating the process of choosing K is an important research topic for future works.

We use a finite mixture model in this study. An alternative method that we plan to explore is the infinite mixture model, such as a nonparametric Bayesian approach, which has been used previously in genomics research [17].

The proposed method has been implemented in an easy-to-use R package `giClust`. We expect the method and the software tool to provide an easy way for biologists to explore their ChIP-seq results and compare with public datasets.

Acknowledgments

Hao Wu is partially funded by PHS Grant UL1 R025008 from the Clinical and Translational Science Award program, National Institute of Health, National Center for Research

Resources. Zhaohui S. Qin is partially funded by NIH Grant R01HG005119.

References

[1] A. P. Bird, "CpG-rich islands and the function of DNA methylation," *Nature*, vol. 321, no. 6067, pp. 209–213, 1986.

[2] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[3] X. Chen, H. Xu, P. Yuan et al., "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells," *Cell*, vol. 133, no. 6, pp. 1106–1117, 2008.

[4] A. Favorov, L. Mularoni, L. Cope et al., "Exploring massive, genome scale datasets with the genometricorr package," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002529, 2012.

[5] M. Chikina and O. Troyanskaya, "An effective statistical evaluation of ChIPseq dataset similarity," *Bioinformatics*, vol. 28, no. 5, pp. 607–613, 2012.

[6] A. Q. Fu and B. Adryan, "Scoring overlapping and adjacent signals from genome-wide ChIP and DamID assays," *Molecular BioSystems*, vol. 5, no. 12, pp. 1429–1438, 2009.

[7] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization," *Nature Methods*, vol. 9, no. 3, pp. 215–216, 2012.

[8] X. Zeng, R. Sanalkumar, E. Bresnick, H. Li, Q. Change, and S. Keles, "jMOSAiCS: joint analysis of multiple ChIP-seq datasets," *Genome Biology*, vol. 14, article R38, 2013.

[9] A. Doi, I. H. Park, B. Wen et al., "Differential methylation of tissue-and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts," *Nature Genetics*, vol. 41, no. 12, pp. 1350–1353, 2009.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, pp. 1–38, 1977.

[11] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, pp. 461–464, 1978.

[12] R. Giancarlo and F. Utro, "Stability-based model selection for high throughput genomic data: an algorithmic paradigm," *Artificial Immune Systems*, pp. 260–270, 2012.

[13] J. Ernst and M. Kellis, "Discovery and characterization of chromatin states for systematic annotation of the human genome," *Nature Biotechnology*, vol. 28, no. 8, pp. 817–825, 2010.

[14] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[15] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: archive for high-throughput functional genomic data," *Nucleic Acids Research*, vol. 37, no. 1, pp. D885–D890, 2009.

[16] H. Wu and H. Ji, "JAMIE: joint analysis of multiple ChIP-chip experiments," *Bioinformatics*, vol. 26, no. 15, pp. 1864–1870, 2010.

[17] Z. S. Qin, "Clustering microarray gene expression data using weighted Chinese restaurant process," *Bioinformatics*, vol. 22, no. 16, pp. 1988–1997, 2006.