Tsinghua Science and Technology

Volume 19 | Issue 6 Article 1

2014

Opportunities for Computational Techniques for Multi-Omics Integrated Personalized Medicine

Yuan Zhang
Department of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China.

Yue Cheng
Department of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China.

Kebin Jia
Department of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China.

Aidong Zhang
Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260-2500, USA.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/10 pp545-558
Volume 19, Number 6, December 2014

Opportunities for Computational Techniques for Multi-Omics Integrated Personalized Medicine

Yuan Zhang, Yue Cheng, Kebin Jia*, and Aidong Zhang

Abstract: Personalized medicine is defined as "a model of healthcare that is predictive, personalized, preventive, and participator" and has very broad content. With the rapid development of high-throughput technologies, an explosive accumulation of biological information is collected from multiple layers of biological processes, including genomics, transcriptomics, proteomics, metabonomics, and interactomics (omics). Implementing integrative analysis of these multiple omics data is the best way of deriving systematical and comprehensive views of living organisms, achieving better understanding of disease mechanisms, and finding operable personalized health treatments. With the help of computational methods, research in the field of biology and biomedicine has gained tremendous benefits over the past few decades. In the new era of personalized medicine, we will rely more on the assistance of computational analysis. In this paper, we briefly review the generation of multiple omics and their basic characteristics. And then the challenges and opportunities for computational analysis are discussed and some state-of-art analysis methods that were recently proposed by peers for integrative analysis of multiple omics data are reviewed. We foresee that further integrated omics data platform and computational tools would help to translate the biological knowledge to clinical usage and accelerate development of personalized medicine.

Key words: personalized medicine; translational bioinformatics; multi-omics integration

1 Introduction

Personalized medicine, as one of the most important biomedicine developments, has arrived together with the emergence of Translational BioInformatics (TBI) which builds a bridge between clinical practice and fundamental findings of biomedicine[1]. Personalized medicine has a broad and inclusive definition, "a model of healthcare that is predictive, personalized, preventive, and participator", or "P4 Medicine"[2, 3]. Personalized medicine optimizes patient care and improves prescription of treatments with higher likelihood of success by utilizing the patient's own genetic information and personal clinical data[4]. The aim of personalized medicine is to better tailor cost-effective strategies using specific techniques based on individual biological knowledge. It is becoming a hot topic during last five years as shown in Fig. 1, for example, which illustrates the number of papers that are involved in the term of "personalized medicine" and indexed in PubMed. Along with increasing attention that the topic receives, the challenges of translating omics knowledge to real clinical practices have emerged and the assistant of computational methods is in great need.

The genomics era has yielded great advances in deciphering some human diseases. However, the immense complexity of biological mechanistic

• Yuan Zhang, Yue Cheng, and Kebin Jia are with the Department of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China. E-mail: zhangyuan@emails.bjut.edu.cn; ycheng12@emails.bjut.edu.cn; kebinj@bjut.edu.cn.
• Aidong Zhang is with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260-2500, USA. E-mail: azhang@buffalo.edu.
* To whom correspondence should be addressed.
Manuscript received: 2014-06-09; accepted: 2014-06-16

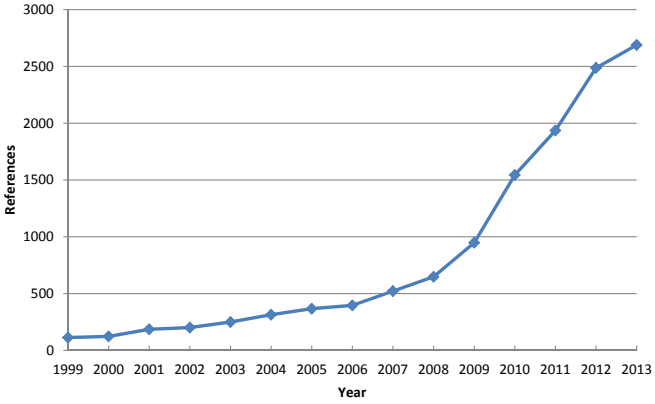


Fig. 1 Number of results found in PubMed for the term "personalized medicine".

continues to challenge researchers' capacity of finding valuable answers for complicate diseases. Researchers began to recognize the erroneous assumption that any given human disease can be distilled down to a single molecular determinant and a single "magic bullet" can be identified as the therapeutic target. Although biologists did find some successful treatments based on this assumption during the past few decades, this assumption has been facing increasingly limitations as witnessed by the lack of growth of new drugs approved for human use, remaining effectively constant at 20 to 25 new drugs per year on average for the past 20 years[5, 6].

Moreover, the genetic factors are not the only reason for bringing on an attack of even a simple and classical Mendelian disorder. People carrying the same mutation are not necessarily to suffer from the same prophetic disease. For example, we have learned that retinoblastoma is caused by mutations in the Rb gene, but not all people who carry this mutation will suffer from retinoblastoma. Even two sisters in the family with the same mutation inherited from their parents can develop different phenotypes and have different effects of disease. The reason underlying here is still not clear, but commonly it is believed that the whole genome and the environmental risk factors should be included into the research. In this regard, there are numerous other inherited diseases in which the same mutation is not expressed in all the individuals who carry it. Disease expression is influenced by disease-modifying genes, post-translational modification of the proteome, and environmental factors[7] and this is called penetrance by geneticists.

Thus, it is time to treat human diseases from a systematical view. With the development of high-throughput biological technology, multiple omic data are captured by biologists, including genomics, transcriptomics, proteomics, metabolomics, and interactomics. These data provide a potential way of conducting comparative analysis from different aspects, performing systematical research based on correlations of heterogeneous biological molecules in network models and seeking more insightful meaning of macromolecules in the context of dynamic progresses which can be modeled out of multiple omic data.

Although omic data are used in both industry and academia in the fields of bioinformatics and translational research, these data are extracted from exclusive experiments which show hardly any correlation among these heterogeneous biological molecules within the system of an organism. Hence, one of the challenges is to match the heterogeneous data and capture the correlations between them. The second challenge is to model a comprehensive representation of multiple omics of individuals using some more intelligent approaches like network modeling which naturally illustrate the relationships of different objects inside a cell. The challenge lies in extracting useful information out of these network models, either static networks or dynamic ones, and identifying the clinical meanings of these critical findings that can lead the researchers to find more practicable health care or biomedicines. The third but not the least important is to handle the big data analysis problem.

This paper will introduce multiple omics that are available for us to dig into and discuss specific computational barriers and current methods of implementing the personalized medicine research with these heterogeneous data. And then some representational applications in this area are introduced

from which we see the opportunities provided in this interdiscipline of molecular biology, clinical biology, and computational science.

2 Materials and Traditional Analysis in Biomedicine Research

For computational study, the following Genomics, Transceiptomics, Proteomics, Metabonomics, and Interactomics (omics) data can be used[8].

2.1 Genomics

Raw genomic data is basically referred to the sequence reads from a specific sequencing technique. The processed genomic data also include the mappings of the sequence reads to a reference genome, the variations detected against the reference genome, and other summarized categories based on the degree of computational modification and integration applied to the raw data[9]. Genetic research is to link each genotype to the specific phenotype to finally extract the reasons of diseases, design accurate genetic medicines which point to the specific genetic disorder, and implement better diagnosis and treatment at the end[10].

During the past several decades, continuous improvements in genomic technology, including Next-Generation Sequencing (NGS)[11,12], have led to a series of breakthroughs in the field of genetic biomedicine, especially to the individual genetic research[13]. NGS, with its unprecedented throughput, scalability, and speed, has greatly facilitated the study of individual genome. NGS has greater coverage of the genome or transcriptome and is more accurate in detection of rare variants, haplotypes, allele usage, insertions/deletions, splice variants, alternative start/stop sites, and RNA editing[14]. With the invent of NGS, the cost of sequencing a whole genome has been greatly pulled down and will reach 1000 USD in the near future under the cooperative effort of researches and industries all over the world[15]. The reduced price makes it possible for clinical use with personalized medicine concerned. It is foreseeable that NGS will be a great promoter for the technology of personalized medicine.

2.2 Transcriptomics

The genome-wide gene expression data is another data used to identify those genes which demonstrate significant changes under the impact of certain experimental conditions, for example, the presence of cancerous tumors[16]. The advent of microarray technology has made it possible to monitor the expression levels of thousands of genes in parallel and the experiment explores those genes which are differentially expressed in one set of samples relative to another, establishing potentially meaningful correlations between genes and specific conditions. Moreover, filtering out the non-differentially expressed genes can reduce the dimensionality of the data set and facilitate further analysis. The representative analysis of gene expression data can be divided into several categories according to different experimental objectives: (1) comparing gene expression profiles in different tissues[17,18]; (2) analyzing gene expression patterns in model systems[19], for example, when we do research about the cell cycle process in yeast Saccharomyces cerevisiae[20]; (3) extracting differential gene expression patterns in disease or pathogens[21]; and (4) studying gene expression in response to drug treatments[22]. Genome-wide gene expression analysis (transcriptomics) can therefore deliver a comprehensive view on all genes active at a given time in a given sample. Consequently, transcriptomics is suitable for the first "round of discovery" in regulatory networks and serves to put proteomic and metabolomic markers into a larger biological perspective.

2.3 Proteomics

The genomic data and transcriptomic data are immensely powerful, yet it is incomplete to deliver personalized medicine if we are not able to measure the relevant molecular phenotype of individuals in real time and over time[23]. Proteins offer the most potential for personalized medicine design as they show the real state of individual health or disease. Since diseases will affect the protein levels inside human body, and also the drug that is used to treat the disease will change the protein expression by different levels, proteomic data are used to analyze the different effects of drugs to individuals[24]. Protein expression can be derived from blood, tissue or other biological samples using techniques like 2D gels, immunoassays or the latest Mass Spectrometry (MS)[25]. Kim et al. [26] proposed a framework of predicting individual drug sensitivity in cancer using proteomic data. In their work, the proteomic information of cancer was collected by Reverse-Phase Protein Array (RPPA) technology and the drug sensitivity, i.e., drug resistance or

intolerance, served as training label in their Augmented Naive Bayesian Classifier (ANBC). Once the ANBC was trained, it could be used to predict the drug sensitivity for another individual. As one of the most instant measurement of health state, protein levels are attracting great attention for drug development and disease diagnose.

2.4 Metabonomics

Metabolic phenotype is the downstream of genetics and environmental effects which can be used to measure and analyze the disease or drug outcomes[27]. Metabonomics profile can be collected by Nuclear Magnetic Resonance (NMR) spectroscopy[28] and mass spectrometry.

The clinicians have been using metabolimics as biomarkers to diagnose and predict diseases[29]. It is also the final aim of metabolimics to identify effective biomarkers. For example, the patients with colorectal cancer[30] and normal samples were studied comparatively by the NMR-based metabolic profiling. They observed upregulated lipid-based resonances in pretreated patients that showed poisoned phenomenon after dosing. Their work shows that it is possible to predict toxicity and design best treatments for a patient by studying the predose patterns of metabolics. Several studies have also investigated the drug efficacy of cancers like the different responses of MCF7 breast cancer cells to treatment with docetaxel, which is used in the treatment of advanced or metastatic breast cancer[31]. Metabolic data have also been used to identify the prognostic biomarkers and stratify disease subtypes like rheumatoid arthritis, diabetes, and Alzheimer's disease.

2.5 Interactomics

Molecules inside cells don't perform their functions alone, but tend to interact with others for proper biological activities[32]. Varying molecular interactions exist among different biomedical families and compose varying interactiomic. Interactions between proteins are described by Protein-Protein Interaction (PPI) data which have been serving as the most important knowledge of studying proteins functions besides protein structure information. Also, protein-DNA interactions are studied to get the regulatory relationships between proteins and genes, and the nodes inside the network include transcription factors, chromatin regulatory proteins, and the targeted genes. The metabolic molecules can also be illustrated in a network where the chemical compounds are converted into each other by enzymes which have to bind their substrates physically. In these networks, all the related biological molecules are represented as vertices and the interactions as edges. Network, or graph, is a very systematical way of illustrating the interactomic information. Using graph theory and network mining technologies we are able to do more systematical analysis with these interactomic data.

The public available datasets for the above omics are shown in Table 1.

3 Computational Integrative Analysis of Multiple Omics in Personalized Medicine

The advanced biological measurements of multiple omics provide tons of data about diseases for the medical paradigm of "measure, diagnose, and treat", yet none of the various human omics are complete to make this challenge on its own because of the noise, incompleteness, and other tentative limitations. It is quite natural to analyze heterogeneous data integratively using either statistical or more intelligent computational methods and conquer the challenges that hinder us from "seeing" deeper biological mechanics of diseases.

In the workflow of data-driven personalized medicine research, the fundamental issue is to estimate the biological features of individual multi-omics data which is also the base of integration analysis of multiple omics, multiple platform, and multiple organisms. The second issue is to create data integration models according to different applications including identifying predictive and prognostic biomarkers for relevant treatments[33], predicting diseases, and recognizing disease states[34].

There are generally two directions of integrative analysis: On one side, we need to design efficient methods to interpret each of these data in their own right; and on the other side, careful integrative statistical or more systematical methods of multi-omics analysis should be designed considering the difficulties of ID conversion, object mapping, heterogeneous data handling and mining, and so on. Moreover, the correlations and causal relationships between different omics layers should also be addressed to get more insightful and systematical understanding of biological progress[35,36].

Table 1a Omics database (Components)

Data types	Online resource	Description	URL
Genomics	The Cancer Genome Atlas (TCGA)	Contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes.	http://www.broadinstitute.org/software/igv/tcga
	1000 Genome	The first project to sequence the genomes of a large number of people, and to provide a comprehensive resource on human genetic variation, including 2577 samples from all over the world.	http://www.1000genomes.org/data
Transcriptomics	Gene Expression Omnibus (GEO)	Microarray and SAGE-based genome-wide expression profiles.	http://www.ncbi.nlm.nih.gov/geo
	Stanford Microarray Database (SMD)	Microarray-based genome-wide expression data.	http://genome-www.stanford.edu/microarray
Proteomics	World-2DPAGE	A public standards-compliant repository for gel-based Proteomics data linked to protein identification published in the literature.	http://us.expasy.org/ch2d/2d-index.html
	The PRoteomics IDEntifications (PRIDE) database	A centralized, standards compliant, public data repository for proteomics data of 1670 projects and 33 230 assays, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.	http://www.ebi.ac.uk/pride/archive/
Metabolomics	The Human Metabolome Database (HMDB)	A freely available electronic database containing detailed information about small molecule metabolites found in the human body.	www.hmdb.ca/

Table 1b Omics database (Interactions)

Data types	Online resource	Description	URL
Protein-DNA interactions	Biomolecular Network Database (BIND)	Published protein-DNA interactions.	http://www.bind.ca/Action/
	Encyclopedia of DNA Elements (ENCODE)	Database of functional elements in human DNA.	http://genome.ucsc.edu/ENCODE/index.html
Protein-protein interactions	Database of Interacting Protein (DIP)	6678 articles for 76 844 protein interactions of 26 453 proteins.	http://dip.doe-mbi.ucla.edu/dip/Stat.cgi
	Biological General Repository for Interaction Datasets (BioGRID) (Till April 12, 2014)	42 396 publications for 737 271 raw protein and genetic interactions from major model organism species.	http://thebiogrid.org/
	InAct (Till April 12, 2014)	12 659 publications for 445 718 protein, compound, nucleic acids and genetic interactions of 82 232 interactors from 32 878 experiments.	http://www.ebi.ac.uk/intact/

3.1 Single-sample estimation

Many biotechnology high-throughput platforms for gene expression profiling, sequencing, and protein expression profiling are invented during the past few years. These platforms are designed to technology-specific biases and produce raw data of distinct distributions. To get an integrative analysis of multiple platform data, the fundamental challenge is to solve the normalization problem[37].

Bengtsson et al. [38] proposed a single-sample multi source normalization method that brought different copy numbers from multiple platforms to the same mean levels. This method assumed a non-linear measurement function among the observed copy numbers. Using a kernel-based estimation procedure, the complete copy numbers at a common set of loci are obtained for the further normalization function estimation. And to model the non-linear normalization

function, a principal curves[39] based on PCA analysis was adopted. Single-sample normalization methods for microarray, qPCR, and RNA-Seq experiments[37, 40, 41] are also developed with great potential of assisting the personalized analysis. Although it is still unproved of the validity of comparing platforms, labs, or algorithms based on these normalizing estimation, this work is an effective attempt.

The single-sample estimation presents another major problem of "large p, small n". In the framework of personalized medicine prediction, techniques of dimensionality reduction, sparsity estimation, rule-based models, and expert systems can best serve the purpose of translational analysis. Some researchers use knowledge-driven strategy and adopt existing datasets from population studies to compensate the insufficient sample information. For example, Liu et al.[42] discussed the possibility of identifying critical transitions as the biomarkers of complex diseases based on a single sample by exploiting existing disease samples.

3.2 Jointly comparative analysis of multiple omics

Tremendous studies have shown the outstanding power of the jointly comparative analysis of multiple omics data than separate analysis[35, 43-45].

Liu et al.[35] proposed a comparative analysis framework of jointly analyzing multiple omics data including the miRNA, gene, and protein expression in which these data were normalized and integrated into a joint matrix and decomposed with factor analysis and linear discriminate analysis methods to extract the useful factors inside the multiple types of information for cancer subtype identification. They reported great improvement of identifying mutations drivers in cancers than the separate analysis with each omics layer or other traditional joint analysis methods like SAM. By modeling multiple omics in microbes, Cotton and Reed[46] proposed a multi-omics based optimization method for estimating the kinetic rate parameters of constraining metabolic fluxes in *Escherichia coli*. Using this integrative but simplified kinetic model, they showed us a more general way of predicting the intracellular fluxes and biomass yield and identifying potential metabolic limitations through the integrated analysis of multi-omics datasets.

In the research of recombinant protein producing, Dietmair et al.[44] implemented a jointly comparison of the combination of transcriptomics, metabolomics, and fluxomics of a recombination protein producing Hek293 cell line and its parental cell line to reveal a surprising observation that the consumption of glucose of the producer cell line is less than the non-producer cell line despite the additional burden of recombinant protein production, and the same happened to the glutamine consumption which was also lower than the non-producer cell line. With the joint comparison of genomics, they also found that the majority of genes involved in oxidative phosphorylation and broad cellular functions were down-regulated in producer cultures. With these findings, the researchers presented the speculation that there must be a large scale of cellular network adaption for the recombinant protein production which made bountiful resources available for this specific progress possible. With the indication of endoplasmic reticulum stress, they suspected that the possible hurdle of recombinant protein production might lie at the point of protein folding and assembly.

Based on these researches, we can say that the jointly comparative analysis of multiple layers of omics provides more insightful knowledge about the diseases or biological progresses and gives us a more comprehensive view of the biological system which, with great potential, will help us to develop more accurate and lower-cost treatments and health care in clinical activities.

3.3 Network-based analysis

Networks are the natural representations of the biological systems inside a cell by illustrating the interaction relationships among the molecules. Network analysis or graph analysis is a systematical way of studying the mechanics of biological progresses within human body[47]. As introduced in Section 2, the interactomics data have been used in the personalized medicine research. However, the current interatomic databases are far from complete for fully modeling the complex biological progresses. For example, there are no genes or other molecules other than proteins that are included in PPI networks. And for the other thing, the current biological networks are generally accumulated from all kinds of experiments, but without temporal information therefore with these networks, it is impossible to describe the dynamic changes of biological processes. In this section, we introduce the challenges and also the opportunities for computational technologies in the content of network modeling of multiple omics analysis.

3.3.1 Heterogeneous component networks

As in the multiple omics data, there are heterogeneous components. One way of representing these omic information is to find the correlations between different networks of single sort of components, and the other is to integrate these components into one comprehensive network which shows the complex relationships among them. We observe tremendous challenges from both perspectives.

For the first line of research, there are many researchers who adopted compensational information to correct their vital ones, like the works that use gene expression to confirm the interactions between proteins[48-50]. The co-expression correlations are computed and added into the networks as the edge weights. Besides, some works have implied that similarly to the genome conservation, the protein interactions have conserved along the evolution to a great extent[51-53]. Hence there were some researchers used interactomic data from other organisms to verify the targeted PPI data[54]. As shown in Fig. 2, Du et al.[54] adopted three auxiliary interaction networks of other species — *C.elegans*, *H.plylori*, and *S.cerecisiae* to denoise the target interaction network. Using the links of auxiliary networks and the similarity of the nodes between the auxiliary networks and target networks, link propagation techniques were performed to predict the links in the target network. The effectiveness of the cross-organism analysis method was verified on known PPI datasets by blocking partial links for prediction.

In Gat-Viks' work[55], they proposed a method of constructing a regulatory network out of multiple biological networks including heterogeneous components (mRNA, proteins, and metabolites). Using lysine biosynthesis pathway as a research case,

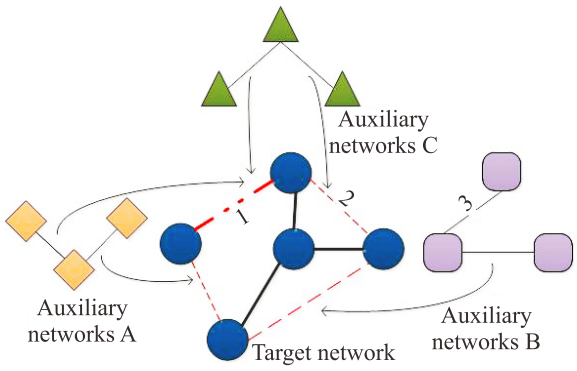


Fig. 2 The structure of mapping the external labeled information to the target network[54].

they claimed that the method provided a prediction methodology in the presence of cycles and a polynomial time, constant factor approximation for learning the regulation of a single entity. In this line of work, another basic process is to match the entities in different omics[56] and some public tools are available for this problem, including the DAVID[57], MatchMiner[58], and other ID mapping tools from public biological databases[59].

For the second line of research, which is more complicated, some works attempted to construct a comprehensive network for heterogeneous components. Yeang and Vingron[60] studied the gene regulation and metabolic reactions in *Escherichia coli* by adding links specifying the feedback control from the substrates of metabolic reactions to enzyme gene expressions. A joint network was constructed in their work which adopted both genes and metablites. They inferred the feedback links between metabolites and transcription factors by fitting the perturbation data in a probabilistic graphical model on one side, and on the other side they directly encoded general hypotheses about metabolic enzyme regulations with enzyme expression data to get the feedback control from metabolites and corresponding genes. The perturbation data included 5 sources covering from gene expression data to metabolics[60]. This work provides a way of building a bridge between the gene regulation and metabolic reaction systems.

There were some other researchers proposing to simultaneously process multiple heterogeneous networks and identify the interested patterns of multi-omics data in a unified fashion. In Mosca and Milanesis' work[61], based on PPI interactions, different views of interactome are constructed by adding different weights to the PPI network and a multi-objective optimization model was developed to identify networks that are optimal according to several scores which make the optimization results of biological meaning, for example, they identified the differentially expressed networks of PPIs that regulated the basal cancer markers which are believed to be related to breast cancer.

3.3.2 Dynamic networks

Biological systems are inherently dynamic in nature. However, most of the public biological networks, like PPI datasets, are simply the aggregation of all the possible interactions occurring in all

examined conditions[62], and the temporal changes of these networks are ignored. In PPI networks, proteins have active forms and perform their functions at specific time courses and environments. Similarly metabolites can be present or absent to certain biological progress. Dynamic biological networks illustrate the real state of molecules inside human body and are the best sources for personalized medicine research. Analysis of dynamic networks falls into several categories similar to static network analysis, including interaction detection also known as dynamic PPI network construction, complex identification, and critical node detection.

Extracting dynamic PPI networks from the known PPI datasets is the basic task for dynamic analysis. Basically, there are two main directions to construct dynamic PPI networks. One is based on the differential co-expression correlations. Researchers[63] have shown that the highly positive co-expressed proteins tend to form most static modules. At the center of these modules there are some hubs with high degree namely "party" hubs, which tend to appear at all of the time points. Meanwhile, there are some lower positive co-expressed proteins which are believed to interact at some particular time points. Among these proteins the hubs are called "date" hubs and are believed to be the cause of dynamic interactions and also may induce aberrant pathways and molecular disorders. Taylor et al.[64] observed a multi-modal distribution of gene expression correlation coefficients using a literature-curated source. The other way to construct dynamic PPI networks is based on expression variance[65]. They determine the peak time points of expression for each protein. If a protein is at its peak point, it is said to be in its active form and can interact with its active neighbors. Based on this assumption, scored gene expression activity is computed using a single threshold[66] or a systematical threshold[67] in literature. In addition, some researchers argue that the gene expression data contain far more noise that will induce unauthentic factors. For example, the genes are sent into a filter that defines a criterion for genes of being dynamic or stable in Xiao et al.'s paper[68], and the stable ones are left out of the subsequent construction of dynamic networks.

Complex detection in dynamic networks is more challenging than the problem in static networks because both the number of proteins and complexes are changing during the biological progress. A heuristic dynamic module detection method was presented by Jin et al.[69], which performs a Depth-First Search (DFS) style enumeration in temporal networks to find the dynamic modules, the connected subgraphs, in both the temporal network and the corresponding second order network. Some other researchers proposed effective algorithms to overtake the challenge of changing number of proteins and complexes across biological processes. One of these works is based on the Bayesian graphical model which incorporates two priors to approximate the changing number of complexes at different times[70]. These researchers showed that dynamic network complexes or modules are more biologically meaningful than static modules by evaluating the functional homogeneity of dynamic modules with Gene Ontology (GO) annotations[69,70]. Another characteristic for the dynamic modules is that a majority of them are very sparse which is very challenging to detect for the topology-based analysis of static network. Moreover, there is an assumption about the smoothness among adjacency networks that is usually utilized in functional module detection for dynamic network[71]. Dynamic network analysis can facilitate the identification of likely important insights that are otherwise easily overlooked.

The other application of dynamic network analysis in personalized medicine research is to identify critical biomarkers using the dynamic properties. Dynamic network biomarkers show higher or lower expression compared with the normal cases, meantime they have time dependent stronger or weaker interactions with their neighbors. It has considered as one of powerful ways to detect the bifurcation of gene or protein interactions, indicating the early change of biomarkers and predicting the occurrence of diseases[72]. Dynamic network biomarkers show the advantage of demonstrating pathophysiological changes at different stages and periods. The disease specificity of dynamic network biomarkers should be validated by the integration with clinical informatics which translates clinical descriptive information on complaints, sign, symptoms, biochemical analyses, imaging, and therapies into the digital data[73]. One of the most challenges is to detect the abnormal changes of network structures. It is assumed that biomarkers are a minority group in networks and this assumption is true in most cases and treats the biomarker detection of dynamic networks as an anomaly detection

problem[74]. Data mining technologies including graphic probabilistic model, clustering, and network comparison, are basic algorithms that can be utilized in the dynamic network mining problem.

3.4 Big data analysis

In the past decade, biologists experienced an era of data "Tsunami" in which the amount of biological data accumulated at a rate that exceeds Moore's Law and the acceleration continues at the scale and at multiple levels, i.e., the multiple omics[75]. With the 1000 USD Genome approaching, it will be quite convenient to get the genome of individual patients[76]. Just a single human genome takes 140 GB in size. Hence, the production of petabytes of omics data is of great challenge to the existing computational technology in the respects of data storage and maintenance, transfer and quality control, especially data integration and data mining analysis[77]. New frameworks and powerful computational tools must be invented to tackling the data mountains.

Newly developed open source software framework for parallel computation, called Hadoop, is one of the powerful platforms to deal with petabytes data[78, 79]. This platform provides a foundation for scaling to petabyte scale data warehouses on Linux clusters, providing fault-tolerant parallelized analysis on such data using a programming style named MapReduce. Except for the criteria about system components, data location, Fault-tolerant, shared-nothing architecture, and the reliability guarantee, the MapReduce paradigm is core technology of realizing parallel computation in Hadoop. There are two steps: Map and Reduce, where each step is done in parallel and operating on sets of key-value pairs. These two steps are separated by the data transfer between nodes in the cluster[80]. In the MapReduce paradigm which is shown in Fig. 3, input data are loaded to the processing cluster in HDFS file management system. And these files are evenly distributed across all nodes. For the mapping phase, each node runs equivalent mapping tasks, which means all mappers are identical and any mapper can process any input file. At the end of Map phase, we get a set of records of key-value pairs which must be exchanged between machines to send all values with the same key to a single reducer. Then in the Reduce phase, a distinct task is performed on a single node. The Reduce stage produces another set of key-value pairs, as final output.

Hadoop supports programming to each computational task's need, which is quite different from MPI-based parallel programming[82]. There are some successful implementations of bioinformatics tools like BLAST and Gene Set Enrichment Analysis (GSEA) in Hadoop. Although Hadoop is designed to run on industry-standard hardware, an ideal cluster configuration is required to get it run stably and efficiently which also means expensive investment. An economical way of implementing the parallel computing is to rent commercial cloud hosting services like Amazon Elastic Compute Cloud (EC2)[83], Google AppEngine[84], and Microsoft Azure[85]. A number of bioinformatics systems have been developed on the top of cloud computing infrastructure. Table 2 shows

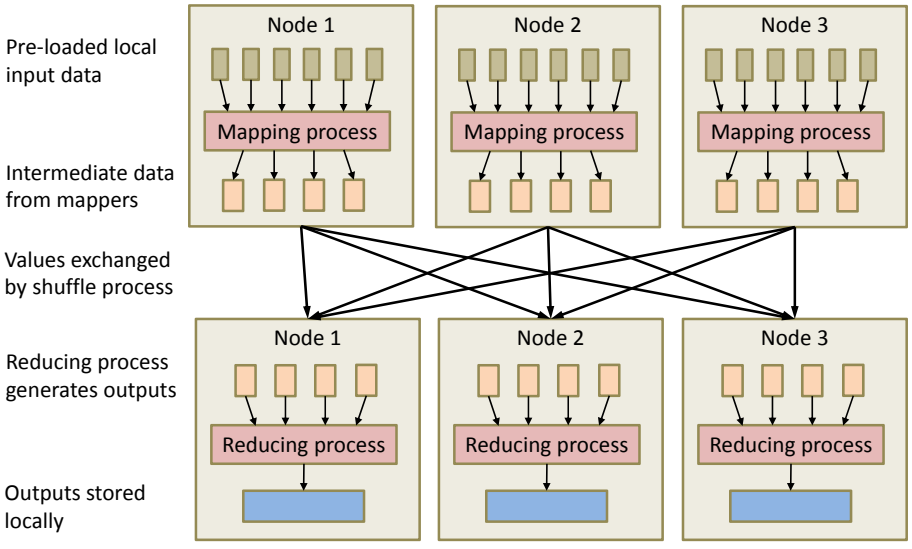


Fig. 3 High-level MapReduce pipeline[81].

Table 2 Current frameworks for big bio-data.

Infrastructure	Description	Platform	Reference
Cloudburst	Map next-generation short read sequencing data to a reference genome for SNP discovery and genotyping	Hadoop	[90]
Crossbow	Whole genome resequencing analysis and SNP genotyping from short reads	EC2	[91]
Contrail	De novo assembly from short sequencing reads		[92]
Myrna	Calculate differential gene expression from large RNA-seq data sets	Elastic MapReduce / a local Hadoop-based cluster	[93]
Biodoop	Three qualitatively different algorithms: BLAST, GSEA and GRAMMAR.	Hadoop	[80]
CloudBLAST	A parallelized version of the NCBI BLAST2 algorithm	Hadoop	[94]
Bionimbus	A cloud for managing, analyzing, and sharing large genomics datasets		[95]
BioVLAB-MMIA	Integrated analysis of microRNA and mRNA expression data	EC2 and S3	[86]

some current frameworks for big bio-data analysis. We see that these frameworks most focus on only one kind of omics, except the BioVLAB-MMIA which compares inversely correlated expression patterns between microRNA and mRNA and is built upon the Open Grid Computing Environments workflow suite tools and Amazon computing cloud resources[86]. More and more institutions are paying attention on using Hadoop on biological data integration, for example, the Environmental Molecular Sciences Laboratory[87], a national user facility located at the U.S. Department of Energy's Pacific Northwest National Laboratory (PNNL), and the U.S. Dept. of Energy[88]. This is merely a beginning of integration analysis of big multi-omics. A wide range of bioinformatics applications maintaining good computational efficiency, scalability, and ease of maintenance are needed in the future. Also facilitated by the programming support characteristic, some advanced data mining and network modeling algorithms can be transformed to parallel fashion and run on MapReduce clusters[89].

4 Conclusions

Over the past few decades, computational methods have greatly helped the analysis of biological information. In the new post-genomic era, computational analysis will play an even vital part in integrating multiple omics data to realize personalized medicine technology. In this paper, we briefly reviewed the generation of multiple omics data and their basic characteristics. And then the challenges and opportunities for computational analysis are discussed and some state-of-art analysis methods of implementing integrative analysis of multiple omics data are reviewed. Under their specific applications, the reviewed computational multi-omics integration analysis provides insightful understanding of diseases or drug reactions, identifies biomarkers and critical targeted functional molecules, and classifies disease subtypes. We foresee that further integrated omics data platform and computational tools would help to translate the biological knowledge to clinical usage and accelerate development of personalized medicine. What is more, although this paper focused on the multiple omics data integration, clinical diagnosis, patient screening, and other prior knowledge about patients can be utilized in a more broader scenario of personalized medicine.

Acknowledgements

This paper was supported by the Project for the Innovation Team of Beijing, the National Natural Science Foundation of China (No. 81370038), the Beijing Natural Science Foundation (No. 7142012), the Science and Technology Project of Beijing Municipal Education Commission (No. km201410005003), the Rixin Fund of Beijing University of Technology (No. 2013-RX-L04) and the Basic Research Fund of Beijing University of Technology. The authors declare that they have no competing interests.

References

[1] C. L. Overby and P. Tarczy-Hornoch, Personalized medicine: Challenges and opportunities for translational bioinformatics, *Per. Med.*, vol. 10, no. 5, pp. 453-462, 2013.

[2] S. Olson, S. H. Beachy, C. F. Giammaria, and A. C. Berger, *Integrating Large-Scale Genomic Information into Clinical Practice: Workshop Summary*. The National Academies Press, 2012.

[3] L. Hood and M. Flores, A personal view on systems

medicine and the emergence of proactive fP4g medicine: Predictive, preventive, personalized and participatory, *New Biotechnology*, vol. 29, no. 6, pp. 613-624, 2012.

[4] I. S. Chan and G. S. Ginsburg, Personalized medicine: Progress and promise, *Annual Review of Genomics and Human Genetics*, vol. 12, no. 1, pp. 217-244, 2011.

[5] B. Munos, Lessons from 60 years of pharmaceutical innovation, *Nature Reviews Drug Discovery*, vol. 8, no. 12, pp. 959-968, 2009.

[6] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, How to improve r&d productivity: The pharmaceutical industry's grand challenge, *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 203-214, 2010.

[7] C. P. Webb and D. M. Cherba, Systems biology of personalized medicine, in *Bioinformatics for Systems Biology*. Springer, 2009, pp. 615-630.

[8] M. Kussmann, F. Raymond, and M. Affolter, Omicsdriven biomarker discovery in nutrition and health, *Journal of Biotechnology*, vol. 124, no. 4, pp. 758-787, 2006.

[9] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson, Making sense of cancer genomic data, *Genes & Development*, vol. 25, no. 6, pp. 534-555, 2011.

[10] Y. Guan, C. L. Ackert-Bicknell, B. Kell, O. G. Troyanskaya, and M. A. Hibbs, Functional genomics complements quantitative genetics in identifying diseasegene associations, *PLoS Computational Biology*, vol. 6, no. 11, p. e1000991, 2010.

[11] J. Shendure and H. Ji, Next-generation dna sequencing, *Nature Biotechnology*, vol. 26, no. 10, pp. 1135-1145, 2008.

[12] M. L. Metzker, Sequencing technologiesthe next generation, *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31-46, 2010.

[13] G. Chen and T. Shi, Next-generation sequencing technologies for personalized medicine: Promising but challenging, *Science China: Life Sciences*, vol. 56, no. 2, pp. 101-103, 2013.

[14] I. Toma, G. St Laurent, and T. A. McCaffrey, Toward knowing the whole human: Next-generation sequencing for personalized medicine, *Personalized Medicine*, vol. 8, no. 4, pp. 483-491, 2011.

[15] H. Li and N. Homer, A survey of sequence alignment algorithms for next-generation sequencing, *Briefings in Bioinformatics*, vol. 11, no. 5, pp. 473-483, 2010.

[16] A. Zhang, *Advanced Analysis of Gene Expression Microarray Data*. World Scientific, 2006.

[17] N. Azad, A. K. V. Iyer, and Y. Rojanasakul, DNA microarrays in drug discovery and development, in *Biopharmaceutical Drug Design and Development*. Springer, 2008, pp. 47-66.

[18] D. Amaratunga, J. Cabrera, and Z. Shkedy, *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*. John Wiley & Sons, 2014.

[19] A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, and P. O. Brown, Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p, *Molecular Biology of the Cell*, vol. 12, no. 10, pp. 2987-3003, 2001.

[20] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273-3297, 1998.

[21] P. A. Bryant, D. Venter, R. Robins-Browne, and N. Curtis, Chips with everything: DNA microarrays in infectious diseases, *The Lancet Infectious Diseases*, vol. 4, no. 2, pp. 100-111, 2004.

[22] C. Debouck and P. N. Goodfellow, DNA microarrays in drug discovery and development, *Nature Genetics*, vol. 21, pp. 48-50, 1999.

[23] F. Steele and L. Gold, Taking measure of personalized medicine: The proteome, *Personalized Medicine*, vol. 10, no. 2, pp. 177-182, 2013.

[24] R. Kellner, Proteomics. Concepts and perspectives, *Fresenius' Journal of Analytical Chemistry*, vol. 366, nos. 6-7, pp. 517-524, 2000.

[25] B.-L. Adam, A. Vlahou, O. J. Semmes, and G. L. Wright Jr, Proteomic approaches to biomarker discovery in prostate and bladder cancers, *Proteomics*, vol. 1, no. 10, pp. 1264-1270, 2001.

[26] D.-C. Kim, X. Wang, C.-R. Yang, and J. Gao, A framework for personalized medicine: Prediction of drug sensitivity in cancer by proteomic profiling, *Proteome Science*, vol. 10, no. Suppl 1, p. S13, 2012.

[27] J. Sun, R. D. Beger, and L. K. Schnackenberg, Metabolomics as a tool for personalizing medicine: 2012 update, *Personalized Medicine*, vol. 10, no. 2, pp. 149-161, 2013.

[28] L. M. Vandersypen, M. Steffen, G. Breyta, C. S. Yannoni, M. H. Sherwood, and I. L. Chuang, Experimental realization of shor's quantum factoring algorithm using nuclear magnetic resonance, *Nature*, vol. 414, no. 6866, pp. 883-887, 2001.

[29] D. L. Mendrick and L. Schnackenberg, Genomic and metabolomic advances in the identification of disease and adverse event biomarkers, *Biomarkers in Medicine*, vol. 3, no. 5, pp. 605-615, 2009.

[30] A. Backshall, R. Sharma, S. J. Clarke, and H. C. Keun, Pharmacometabonomic profiling as a predictor of toxicity in patients with inoperable colorectal cancer treated with capecitabine, *Clinical Cancer Research*, vol. 17, pp. 3019-3028, 2011.

[31] M. Bayet-Robert, D. Morvan, P. Chollet, and C. Barthomeuf, Pharmacometabolomics of docetaxeltreated human mcf7 breast cancer cells provides evidence of varying cellular responses at high and low doses, *Breast Cancer Research and Treatment*, vol. 120, no. 3, pp. 613-626, 2010.

[32] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, et al., Proteome survey reveals modularity of the yeast cell machinery, *Nature*, vol. 440, no. 7084, pp. 631-636, 2006.

[33] S. Matsui, Genomic biomarkers for personalized medicine: Development and validation in clinical studies, *Computational and Mathematical Methods in Medicine*, vol. 2013, p. 865980, 2013.

[34] J. A. Dawson and C. Kendziorski, Survival-supervised latent dirichlet allocation models for genomic analysis of time-to-event outcomes, arXiv preprint arXiv:1202.5999, 2012.

[35] Y. Liu, V. Devescovi, S. Chen, and C. Nardini, Multilevel omic data integration in cancer cell lines: Advanced annotation and emergent properties, *BMC Syst. Biol*., vol. 7, p. 14, 2013.

[36] K. Arakawa and M. Tomita, Merging multiple omics datasets in silico: Statistical analyses and data interpretation, in *Systems Metabolic Engineering*. Springer, 2013, pp. 459-470.

[37] S. R. Piccolo, M. R. Withers, O. E. Francis, A. H. Bild, and W. E. Johnson, Multiplatform single-sample estimates of transcriptional activation, in *Proceedings of the National Academy of Sciences*, vol. 110, no. 44, pp. 17778-17783, 2013.

[38] H. Bengtsson, A. Ray, P. Spellman, and T. P. Speed, A single-sample method for normalizing and combining full resolution copy numbers from multiple platforms, labs and analysis methods, *Bioinformatics*, vol. 25, no. 7, pp. 861-867, 2009.

[39] T. Hastie and W. Stuetzle, Principal curves, *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502-516, 1989.

[40] S. R. Piccolo, Y. Sun, J. D. Campbell, M. E. Lenburg, A. H. Bild, and W. E. Johnson, A single-sample microarray normalization method to facilitate personalized-medicine workflows, *Genomics*, vol. 100, no. 6, pp. 337-344, 2012.

[41] K.-A. Lê Cao, F. Rohart, L. McHugh, O. Korn, and C. A. Wells, Yugene: A simple approach to scale gene expression data derived from different platforms for integrated analyses, *Genomics*, vol. 103, no. 4, pp. 239-251, 2014.

[42] R. Liu, X. Yu, X. Liu, D. Xu, K. Aihara, and L. Chen, Identifying critical transitions of complex diseases based on a single sample, *Bioinformatics*, p. btu084, 2014.

[43] S. H. Yoon, M.-J. Han, H. Jeong, C. H. Lee, X.-X. Xia, D.-H. Lee, J. H. Shim, S. Y. Lee, T. K. Oh, and J. F. Kim, Comparative multi-omics systems analysis of Escherichia coli strains b and k-12, *Genome Biol*., vol. 13, no. 5, p. R37, 2012.

[44] S. Dietmair, M. P. Hodson, L.-E. Quek, N. E. Timmins, P. Gray, and L. K. Nielsen, A multi-omics analysis of recombinant protein production in hek293 cells, *PloS One*, vol. 7, no. 8, p. e43394, 2012.

[45] T. Tebaldi, A. Re, G. Viero, I. Pegoretti, A. Passerini, E. Blanzieri, and A. Quattrone, Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells, *BMC Genomics*, vol. 13, no. 1, p. 220, 2012.

[46] C. Cotten and J. L. Reed, Mechanistic analysis of multiomics datasets to generate kinetic parameters for constraintbased metabolic models, *BMC Bioinformatics*, vol. 14, no. 1, p. 32, 2013.

[47] W. Zhang, F. Li, and L. Nie, Integrating multiple omics analysis for microbial biology: Application and methodologies, *Microbiology*, vol. 156, no. 2, pp. 287-301, 2010.

[48] J. Chen and B. Yuan, Detecting functional modules in the yeast protein-protein interaction network, *Bioinformatics*, vol. 22, no. 18, pp. 2283-2290, 2006.

[49] Y.-R. Cho, W. Hwang, and A. Zhang, Identification of overlapping functional modules in protein interaction networks: Information flow-based approach, in *ICDM Workshops 2006, Sixth IEEE International Conference on*, 2006, pp. 147-152.

[50] M. Li, X. Wu, J. Wang, and Y. Pan, Towards the identification of protein complexes and functional modules by integrating ppi network and gene expression data, *BMC Bioinformatics*, vol. 13, no. 1, p. 109, 2012.

[51] G. Cesareni, A. Ceol, C. Gavrila, L. M. Palazzi, M. Persico, and M. V. Schneider, Comparative interactomics, *FEBS Letters*, vol. 579, no. 8, pp. 1828-1833, 2005.

[52] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proceedings of the National Academy of Sciences*, vol. 96, no. 8, pp. 4285-4288, 1999.

[53] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, Detecting protein function and protein-protein interactions from genome sequences, *Science*, vol. 285, no. 5428, pp. 751-753, 1999.

[54] N. Du, J. Gao, V. Gopalakrishnan, and A. Zhang, De-noise biological network from heterogeneous sources via link propagation, in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, 2012, pp. 1-6.

[55] I. Gat-Viks, A. Tanay, and R. Shamir, Modeling and analysis of heterogeneous regulation in biological networks, *Journal of Computational Biology*, vol. 11, no. 6, pp. 1034-1049, 2004.

[56] K. Shimizu, Toward systematic metabolic engineering based on the analysis of metabolic regulation by the integration of different levels of information, *Biochemical Engineering Journal*, vol. 46, no. 3, pp. 235-251, 2009.

[57] D. W. Huang, B. T. Sherman, and R. A. Lempicki, Systematic and integrative analysis of large gene lists using david bioinformatics resources, *Nature Protocols*, vol. 4, no. 1, pp. 44-57, 2008.

[58] K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay, and J. Weinstein, Matchminer: A tool for batch navigation among gene and gene product identifiers, *Genome Biol*., vol. 4, no. 4, p. R27, 2003.

[59] S. S. Chavan, J. D. Shaughnessy Jr, and R. D. Edmondson, Overview of biological database mapping services for interoperation between different'omics' datasets, *Human Genomics*, vol. 5, no. 6, p. 703, 2011.

[60] C.-H. Yeang and M. Vingron, A joint model of regulatory and metabolic networks, *BMC Bioinformatics*, vol. 7, no. 1, p. 332, 2006.

[61] E. Mosca and L. Milanesi, Network-based analysis of omics with multi-objective optimization, *Molecular BioSystems*, vol. 9, no. 12, pp. 2971-2980, 2013.

[62] N. Du, Y. Zhang, K. Li, J. Gao, S. D. Mahajan, B. B. Nair, S. A. Schwartz, and A. Zhang, Evolutionary analysis of functional modules in dynamic ppi networks, in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2012, pp. 250-257.

[63] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, et al., Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, vol. 430, no. 6995, pp. 88-93, 2004.

[64] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, Dynamic modularity in protein interaction networks predicts breast cancer outcome, *Nature Biotechnology*, vol. 27, no. 2, pp. 199-204, 2009.

[65] S. Srihari and H. W. Leong, Temporal dynamics of protein complexes in ppi networks: A case study using yeast cell cycle dynamics, *BMC Bioinformatics*, vol. 13, no. Suppl 17, p. S16, 2012.

[66] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, A comparison of the functional modules identified from time course and static ppi network data, *BMC Bioinformatics*, vol. 12, no. 1, p. 339, 2011.

[67] J. Wang, X. Peng, M. Li, and Y. Pan, Construction and application of dynamic protein interaction network based on time course gene expression data, *Proteomics*, vol. 13, no. 2, pp. 301-312, 2013.

[68] Q. Xiao, J. Wang, X. Peng, and F.-X. Wu, Detecting protein complexes from active protein interaction networks constructed with dynamic gene expression profiles, *Proteome Science*, vol. 11, no. Suppl 1, p. S20, 2013.

[69] R. Jin, S. Mccallen, and C. Liu, Identify dynamic network modules with temporal and spatial constraints, *Pac. Symp. Biocomput.*, vol. 14, pp. 203-214, 2009.

[70] Y. Zhang, N. Du, K. Li, K. Jia, and A. Zhang, Co-regulated protein functional modules with varying activities in dynamic ppi networks, *Tsinghua Science and Technology*, vol. 18, no. 5, pp. 530-540, 2013.

[71] B. Chen, W. Fan, J. Liu, and F.-X. Wu, Identifying protein complexes and functional moduleslfrom static ppi networks to dynamic ppi networks, *Briefings in Bioinformatics*, p. bbt039, 2013.

[72] X. Wang, Role of clinical bioinformatics in the development of network-based biomarkers., *J. Clinical Bioinformatics*, vol. 1, p. 28, 2011.

[73] L. Chen, R. Liu, Z.-P. Liu, M. Li, and K. Aihara, Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers, *Scientific Reports*, vol. 29, no. 2, p. 342, 2012.

[74] Y. Zhang, N. Du, K. Li, J. Feng, K. Jia, and A. Zhang, msidbn: A method of identifying critical proteins in dynamic ppi networks, *BioMed Research International*, vol. 2014, 2014.

[75] V. Marx, Biology: The big challenges of big data, *Nature*, vol. 498, no. 7453, pp. 255-260, 2013.

[76] G. M. Church, Genomes for all, *Scientific American*, vol. 294, no. 1, pp. 46-54, 2006.

[77] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al., Accurate whole human genome sequencing using reversible terminator chemistry, *Nature*, vol. 456, no. 7218, pp. 53-59, 2008.

[78] C. Lam, *Hadoop in Action*. Manning Publications Co., 2010.

[79] R. C. Taylor, An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics, *BMC Bioinformatics*, vol. 11, no. Suppl 12, p. S1, 2010.

[80] S. Leo, F. Santoni, and G. Zanetti, Biodoop: Bioinformatics on hadoop, in *ICPPW'09, International Conference on*, 2009, pp. 415-422.

[81] High-level mapreduce pipeline, https://developer.yahoo.com/hadoop/tutorial/module4.html.

[82] X. Qiu, J. Ekanayake, S. Beason, T. Gunarathne, G. Fox, R. Barga, and D. Gannon, Cloud technologies for bioinformatics applications, in *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers*, 2009, p. 6.

[83] Ec2, http://aws.amazon.com/ec2, 2014.

[84] Google appengine, http://code.google.com/appengine/, 2014.

[85] Microsoft azure, http://www.microsoft.com/azure, 2014.

[86] H. Lee, Y. Yang, H. Chae, S. Nam, D. Choi, P. Tangchaisin, C. Herath, S. Marru, K. P. Nephew, and S. Kim, Biovlab-mmia: A cloud environment for microrna and mrna integrated analysis (mmia) on amazon ec2, *NanoBioscience, IEEE Transactions on*, vol. 11, no. 3, pp. 266-272, 2012.

[87] The environmental molecular sciences laboratory, http://www.nersc.gov/users/computational-systems/, 2014.

[88] The US Eept. of Energy, http://genomicscience.energy.gov/compbio/, 2014.

[89] H. Xiao, Towards parallel and distributed computing in large-scale data mining: A survey, Technical University of Munich, Tech. Rep, 2010.

[90] M. C. Schatz, Cloudburst: Highly sensitive read mapping with mapreduce, *Bioinformatics*, vol. 25, no. 11, pp. 1363-1369, 2009.

[91] Crossbow, http://bowtiebio.sourceforge.net/crossbow/index.shtml, 2014.

[92] Contrail, http://sourceforge.net/apps/mediawiki/contrailbio/index.php?title=contrail, 2014.

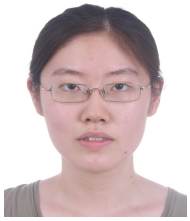
[93] Myrna, http://bowtie-bio.sourceforge.net/myrna/index.shtml, 2014.

[94] A. Matsunaga, M. Tsugawa, and J. Fortes, Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications, in *eScience'08,*

IEEE Fourth International Conference on, 2008, pp. 222-229.

[95] A. P. Heath, M. Greenway, R. Powell, J. Spring, R. Suarez, D. Hanley, C. Bandlamudi, M. E. McNerney, K. P. White, and R. L. Grossman, Bionimbus: A cloud for managing, analyzing and sharing large genomics datasets, *Journal of the American Medical Informatics Association*, doi: 10.1136/amiajnl-2013-002155, 2014.

Yuan Zhang is a PhD candidate in Beijing University of Technology. She received her MS degree and BS degree from Beijing University of Technology in 2010 and 2007, respectively. She is also a joint PhD student of the State University of New York at Buffalo. Her research interestes include data mining and network modeling.



Yue Cheng is a PhD candidate in Beijing University of Technology. She received her BS degree from Beijing University of Technology in 2012. Her research interests include network modeling and data mining.

Kebin Jia is Professor and Dean in Department of Electronic Information and Control Engineering in Beijing University of Technology. He received his PhD degree and MS degree from University of Science and Technology of China in 1998 and 1990, respectively. His research interests include multimedia and database systems, 3D video coding, data mining, and pattern recognition. He has published more than 200 research publications in these areas. He is a senior member of The Chinese Institute of Electronics and member of Audio Video Coding Standard Workgroup of China. He has served as PI for more than 15 research projects from The National Natural Science Foundation of China (NSFC), 973 National Basic Research Program, and 863 Program. Dr. Jia has published two books about digital medical systems and video coding.

Aidong Zhang is UB Distinguished Professor and Chair in the Department of Computer Science and Engineering at State University of New York at Buffalo. She received her PhD degree from Purdue University in 1994. Her research interests include bioinformatics, data mining, multimedia and database systems, and content-based image retrieval. She is an author of over 200 research publications in these areas. She has chaired or served on over 100 program committees of international conferences and workshops, and currently serves several journal editorial boards. She has published two books: *Protein Interaction Networks: Computational Analysis* (Cambridge University Press, 2009) and *Advanced Analysis of Gene Expression Microarray Data* (World Scientific Publishing Co., Inc. 2006). Dr. Zhang is a recipient of the National Science Foundation CAREER award and State University of New York (SUNY) Chancellor's Research Recognition award. Dr. Zhang is an IEEE Fellow.