



GUEST EDITORIAL

Integrative data mining in systems biology: from text to network mining

1. Introduction

Systems biology is a new field that aims to gain system-level understanding of complex functions of biological systems and biological processes, ranging from molecules to cells, tissues, or entire organisms. From the systems science point of view, many important properties of a complex system emerge from the interaction of system components and, therefore, rather than investigating the characteristics of isolated system parts separately, systems biologists focus on discovering emergent properties and functions that do not appear in individual components, but are driven by the interactions among all the system components or component groups [1,2].

A general framework of systems biology can be described by three main steps: (1) identifying elements/components of the system; (2) describing the system using connective networks, in which nodes represent the system components and edges represent interactions between nodes. The network describes the functional relationship among the system components, and the interactions ultimately determine an organism's behavior and functions; (3) gaining insights into emergent properties of biological systems by means of analyzing structural properties and dynamics of the network.

Advances of high-throughput '-omics' technologies, such as genomic sequences gathered by the Human Genome Project [3,4], gene expression data from microarray experiments [5,6], proteome databases [7,8] and protein interaction databases [9,10], together with large volume of digital textual documents such as the PubMed [11], have created an unprecedented opportunity to apply computational techniques/tools for a comprehensive study of the structure and dynamics of system components, and thus provides a foundation to systems biology.

A variety of data mining and machine learning techniques have been developed during recent years aiming at extracting information/knowledge/patterns from large volumes of data. Due to the useful capability of processing large datasets and extracting hidden patterns, data mining and machine learning techniques have been applied to bioinformatics for managing the data, identifying the structure and functions of system components/elements (genes and proteins), and their functional relationships. However, the traditional data mining techniques, when working for systems biology, are facing many new challenges. A fundamental issue is that the biomedical data repositories are formed with various (diverse) technologies and are normally presented in heterogeneous and unstructured forms. The ability to automatically and effectively extract, integrate, understand, and make use of information embedded in such heterogeneous unstructured data remains a challenging task. We are looking forward to new data mining techniques that can assist systems biologists to direct the whole investigation process from information gathering, analysis and interpretation to incrementally improve our understanding and eventually gain a panorama of the biological systems.

2. Aim of this special issue

This special issue aims at contributing to narrow the gap between methodological advancements in data mining and scientific investigations in systems biology. It is our view that integrative data mining (the integration of heterogeneous data, information and knowledge for the generation of higher level knowledge and new testable hypotheses) will provide effectively tools for the better understanding of behavior and functions of genes and proteins within the context of biological systems.

This special issue is particularly interested in newly development of computational approaches for the construction, refinement and analysis of biological networks, i.e. (1) the construction and enhancement of biological networks; (2) the analysis of networks constructed; (3) the use of networks to interpret dynamics and functions of interacting genes and proteins.

3. Systems biology: from text mining to network mining

This issue attracted 18 research submissions covering a wide range of topics related with systems biology. Seven research papers are selected for publication to cover the core topics from text mining to network mining and applications.

3.1. Text mining and knowledge integration

Text mining from the massive accumulation of biomedical publications has been suggested to have high potential for discovering knowledge buried in the literature. Despite the sometimes conflicting views, traditional Chinese medicine (TCM) is widely regarded as a promising alternative medicine and its holistic way of studying patients and individuals makes it become a natural objective in current systems biology investigations. Li et al. [12] has shown that improved understanding of the ZHENG concept (particular patterns of symptoms, syndromes or phenotypes) in the centre of TCM theories could be reached from systematically mining of existing Western biomedical literature. This special issue collects one interesting paper [13] which deals with the integration of two independent and complementary literature databases: the TCM literature database and the MEDLINE bibliographic databases. Although the scientific observations are still preliminary, the paper presents an interesting approach to integrate TCM literature and modern biomedical literature data to discover novel gene networks and functional knowledge of genes.

Two most important genomic data resources for systems biology are the gene ontology (GO) [14] and the publicly available microarray databases such as the Stanford Microarray Database (SMD) [5] and the NCBI Expression Omnibus [6]. The GO database is intended to provide a controlled vocabulary describing attributes of genes and gene products, and has become a routine resource for functional analysis. Due to the many-to-many relationship between GO terms and genes, extracting significant GO terms relevant to given genes is one of very valuable but

challenging tasks. Paper [15] presents one useful approach that takes multiple gene lists as input, and outputs the GO terms that are enriched by the input gene list(s). The proposed method has been evaluated by time course gene expression data of embryonic stem cells. Five gene clusters have been generated through clustering analysis, representing respectively up-regulation, down-regulation and other novel patterns in the differentiation time course. Based on the gene lists associated with these five gene clusters, “cell adhesion” and “muscle contraction” were identified as the significant GO terms for the up-regulated cluster, and “amino acids metabolism” as a significant GO term for the down-regulated gene cluster. Furthermore, a number of GO terms related to RNA processing and RNA transport have been also identified as significant terms to a cluster that is up-regulated in both early and late time points. This suggests that genes for RNA processing and genes for RNA transport are co-regulated in the differentiation process of embryonic stem cells [15].

3.2. Biological network mining: from network construction to modularity analysis

Much attention has been received during the past years for investigating the properties of biological networks such as the protein interaction networks, and network biology now has become one of the main topics in systems biology [16]. Researchers are making progress toward understanding the organizing principles that govern the formation and evolution of complex biological networks [17]. Considered as a major challenge in systems biology, predicting network behaviors and functions requires the identification of functionally and statistically significant sub-network. To understand their structural organizing principles and evolutionary mechanisms, the concept of ‘network motif’ has been proposed [18], and previous research has shown that a few significant network motifs with regular structures has been identified in several biological systems such as *Escherichia coli* [19].

In this issue, the paper [20] defines the so-called ‘bridge motifs’ as composed of at least one weak link, and ‘brick motifs’ as composed of strong links only. An algorithm is presented in [20] for performing two simultaneous tasks: detecting global statistical features and local connection structures of networks, and extracting and locating functionally and statistically significant network motifs. It has been observed that most motifs in genetic networks belong to bridge motifs, inferring that weak-tie links provide unique paths for signal control exert

significant impacts on the signal processing function of transcription networks, and the authors argue that bridge and brick motifs would provide a useful approach to capture global and local views of complex networks and thus provide an effective approach to analyze functions, behaviors, and similarities of networks.

Another paper [21] collected in this issue concerns the analysis of protein–protein-interaction (PPI) networks. A suite of graph-mining techniques is presented for the identification of important structures within protein–protein-interaction (PPI) networks. The centrality measures betweenness, closeness and degree are employed as a means of identifying essential proteins, while the paper also demonstrates that the rational Erdos numbers can be used to identify collaborating proteins based solely upon network structure. On the other hand, this study demonstrated that by using dynamic cut-off limits, the collaboration subgraphs can be generated for each protein, and by graph containment the subgraphs linking protein complex can be extracted. In addition, this study showed that the PPI network relating to human diabetes, built from data collated from the Human Protein Reference Database (HPRD) [22], is a scale-free, small-world graph with a power-law degree distribution of interactions on nodes.

3.3. Dynamics modelling

Another important research subject of systems biology is to understand the dynamics of biological (sub-)systems. This issue collects one research paper [23] concerning the motility and chemotaxis of bacteria cells as they play an important role in the virulence of pathogenic bacteria, such as escaping from host immune responses. *E. coli* chemotaxis provides a well-characterized model system for the bacterial chemotaxis network, and is featured by its signal amplification and robustly accurate adaptation. Although recent studies with models considering the effects of receptors have suggested possible mechanisms for signal amplification, and precise adaptation to aspartate has been explained by conventional kinetic models, the adaptation behavior of models incorporating the effects of other receptors remains unclear. The study presented in [23] investigates how receptor crosstalk affects the minimization of adaptation error and compares models in which the contribution of other receptors varied. Results suggest that accurate adaptation is maintained through the control of both the interaction of cytoplasmic Che proteins and the activity of the receptor complex.

3.4. Algorithms for integrative data mining

This issue also selects two papers presenting new algorithms for integrative data mining. Paper [24] presents a combinational method for the operon prediction, which appears as a critical issue to the reconstruction of genome regulatory networks in prokaryotes. In the literature multiple genome features have been used for predicting operons; however, they are usually dealt with using only single method. The method presented in [24] is characterized by a so-called ‘multi-approach guided’ genetic algorithm, i.e. involving diverse methods to pre-process different genome features in order for extracting their unique characteristics. The experimental results, examined on *E. coli* K12 genome, *Bacillus subtilis* genome, and *Pseudomonas aeruginosa* PAO1 genome, demonstrate it as a promising and effective approach.

Paper [25] presents another integrative data mining techniques, called multiple kernel support vector machine (MK-SVM), which incorporates feature selection and rule extraction steps in order for extracting the appropriate knowledge from the gene expression data. The feature selection task is translated into a model selection problem of SVM, and rule extraction is performed based on the separating hyperplane and support vectors. The method has been evaluated on two public datasets and the results showed MK-SVM achieves promising results in terms of classification accuracy, and more importantly, the rules extracted can improve the comprehensibility of the system and therefore help understanding the biology of the studied diseases like cancers.

4. Systems biology: the blind men and the elephant!

Systems biology is a very broad field and a dynamically developing discipline. With the current availability of the data and technology, the task of understanding a biological system is still like the task for ‘six blind men to learn what an elephant looks like’. Some investigators focus on the aspects they are working on by further improving the methods for the observations and experiments, while others are trying new angles to touch different part of the elephant in order to get different figures. Different from those, the systems biologists are more interested in capturing the whole picture of system by developing new techniques or using existing methods for integrating the diverse views obtained from different angles. From the broad

spectrum of systems biology studies, this issue collected seven papers covering different aspects on this direction. It is true that we are still very far from forming a whole picture of a biological system; however, we believe that, only when the information obtained can be utilized in a systematic manner, every touch at a new angle and every improvement in the touching at existing angles will provide incremental contributions to narrow the gap. Integrative data mining is hoping to play important roles in this effort.

Acknowledgements

We would like to thank all external reviewers for their valuable contributions to the specific issues.

References

- [1] Kitano H. Systems biology: a brief overview. *Science* 2002;295(5560):1662–4.
- [2] Aderem A. Systems biology: its practice and challenges. *Cell* 2005;121(4):511–3.
- [3] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001;291(5507):1304–51.
- [4] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931–45.
- [5] Stanford Microarray Database (SMD). <http://genome-www5.stanford.edu/> (accessed: 20 July, 2007).
- [6] NCBI Expression Omnibus. <http://www.ncbi.nlm.nih.gov/geo/> (accessed: 20 July, 2007).
- [7] Prince JT, Carlson MW, Wang R, Lu P, Marcotte EM. The need for a public proteomics repository. *Nature Biotechnol* 2004; 22(4):471–2.
- [8] Open proteomics data. <http://apropos.icmb.utexas.edu/OPD/> (accessed: 20 July, 2007).
- [9] Database of Interacting Proteins (DIP). <http://dip.doe-mbi.ucla.edu/> (accessed: 20 July, 2007).
- [10] EMBL-EBI Protein Interact Database (IntAct). <http://www.ebi.ac.uk/intact/site/index.jsf> (Accessed: 20 July, 2007).
- [11] PubMed. <http://www.pubmed.gov/> (accessed: 20 July, 2007).
- [12] Li S, Zhang ZQ, Wu LJ, Zhang XG, Li YD, Wang YY. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. *Inst Eng Technol J Syst Biol* 2007;1(1):51–60.
- [13] Zhou X, Liu B, Wu Z, Feng Y. Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. *Artif Intell Med* 2007;41:87–104.
- [14] The gene ontology. <http://www.geneontology.org/> (accessed: 20 July, 2007).
- [15] Xie D, Zhong S. Gene ontology analysis in multiple gene clusters under multiple hypothesis testing framework. *Artif Intell Med* 2007;41:105–15.
- [16] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5(2): 101–13.
- [17] Huang S. Back to the biology in systems biology: what can we learn from biomolecular networks? *Brief Funct Genomic Proteomic* 2004;2(4):279–97.
- [18] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science* 2002;298:824–7.
- [19] Shen-Orr S, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 2002;31:64–8.
- [20] Huang C-Y, Cheng C-Y, Sun C-T. Bridge and brick network motifs: identifying significant building blocks from complex biological systems. *Artif Intell Med* 2007;41:117–27.
- [21] McGarry K, Chambers J, Oatley G. A multi-layered approach to protein data integration for diabetes research. *Artif Intell Med* 2007;41:129–43.
- [22] Human Protein Reference Database (HPRD). <http://www.hprd.org/> (accessed: 20 July, 2007).
- [23] Matsuzaki Y, Kikuchi S, Tomita M. Robust effects of Tsr–CheBp and CheA–CheYp affinity in bacterial chemotaxis. *Artif Intell Med* 2007;41:145–50.
- [24] Wang SQ, Wang Y, Du W, Sun FX, Wang XM, Zhou CG, Liang YC. A multi-approaches-guided genetic algorithm with application to operon prediction. *Artif Intell Med* 2007;41: 151–9.
- [25] Chen Z, Li J, Wei L. A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artif Intell Med* 2007;41:161–75.

Yonghong Peng*

*Department of Computing,
University of Bradford,
West Yorkshire BD7 1DP, United Kingdom*

Xuegong Zhang

*MOE Key Laboratory of Bioinformatics and
Department of Automation,
Tsinghua University,
Beijing 100084, China*

*Corresponding author.

Tel.: +44 1274 233963; fax: +44 1274 233920

E-mail address: y.h.peng@bradford.ac.uk

(Y. Peng)