The First International Symposium on Optimization and Systems Biology (OSB 2007)

Abstracts

August 8-10, 2007
Beijing, China

A Systems-Biology View to Diabetes

Jia-Rui Wu[1,2,*]

[1] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology,
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
[2] Department of Systems Biology, University of Science and Technology of China

Now Life science has come to a new era, the post-genome era, with the accomplishment of "Human Genome Project". In the post-genome era, the "big sciences" such as genomics, proteomics and metabolomics (so-called "omics") gradually become popular research ways to provide global pictures of cells or organisms, although the classical experimental biology (small sciences) such as molecular biology or cell biology is still the mainstream in life sciences.

Systems biology is a newly born discipline in the post-genome era, which integrates the research strategy of classical experimental biology with the research strategy of "omics". Systems biology is also a new interdisciplinary frontier based mainly on the integration of the "wet" experiments such as molecular biology or "omics" with the "dry" experiments such as bioinformatics and computational biology. Technology of systems biology includes the "omics" platforms such as proteomics-platform and the theoretical platforms for computing and modeling.

These properties have made systems biology to be a powerful analytical tool to reveal the complex diseases such as diabetes. Diabetes remains one of the most frequent progressing diseases in the world. The causes of the initiation and development of diabetes involve genetic factors, environment factors and the interaction of these two kinds of factors. Furthermore, more and more studies have shown that the gene- and protein-networks play important roles for the molecular mechanisms of diabetes. Therefore, it is clear that the methodology and techniques of system biology must be applied for analyzing the molecular mechanisms of diabetes, whereas the classical experimental biology that focuses on individual genes or proteins is not good enough to solve such problems.

Our ongoing project is to identify biomarkers in order to detect the progress stages of the disease. As a metabolic disease, diabetes-related proteins tend to enter the circulation system, thus the biomarker discovery in plasma has the potential to play an important role in the early detection and treatment of diabetes. Protein analysis of plasma is a formidable challenge, due to its huge complexity and dynamic range. We have developed systematic approaches based on proteomics and bioinformatics to analyze human normal and diabetic serum. This qualitative and quantitative proteomic analysis of serum-proteins provides a snapshot of the protein expression state in normal and diabetes serum and should be very helpful to identify the biomarker of diabetes.

[*] Email: wujr@sibs.ac.cn

An LP-based Approach towards Detecting Community Structures in Networks

Andreas Dress[*]

Department of Combinatorics and Geometry (DCG)
CAS-MPG Partner Institute for Computational Biology (PICB)
Shanghai Institutes for Biological Sciences (SIBS)
Chinese Academy of Sciences (CAS)

I'll present an approach to studying the community structures of networks by using linear programming (LP) developed jointly with William Y.C.Chen and Winking Q.Yu from the Center of Combinatorics at Nankai University. Starting with a network in terms of (a) a collection of nodes and (b) a collection of edges connecting some of these nodes, we use a new LP-based method for simultaneously (i) finding, at minimal cost, a second edge set by deleting existing and inserting additional edges so that the network becomes a disjoint union of cliques and (ii) appropriately calibrating the costs for doing so. We provide examples that suggest that, in practice, this approach provides a surprisingly good strategy for detecting community structures in given networks.

[*]Email: andreas@picb.ac.cn

Computational Systems Biology

Satoru Miyano[*]

Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan

One of the key issues for exploring systems biology is development of computational tools and capabilities which enable us to understand complex biological systems. We have been taking two approaches to this issue.

The first is a series of computational methods based on Bayesian networks and nonparametric regression for mining gene networks from microarray gene expression data. These computational methods for computing gene networks were applied for searching drug target pathways. For a given drug, our strategy assumes two kinds of microarray gene expression data: One is time-course gene expression data for the drug responses. The other is a set of gene expression data obtained by knock-downs of several hundreds of carefully selected genes (one knock-down for each microarray measurement). We prepared more than 350 novel gene knock-downs for HUVEC by using siRNA and some time-course drug response gene expression data. From these data, we computed gene networks of 1000 genes by intensively using the supercomputer system at our Human Genome Center of University of Tokyo. We show how we can explore these computed networks for searching drug target genes and hubs in the networks.

The second is our development of a software tool Cell Illustrator (http://www.gene-networks.com/). This software aims at describing and simulating structurally complex dynamic causal interactions and processes such as metabolic pathways, signal transduction cascades, gene regulations. It employs the architecture based on the notion of Hybrid Functional Petri Net with extension (HFPNe). Simultaneously, we have been developing an XML format Cell System Markup Language CSML (http://www.csml.org/) for describing biological systems with dynamics and ontology (Cell System Ontology). The newest version CSML 3.0 covers widely used data formats and applications, e.g. CellML 1.0, SBML 2.0, BioPAX, and Cytoscape. Since Cell Illustrator equips biology-oriented sophisticated GUIs, we can make modeling of very complex biological processes like with a drawing tool. We used this tool to explore the HUVEC gene networks for drug target pathway discovery. Cell Illustrator is also used for developing a simulatable EGF receptor signal transduction pathway model (EFGR model) based on the biological data and knowledge from the literature. Then, by using some quantitative time-course proteomic data produced with the recent tandem mass spectrometry coupled with liquid chromatography (LC/MS/MS) technology, we semi-automatically constructed a well-tuned EGFR model with the data assimilation technology which we developed for data-driven computational modeling of biological systems. We have recently developed a software tool AYUMS for automatic quantitation of the proteome by LC/MS/MS which shall speed up the analysis process.

[*] Email: miyano@ims.u-tokyo.ac.jp

Computational Approaches to Gene Regulation

Jun Liu[*]

Department of Statistics, Harvard University, Cambridge, MA 02138

Understanding how genes are regulated in various circumstances (e.g., heatshock, starvation, etc.) is a central problem in molecular biology. The adoption of large-scale biological data generation techniques such as the mRNA microarrays has enabled researchers to tackle the gene regulation problem in a global way. I will survey some computational and statistical strategies developed by our group on how to effectively use the gene upstream sequence information in junction with mRNA expression microarray data to dissect the gene regulatory network. I will describe in detail a study of RacA binding activities in Bacillus Subtilis, explaining how statistical approaches helped the biologists discover RacA's binding sites. I will describe a new dimension reduction technique that has been applied successfully to our gene regulation studies and show a cute theorem supporting the technique.

[*] Email: jliu@stat.harvard.edu

SIMPLE GENE SELECTION METHOD FOR MICROARRAY DATA ANALYSIS

Michael K. Ng Eric S. Fung

Centre for Mathematical Imaging and Vision and Department of Mathematics
Hong Kong Baptist University, Kowloon Tong, Hong Kong.

Background

Gene selection algorithms for patient and normal classification, based on the expression of a small number of non-redundant but relevant genes, become an important topic in the development of microarray technology. However, the selection of a suitable set of biomarkers over thousands of genes in microarray data is complicated. The main challenge is that the number of selected genes should be sufficiently small and relevant in cancer or disease classification, in order to allow medical diagnosis or identify genes involved in cancer-specific regulatory pathways in regular laboratories. One of the promising approaches in microarray binary label data classification is the linear discriminant analysis. The fundamental objective is to construct a projection vector u for projecting a high-dimensional sample data into one-dimensional space. The classification can then be done by the projection value discrepancy of the two class data. Currently, there are several well-known machine learning methods for gene classification analysis. For example, gene-based [7, 11], mutual information [1], sparse logistic regression [13], the relevance vector machine [9], Gaussian process models [3] and simple decision rules [12] are proposed for discovering biomarkers in microarray gene expression data. Support vector machines [2, 4, 6, 8, 10] are also developed in applications of bioinformatics. Fung and Ng [5] developed a Fisher-type discriminant method for microarray data classification. The selection criterion is based on the estimated weight values to identify the subsets of relevant genes for categorizing patient and normal samples. This is achieved by including the weight sparsity in the Fisher objective function that is minimized in the discriminant process:

$$\|S_w u - z\|_2^2 + \alpha \|u\|_1.$$

Here S_w is the within-class scatter matrix of the samples in a microarray data and z comes from the between-class scatter matrix, u is the projection vector and α is the regularization parameter to control the sparsity of u. Each entry of u represents a weight for each gene.

Results

The main contribution in this poster is to propose simple gene selection criterions to improve the classification performance based on the estimated model parameters in [5]. In the numerical results, we use leave-one-out cross validation to evaluate the performance of each model. We notice that in their discriminant process, their gene selection criterion is only based on the weight value of u_i and this weighting in the samples can be automatically computed by a l_2-l_1 norm minimization method. In the new method, we also use the normalized normal sample means $\bar{n}_i$ and normalized cancer sample means, $\bar{c}_i$ to identify the subsets of relevant genes that categorize patient and normal samples. Here, we perform three different gene selection approaches and their gene selection criterions are based on the weight of (1) $|u_i|$, (2) $|\bar{c}_i - \bar{n}_i|$ and (3) $|u_i(\bar{c} - \bar{n}_i)|$. In each selection procedure, once a particular gene j is selected, the corresponding weighting u_j of gene j will be used for projection. Thus, if there are f genes are selected as biomarkers, the new projection vector $\hat{u}$ possesses exactly f corresponding non-zero entries.

Apart from the above three selection methods, we perform a simple K-NN (K-nearest neighborhood) classification on the dataset for gene selection. The gene selection criterion is also based

Method	k	α	Accuracy in%	Cross Entropy	No. of genes
Fisher-type					
$\lvert u_i \rvert$	5	1	82.26	0.43	4
$\lvert \bar{c}_i - \bar{n}_i \rvert$	5	0.1	85.48	0.29	3
$\lvert u_i(\bar{c}_i - \bar{n}_i) \rvert$	7	0.1	87.10	0.39	2
Norm-1					
$\lvert \bar{c}_i - \bar{n}_i \rvert$	3	-	88.71	0.25	36

Table 1: The average classification results on colon cancer data.

on the value of $|\bar{c}_i - \bar{n}_i|$, the main different is that the discrepancy measurement is determined by the vector norm-1 distance of the selected genes rather than the projection value. The results are shown in Table 1.

In three different selection criterions on Fisher-type method, the case of $|u_i(\bar{c}_i - \bar{n}_i)|$ obtains the best classification accuracy and the minimal number of selected genes, while the case of $|u_i(\bar{c}_i - \bar{n}_i)|$ gives the lowest cross entropy value. Here, we observe that the gene selection criterion proposed in [5] give the worst classification performance compare with the other two selection methods. With comparing the results with norm-1 method, although the lowest value of the cross entropy is obtained in the simple norm-1 approach, the number of genes selected are much higher than the other cases and the classification accuracy in the case of $|u_i(\bar{c}_i - \bar{n}_i)|$ is very close to norm-1 approach.

Conclusions:

In this study, we demonstrate the gene selection's ability and the computational effectiveness of the proposed method. The experiments on colon cancer dataset have shown that the new gene selection method can generate classification results in a competitive manner than the previous criterion. Furthermore, the set of the selected genes obtained from our model is very compact; it highly reduces the redundancy among genes.

References

[1] Ben-Dor A., Friedman N. and Yakhini Z., *Scoring Genes for Relevance*, Agilent Technologies Technical Report AGL-2000-13, 2000.

[2] Brown M. P. S., Grundy W. N., Lin D., Cristianini N., Sugnet C. W., Furey T. S., Ares M. and Haussler Jr. D., *Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines*, Proc. Nat. Acad. Sci. USA, 97, No. 1, pp. 262-267, 2000.

[3] Chu W., Ghahramani Z., Falciani F., and Wild L., *Biomarker Discovery in Microarray Gene Expression Data with Gaussian Processes*, Bioinformatics, 21, pp. 3385-3393.

[4] Cristianini N. and Shawe-Taylor J., *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[5] Fung E. and Ng M., *On Sparse Fisher Discriminant Method for Microarray Data Analysis*, submitted to InCoB2007.

[6] Furey T. S., Cristianini N., Duffy N., Bednarski D. W., Schummer and Haussler D., *Support vector machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data*, Bioinformatics, 16, pp. 906-914, 2000.

[7] Golub T. R., Slonim D. K., Tamayo P., Gaasenbeek C. H. M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D. and Lander E. S., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science, 286, pp. 531-537, 1999.

[8] Guyon I., Weston J., Barnhill S. M. D. and Vapnik V., *Gene Selection for Cancer Classification Using Support Vector Machines*, Machine learning, 46, pp. 389-422, 2002.

[9] Li Y., Campbell C. and Tipping M., *Bayesian automatic relevance determination algorithms for classifying gene expression data*, Bioinformatics, 18, pp. 1332-1339, 2002.

[10] Mukherjee S., Tamayo P., Slonim D., Verri A., Golub T., Mesirov J. P. and Poggio T., *Support Vector Machine Classification of Microarray Data*, A.T. Memo No. 1677, C.B.C.L. Paper No. 183, MIT, 1998.

[11] Pavlidis P., Weston J., Cai J. and Grundy W. N., *Gene Functional Classification from Heterogeneous*, In RECOMB 2001: Proceedings of the fifth International Conference on Computational Molecular Biology. pp. 242-248, 2001.

[12] Tan C., Naiman Q., Xu L., Winslow L. and Geman D., *Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles*, Bioinformatics, 21, pp. 3896-3904.

[13] Shevade S. and Keerthi S., *A simple and efficient algorithm for gene selection using sparse logistic regression*, Bioinformatics, 19, pp. 2246-2253, 2003.

Application of Self-Organizing Map to Structure-Odour Relationships

Koki Sato[1] Atsushi Miyazaki[1] Shin-ya Takane[1]
John B. O. Mitchell[2]

[1] Osaka Sangyo University
[2] University of Cambridge

We have previously studied structure-odour relationship analyses using hierarchical clustering on a diverse dataset of 47 molecules. These molecules were divided into seven odour categories: ambergris, bitter almond, camphoraceous, rose, jasmine, muguet, and musk. We used the alignment-independent QSAR descriptor EVA and compared with those of another kind of descriptor, UNITY 2D fingerprint. The dendrograms produced by EVA consistently outperformed those from UNITY 2D in reproducing the experimental odour classifications of these 47 molecules.

In this study, we used the Kohonen's self-organizing map (SOM) neural network to classify the same dataset. It is also unsupervised learning method. Unlike hierarchical clustering techniques, however, SOM can find useful clusters / segmentations intuitively, remaking visually understandable maps. The result of the classification of the dataset (47 molecules) is in good agreement with the previous hierarchical clustering study. We also carried out other dataset composed of only nitro and / or macro musks.

Emerging Collective Behaviors of Animal Groups

Jinhu Lu

Academy of Mathematics and Systems Science, Chinese Academy of Sciences
E-mail: jhlu@iss.ac.cn

Many animal groups routinely make consensus decisions jointly with all group members. For example, a swarm of honeybees searching a new nest site, a flock of birds deciding when to leave a foraging patch, or a group of ungulates, primates, and fishes selecting where to travel after a rest period. Here we build a novel model merging the locally neighboring reciprocal action and alignment together to investigate the mechanisms of consensus decision-making and its robustness. Our model reveals that the shapes of the coherent flocks are limited in a common narrow interval for different group sizes and information structures, that is, the average elongation of a coherent flock is approximately varying from; the larger the proportion of informed individuals the easier to reach a consensus decision for a group with a little conflict of interest between informed individuals, however, the larger the proportion of informed individuals the more difficult to reach a consensus decision for a group with a significant conflict; the larger the group the easier to reach a consensus decision for a group with a little conflict of interest between informed individuals, on the contrary, the larger the group the more difficult to reach a consensus decision for a group with a significant conflict; the larger the difference between the numbers of the informed individuals the easier to reach a consensus decision for a group with a fixed total number of informed individuals; the alignment ratio and weights of preferred goals of informed individuals of the group are trade-offs between accuracy and split ratio for consensus decision-making. Moreover, the coherent groups keep a fixed shape and their average maximum move distances are linearly increasing as the distance between two information sources increases. In particular, when the information source is suddenly changed, a coherent group will collectively select the exact direction of the changed information sources in a short response time providing that there are enough informed individuals. The smaller the group the higher the average accuracy to collect the exact direction of the changed informed sources for a given proportion of informed individuals. Furthermore, the coherent groups display a surprising degree of tolerance against errors, however, they simultaneously show an extremely fragile to attacks (that is, to the suddenly big changes of the information sources). Our model and approach discover some novel phenomena and also reveal some underlying mechanisms of the consensus decision-making and its robustness in biological systems.

References

[1]J. Lu, I. D. Couzin, J. Liu and S. A. Levin. Emerging collective behaviors of animal groups,

preprint, 2006.

[2]J. Lu, X. Yu, G. Chen and D. Cheng. Characterizing the synchronizability of small-world dynamical networks. IEEE Transactions on Circuits and Systems I, 2004, 51(4): 787-796.

[3]J. Lu and G. Chen. A time-varying complex dynamical network model and its controlled synchronization criteria. IEEE Transactions on Automatic Control, 2005, 50(6): 841-846.

[4] J. Zhou, J. A. Lu and J. Lu. Adaptive synchronization of an uncertain complex dynamical network. IEEE Transactions on Automatic Control, 2006, 51(4): 652-656.