Meeting report

Mining the genome and regulatory networks

Anatolij P Potapov* and Edgar Wingender*†

Addresses: *Department of Bioinformatics, University of Göttingen, Goldschmidtstrasse 1, D-37077 Göttingen, Germany. †BIOBASE GmbH, Halchterschestrasse 33, D-38304 Wolfenbüttel, Germany.

Correspondence: Anatolij P Potapov. Email: anatolij.potapov@bioinf.med.uni-goettingen.de

Published: 22 March 2006

Genome ***Biology*** 2006, **7:**309 (doi:10.1186/gb-2006-7-3-309)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/3/309

A report on the 16th International Conference on Genome Informatics (GIW 2005), Yokohama, Japan, 19-21 December 2005.

The conference on genome informatics held in Yokohama at the end of last year provided an international forum for disseminating the latest developments and applications in advanced computational methods that can be used for solving several biological problems. In bioinformatics and systems biology the topics covered included the prediction of regulatory networks, microarray data analysis, and graph mining for structural biological data.

Modules and networks

Biological processes in microbial cells are known to be carried out through interactions between multiple 'functional modules' that serve as the basic building blocks of complex biological networks. At the gene level, for example, a module is defined as a set of genes that can be grouped according to their biological function in a pathway or process, and in many cases modular structure and module components are highly conserved. Ying Xu (University of Georgia, Atlanta, USA) presented a computational method for predicting modules in microbial genomes that is based on the fact that neighboring genes in prokaryotic genomes are likely to be functionally related, as a result of the operon organization. Using 224 microbial genomes from 175 different species, Xu and colleagues quantified the functional relatedness among genes on the basis of their presence (or absence) and relative proximity in different genomes, and obtained a gene network in which all possible pairs of genes had a score representing functional relatedness. By applying a threshold-based clustering algorithm to this network, numerous functional modules were obtained. The predicted modules were statistically and biologically significant, and the genes of a module were more likely to share Gene Ontology 'biological process' terms than terms relating to 'molecular function' or 'cellular component'. These predictions could be used to guide experimental designs for investigating particular biological processes and might also provide a basis for further prediction of detailed gene functions and biological networks.

Turning to protein-protein interactions, recent technological advances have made available large datasets of interactions between individual proteins. There is, however, still a lack of experimental data on the larger protein complexes through which many proteins carry out their biological function. With the aim of predicting protein complexes, Xiao-Li Li (Institute for Infocomm Research, Singapore) presented a novel graph-mining algorithm for protein-protein interaction data to detect the highly connected regions (dense neighborhoods) in an interaction graph, which may correspond to protein complexes. The algorithm first locates local cliques, complete subgraphs in which any two vertices are connected by an edge, for each graph vertex (a vertex represents a protein), and then merges the detected local cliques according to their affinity to form maximal dense neighborhoods. This method differs from others in basing the predictions on dense neighborhoods rather than on cliques. As there is no requirement for the dense graphs to be fully connected, however, this algorithm is less sensitive to the 'incompleteness' of protein-interaction data than are conventional clique-detection methods. Nevertheless, compared with existing techniques some of the complexes predicted by the dense-neighborhood approach match or overlap significantly better with known protein complexes in the Munich Information Center for Protein Sequences (MIPS) benchmark database [http://mips.gsf.de].

Various mathematical models describing gene regulatory networks, and algorithms for network reconstruction from

experimental data, have been proposed in recent years. Nevertheless, the theory and practice of computational network inference is not completely established. There is still a need for suitable models of gene regulation that are sufficiently accurate in representing biological reality. Dace Ruklisa (University of Latvia, Riga, Latvia) in collaboration with Alvis Brazma's group (European Bioinformatics Institute, Cambridge, UK) discussed the reconstruction of gene regulatory networks using the finite state linear model (FSLM) proposed recently by Brazma and Thomas Schlitt. This model combines both discrete and continuous aspects of gene regulation. The continuous part considers the concentration of proteins, some of which are transcription factors regulating the activity of a gene and others of which are the products of gene expression. The states of promoter regions are modeled with discrete components. The expression level of a gene depends on the state of a promoter and can be in a finite number of levels. The expression level defines whether the concentration of the corresponding protein is increasing or decreasing. Proteins either bind to or dissociate from a promoter according to certain thresholds, thus influencing the state of each promoter. Ruklisa and colleagues analyzed several theoretical properties of FSLM and showed that, generally, the question of whether a particular gene will reach an active state is algorithmically unsolvable. This imposes some practical difficulties in reverse engineering of FSLM networks. Simulation experiments made by the authors show, however, that many, although not all by far, random networks might exhibit periodic behavior. Thus, FSLM can be considered as a good compromise that is adequate enough in representing biological reality. An efficient time algorithm has been proposed for reconstruction of FSLM networks from experimental data.

Making the most of data

DNA microarray experiments to measure gene expression are often conducted in triplicate to generate a mean and standard deviation. When analyzing gene-expression data, however, traditional clustering methods usually focus on mean or median values, rather than on the original data, thus ignoring the deviation values. Shinya Matsumoto (NCR Japan, Tokyo, Japan) proposed a clustering method that allows the use of each of the triplicate datasets as a probability distribution function, thereby avoiding the step of pooling data points into a median or mean. This method might be helpful in unsupervised clustering of the data from DNA microarrays, and it could also be used to treat other types of data that retain error or variation information.

Curators of databases also face the problem of making sense of mountains of data. The exponentially growing number of publications has led to the need to automate database entry and annotation. Text-mining techniques and suitable machine-learning algorithms can significantly reduce the workload of curators and increase the efficiency of annotation. Oliver Miotto (National University of Singapore) proposed a new method for supporting curators by means of document categorization, which can be 'trained' by the curator to select the most relevant documents. This approach is attractive as it does not require domain-specific programming or a knowledge of linguistics. To demonstrate its feasibility, Miotto described a prototype application designed to identify PubMed abstracts that contain allergen cross-reactivity information. He and his colleagues tested the performance of two different classifier algorithms (CART and ANN), applied to both composite and single-word features, using several feature-scoring functions. Both classifiers exceeded the performance targets, the ANN classifier yielding the best results: it delivered around 80% recall and as high as 94% precision in a standard implementation.

The conference revealed many new applications and indicated directions of future research. It has shown that the specific bioinformatics approaches to systems biology have become an increasingly important issue for the bioinformaticians who gathered in Yokohama. It will be very interesting to see the dynamic progress to be expected in this area at this year's conference in December in Yokohama.