Meeting report

Genome informatics for data-driven biology

Kenta Nakai* and Jean-Philippe Vert†

Addresses: *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan. †Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan.

Correspondence: Kenta Nakai. E-mail: knakai@ims.u-tokyo.ac.jp

Published: 27 March 2002

Genome ***Biology*** 2002, **3(4)**:reports4010.1–4010.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/4/reports/4010

A report on the 12th International Conference on Genome Informatics, Tokyo, Japan, 17-19 December 2001.

Genome informatics as a field encompasses the various methods and algorithms for analyzing and extracting biologically relevant information from the rapidly growing biological - and especially genome sequence - databases. This leads to a new data-driven research paradigm for post-genomic biomedical research, which Charles (Chip) Lawrence (Wadsworth Center, Albany, USA), speaking at the meeting, claimed is replacing the traditional hypothesis-driven paradigm in which experiments are carefully designed to address a specific prior hypothesis. To illustrate this idea, Lawrence described an automatic procedure for finding putative transcription-factor binding sites in *Escherichia coli* by comparing the non-coding regions near homologous genes in several other proteobacteria genomes and searching for conserved motifs. The relative accuracy of this prediction method enabled his group to formulate hypotheses about new binding sites, which could then be successfully checked through experiments, such as the binding of a putative transcription factor YijC to a predicted site upstream of the *FabA* gene.

Many other talks at the conference highlighted progress in data-driven research paradigms. For example, Limsoon Wong (Kent Ridge Digital Labs, Singapore) presented a novel method for finding simple rules ('emerging patterns') that best distinguish normal and cancerous tissues from microarray data. Here is a sample of the rules he found:

{*gene*(K03001) ≥ 89.20} and {*gene*(R76254) ≥ 127.16} and {*gene*(D31767) ≥ 63.03},

where *gene*(X) represents the expression value of gene X. This rule turned out to hold in 75% of cancer cells but in 0% in normal tissues. Essentially, he developed a novel method ('entropy-oriented discretization method') to automatically find such rules, including the determination of the threshold values. His method showed better performance than previously reported methods of microarray analysis; furthermore, the method seems to be applicable to many other problems, including the prediction of protein subcellular localization.

Comparing gene networks

Several relationships can be formulated between genes in a given organism, such as those based on structural or functional similarity. Moreover, different technologies can be used to systematically check the degree of similarity between any two genes, including sequence comparison, microarray data to track co-regulated genes, pathway or ontology databases for detecting functional similarities, and two-hybrid systems to find interacting proteins. Several talks were devoted to comparing these types of relationships.

Akihiro Nakaya and Minoru Kanehisa (Kyoto University Bioinformatics Center, Japan) proposed a general algorithm for extracting correlated gene clusters by comparing the graphs obtained when each gene is considered as a node and edges are created between genes with a particular relationship; these relationships include sequence similarity, gene interaction, co-expression or participation in a common pathway. Roughly speaking, the idea behind the algorithm is to define a distance between any two genes as the sum of their distances in different graphs, and then to use this distance to cluster the genes. In one of many examples, Nakaya showed that by comparing three gene networks of *E. coli* on the basis of positional, functional and three-dimensional similarity relations, he was able automatically to find a set of three genes, which simultaneously participate in the biotin metabolism pathway, that are close to each other in the *E. coli* genome and belong to the same structural family.

David Eisenberg (University of California, Los Angeles, USA) reviewed two other types of relationship between proteins: the 'Rosetta stone' relationship, based on the detection of fused domains in other genomes, and the 'phylogenetic profile' relationship, based on the correlated occurrence of both genes of a pair in different genomes. He pointed out the surprisingly low correlation between proteins with different types of functional relationships, such as the proteins related by co-expression analysis of microarray experiments, two-hybrid experiments or phylogenetic profiles. This observation was also apparent in the talk by Terry Gaasterland (Rockefeller University, New York, USA), who systematically compared the co-expression of Rosetta stone protein pairs with that of randomly selected gene pairs in the genome of *Saccharomyces cerevisiae*. In spite of a statistically significant difference between the expression-profile distance distributions of the two classes of gene pairs, the overlap is such that protein-fusion events can barely help to predict co-expression patterns. She also observed a non-zero, but low, correlation between Rosetta stone pairs and functional relationship, using pathway identifiers obtained from the Gene Ontology terms database [http://www.geneontology.org] or the Kyoto encyclopedia of genes and genomes (KEGG) database [http://www.genome.ad.jp/kegg/].

Pathway simulation

The simulation of biological pathways, or even of an entire molecular network within a cell (as in Masaru Tomita's 'E-cell' project [http://e-cell.org], where all biochemical and genetic processes within the whole cell are modeled) is regarded as a new frontier in bioinformatics. In making realistic simulations, however, the lack of experimentally determined parameters is a serious problem. One important variable to describe the status within the cell is the metabolic flux distribution (MFD) vector. It is difficult to compute MFD from mass spectroscopy (MS) or nuclear magnetic resonance (NMR) data, however, partly because the computation becomes too heavy. Marcos Araúzo-Bravo (Kyushu Institute of Technology, Japan) proposed a simple method for such a calculation. Although his method is still only applicable to small data sets, such studies will encourage so-called metabolome analyses in the near future.

Hiroshi Matsuno (Yamaguchi University, Japan), in a collaborative work with the laboratory of Satoru Miyano (University of Tokyo, Japan), presented recent advances in their Genomic Object Net software tool for modeling and simulating biological pathways. The simulation mechanism is based on hybrid functional Petri nets, which are a class of networks useful for representing complex interactions between several variables and simulate their evolution over a period of time. In the context of biological pathways, hybrid functional Petri nets were shown to represent intuitively, and simulate naturally, typical pathways of gene regulation, metabolism and signal transduction. The emphasis was placed on removing elements that are biologically irrelevant, in order to make the software tool acceptable in biology and medicine. Matsuno explained how pathways can be represented as documents in the widely used web language XML, and he presented several animated simulations of the gene regulatory network of bacteriophage λ.

Algorithm development

As an important theme of the conference, the development of basic algorithms was key - in addition to state-of-the-art processing pipelines for experimental data. Gene Myers (Celera Genomics Inc., Rockville, USA) addressed the problem of separating a set of multiply-aligned sequences into two closely related subsets. Such a problem is important in phylogenetic analysis or in DNA sequence assembly, where the distinction between real sequence differences and sequencing errors is important. So far, such problems have been dealt with by heuristic methods, but Myers proposed two optimal branch-and-bound algorithms. The branch-and-bound algorithm is a standard strategy to solve optimization problems, where a solution that gives the maximum (or minimum) value of a given objective function is sought. Roughly speaking, its basic idea is to divide the massive search possibilities into smaller parts ('branching') and to try to 'prune branches' by estimating the 'bound' of possible values in each part. If the estimation is effective enough, the search space would be greatly reduced. Thus, the efficiency of the algorithm essentially depends on how to make effective estimation. Myers introduced two kinds of bound function and examined their efficiencies both theoretically and empirically. It will be intriguing to see whether such algorithms actually outperform heuristic methods in real data processing.

The 'best paper' award was won by Richard Lathrop (University of California, Irvine, USA) and his colleagues, who developed a rather general algorithm, a multi-queue variant of the branch-and-bound search algorithm. Their algorithm has many favorable properties, including its completeness and lack of redundancy. They applied their idea to several problems including *ab initio* protein backbone prediction and protein-DNA binding motif discovery. Although the authors admit their results are still preliminary, they are encouraging. Further improvements of the algorithm may make it possible to challenge quite difficult biological problems, such as searching the conformational space of proteins, in a reasonable way.

In 12 years the Genome Informatics meeting has grown from a workshop to an international conference (abstracts from the conference are included in Medline and full texts are freely available online from the Japanese Society for Bioinformatics [http://www.jsbi.org/journal.html]). In that time the role of genome informatics has moved from theory to practice. New methods and algorithms will continue to enable a more systematic understanding of biological data.

Acknowledgements
We thank Hideo Matsuda for critically reading the manuscript.