Meeting report

The intelligence in developing systems for molecular biology

S Cenk Sahinalp

Address: School of Computing Science, Simon Fraser University, University Drive, Burnaby, BC, Canada V5A 1S6.
Email: cenk@cs.sfu.ca

Published: 31 January 2007

Genome ***Biology*** 2007, **8:**301 (doi:10.1186/gb-2007-8-1-301)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/1/301

A report on the 14th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Fortaleza, Brazil, 6-10 August 2006.

The 900 or so participants at the Annual International Conference on Intelligent Systems for Molecular Biology last August were treated to talks on topics ranging from sequence analysis, structural bioinformatics, and comparative genomics through to proteomics and systems biology. It was evident that interest in RNA, especially non-coding RNA (ncRNA), is growing, with quite a few talks on locating and predicting the structure of small (and not so small) ncRNAs. As well as such relatively new topics, the classic problem of discovering sequence motifs and assessing their significance seems to be re-emerging, especially in the context of new applications. As the biological problems scientists aim to address become more complex, the mathematical principles and computational tools being developed to solve them must become more sophisticated. The conference showed that not only are computer science and mathematics being applied to solving key problems in molecular biology, but these problems are inspiring the development of new computer science, and, to a certain degree, new mathematics.

Sequences and statistics

Sequence analysis was still the theme running through most talks. Its application outside DNA and proteins was illustrated by Kiyoko Aoki-Kinoshita (Kyoto University, Japan), who described motif discovery in carbohydrate sugar chains (glycans), the third major class of macromolecules. Starting from a single monosaccharide, many glycans have a tree-like structure consisting of branching chains with various combinations of monosaccharides. Aoki-Kinoshita described a profile Markov model using a probabilistic sibling-dependent tree (PST) that aims to recognize glycan motifs, which are basically paths on their tree representation. The model has been tested successfully on both synthetic glycans and glycan data from the KEGG GLYCAN database, accessed from [http://www.genome.jp/kegg/glycan].

Eugene Fratkin (Stanford University, Palo Alto, USA) described a combinatorial technique for finding motifs. Combinatorial techniques, unlike commonly used machine learning techniques, are based on a branch of mathematics called combinatorics (graph theory is part of combinatorics). The method, appropriately named MotifCut, can be accessed at [http://motifcut.stanford.edu] and is a graph-theoretical approach to the problem which, through an optimization method called convex optimization, can be solved in polynomial time. The main idea of MotifCut is to build a graph in which the vertices represent all sequences of a given length (k-mers) in the input sequences and the edges represent the degree of sequence similarity. In this graph, a motif is defined as the maximum density subgraph; that is, a set of k-mers that have the most highly weighted edges between each pair. The dense subgraph is computed by iterative application of the classic min-cut algorithm (hence the name MotifCut) of Gallo and colleagues (1989). Uri Keich (Cornell University, Ithaca, USA) introduced a new optimization function to improve the ability of the Gibbs sampling algorithm to discover motifs, especially weak motifs. Keich showed that relying on entropy scores and their E-values when finding weak motifs by Gibbs sampling can lead to undesirable results. As an alternative, he suggested using the incomplete likelihood ratio as a scoring function, which performs much better on the famed 'implanted motif' problem. The implanted motif finding problem is an artificial problem in which a motif of a given length (say 17 nucleotides) is randomly implanted in a number of genome sequences (say five); each implantation differs from others in at most a fixed number of locations (for example, three). Knowing the length of the motif, and the differences between the occurrences of the motif, a motif finder is supposed to find the motif exactly.

The problem of counting the occurrences of a position weight matrix in a DNA sequence has applications in *cis*-regulatory analysis. Saurabh Sinha (University of Illinois, Urbana-Champaign, USA) described a probabilistic scoring method to solve this problem in a statistically sound framework. He also described a local search technique to solve the discriminative motif-finding problem; that is, how to find position weight matrices that have high counts in one set of sequences and low counts in another set.

Also addressing fundamental statistical questions in bioinformatics, Karsten Borgwardt (University of Munich, Germany) introduced a test for determining whether two sets of biological observations have been generated by the same probability distribution. This involves a 'kernel'-based statistical test, which compares the maximum discrepancy between the means of a set of functions. A discrepancy between the means of any member of a kernel-function class in the two observations implies a difference in the distributions that must have generated them. The test has been applied to various tasks, such as microarray data comparison, cancer diagnosis and classification of protein function.

One very important and timely problem in sequence analysis was discussed by Tien-Ho Lin (Carnegie Mellon University, Pittsburgh, USA) - the identification of victims in a mass disaster using DNA fingerprints. In such a situation, hundreds of samples are taken from remains that must be matched to the pedigrees of the victims' surviving relatives, and the DNA is also degraded by heat and exposure. Lin described a very interesting probabilistic framework for clustering samples while eliminating implausible sample-pedigree pairings. This framework handles both degraded samples (missing values) and experimental errors in producing and/or reading a genotype.

Lutz Krause (Bielefeld University, Bielefeld, Germany) described the application of the powerful pyrosequencing-based technology (developed by the company 454 Life Sciences and now marketed by Roche Diagnostics) to explore the genomes of organisms that are difficult to culture by conventional means, and which can be studied only through DNA extracted directly from environmental sources. Krause described the development of a new gene-finding algorithm that aims to address the problems in identifying genes from this DNA, namely the short lengths of the contigs and the existence of in-frame stop codons and frameshifts, which arise due to poor sequence quality in DNA extracted from environmental sources.

Exploring gene expression

A popular theme in the contributions on transcriptomics was novel motif-discovery and modeling algorithms for transcription factor binding sites. Barret Foat (Columbia University, New York, USA) described a new algorithm, MatrixREDUCE, to model transcription factor binding sites. MatrixREDUCE can be found at [http://bussemaker.bio.columbia.edu/software/MatrixREDUCE]. The algorithm uses genome-wide occupancy data for a transcription factor and the associated nucleotide sequences to discover the sequence-specific binding affinity of the factor.

Yong Lu (Carnegie Mellon University, Pittsburgh, USA) described the identification of cycling (self-regulatory) genes from gene-expression data. The idea is to combine microarray data from multiple species with sequence information in a graph-theoretical framework in which each gene is represented by a node and each edge represents sequence similarity. Starting from the measured expression values for each species, a 'belief propagation' machine learning approach is used to determine a posterior score, indicating expression, for genes, which is then used to determine a new set of cycling genes from each species.

Gene-expression profiling is commonly used as a tool for identifying genes that are important for the development and maintenance of different cell types. Yuan Qi (Massachusetts Institute of Technology, Cambridge, USA) described work aimed at detecting relevant genes from a large set of expression profiles via a novel Bayesian, 'semi-supervised' clustering method called BGEN. This new method trains a kernel classifier based on labeled and unlabeled gene-expression examples. The semi-supervised trained classifier can then be used to efficiently classify the remaining genes in the dataset.

RNA bioinformatics and structural informatics

The importance of ncRNAs was recognized in 2006 by the award of the Nobel prize for Physiology or Medicine for work on RNA interference (RNAi), and interest in ncRNAs was clear in the number and quality of talks on this topic at the meeting. One theme was the detection of potential ncRNAs in genome sequences. Shaujie Zhang (University of California, San Diego, USA) introduced a framework for constructing and comparing sequence-based ncRNA filters. The use of this framework gives rise to a new formulation of the covariance model, which, in turn, speeds up the alignment of the potential RNA sequence with the model and thus gives a much faster ncRNA filter than the available alternatives. Unlike short interfering RNAs (siRNAs) and micro RNAs (miRNAs), there are no current effective computational and experimental screening methods for the class of ncRNAs known as small modulatory RNAs (smRNAs). These are a novel class of small (approximately 20 base pair) RNAs that are double-stranded, exist in the cell nucleus, and do not code for proteins. Despite their very small size, smRNAs perform a major role in the differentiation of neural stem cells to neurons. There are currently no screening methods for them. Neil Jones (University of California, San Diego, USA) addressed this question and described a graph-theoretical discovery method for long and highly similar motifs

through a comparative genomics approach that does not require an alignment of orthologous upstream regions (which do not align well); which can be accessed at [http://www.cse.ucsd.edu/groups/bioinformatics].

At present, RNA structure prediction is based on thermodynamic models. Chuong Do (Stanford University, Palo Alto, USA) described a computational alternative to these models that derives RNA-folding parameters through statistical learning tools. The computational tool developed, called Contrafold and accessible at [http://contra.stanford.edu/contrafold/], is based on conditional log-linear models, a class of probabilistic models that generalize stochastic context-free grammars. By providing a means of distinguishing RNA stems of different lengths, Contrafold can predict the secondary structure of treacherous RNA sequences, such as 5S rRNA, much more accurately than the thermodynamic models.

Structural-similarity searching among small molecules is a standard tool in molecular classification and *in silico* drug discovery, and public databases of such information are now being developed. I described our team's work on a novel k-nearest-neighbor search method for structural similarity and classification of small molecules, represented by arrays of chemical descriptors. This is aimed at finding the best methods to separate molecules that exhibit a given activity from those that do not. We have shown how to compute a weighted Minkowski distance, which aims to show how similar the molecules are in terms of the bioactivity in question, on the descriptor arrays for the best separation through a linear programming formulation. I also described a data structure that exploits all available memory to search for all similar small molecules to a query molecule through a distance-based approach.

Visualizing systems biology

A common theme in contributions on systems biology was the integration of various data sources for visualizing, inferring the topologies of, or understanding the dynamics of networks and subnetworks. Using genotype information, gene expression, protein-protein interaction, protein phosphorylation and transcription-factor-binding information, Zhidong Tu (University of Southern California, Los Angeles, USA) described ways of showing which genes control the expression levels of a specific gene. He described a stochastic algorithm that infers the causal genes and identifies significant pathways on the expression network where each node is either a protein or a transcription factor.

Yanay Ofran (Columbia University, New York, USA) introduced a new platform for integrating molecular data and insights about the qualities of individual proteins in a network visualizer, which goes beyond the traditional topology-oriented presentation. The platform generates networks on the macro systems level and analyzes the molecular characteristics of each protein on the micro level at the same time. It also annotates the function and subcellular localization of each protein and displays the process on an image of a cell. Adrien Faure (Institut de Biologie du Developpement de Narseille-Luminy, France) aims to understand the dynamics of a regulatory network by treating it as a Boolean logic circuit that can work synchronously or asynchronously. The idea makes a lot of sense, as most of the available data on regulation are qualitative. Faure showed how this general approach can be applied to test some of the dynamical properties of the mammalian cell.

Cells need to adapt the activity levels of metabolic functions to changes in the environment. Jose Nacher (Kyoto University, Kyoto, Japan) explored the connections between the gene-expression response to external changes and the induction or repression of specific metabolic functions. His team has analyzed the transcriptional response of *Saccharomyces cerevisiae* to different stress conditions or stress signals. These signal-induced expression data are then integrated with structural data about the yeast network and the topological properties of the induced or repressed subnetworks are analyzed. These subnetworks turn out to be quite different from random networks; for example, their degree of distribution, the number of vertices with a specific number of neighbors, seems to have a heavy tail, indicating few nodes with many neighbors.

Mustafa Kirac (Case Western Reserve University, Cleveland, USA) addressed the question of automatic assignment of Gene Ontology (GO) annotations to partially annotated proteins through a data mining approach. The most accurate protein annotations are currently provided by curators, but the possibility of automatically assigning annotations through mining of protein-protein interaction networks is appealing. Kirac showed how to compute the probabilistic relationships between GO annotations of proteins and assign highly correlated GO terms of annotated proteins to non-annotated proteins in the target set to achieve a prediction accuracy of up to 81%.

The meeting showed how much bioinformatics has matured in the past few years. The computational tools for what can now be considered as 'classic' bioinformatics problems, such as motif discovery and RNA structure prediction, now have much more solid foundations. The need for depth in developing both mathematical models and algorithm tools is very evident for these problems, and their application is also being broadened. As many of the talks, especially in systems biology, showed, new problems are emerging very rapidly, requiring development of new computational tools that need to integrate various types of data. These are all signs that bioinformatics is maturing into an independent scientific field with considerable depth and breadth.

Acknowledgements

I thank the members of SFU Lab for Computational Biology, in particular Emre Karakoc, Rahaleh Salari, Cagri Aksay and Fereydoun Hormozdiari, for their help.