Meeting report

Understanding biology through intelligent systems

Igor Jurisica*†¶ and Dennis A Wigle‡§¥

Addresses: Departments of *Computer Science, †Medical Biophysics, ‡Surgery and §Medical Genetics and Microbiology, University of Toronto, Toronto, Ontario, M5S 1A8, Canada. ¶Division of Cancer Informatics, Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario, M5G 2M9, Canada. ¥Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario, M5G 2X2, Canada.

Correspondence: Igor Jurisica. E-mail: ij@uhnres.utoronto.ca

Published: 24 October 2002

Genome ***Biology*** 2002, **3(11)**:reports4036.1–4036.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/11/reports/4036

A report on the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB), Edmonton, Canada, 3-7 August 2002.

Summer 2002 in Edmonton was a computer science hot spot as at least six major computational conferences were held in the town during July and August. The International Conference on Intelligent Systems for Molecular Biology (ISMB) is one of three major conferences focusing on computational biology, the other two being the Pacific Symposium on Biocomputing and RECOMB. Since the first conference in 1993, the International Society for Computational Biology has organized the ISMB meeting for advancing the scientific understanding of living systems through computation, with this year's being the largest ISMB conference yet held. Some of the key themes of the conference covered in this report include sequence analysis, processing microarray data, genome sequence annotation, predicting protein structure, and integrating data from different sources.

Sequence analysis

Sequence searches and comparisons are the infrastructure of many bioinformatic efforts and were the topic of a significant number of presentations at ISMB 2002. Stephen Altschul (National Center for Biotechnology Information, Bethesda, USA), the original developer of the BLAST algorithm, presented the history and commented on future trends in statistical methods for assessing sequence similarity. The main focus was on improving sensitivity and specificity of sequence similarity searches, most importantly by using a position-specific scoring system and amino-acid composition-based statistics. He also discussed ideas that did not yield expected improvements. Ford Doolittle (Dalhousie University, Halifax, Canada) further highlighted the application of such comparisons. His keynote address focused on phylogenetic classification and the 'tree of life' derived from sequence data, mainly highlighting archaebacterial and primitive eukaryotic genomes. The idea is to use a 'universal tree of life' as a 'natural' hierarchical classification of all living organisms, but there is evidence that most archaeal and bacterial genomes contain genes from multiple sources. This may be explained by lateral gene transfer, in which case one may not want or be able to define an absolute tree of life.

The concept of using statistical algorithms to find unusual patterns in the composition of unknown proteins has also been explored by Michael J. Wise (Cambridge University, UK), who introduced a set of tools called POPP for clustering proteins using peptide probability profiles. This approach is effective because of the non-random nature of protein sequences, especially in regions such as catalytic domains.

The importance of selecting proper representations of information for analysis was highlighted by Isidore Rigoutsos (IBM TJ Watson Research Center, Yorktown Heights, USA) in his keynote presentation. As he pointed out, the success of any attempt to find significant motifs within a sequence - pattern discovery process - is strongly dependent on effective representation of the sequence in the first place. Rigoutsos's approach to searching systematically for motifs in full genome sequences using the Teiresias algorithm [http://cbcsrv.watson.ibm.com/Tspd.html] has led to the development of a commercially available Bio-Dictionary [http://www.research.ibm.com/bioinformatics/metadata.phtml.html] - a collection of recurrent amino-acid combinations that completely cover 'sequence space', the biggest possible collection of amino-acid sequences that have been found in known proteins. It has been shown that the motifs contained

in Bio-Dictionary can capture both functional and structural signals that have been re-used during evolution both within and across families of related proteins. Seventeen individual dictionaries have already been compiled for complete genome sequences and made publicly available, and these should greatly facilitate comparative genomics and other studies.

Microarray data

Image analysis, pattern discovery, and data mining and interpretation were some of the frequently addressed topics this year. It is apparent that the field is still rapidly evolving. Unfortunately, our rush to obtain biologically or clinically relevant results has often led to ignoring systematic analysis and treatment of errors that are abundant in microarray data processing. A few notable exceptions included posters by Andrew Goryachev (GeneData AG, Basel, Switzerland), who introduced a statistical approach for quality assessment and correction of gene-expression data and presented Expressionist Refiner, a tool for systematically extracting true expression values from raw microarray data. Marlena Maziarz (University of Toronto, Canada) presented a system for assessing the quality of microarray-data image analysis. The focus is on automatically and objectively identifying artifacts in microarray images for each spot, showing the effect of their existence on analysis of microarray data, and suggesting approaches to minimize their impact. The main goal is to allow automated spot-quality assessment and thus classification, but an interesting by-product is a comparison of the advantages and disadvantages of existing commercial and public-domain packages for image analysis.

Given that there are now many data-mining and pattern-discovery approaches available for the analysis of microarray data, it is possible systematically to compare their benefits and drawbacks, which may lead to more powerful hybrid analysis methods. Michael de Hoon (University of Tokyo, Japan) presented a first step in this direction by implementing and comparing the performance of several clustering algorithms. The main result is not surprising - different clustering algorithms produce different results. Thus, one should verify results by applying multiple analysis methods to a given dataset. It is also apparent that changes in the implementation of a single basic algorithm can produce different results. This makes the effort of the Bioinformatics Open Software Consortium [http://open-bio.org/] - a non profit, volunteer-run organization focused on supporting open source programming in bioinformatics - even more important. In addition to improving a specific algorithm, the combination of existing diverse approaches may give us improved analysis and thus better, more biologically relevant, results.

Annotation

Annotation of high-throughput data is indispensable for improving the interpretation and integration of information. Annotation systems must rely heavily on natural language processing algorithms. One such system, using lexical analysis of the SWISS-PROT database [http://www.expasy.ch/sprot/], was presented by Rajesh Nair and Burkhard Rost (Columbia University, New York, USA) with the goal of inferring subcellular localization. Their LOCkey system [http://cubic.bioc.columbia.edu/services/LOCkey] has been successfully applied to the annotation of the predicted proteomes of five entirely sequenced genomes, with more than 82% accuracy in predicting subcellular localization. Michael Krauthammer (Columbia University, New York, USA) presented another systematic analysis of textual information. The GeneWays system [http://genome6.cpmc.columbia.edu/~krautham/geneways] was used to search 50,000 research articles in molecular biology as a way of inferring differences about 'true statements' (as inferred computationally from available evidence) and statements accepted by the community. The work proposes a stochastic model that describes the process of generating and propagating knowledge about molecular interactions through scientific publications. One has to be careful, however, during the interpretation and use of results from mining the text of the available literature, as most of the information that is being text-mined is heavily biased and can be incorrect because inferences are extended beyond the context of the discovery, because of poor experimental results, bad design, or even because of simple typographic errors. Diverse teams looking at the quality of protein-protein interaction data have already highlighted some of these issues.

Systematically collecting available information for each organism in conjunction with annotating full genomes is a valuable approach. At the forefront of this task is a group at the Stanford Research Institute (USA) led by Peter Karp, who introduced the BioCyc collection of pathway and genome databases [http://biocyc.org], with each database encompassing a single organism. This effort clearly shows the need for, and advantages of, a distributed biological knowledge-management system that allows users both to query the database and to enter information into it, as no single group has all the information about any single organism, let alone about multiple species. Visualizing microarray data by overlaying the gene-expression data on top of known pathways provides a powerful approach to data interpretation.

Predicting protein structure

Protein structure prediction is one of the most active and fruitful areas of bioinformatics, as most disease processes and treatments are manifest at the protein level. Interest in this field has been fueled by the rapid progress in determining protein sequences from the starting point of genomic data. The importance of structure prediction lies in the fact that knowing a protein's structure generally contributes to a greater understanding of its function. There are three main approaches to structural prediction: comparative

modeling, threading, and *ab initio* prediction. The first of these, comparative modeling, exploits the fact that evolutionarily related proteins with similar sequences (measured by the percentage of identical residues at each position based on an optimal structural superposition) often have similar structures. The second approach, threading, compares a target sequence against a library of structural templates, producing a list of ranked scores. The fold with the best score is assumed to be the one adopted by the sequence of interest. Finally, *ab initio* prediction of protein structure consists of modeling all the energetics involved in the process of folding and then determining the structure with the lowest free energy, which is assumed to be the native structure.

Although protein structure prediction is generally not yet accurate enough to directly assist in drug design, models produced by prediction algorithms are of sufficient quality to be used to understand and test hypotheses about biological function. A hybrid approach comprising comparative modeling plus threading uses the I-SITES library [http://isites.bio.rpi.edu/] of sequence-structure motifs, the HMMSTR model [http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.html] for local structure in proteins and ROSETTA, the Monte Carlo fragment-insertion method for protein tertiary structure prediction (presented by Christopher Bystroff, Rensslaer Polytechnic Institute, Troy, USA). Validating the system on 40 protein sequence targets, 31 predictions of secondary structure achieved 73% overall accuracy. The 40 proteins used were the targets selected for the 'blind' structure-prediction exercise CASP4 (the fourth community-wide experiment on the critical assessment of protein structure prediction). Pier-Luigi Martelli (University of Bologna, Italy) used a comparative method, namely a hidden Markov model (HMM), to predict β-barrel membrane proteins. By using a dynamic programming algorithm, the model achieved 82% accuracy per residue tested, and the system predicted seven out of twelve topological models included in the test set. An intermediary step in protein structure prediction, namely prediction of maps of the contacts between residues in the protein, or contact maps, was shown by Gianluca Pollastri (University of California, Irvine, USA) to be improved by a hybrid recurrent neural network and HMM approach.

Data integration

Paul Gilna (Los Alamos National Laboratory, USA) suggested earlier this year during a Bioinformatics Workshop at the US National Institutes of Health that currently the three most important aspects of bioinformatics are integration, integration and integration. Three main aspects of integration were pursued by people presenting at the meeting who share this view: integration of different tools and approaches, integration of a single type of data, and integration of diverse data types.

One pitfall in dealing with high-throughput biological data is ascribing too much meaning to individual data points. Many high-throughput datasets, whether from gene-expression profiles or protein-protein interactions, contain noise that can prevent reliable conclusions for specific genes or proteins. Estimates of the error rate in existing protein-interaction datasets run as high as 30%. Although it has been speculated that more meaningful hypotheses might be formulated by integrating the data from diverse functional genomic and proteomic projects, it has until recently been unclear to what extent such data can be correlated and thus how integration can be achieved. Some of the more promising integration strategies begin with the concept of integrating orthogonal (or interdependent) datasets, such as the same kind of information from different platforms. One example would be interaction data from phage display and two-hybrid approaches; other strategies begin with the integration of data of completely different forms - for example, gene expression data with protein-interaction data.

Integration of gene-expression and protein-interaction data was the topic of several presentations at the meeting. Two main themes were the quantification of interactions by providing weight/distance from gene-expression data - not taking interactions as binary relations, but rather as weighted relations, using information from gene-expression data - and determining the quality or reliability of protein-interaction data. It is apparent that the field is moving in leaps and bounds, judging by the progress since the beginning of this year. Several new papers have appeared that increase the cumulative yeast protein-interaction dataset that is publicly available to about 80,000 interactions. Using this information, Trey Ideker (Whitehead Institute, Cambridge, USA) presented an approach that integrates yeast gene-expression data with protein-protein and protein-DNA interaction data to predict regulatory and signaling subnetworks. The main goal is to create concrete hypotheses that can be further verified experimentally. A possible approach to dealing with all the complexities of high-throughput data, and data integration, as well as their use for prediction, was nicely presented by Dana Pe'er (Hebrew University, Jerusalem, Israel), who described a system for the efficient prediction of regulatory sets of genes. The Minreg system uses genome-wide measurements to predict a small set of global active regulators. The predicted regulatory model in *S. cerevisiae* has been cross-validated, and selected predictions have been further subjected to biological analysis. We can expect more computationally generated hypotheses from high-throughput data in the near future.

The coming year of intelligent biology

As Barry Honig (Columbia University, New York, USA) suggested in his keynote speech, we should start focusing on building 'systems for intelligent biologists'. His notion of integrating methods has expanded across multiple disciplines as

he attempts to combine bioinformatics and biophysics to understand protein structure and function. This theme is likely to mature significantly over the coming year, as investigators have more time to process the flood of high-throughput data becoming available and to apply ever more novel approaches. The next meeting will be held in Brisbane, Australia, June 29-July 3, 2003. If this year's conference is an indicator of a trend, then the next meeting will be bigger and better still.