Meeting report

Putting the 'bio' into bioinformatics

Olga G Troyanskaya

Address: Department of Computer Science and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Washington Road, Princeton, NJ 08544, USA. E-mail: ogt@cs.princeton.edu

Published: 29 September 2005

Genome ***Biology*** 2005, **6:**351 (doi:10.1186/gb-2005-6-10-351)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/10/351

A report on the 13th Annual Conference on Intelligent Systems for Molecular Biology (ISMB), Detroit, USA, 25-29 June 2005.

The annual meeting on computational methods for molecular biology brought together 1,731 attendees and covered a diversity of topics from sequence analysis and text mining to structural bioinformatics and pathway prediction. This year saw an increased emphasis on the biological problems that bioinformatic methods are being developed to solve; in addition to many novel developments in traditional areas of bioinformatics, a substantial number of talks focused on integrative approaches, pathway analysis, and comparative genomics. Also on the menu this year were ways of making bioinformatic methods more 'data-centric' and how to make new technologies easily accessible to biologists.

Bioinformatics for biology: from data to results

Numerous presentations reflected the trend for bioinformatic studies to include new biological findings in addition to innovative methods. This mirrors the general trend in the bioinformatics community, as reflected in the recent launch of *PLoS Computational Biology*, which emphasizes the biological results of computational methods, as the official journal of the International Society for Computational Biology (ISCB). The use of computational methods to solve specific biological problems was highlighted in talks such as that of Yoonsoo Hahn (National Cancer Institute, Bethesda, USA), who described the use of comparative analysis of the human and the unfinished chimpanzee genome sequences to identify potential human-specific frameshift mutations that occurred after the divergence of human and chimpanzee. Pavel Pevsner (University of California, San Diego, USA) presented new evidence for rearrangement hot-spots in mammalian genomes, supporting a model of chromosome evolution in which rearrangement breakpoints are much more likely to occur in relatively short fragile areas of the chromosomes. In the area of gene regulation, Wei Li (Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, USA) reported a new method based on hidden Markov models (HMMs) for analyzing chromatin immunoprecipitation-microarray experiments (ChIP-chip) based on tiling arrays, and its use to identify binding sites for the transcription factor p53.

The closer integration of bioinformatics and biology is also reflected in methods that incorporate known biological information into bioinformatic analyses. One such approach was highlighted in a keynote lecture by Jill Mesirov (Broad Institute and Massachusetts Institute of Technology, Cambridge, USA). She described the use of biological information from curated databases to define groups of genes that participate in the same or related processes. Her group (together with collaborators) then examined the expression behavior of these groups of genes using a 'gene set enrichment analysis' method that they have developed and that involves both experimental and computational analysis, and identified genes that link exercise and the metabolism of simple sugars; interestingly, these genes are expressed at lower levels in people with type 2 diabetes. The growing interest in this type of approach has led to an increasing need for methods that extract information automatically from the biological literature. One of these was described by Zhenzhen Kou (Carnegie Mellon University, Pittsburgh, USA), who reported a new learning method, dictionary hidden Markov models, for dictionary-based extraction of protein names from the biological literature.

In his keynote talk, Satoru Miyano (Human Genome Center, University of Tokyo, Japan) addressed the issue of making the advanced algorithms developed by bioinformaticians easily accessible to biological researchers. He emphasized the need for advanced computational methods to have user interfaces that can be intuitively understood by biologists and to include effective visualization of results. Such user-friendly

algorithm implementations should be freely available for download and use. The numerous software demonstrations at the conference demonstrated that user-friendly implementations of bioinformatic algorithms are becoming more commonplace, and increasing numbers of these packages are indeed freely distributed as open source software.

Data-centric approaches

In addition to becoming more focused on biology, bioinformatics is becoming increasingly data-centric in that the biological data themselves are crucial in the development of the method or algorithm. This emphasis on data requires published data to be freely available in integrated databases, the development of methods tailored toward unique characteristics of the data in hand, and thorough evaluation of computational methods on real biological data. Ewan Birney (European Bioinformatics Institute, Hinxton, UK) emphasized the central role of biological data in bioinformatics in his ISCB Overton Prize lecture. He highlighted the importance of databases in bioinformatics and emphasized the need for more research on databases and more open data sharing.

Work on ontologies and databases was well represented and included technologies for the creation, analysis, visualization, and integration of ontologies and databases. Kei-Hoi Cheung (Yale University, New Haven, USA) presented a standard based on a resource description framework (RDF) for the integration of genomic databases into a data warehouse. A prototype application of this system, called YeastHub, incorporates a variety of yeast data and allows RDF-based queries.

Data-centric approaches do not stop with data storage and sharing, however. Analysis methods are now being created with specific data properties in mind. To take one example, the emphasis is moving from general gene-expression analysis tools to tools specialized for particular tasks or particular types of expression data, such as the clustering algorithm for short time-series microarray data presented by Jason Ernst (Carnegie Mellon University, Pittsburgh, USA). More general approaches to data analysis can also provide an effective and robust solution. In regard to sequence analysis, new techniques were presented for long-standing challenges such as the identification of repeats, exon detection, and homolog analysis. Sequence-based techniques are increasingly being used in functional genomics for predicting molecular function and identifying regulatory motifs. In one such study, Tali Sadka (The Hebrew University, Jerusalem, Israel) used the amino-acid composition of transmembrane domains to assign proteins to their functional family with high accuracy.

Integrative technologies

At a time when increasing amounts and types of high-throughput biological data are being generated, integrative bioinformatic technologies that can combine information from multiple experimental methods and diverse organisms are becoming essential. The numerous data-integration algorithms presented at the meeting illustrated the variety of areas in which combined analysis of diverse data sources can lead to valuable advances. In functional genomics, for example, Asa Ben-Hur (University of Washington, Seattle, USA) introduced a kernel method, which uses a kernel function to implicitly transform data into a higher-dimensional feature space, for predicting physical interactions between proteins on the basis of a combination of protein sequences, Gene Ontology annotations, homology information, and local properties of the protein-protein interaction network. Elena Nabieva (Princeton University, USA) presented an algorithm based on network flow that exploits the structure of protein-interaction maps constructed from different types of genomic data to predict protein function. She described how the performance of this algorithm is substantially improved by considering multiple data sources combined in a weighted interaction network.

Going beyond studies of a single organism, several approaches incorporated phylogenetic information into analyses. Some of these methods focused on problems in comparative genomics, including phylogenetic tree construction and detection of co-evolving genomic sites. Matthew Dimmic (University of Copenhagen, Denmark) introduced a Bayesian phylogenetic approach for the detection of coevolving amino-acid residues in protein families. This method can provide information about interacting sites on proteins: when it was applied to eukaryotic phosphoglycerate kinase family proteins, interdomain site contacts were found to have coevolved significantly more frequently than non-contact sites. Others focused on using information about homology to address a more general set of problems. Mary Dolan (Jackson Laboratory, Bar Harbor, USA) presented a general method for evaluating the consistency of Gene Ontology protein annotations and demonstrated its application by comparing mouse and human homolog annotations. Raja Jothi (National Library of Medicine, Bethesda, USA) predicted protein-protein interactions based on protein co-evolution; for this, he and his colleagues have developed a new method for identifying the best superposition of the corresponding evolutionary trees based on tree automorphism groups (tree structures with one-to-one mapping of both nodes and edges).

One of the forthcoming challenges for the integrative approach will be to combine biological information at different levels of resolution. The Physiome project, described by Peter Hunter (University of Auckland, New Zealand) in a keynote address, is attempting to develop an infrastructure for computational physiology that will integrate genomic, proteomic, morphological and physiological information across different time scales and levels of spatial organization to provide the 'physiome' - the quantitative and integrated

description of the functional behavior of the physiological state of an individual or species. The heart physiome project is currently constructing integrated models from the molecular level all the way to the whole-organ scale, some functioning on the microsecond timescale and others changing slowly throughout the human lifetime.

The meeting clearly showed how state-of-the-art bioinformatic technologies are making a significant contribution to solving important biological problems. Much progress is being made, both in traditional areas of research and in new directions. Many challenges still lie ahead, however - challenges that promise an exciting future for bioinformatics as an integral part of systems-level biology.

Acknowledgements

O.G.T. is partially supported by NIH grant RO1 GM071966, NSF grant 0406415 (to Kai Li), and NIGMS Center of Excellence grant P50 GM071508 to David Botstein. Matthew Hibbs and Chad Myers have contributed to this report with many helpful discussions.