Meeting report

Computational biology touches all bases

Susmita Datta and Somnath Datta

Address: Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40292, USA.

Correspondence: Susmita Datta. Email: susmita.datta@louisville.edu

Published: 17 February 2009

Genome ***Biology*** 2009, **10:**303 (doi:10.1186/gb-2009-10-2-303)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/2/303

A report of the 6th Annual Rocky Mountain Bioinformatics Conference, Aspen, USA, 4-7 December 2008.

The sixth Rocky Mountain bioinformatics conference, held under the auspices of the International Society of Computational Biology, attracted an international audience that ranged from biologists to statisticians to physicists. A diverse range of topics were addressed, including proteomic biomarker discovery, biological text mining, systems modeling, genomic and protein-protein interaction networks, medical informatics and computational bioethics. Abstracts of the talks are available at [http://www.iscb.org/cms_addon/conferences/rocky08/agenda.php#allen] and poster abstracts are available at [http://www.iscb.org/cms_addon/conferences/rocky08/pdf/RockyPosters2008Nov13.pdf]. Here we focus on some of the highlights of the meeting.

Reconstruction and analysis of genetic networks

Computational approaches to understanding the topology of genomic and proteomic networks have sparked considerable interest in recent years. Analysis of a single gene or protein in the context of a complex biological function cannot provide a systematic understanding of the interrelated actions and interactions of a cohort of genes or proteins. But the experimental detection of gene-gene or protein-protein interactions using high-throughput technology is labor-intensive, expensive and also not always reproducible. So aiding experimentation with computational approaches provides an efficient alternative.

Andrea Califano (Columbia University, New York, USA) presented an overview of the construction of association networks and identification of 'master regulatory modules' affected by biological processes such as tumor development. He discussed the use of ARACNE (algorithm for the reconstruction of accurate cellular networks) [http://amdec-bioinfo.cu-genome.org/html/ARACNE.htm] to infer gene coexpression association networks from microarray gene-expression data. ARACNE uses a pairwise scoring mechanism that has its roots in information theory. Given two genes, a mutual information score is calculated based on their mRNA abundances. Califano described the use of ARACNE and master regulator analysis to predict regulatory interactions in human B cells that control the formation of germinal centers, and also to identify the genes responsible for the initiation and maintenance of the mesenchymal phenotype of glioblastoma multiforme.

Graph-theoretic methods for the construction of protein-protein interaction networks were discussed by Suzanne Gallagher (University of Colorado, Boulder, USA). Her rationale was to identify protein complexes or groups of proteins binding together to perform a specific task. Her team has surveyed the topological properties of known protein complexes in isolation and in the context of high-throughput data to aid in the prediction of unknown protein-protein interactions. Also addressing the question of inference of protein-protein interaction networks, Todd Gibson (University of Colorado, Denver, USA) pointed out that the stationary steady-state topological information on protein complexes may not be enough to construct a network. Instead, it might be necessary to use evolutionary snapshots of the genes or proteins gained from a phylogenetic tree. Anis Karimpour-Fard (University of Colorado, Denver, USA) echoed the same sentiment, showing that phylogenetic profiles (co-conservation information) of genes provide better information for the construction of protein-protein interaction networks. In addition, the connectivity of a network constructed this way is more informative for predicting protein function.

Michael Verdicchio (Arizona State University, Tempe, USA) described an assessment of two competing algorithms - mIC

(modified inductive causation) and mIC-CoD (modified inductive causation - coefficient of determination) - to predict causal relationships between multiple genes. Both these algorithms are modified versions of the original Pearl's inductive causation (IC) algorithm. In these algorithms, he used partial prior knowledge about the topological ordering of the genes in the context of interactions. The mIC_CoD algorithm optimizes the algorithm by considering the number of causal parents using the coefficient of determination. Both algorithms proved superior in sensitivity and specificity to a Bayesian network inference method when tested on simulated and real microarray data.

Vincent VanBuren (Texas A&M Health Science Center, College Station, USA) described the prediction and visualization of a group of interactive genes using 'guilt by association' with a particular gene of interest in the context of microarrray studies. His team calculated a large correlation matrix of gene pairs in a heterogeneous data set built from 2,145 publicly available microarray samples. An association network was then constructed based on pairs with correlation coefficients exceeding a certain threshold. In addition, they have developed a web-based visualization tool for the networks called StarNet [http://vanburenlab.medicine.tamhsc.edu/starnet.html]. Vanburen also discussed Cognoscente, a visual database of existing interactions for more than 300,000 unique gene ID pairs that he and colleagues have developed. The unique feature of this database is that users can also add novel interactions to it, provided the interaction is published with a PubMed ID. This database currently has entries for 20 different species.

As well as network reconstruction, researchers are interested in detecting genes (or modules of genes) or proteins that are altered in different biological conditions. One of us (Susmita Datta) presented a new statistical method for constructing coexpression association or interaction networks from microarray gene-expression data, by fitting multiple partial least squares (PLS) models to the data (the association/interaction scores obtained). Formal statistical tests that can be used to detect a change in connectivity of a gene or set of genes, or to detect a difference in global modular structure between two networks were also discussed. The R-code for network construction using PLS can be accessed at [http://www.susmitadatta.org/Supp/GeneNet/supp.htm].

Visualization and prediction

Tools and techniques for a visual analysis of high-dimensional datasets are important components of bioinformatics research. Chris Shaw (Simon Fraser University, Vancouver, Canada) introduced the stereoscopic field analyzer (SFA), a glyph-based visualization software system that has the capability for interactive visualization, manipulation and exploration of multivariate, regular and irregular volumetric data. This technique has the potential to visualize three-dimensional shape and interaction data more clearly. Shaw and his group have used it to visualize multivariate time-varying flow-cytometry data.

State-of-the-art prediction of biological phenomena using simulation models, computational methods, large databases and novel experimental data was well represented at the meeting. Tejaswi Gowda (Arizona State University, Tempe, USA) described the use of threshold logic to construct gene-regulatory models for anterior-posterior segmentation and dorsal-ventral germ layer formation in fruit fly embryos. The steady-state predictions from these models were in good agreement with normal development. These models were also useful in studying the effects of under- and over-expression of particular genes. Todd Castoe (University of Colorado Health Sciences Center, Aurora, USA) presented methods for inferring the evolutionary history of protein sequences and repeated DNA elements using partial genome sequence sampling (random shotgun samples and transcriptome sequencing). As one does not need to observe the entire genome, these methods are a cost-effective means of exploring the new generation of vertebrate genomes.

Elizabeth Siewert (University of Colorado, Denver, USA) described the use of multivariate regression techniques to predict transcription factor binding sites from cross-species expression and sequence data. An exciting feature of this work is that she was able to use additional information from related species and exploit such correlations to strengthen the predictive ability of the model, and she showed that her approach improved the accuracy of prediction compared with earlier methods using data from a single species. Yiqiang Zhao (Indiana University, Indianapolis, USA) has used classification methods to differentiate regulatory single-nucleotide polymorphisms (SNPs) from non-functional SNPs in putative transcription regulatory regions. Gene-expression features, codon usage and functional complexity information were used to build the classifier. In particular, Zhao reported that the distance from the SNP to the transcription start site turned out to be an important variable for predicting regulatory SNPs. This work is a good example of how correlated the various characteristics of genomic elements really are, and how one should utilize information from multiple fronts in building such models.

Computational bioethics

An interactive panel-style workshop session on computational ethics conducted by Lawrence Hunter and Mark Yarborough (University of Colorado, Denver, USA) proved interesting and lively, covering ethical issues that are unique to computational scientists compared with bench and clinical scientists. Computational biologists dealing with computerized medical records and linking them with the various genomic and proteomic databases need to be acutely aware of data privacy, security and intellectual property

issues. These issues will become even more relevant in the future as medical informatics, computational biology and bioinformatics become more widely used in medical research.

Small-scale conferences such as Rocky 2008 are extremely important for driving the field forward and the work presented resulted in many enlightening discussions. We look forward with anticipation to Rocky 2009.

Acknowledgements

Susmita Datta's attendance at the meeting was supported by a grant from the National Science Foundation. Somnath Datta's attendance at the meeting was supported by funds from Elsevier and the National Science Foundation.