Meeting report

Bioinformatics meets systems biology

Carlos Salazar, Jana Schütze and Oliver Ebenhöh

Address: Theoretical Biophysics, Institute for Biology, Humboldt University, Invalidenstrasse 42, 10115 Berlin, Germany.

Correspondence: Carlos Salazar. Email: carlos.salazar@rz.hu-berlin.de

Published: 31 January 2006

Genome ***Biology*** 2006, **7:**303 (doi:10.1186/gb-2006-7-1-303)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/1/303

© 2006 BioMed Central Ltd

A report on the Fifth International Workshop on Bioinformatics and Systems Biology, Berlin, Germany, 22-25 August 2005.

The efficient integration of bioinformatics and systems biology requires worldwide cooperation not only in the research of senior scientists but also in the research training of young scientists. To this end, a student-focused workshop on bioinformatics and systems biology [http://www.biologie.hu-berlin.de/gk/ibsb2005] was held last August at Humboldt University in Berlin, Germany. This was the fifth annual workshop held as part of a research collaboration between the Bioinformatics Program of Boston University in the USA, the Bioinformatics Center of Kyoto University in Japan, and the Berlin-located graduate program 'Dynamics and Evolution of Cellular and Macromolecular Processes'. This time the meeting had two main themes - the integration of genomic and chemical information in the analysis of the dynamics and topology of cellular regulatory networks, and the development of more accurate computational tools for the analysis of gene expression and the prediction of transcription-factor binding sites. Full papers accepted for the fifth workshop have been published in the Genome Informatics Series of the Japanese Society of Bioinformatics, edited by Satoru Miyano (University of Tokyo, Japan) [http://www.jsbi.org/journal/GI16_1.html].

From traditional genomics to chemical genomics

Trends in genome biology and bioinformatics were highlighted in the opening talk by Minoru Kanehisa (Kyoto University Bioinformatics Center, Japan), whose group is responsible for the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [http://www.genome.ad.jp/kegg]. This stores molecular interaction networks and graphics, including metabolic pathways, regulatory pathways and molecular complexes. Kanehisa emphasized the importance of an integrated analysis of genomic and chemical information to predict the complete functional behaviors of cells, organisms and ecosystems. While traditional genomics and other 'omics' have contributed to our knowledge of the genes and proteins that make up a biological system, new chemical genomics initiatives will give us a glimpse of the compounds and reactions that exists as an interface between a biological system and its environment. Kanehisa reported the recent release of databases of chemical information, such as GLYCAN [http://www.genome.jp/kegg/glycan] for complex carbohydrate structures, and DRUG [http://www.genome.jp/kegg/drug] for structures of clinically relevant compounds, which have been inserted into the composite database KEGG LIGAND [http://www.genome.jp/ligand].

Computational tools have been developed for chemical genomics, including graphic-based methods for analyzing chemical compounds and reactions. In this regard, Kosuke Hashimoto (Kyoto University) introduced the 'composite structure map' (CSM) [http://www.genome.jp/kegg-bin/draw_csm] that allows a global analysis of carbohydrate structures. Using the KEGG GLYCAN database, he could represent all possible variations of these structures in a single structure called a variation tree. The CSM tool integrates the variation trees with a list mapping glycosyltransferases to their catalyzing glycosidic linkages, creating a bridge between carbohydrate structures and functions. With such powerful tools, biologists now have access to a wide spectrum of methods for investigating how the genomic and chemical organization of organisms governs their cellular behavior.

The development of bioinformatics tools that link the chemical and genomic spaces was vividly demonstrated by Yoshinori Tamada (also at Kyoto University), who presented a novel computational method for identifying the effects of drugs on genes and their regulatory relationship. In the first

step, a gene network is reconstructed from microarray gene-expression data generated from single-gene disruptions. To estimate the structure of the gene network, Tamada and colleagues use a Bayesian network with nonparametric regression. Then the estimated continuous Bayesian model is converted into a discrete Bayesian model to compute the probabilities of gene expression in the drug-response data for every time point. The time-dependent relationships among the estimated drug-affected genes are used to reduce the number of falsely identified drug-affected pathways.

The need for efficient algorithms analyzing chemical information, such as the detection of pattern in carbohydrate structures, was addressed by Lidio Carvalho-Meireles (also at Kyoto University). A variety of computational methods have been developed for sequence analysis, whereas only a few methods are available for tree-structured data. Carvalho-Meireles presented an innovative approach to detecting tree patterns in a database of rooted unordered labeled trees. By analogy with the concept of a sequence motif and profile, he defined a tree pattern as a tree motif and profile - that is, a tree with associated position-specific label probabilities. His algorithm enumerates different tree topologies and subsequently identifies motifs using Gibbs sampling. It has been used to detect tree motifs within the GLYCAN database.

Computational molecular biology

There are still challenging computational problems in the reliable detection of certain types of sites in genomic DNA. Inverted repeats in DNA consist of two sequences separated by a spacer region, with one sequence being inverted and complementary to the other, and appear to play an important role in DNA replication and the generation of genomic instability. Gary Benson (Boston University, USA) presented recent work from his group on a program, Inverted Repeats Finder (IRF), for detecting approximate inverted repeats in long genomic sequences. Candidate inverted repeats are detected by finding short, exact reverse-complement matches of four to seven nucleotides between non-overlapping fragments of a sequence. The program has been successfully applied to the detection of inverted repeats in the human genome.

The focus of the keynote lecture by Zhiping Weng (also from Boston University) was the computational analysis of transcription-factor binding in the human genome. An unprecedented opportunity for investigating the binding of the cell-cycle regulatory protein p53 in the human genome has been provided by Weng's collaborators at the Genome Institute of Singapore (GIS) who have mapped p53 binding throughout the whole genome in the human cancer cell line HCT116 using chromatin immunoprecipitation (ChIP) coupled with paired-end di-tag (PET) sequencing. By combining information from ChIP-PET and previously characterized p53 sites, Weng, together with scientists at GIS, has developed a computational analysis for a precise and unbiased global mapping of p53 binding sites. Experimental and statistical verification have shown that overlapping PET clusters resulting from p53 ChIP DNA fragments define p53-binding loci with high specificity. From this information, they have also discovered previously unidentified p53 target genes implicated in novel aspects of p53 functions.

Despite considerable effort by theoretical biologists, most purely computational techniques for the prediction of transcription-factor binding sites are unsatisfactory. A key problem appears to be that many positions within the binding sites are not conserved in terms of sequence. Heather Burden (Boston University) hypothesized that such positions contain structural codes that are essential for recognition by the appropriate transcription factors. The structural codes can be defined by base-pair step parameters that describe the relative displacement and orientation of two adjacent base pairs in a nucleic acid structure. She described a method, called identification of conserved structural features (ICSF) [http://zlab.bu.edu/ICSF], that uses base-pair step parameters obtained from a collection of high-resolution DNA crystal structures to discover structural conservation in the sequentially degenerate areas within a binding site. By focusing her study on the Jaspar database [http://jaspar.cgb.ki.se], she found that one third of the binding sites contain this structural conservation.

The integration of different types of genomic information to identify transcription-factor binding sites computationally was discussed by Dustin Holloway (also at Boston University). Such binding sites in gene promoter regions are often predicted using position-specific scoring matrices, which summarize the sequence patterns of experimentally determined sites. Holloway is attempting to reduce the high number of false-positive sites predicted by these scoring matrices. His method is based on the integration of various types of genomic data, such as binding-site degeneracy and conservation, phylogenetic profiling, binding-site clustering and gene-expression profiles, by overlapping the datasets using a Bayesian allocation procedure and support vector machine classification.

The importance of the statistical analysis of microarray data was stressed by Gyan Bhanot (Institute for Advanced Study, Princeton, USA), who pointed out the necessity for robust classification models in order to make cancer diagnoses from microarray data. He presented a classification method that was originally developed for phenotype identification from mass spectrometry data. It uses a robust multivariate gene selection procedure and combines the results of several machine-learning tools on raw and partly analyzed data to produce an accurate meta-classifier. Of particular importance is that this method is independent of the specific analysis technique and can combine data obtained in different laboratories.

Dynamics and topology of cellular networks

The emphasis of the research groups from Berlin that participated in the meeting lies in the investigation of the kinetic behavior and architecture of metabolic and regulatory networks. The mutual benefits of a collaboration between experimental and theoretical research was illustrated in the presentations of Uwe Vinkemeier (Institute for Molecular Pharmacology, Berlin, Germany) and Thomas Höfer (Humboldt University, Berlin, Germany). Vinkemeier described a thorough experimental analysis of the STAT1 signaling system, focusing on the nucleocytoplasmic cycling and transcriptional regulation of STAT1, while Höfer has developed a mathematical model of the interferon/STAT1 pathway that is consistent with these experiments. The model shows that hitherto rather unexplored processes, such as the dephosphorylation of STAT1 and its nuclear export, can regulate the expression of STAT1 target genes.

A successful application of theoretical methods to clinical research was described by Branka Čajavec (also at the Humboldt University). She presented a mathematical model for the molecular processes involved in Huntington's disease, a progressive degenerative brain disorder caused by a mutation in the protein Huntingtin. In particular, the model was used to analyze the dynamic behavior of the protease caspase-2 and the release and aggregation of the mutant forms of Huntingtin. The results generated by the model help to provide insight into the molecular steps involved in the development of this disease.

The importance of microarray data for understanding large-scale interaction networks was stressed by Martin Vingron (Max Planck Institute for Molecular Genetics, Berlin, Germany). He compared microarray experiments on the cell cycles of three model eukaryotes, namely budding yeast, fission yeast and human cells. A subset of orthologous genes with cyclic expression patterns was determined, giving a hint about which events during the cell cycle are conserved in all eukaryotes.

To gain insight into the structural design, dynamics and functional properties of large-scale interaction networks is a major goal of systems biology. Thomas Manke (also at the Max Planck Institute for Molecular Genetics) presented a topological analysis of protein-interaction networks based on the concept of network entropy, a measure of the complexity of the wiring. He showed that nodes with a high contribution to entropy are generally associated with elements of functional importance such as proteins essential to survival.

The structural design and the dynamical properties of a protein kinase network derived from the Transpath database [http://www.biobase.de/pages/products/transpath.html] have been investigated by Bernd Binder (Humboldt University) in an approach to understanding the functioning of large signal-transduction networks. On comparing the Transpath network with random networks, he observed that it exhibits special features that might be the result of natural selection during its evolution. In particular, input kinases and output kinases are generally connected by the shortest signaling routes and the Transpath network contains no cycles whereas they generally appear in random networks of the same size. Binder introduced a measure for quantifying the strength of cross-talk between different signaling routes with which he could characterize the cross-talk spectrum of the Transpath network.

Thomas Handorf (also at the Humboldt University) introduced the newly developed method of network expansion and the concept of 'scopes' to analyze large-scale metabolic networks [http://scopes.biologie.hu-berlin.de], making use of the KEGG database. The scope of a metabolic network is defined as its capacity to synthesize a wide variety of different metabolites when it is provided with a few small chemical substances as external resources. Using this method, a hierarchical structuring of metabolism has been revealed, and it was shown that networks with a large scope, that is, with a high synthesizing capacity, also show a high degree of robustness in the face of structural changes. One of us (O.E.) described the application of this method to compare the carbon-utilization spectra of 178 organisms available in KEGG from three main groups - eukaryotes, bacteria and archaea - through a comparison of their metabolic networks. Together, these sorts of investigations provide ideas for investigating how the structure of a network is responsible for its functional behavior and may give valuable hints on the evolution of metabolism.

A special aspect of these workshops is that many of the talks are given by postgraduate students from the participating research groups. At the end of the 2005 meeting, the three participating institutions announced the intention to establish a common program of postgraduate education and research, which will help to increase collaboration by providing a framework for the exchange of doctoral students and joint supervision of PhD theses. The sixth workshop is scheduled for the summer of 2006 in Boston and is likely to facilitate the interactions between the three participating universities even further.