Meeting report

# *In silico* meets *in vivo*

Zhiping Weng*†‡§ and Roderic Guigó¶¥

Addresses: *Bioinformatics Program, †Biomedical Engineering Department, Boston University, Cummington Street, Boston, MA 02215, USA. ‡Program in Bioinformatics and Integrative Biology, §Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA. ¶Center for Genomic Regulation, C/Dr. Aiguader 88, 08005 Barcelona, Spain. ¥Research Group on BioMedical Informatics, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08022 Barcelona, Spain.

Correspondence: Roderic Guigo. Email address: roderic.guigo@crg.es

Published: 25 February 2008

*Genome* ***Biology*** 2008, **9**:302 (doi:10.1186/gb-2008-9-2-302)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/2/302

© 2008 BioMed Central Ltd

A report of the 6th Georgia Tech-Oak Ridge National Lab International Conference on Bioinformatics '*In silico* Biology: Gene Discovery and Systems Genomics', Atlanta, USA, 15-17 November, 2007.

Technological developments have had a profound impact on biology during the past decade, spectacularly augmenting our ability to survey and interrogate biological phenomena. In particular, they have increased capacity for data generation by several orders of magnitude and made computation a necessary partner of biology. The sixth meeting in the biennial series of bioinformatics conferences co-sponsored by Georgia Institute of Technology in Atlanta and the Oak Ridge National Laboratory addressed the challenges that this technology-driven avalanche of data pose to bioinformatics - increasing the complexity of long-standing problems and creating new ones.

## Genome alignment and gene prediction

Sequence alignment is unquestionably one of the 'founding problems' in bioinformatics. The availability of sequenced genomes of many species has highlighted the need for methods of making reliable multiple alignments of whole genomes. The alignment of entire genome sequences is much harder to achieve than the alignment of amino-acid sequences of individual proteins, because of the much longer sequences involved (ranging from megabases to tens of megabases), complex evolutionary relationships among the genomes (such as duplications, deletions and translocations) and heterogeneous mutation rates along the sequence. Different methods often produce discrepant alignments with the same set of genomic sequences, and Martin Tompa (University of Washington, Seattle, USA) has attempted to navigate through this complexity. Instead of proposing yet another method for multiple sequence alignment, he presented an approach to evaluating the quality of a given multiple alignment. This is a seemingly more modest goal; he was, however, able to identify high-quality and reliable regions in the multiple alignment, which is very important because downstream comparative genome analysis is compromised by incorrect alignments. Tompa presented data showing that about 10% of the positions in multiple alignments of the human genome with other vertebrate genomes - a widely used technique in comparative genomic studies - are likely to be incorrect.

Gene prediction in genomic sequences presents similar problems. Current methods for predicting the exonic structures of protein-coding genes from genomic sequences are generally based on computational models that capture our understanding of the way proteins are encoded in genomes. However, recent surveys of the transcriptional activity of the human genome, made possible by advances in microarray and high-throughput sequencing technologies, are challenging the very notion of a protein-coding gene. One of us (RG) presented the results of one such survey, using genome-wide tiling arrays, that indicates, contrary to common belief, that most protein-coding genes in humans occupy large portions of genomic space and their boundaries are quite diffuse, exhibiting extensive overlaps with neighboring genes. Similarly, Steven Salzberg (University of Maryland, College Park, USA) reported that gene overlaps are much more prevalent than previously anticipated in prokaryotic genomes and presented an evolutionary model to explain their retention. He also described extensions to his gene-prediction work using the program Glimmer [http://www.cbcb.umd.edu/software/glimmer] that attempt to cope with such complexity.

The avalanche of data is, of course, beneficial in general. Mario Stanke (University of Göttingen, Germany) presented a method by which data generated to identify genes in one genome can be used to identify genes in another. In this approach, cDNA sequences of one species - assumed to correspond to protein-coding mRNAs - are aligned to its genome, and then the alignment is mapped, via synteny, to the genome that needs to be annotated, using the program Augustus [http://augustus.gobics.de]. Surprisingly, this approach worked better than the direct alignment of 'non-native' cDNAs to a genome in need of annotation, which has been the typical approach.

Even with the newest rapid-sequencing techniques, the cost of coverage required for a complete assembly of a genome is still prohibitive and many genomes are sequenced to only three- to five-fold coverage, resulting in fragmentary genome sequences even when there is no misassembly. Tatiana Tatusova (National Institutes of Health, Bethesda, USA) illustrated the dangers of assuming that biologically meaningful data can be obtained from such draft genomes using standard computational approaches. She showed, for instance, that low genome sequence coverage correlates with frameshifting disrupting the inferred protein sequence. She illustrated this by comparing annotations in the cow genome (less than 5x coverage) with the equivalent in the mouse genome (more than 5x coverage).

Phylogenetics in particular has felt the impact of the avalanche of genomic data. As more sequences become available for a larger number of species, building phylogenetic trees becomes computationally more demanding. Although algorithms are being developed to minimize computing time and memory, even computationally savvy biologists still need assistance in selecting the most appropriate algorithms and running them. Jean-Michel Claverie (University of the Mediterranean, Marseilles, France) has recognized this challenge and is one of the leaders of the Phylogeny.fr project. The goal of this project is to provide state-of-the-art algorithms for phylogenetic reconstruction in an integrated manner with a user-friendly interface. These algorithms are accessible to experimental biologists as a web server [http://www.Phylogeny.fr], providing the computational resources required to analyze larger datasets. Our personal experience with the tools generated within this project is very positive.

## From sequence to function

Improved techniques have also led to a surge of data identifying regions in metazoan genomes that are bound by regulatory proteins or bear epigenetic marks that influence transcriptional regulation. One of us (ZW) presented an integrative analysis of open-chromatin (DNase-chip) and chromatin immunoprecipitation followed by microarray (ChIP-chip) data in several human cell lines, concluding that most ubiquitous open chromatin regions belong to two types: promoters for housekeeping genes or insulators bound by a protein named CTCF. On the other hand, cell-type-specific open chromatin regions are decorated with a large number of epigenetic marks, are bound by enhancer-binding proteins, and harbor motifs recognized by transcriptional factors specific to the corresponding cell type. Martin Vingron (Max Planck Institute for Molecular Genetics, Berlin, Germany) reported an affinity-based model for predicting binding sites for transcription factors in DNA regions detected by ChIP-chip. His model uses a sophisticated normalization scheme such that the binding-site scores of different transcription factors can be directly compared. Jun Liu (Harvard University, Cambridge, USA) presented a multivariate regression approach for predicting expression patterns from promoter sequences. He concluded that this approach does not over-fit the data and hence is more accurate than a previously implemented Bayesian network method. These studies illustrate new methodological developments driven by the availability of new types and large amounts of data. Martha Bulyk (Harvard Medical School, Boston, USA) updated the meeting on her experimental work characterizing the binding properties of recombinant transcription factors. This study aims to characterize the properties of the DNA-binding domains of hundreds of transcription factors and will have an impact on most computational algorithms that use libraries of such motifs. Bulyk showed how such well-characterized motifs can be used to find *cis*-regulatory modules in promoters. Soojin Yi (Georgia Institute of Technology, Atlanta, USA) presented an analysis of the methylation and evolution of CpG-rich and CpG-poor promoters. She proposed that the evolution of CpG islands was associated with promoters and that this was a unique feature of vertebrate development.

Statistical genetics has not traditionally been considered an area of bioinformatics, but has attracted the interest of many bioinformaticians over the past few years as a result of the availability of genome-wide data. New sequencing technologies and new ways to classify clinical populations have resulted in whole genome sequences being produced for many well phenotyped individuals, which will greatly facilitate the search for genes underlying human phenotypes, and hence diseases. The identification of rare variants from association studies, on the other hand, requires genotyping data from large populations. Shamil Sunyaev (Harvard Medical School) is addressing this problem, and presented a theoretical study on how many genome-sequenced and phenotyped individuals are required to achieve this goal. He concluded that, whereas genome-wide analysis of rare coding variation in individuals at phenotypic extremes will provide a powerful tool for discovery of new gene-phenotype associations, these analyses are likely to require sequencing of very large population panels exceeding 10,000 individuals.

New areas more recently incorporated into bioinformatics, such as systems biology and chemical genomics, were also discussed at the meeting. System-wide modeling, for instance, is starting to incorporate genomic data, as in the work of James Galagan (Broad Institute, Cambridge, USA). He reported on the metabolic modeling of the tuberculosis bacterium (*Mycobacterium tuberculosis*) using the flux-balance approach but incorporating gene-expression data. The goal is to identify bottlenecks in the metabolic pathway that could be used to aid elimination of the mycobacteria. Joel Bader (Johns Hopkins University, Baltimore, USA) presented a method for delineating genome-wide networks based on graph diffusion kernels or clustering/segmentation. This method can reveal the most salient modules of a complex network and aid focused follow-up experimentation. Minoru Kanehisa (Kyoto University, Kyoto, Japan) reported on the efforts of the Kyoto Encyclopedia of Genes and Genomes (KEGG) consortium to integrate information on the genomic space (including sequence, transcription and proteome information) with information on the chemical space.

MicroRNAs (miRNAs) and other noncoding RNAs are another new area recently incorporated into bioinformatics. Anders Krogh (University of Copenhagen, Copenhagen, Denmark) presented a novel iterative method, MASTA, for simultaneous structure prediction and multiple alignment of noncoding RNAs. Artemis Hatzigeorgiou (Institute of Molecular Oncology, Varkiza, Greece) also talked about bioinformatics identification of miRNAS and miRNA targets over a wide range of organisms. Hanah Margalit (Hebrew University, Jerusalem, Israel) described a new method for identifying targets of viral miRNAs, which was then used to search transcribed regions of the human genome to find those genes targeted by cytomegalovirus miRNA, which may possibly be involved in the way the virus circumvents the immune system. Indeed, among the miRNA targets that Margalit and her colleagues identified is a gene that encodes a stress-induced ligand recognized by natural killer (NK) cells, and which is critical to the killing of virus-infected cells by the NK cells. Margalit's talk also highlighted a new direction in bioinformatics: tighter integration of biology and computation. In collaboration with immunologists and pathologists, her computational team was able to prove downregulation of production of the NK-cell ligand by the viral miRNA and, thus, a direct effect of a viral miRNA on the host immune system. Olga Troyanskaya (Princeton University, Princeton, USA) combines computation and experimentation on a large scale: her lab has developed several methods for predicting gene function, which they have used to predict the functions of hundreds of mitochondrial genes. The predictions are followed up by experimental testing and the results are used to improve the computational methods.

The past ten years have witnessed tremendous growth in bioinformatics. It is now an established area and has made a large impact on our understanding of biology and medicine. Experimentalists are becoming more versed in using routine bioinformatics tools and some bioinformaticians are picking up experimental techniques to test their own predictions. The overarching theme is a tight coupling between computation and experimentation in collaborations and consortia, and the emergence of a new generation of scientists skilled in both. These are exciting times for biology.