Meeting report

Discovery and hypothesis generation through bioinformatics

Joaquín Dopazo* and Patrick Aloy†

Addresses: *Bioinformatics Department, CIPF, Autopista del Saler 16, E-46013 Valencia, Spain. †EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

Correspondence: Joaquín Dopazo. Email: jdopazo@ochoa.fib.es

Published: 27 February 2006

Genome **Biology** 2006, **7:**307 (doi:10.1186/gb-2006-7-2-307)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/2/307

A report on the 4th European Conference on Computational Biology and the 6th Spanish Annual Meeting on Bioinformatics, Madrid, Spain, 28 September-1 October 2005.

A combined meeting including the European Conference on Computational Biology (ECCB) and the annual Spanish meeting on bioinformatics brought more than 700 bioinformaticians to Madrid last September. The conference covered both classic and state-of-the-art bioinformatics and computational biology topics, such as protein structure, genes and genomes, phylogenetics, text mining, microarrays, polymorphisms and systems biology. More technical topics such as algorithms and databases were also addressed. A students' symposium, where more than 20 high-quality presentations were given by graduate students and postdocs, proved very popular.

Given the vast amount of genomic data now available, the biomedical sciences are slowly but steadily moving towards a more computational perspective. Along these lines, Jean-Michel Claverie (Centre National de la Recherche Scientifique, Marseille, France), one of the keynote speakers, encouraged the use of bioinformatics as an efficient approach to discovery and to generation of hypotheses, in contrast to the typical role attributed to it by experimentalists (and assumed by a number of computer scientists) of a simple system to manage biological data. He remarked that computational biology seems to be the natural way to systems biology, although a note of caution was necessary because we still do not know many of the parts, the functions, and the relationships of the whole system we want to model.

A description of what we do know about higher eukaryote genomes was given by Ewan Birney (European Bioinformatics Institute, Hinxton, UK), head of the Ensembl project [http://www.ensembl.org]. He presented an interesting historical perspective on how the number of genes in the human genome has shrunk whereas the number of exons and transcripts has increased as prediction methods improve and more experimental information has become available. Birney also pointed to the ENCODE project [http://www.genome.gov/10005107], in which 44 regions covering a total of 30 Mbp (1% of the human genome) are being studied in great detail, as an essential source of information that can help to improve prediction methods and our comprehension of the architecture of the genome and transcriptome.

In a keynote lecture, Chris Sander (Memorial Sloan-Kettering Cancer Center, New York, USA) put forward his vision of the integration of predictions from computational biology, supported by information systems and available data, with experiments as a way to understanding biological systems. He also commented on various initiatives aimed at organizing information in biomolecular networks, including common standard formats, such as the ones put forward by Biopax [http://www.biopax.org]; representation and integration of relationships, such as the ones developed by Cytoscape [http://www.cytoscape.org]; and the role played by RNA and, in particular, microRNAs (see, for example, the microRNA targets listed on the Memorial Sloan-Kettering Cancer Center's Computational Biology Center website [http://www.microrna.org]). Sander wound up his talk with the observation that evolution also needs to be integrated in this systems biology framework.

Ana María Rojas (National Center of Biotechnology, Madrid, Spain) described a real-life collaboration between experimental and computational groups in which a computational approach using phylogeny and structural analysis was used to identify amino-acid residues required for the dimerization of chemokine receptors, followed by experimental corroboration using fluorescence resonance energy (FRET).

Mining the transcriptome

Results from the transcriptome were probably the most surprising and unexpected among all the 'omics' discussed at the conference. The study of the transcriptome is rapidly moving from the coding regions of the genome to the noncoding regions. In his keynote lecture, Tom Gingeras (Affymetrix, Santa Clara, USA) revealed an unexpected landscape of transcription outside coding regions of the ENCODE project; this transcription apparently has a biological role, although what this role is remains unknown. He urged a change from a protein-coding view to a broader transcript-centric view to address problems in cell biology. In this regard, Sungroh Yoon (Stanford University, USA) presented a new method for predicting groups of microRNAs and genes that potentially participate cooperatively in post-transcriptional gene regulation via the RNA-interference pathway.

Another proof of the attraction the RNA world exerts on computational biologists was the session on microarrays, which have passed in only a few years from being a novelty to being a classic topic at any bioinformatics conference. Nevertheless, microarray technology is still far from being a settled discipline, and there were many controversial discussions of how data should be analyzed. How to select the genes that are differentially expressed between experiments is a problem that has long defied easy resolution. Sach Mukherjee (University of California, Berkeley, USA) proposed a new data-adaptive test for differential expression in which test statistics were learned from the data using the notion of reproducibility. The test shows superior performance to t-tests and other similar alternatives. Claus Vogl (Graz University of Technology, Austria) addressed another classic topic in microarray analysis - clustering. He presented a new model-based method for clustering gene-expression profiles, in which missing value imputation and estimation of the number of clusters are built-in features.

The added value of three-dimensional structures

The functional mechanisms of most biochemical and cellular processes are, to a great extent, determined by the three-dimensional structures of the proteins and nucleic acids involved. The prediction of protein structure from sequence information has thus been one of the major goals of computational biologists in the past decades. Zhiping Weng (Boston University, USA) presented a dictionary of super-secondary structures that, when correctly combined, are able to describe a substantial portion of all known protein folds. This repertoire of structural fragments can be used as the starting point for building three-dimensional models of full-length proteins and for understanding how protein folds have evolved.

As computational biologists know, however, it is fairly easy to make predictions such as three-dimensional models; the difficulty arises when these predictions need to be accurate. It is thus crucial to be able to assess the quality of theoretical models. Alejandro Giorgetti (University of Rome, Italy) has investigated the relationship between the quality of homology models and their usefulness to solve the phasing in X-ray crystallography by molecular replacement. He showed how small changes in the model could make a big difference and suggested that the modeling community should focus on improving their models over the best available templates. In his keynote lecture, Temple Smith (Boston University) stressed the central role played by structural bioinformatics, as it is three-dimensional structures that will ultimately provide the molecular details needed to understand biological processes.

Structure-based approaches are also being developed to tackle cell-biology problems such as the regulation of gene expression through small noncoding RNAs. Oranit Dror (Tel Aviv University, Israel) presented a novel method for comparing and analyzing nucleic acid three-dimensional structure. This approach worked equally well for large RNA folds (such as those in the ribosome) and for short local tertiary motifs (such as those in microRNAs).

Systems biology and evolution

After decades of gene-centric approaches to biology, systems biology, which tries to understand the whole as something more than the simple sum of its parts, is becoming increasingly popular. The study of regulatory processes and of the interactions among genes and their gene products are key to understanding how biological systems work. Inferences drawn from gene regulatory networks and protein-protein interactions in yeast, using a unification of Bayesian networks and Markov networks, allowed Satoru Miyano (Human Genome Center, Tokyo, Japan) to predict roles for genes of unknown function. Miyano described how this improved reconstruction of genes and protein networks allows the discovery of false positives in high-throughput data such as yeast two-hybrid data. The study of coexpression modules, constituted by groups of coexpressed genes identified in multiple experiments was addressed by Dmitriy Leyfer (Gene Network Science, Ithaca, USA) by means of a new approach that simultaneously identifies the number and sizes of such modules.

Nature has been carrying out gene knockout 'experiments' for millions of years and the results can be read in the sequences of living organisms. Computational biologists have learned this lesson and use evolution extensively as a tool for prediction and hypothesis generation. There appears to be a general trend towards the use of large-scale phylogenies or single-nucleotide polymorphism (SNP) analysis. For example, Toni Gabaldón (Centro de Investigación Príncipe

Felipe, Valencia, Spain) described a comprehensive large-scale phylogenetic analysis of eukaryotic and prokaryotic genes, which has identified and defined orthologous groups of genes related to the endosymbiosis of proto-mitochondria, monitored their subsequent losses in eukaryotic lineages and predicted functional interactions among them. With regard to human polymorphisms, Tomás Marqués-Bonet (Universitat Pompeu Fabra, Barcelona, Spain) presented a heuristic method that allows proper multiple-testing adjustment for whole-genome scans in which a sliding window is used to locate potentially interesting candidate regions to be associated with the trait under investigation.

The poster presentations at the meeting can be considered a thermometer of the interests of computational biologists. Figure 1 shows the distribution of the 348 posters presented across the different topics covered by the conference. Clearly, protein structure is still the most popular among European computational biologists, followed by genes and genomes. Algorithms and databases, taken together, still occupy third place. Systems biology and microarrays are becoming consolidated as two driving forces of today's computational biology. In contrast, interest in phylogenetics seems to have shrunk in comparison with other conferences. This is a false impression, however, as phylogeny, and evolution in general, is embedded in many applications in other topics. Text mining, SNPs and polymorphisms still have a relatively small presence in computational biology. Their influence could be considerable, however; for example, Robert Hoffmann (Centro Nacional de Biotecnología-CSIC, Madrid, Spain) presented iHop, a web-based text-mining tool that uses genes as hyperlinks between sentences and which makes PubMed into a navigable resource.

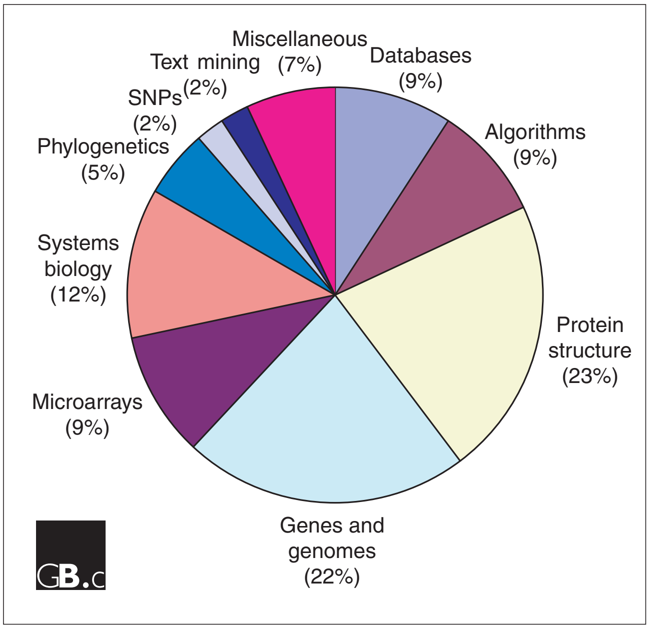

Figure 1
Distribution of the 348 posters presented at ECCB05 among the different themes of the conference.

We will see whether these trends still hold at the next ECCB [http://www.eccb06.org], which will take place in Israel. More detailed information on the 2005 conference can be obtained from the conference website [http://www.eccb05.org] or the Spanish National Bioinformatics Institute (INB) webpage [http://www.inab.org].