Meeting report
From sequence to structure to networks

Nir Yosef* and Lukas Käll†

Addresses: *School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. †Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholms Universitet, Stockholm, Sweden.

Correspondence: Lukas Käll. Email: lukas.kall@cbr.su.se

A report on the 7th European Conference on Computational Biology (ECCB), Cagliari, Italy, 22-26 September 2008.

The diversity of modern biology and the increasing interest of the computer-science community in the field provide the computational biologist with a wide selection of research topics to choose from. This was well reflected at the recent European Conference on Computational Biology in Cagliari. Topics ranged from sequences and structures to systems biology and phylogeny, from DNA to mRNA to proteins. Abstracts of the meeting are published in *Bioinformatics* (15 August 2008) and are freely available online at [http://bioinformatics.oxfordjournals.org/content/vol24/issue16/index.dtl]. Here we focus on three cherry-picked areas: algorithms for next-generation sequencing, structural bioinformatics and systems biology.

Adapting sequence analysis methods to new sequencing technologies

The recent development of so-called next-generation sequencing - that is, massively parallel DNA/RNA sequencing technology - was primarily motivated by the immense cost of genome sequencing, and promises to reduce this cost by orders of magnitude relative to traditional Sanger sequencing. These new technologies do, however, have certain shortcomings. The most striking is the significantly shorter sequence read lengths, which call for new or adapted computational methods for sequence analysis. Daniel Dalevi (Lawrence Berkley National Laboratory, Berkeley, USA) talked about annotation of such short sequence reads in metagenome datasets generated by the 454 Life Sciences pyrosequencing platform. He compared a number of methods for estimating the abundance of protein families and the existence of different taxa in a given microbial community. Using simulated metagenome datasets as a benchmark, he argued that replacing a short sequence read with a full-length protein or set of proteins that have high local similarity with the short read (the so called 'proxygene' or 'proxygene cluster') leads to improved identification of the functional and taxonomic content of the community.

Fabio de Bona (Friedrich Miescher Laboratory, Max Planck Society, Tübingen, Germany) presented an extension to the well known Smith-Waterman alignment algorithm, adjusted for short sequence reads. The algorithm, dubbed QPALMA, incorporates information about the read quality as well as splice-site prediction and an intron model. A support vector machine classifier was trained to optimally weight the different types of information. The algorithm was benchmarked on a large set of *Arabidopsis thaliana* reads from an Illumina Genome Analyzer, and a dramatic reduction in the error rate was found when information on read quality, intron length and splice-site prediction was included.

Computational insights into protein structures

One of the challenges of structural bioinformatics is to predict the function of a protein given its structure. This is most efficiently done when we can infer function from homologous structures with known function. One of the most widely used ways to detect structural homology is protein structure alignment. This has classically been done by measuring atom distances in rigid-body superimposition of the compared structures; however, methods allowing some structural flexibility were recently shown to be more accurate in picking up distant homologies. Gergely Csaba (Ludwig-Maximilians-University, Munich, Germany) presented what he called a phenotypic plasticity method (PPM) for structure alignments that borrows a couple of concepts from sequence alignments. Like sequence alignments that estimate the evolutionary cost of altering one sequence into the other sequence by mutations, insertions and deletions, PPM tries to estimate the cost of altering one structure into

the structure being compared. The ability to detect distant homologies on a test set of some 5,000 protein domains was better than with existing structure alignment methods.

A popular experimental technique for detecting functionally relevant residues is to test whether the function of the protein is altered by replacement of the selected residue with alanine (alanine scans). Yana Bromberg (Columbia University, New York, USA) argued that this could be done *in silico* using an artificial neural network, SNAP, which is aimed at detecting functional changes due to single-position mutations. The prediction method was benchmarked on a curated set of experimental alanine scans to cover about two-thirds of all residues that alter the protein's function while allowing one-third of the predictions to be false. Bromberg also reported that single-residue replacement with alanine is, according to her *in silico* method, as informative as the average result of sequential replacement of the residue with the other possible amino acids.

An alternative approach to inferring protein function is to try to predict its interaction partners. Rafael Najmanovich (European Bioinformatics Institute, Hinxton, UK) discussed the importance of knowing the location of the binding site when discriminating between different potential ligands. The presented algorithm, IsoCleft, aims at discriminating between potential ligands using a graph-matching approach. When the exact location of the binding site is given, the method is (expectedly) reasonably accurate. Unfortunately, the performance of the approach drops off quickly when uncertainty about binding-site location is artificially introduced. But if information about the geometry of the cleft surrounding the binding site is included in the model, then this drop in accuracy is significantly reduced. Najomanovich therefore reasoned that, even though information about the geometry of the cleft does not help in discriminating cases where we know the exact location of the binding site, such information will help when we have less detailed information about the location of the binding site. Maybe the most surprising conclusion from the study is that including highly conserved cleft atoms does not seem to contribute more than non-conserved cleft atoms to the overall performance of the algorithm. Najmanovich explained that atoms might be conserved for reasons other than satisfying the constraints posed by the ligand. The topic of inferring protein function by predicting interaction partners leads us to our third area of focus.

From interactions to networks

A holistic perspective on the analysis of biological systems - the systems-biology approach - is now one of the most popular fields in computational biology. Researchers in this field observe proteins and genes as parts of interacting sets and make extensive use of data on physical protein-protein interactions, transcriptional regulatory interactions, enzyme-

substrate relations and more. In addition to the ISOclef method, described above, several other talks were dedicated to elucidating interactions of various types between proteins or genes.

Chia-Ying Yang (National Taiwan University, Taipei, Taiwan) presented the PhosphoPoint database of phosphorylation in humans. This database integrates protein kinases and phosphorylation sites with a large collection of protein-protein interactions, aiming to provide researchers with a comprehensive repository of information on the human kinome. Yang and colleagues also utilized gene-expression profiles and similarities in gene ontology (GO) annotation to highlight over 180 new potential kinase-substrate pairs, which are available as part of the database. The relation between GO similarity and the interactions between proteins or protein domains was also discussed by Jayesh Pandey (Purdue University, West Lafayette, USA), who described a new method for measuring functional similarity using GO annotations and investigated how this similarity is correlated with topological proximity in protein-protein and domain-domain interaction networks. His new scoring method explicitly accounts for the specificity of the compared annotations (rather than considering only their common ancestor in the ontology) and for multiple annotations (rather than simply taking the average or maximum), and produced improved results for both proteins and domains.

Transcriptional regulatory interactions also received their share of attention. Tristan Mary-Huard (UMR AgroParisTech/INRA, Paris, France) described a way of efficiently identifying these interactions using data from chromatin immunoprecipitation/DNA microarray (ChIP-chip) experiments. He argued that a two-component mixture model could be used, where one component is fitted to the distribution of microarray signal from immunoprecipitated transcription factor binding sites and one component to the signal from other regions. This achieves better discrimination between binding and non-binding sites as well as better statistical measures than existing methods.

Fantine Mordelet (Ecole des Mines de Paris, France) presented a method for predicting transcriptional regulatory interactions from gene-expression data. The bioinformatics literature includes a large number of methods for transcription network inference based on correlations in expression, including Bayesian networks, methods based on information theory, and more. Surprisingly, as Mordelet explained, none of these seem to have employed supervised learning - that is, used the increasing amount of data from experimentally verified interactions. Using a set of validated transcriptional regulatory interactions from *Escherichia coli* in conjunction with a large compendium of gene-expression data, Mordelet has constructed a supervised model (using support vector machines) that predicts regulation on the basis of similarity in gene-expression profiles. The main assumption

of her method is that the transcription factor of interest has at least a few known direct targets. On the one hand, this is quite a limiting assumption, which prohibits the analysis of new transcription factors. On the other hand, where some of the targets of the transcription factor are known, the method gives a most impressive improvement in accuracy.

Gene-expression data, be they time series or compendia of various experiments, are often used for modeling and understanding the logical structure and/or the dynamics of gene regulatory networks. One key problem in inferring such systems is to correctly account for latent variables (that is, components that are controlled outside of the subsystem being modeled and that are often not directly measurable). Neil Lawrence (University of Manchester, UK) described a method for inferring transcription factor activity, which is typically very hard to discern experimentally, by repre-senting the activity as a latent variable in a dynamic model of the expression of the factor's downstream genes. The method uses Gaussian processes to define priors over the time-continuous activity of the transcription factor, and a set of differential equations (linear and nonlinear) to model its effects on gene expression. The parameters of this probabi-listic model are fitted according to the observed time-series of mRNA concentrations, and were then used to estimate the unobserved activities of a number of transcription factors, including P53 (humans), and LexA (*E. coli*).

Although biochemical network models of this kind are often very comprehensive, they are confined to a small scale, primarily because of the lack of sufficient data. Aitor González (Kyoto University, Japan) described such a fine-grained model of the logic of the Hedgehog (Hh) signaling pathway in *Drosophila melanogaster* and its role in wing development. The model was constructed manually and comprises seven genes and four cell types. Interactions between genes were collected from the literature and trans-lated into a set of logical rules describing the effect of one gene on the activity of the others. Through simulations of the dynamics under various perturbation regimes, this model was shown to provide novel insights regarding several, less understood, aspects of this system.

Invigorating poster sessions, talks and keynote speeches at the meeting covered many perspectives in the growing discipline of computational biology. We look forward to the next ECCB meeting, which will be co-located with Intelligent Systems for Molecular Biology (ISMB) in Stockholm in 2009.