Meeting report

Evidence for intelligent (algorithm) design

Balaji S Srinivasan*[†], Chuong B Do[‡] and Serafim Batzoglou[‡]

Addresses: *Department of Electrical Engineering, Stanford University, Stanford CA 94305, USA. [†]Department of Developmental Biology, Stanford University, Stanford CA 94305, USA. [‡]Department of Computer Science, Stanford University, Stanford CA 94305, USA.

Correspondence: Serafim Batzoglou. Email: serafim@stanford.edu

Published: 25 July 2006

Genome ***Biology*** 2006, **7:**322 (doi:10.1186/gb-2006-7-7-322)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/7/322

A report on the 10th annual Research in Computational Molecular Biology (RECOMB) Conference, Venice, Italy, 2-5 April 2006.

More than 700 computational biologists convened in beautiful Venice in early April for RECOMB 2006, the 10th annual Conference on Research in Computational Molecular Biology. After 40 talks, 6 keynote lectures, 180 posters, and at least two cameos by the Riemann zeta function, several emerging trends in computational biology are apparent.

First, there has been a strong shift towards empirical studies of molecular evolution and variation, with approximately 25% of the papers in this broad area. We expect that this number can only increase in the near future, given the ENCODE project [http://www.genome.gov/10005107] and the forthcoming release of several new eukaryotic genomes.

Second, there is a resurgence of interest in two of the oldest problems in computational biology: RNA folding and protein sequence alignment. The interest in noncoding RNAs (ncRNAs) is driven by experiment: recent work on RNA interference (RNAi), microRNAs, ribozymes, and the rest of the 'modern RNA world' has once again stimulated interest in the classical problems of ncRNA identification and fold prediction. Advances in protein sequence alignment draw on the development of new algorithmic and machine-learning techniques for principled estimation of gap penalties (the penalty for inserting a gap in the alignment to improve it) and the rigorous incorporation of non-local similarity measures that move beyond residue-residue similarity.

Interest in classic areas such as protein structure and folding remains strong, with several papers tacitly or explicitly motivated by the coming flood of data promised by structural genomics. Many of the other mainstays of computational biology were also represented at the conference, including old favorites such as expression analysis and genome evolution, as well as the newer areas of data integration and network alignment. Notable by their absence were papers on genome assembly and human single-nucleotide polymorphism (SNP) variation; this is likely to be a fluke rather than a trend, however, given the impending deluge of data from high-throughput sequencing and resequencing projects. We have selected a few of the talks that particularly caught our eye out of the many excellent ones given at the conference.

Focus on ncRNA folding

One of the highlights of the conference was the demonstration by Ydo Wexler (Technion-Israel Institute of Technology, Haifa, Israel) of a quadratic time algorithm for RNA folding, a result that deservedly won a special mention award. For several decades, RNA folding algorithms had running times that scaled with at least the cube of the length of the RNA sequence. This $O(L^3)$ time complexity worsens further if pseudoknots are involved in the folding model. By combining a simple 'triangle inequality' heuristic with empirical validation of the 'polymer zeta' behavior of RNA folding, Wexler and colleagues developed an $O(L^2)$ average time algorithm for RNA folding, a result which makes high-throughput ncRNA prediction far more feasible.

The pseudoknot, a fold comprising two or more helical segments connected by single-stranded loops, was the subject of a talk by Banu Dost (University of California, San Diego, USA), who presented a new algorithm for aligning a subset of ncRNAs with computationally tractable pseudoknots to a database of known ncRNA sequences. Sequences that interact to serve a structural or biochemical function often show coevolution. In this regard, Jeremy Darot (University of Cambridge, UK) presented a general probabilistic graphical model for detecting interdependent evolution between sites in nucleic acid and protein sequences, which he applied to the problem of identifying secondary and tertiary structure interactions in tRNA.

Interaction networks and microarray analysis

The broad area of functional genomics encompasses methods for the prediction of gene function and interaction. Talks covered the inference and comparison of protein-interaction networks, and the statistical issues associated with the detection of functional enrichment in microarray data. One of us (B.S.S) described an algorithm for integrating a number of different predictors of protein interaction without making assumptions about statistical dependence. He showed that this approach revealed hidden interactions that would not have been found without data integration, and used the method to produce probabilistic protein-interaction networks for 11 microbes. Benny Chor (Tel-Aviv University, Tel-Aviv, Israel) presented work on graphs of metabolic reactions from different species, showing that a taxonomy inferred from network-based characters corresponded fairly well to the known consensus phylogeny.

The problem of comparing large collections of networks from different species motivates work on network alignment, whose goal is to detect conserved modules between networks. By analogy with the existing theory for sequence alignment, Mehmet Koyutürk (Purdue University, West Lafayette, USA) presented an asymptotic theory for estimating the statistical significance of network alignments, with respect to certain classes of large random networks. Developing a version of this theory applicable to alignments of few proteins, which are more common in practice, is an open problem.

Steffen Grossmann (Max Planck Institute for Molecular Genetics, Berlin, Germany) presented an improved statistic for estimating the functional enrichment of gene sets based on Gene Ontology (GO) that takes account of the complex parent-child dependencies in the GO hierarchy (this statistic is implemented in the Ontologizer package available at [http://www.charite.de/ch/medgen/ontologizer]). Stefanie Scheid (Max Planck Institute for Molecular Genetics) presented a novel permutation-filtering technique for the detection of differentially expressed genes in microarray analysis. Her method filters the results of a naive data permutation to estimate a more accurate null distribution, and her work is implemented in the Twilight package available online [http://www.bioconductor.org].

Parameter estimation in protein sequence alignment

Two speakers addressed the issue of estimating parameters such as substitution scores and gap penalties for protein sequence alignment. John Kececioglu (University of Arizona, Tucson, USA) provided a solution to the 'inverse sequence alignment' problem, where one estimates parameter values from a training set of alignments. He described a linear programming algorithm for determining a set of alignment parameters under which every example alignment in a given training set is guaranteed to be nearly optimal with respect to that parameter set. The algorithm can learn both residue substitution and gap scores simultaneously, and it will be interesting to see how the resulting parameters perform when used to make new alignments.

One of us (C.B.D) introduced pair-conditional random fields for incorporating non-local sequence similarities (such as hydropathy) into the alignment scoring framework. As such similarities are functions of peptide windows of variable length rather than of individual residues, they cannot easily be incorporated into standard methods based on hidden Markov models (HMMs) for sequence alignment without heuristics. The resulting algorithm, CONTRAlign (source code available online [http://contra.stanford.edu/contralign]), achieves the highest cross-validated pairwise protein alignment accuracies to date.

Protein structure, dynamics and identification

Perhaps the biggest obstacle to deriving insights from protein structure is the sheer size and complexity of a typical polypeptide. Addressing the problem of protein structure alignment, Wei Xie (University of Illinois at Urbana-Champaign, USA) and Jinbo Xu (Massachusetts Institute of Technology, Cambridge, USA) manage this complexity by focusing on maps of intra-protein contacts. Xie presented work on aligning structures by overlapping their contact maps, by developing a brand-and-reduce algorithm that allows rapid superposition of structurally homologous proteins. By analogy with sequence alignment, Xu presented a polynomial time-parametric algorithm for aligning a protein represented by a contact map to another protein represented by a contact map or an interatomic distance matrix.

Many scientists are interested not just in alignments of stable protein structures, but also in the dynamics of the folding process. Chakra Chennubhotla (University of Pittsburgh, Pittsburgh, USA) reduced the complexity of an all-atom protein simulation by calculating a low-rank, eigenmode-based approximation to the molecular dynamics that is designed to preserve certain stochastic properties of the original protein. Shawna Thomas (Texas A&M University, College Station, USA) took a technically different but conceptually similar approach by approximating a protein as a chain of rigid bodies and then sampling its conformation space with a probabilistic roadmap method imported from motion planning for robotics (for further details, see the parasol website [http://parasol.tamu.edu/foldingserver]). The 'roadmap' in the protein context contains thousands of feasible folding pathways.

The ultimate purpose of protein-folding simulation (as distinct from protein-structure prediction) is to use the observed dynamics to yield insight into aspects of protein biochemistry, such as cooperativity or macromolecular assembly. To this end, Tsung-Han Chiang (National University of Singapore, Singapore) showed that the probabilistic roadmap framework can be used to calculate the probability of proper folding from any given protein conformation, and then to estimate protein-folding rates.

Two speakers addressed problems of fast protein identification by clever hashing methods. Brian Chen in collaboration with Viacheslav Fofanov (both from Rice University, Houston, Texas, USA) showed that one can use geometric hashing techniques to speed up the identification of three-dimensional structural motifs in functionally uncharacterized proteins of known structure. In a different problem domain, Nuno Bandeira (University of California, San Diego, USA) demonstrated a rapid hashing algorithm for identifying proteins from tandem mass spectrometry (MS/MS) spectra. The input protein sample is split into two groups, chemical modifications are applied to one group, and spectra are obtained for both groups. Bandeira showed how using correlations between the two spectra greatly reduces the noise of protein identification.

Reconstructing the past

Talks on evolution and phylogenetics included richer models of sequence evolution, new methods for tree building, and applications of molecular evolution to questions in functional genomics. On the topic of richer models for deducing phylogeny from sequences, Yun Song (University of California, Davis, USA) described a method for including gene conversion in reconstructions of SNP phylogenies (software available online [http://www.cs.ucdavis.edu/~gusfield]); existing methods typically incorporate only point mutation and recombination as possible events. Sagi Snir (University of California, Berkeley, USA) presented work on the inference of micro-indel events (insertions and/or deletions) from multiple sequence alignments; the method has a time-complexity that is exponential in the number of species, but is linear in terms of sequence length. Miklós Csürös (Université de Montréal, Montreal, Canada) dealt with gene evolution. He described a parametric model for gene family evolution that models gene duplication, gene loss, and (most significantly) horizontal gene transfer.

With respect to the general problem of building trees from data, Constantinos Daskalakis (University of California, Berkeley, USA) described an algorithm for calculating phylogenies from distance matrices (compiled from the differences between sequences), which compares favorably to neighbor-joining on specific examples, without requiring strong assumptions about possible model tree topologies. Adam Siepel (University of California, Santa Cruz, USA) addressed the problem of using molecular evolution to detect functional elements in genomic sequences. He has extended the phastCons program, a phylogenetic HMM model for segmenting a genomic sequence into conserved and nonconserved regions, by introducing lineage-specific models which allow for simple gains or losses of constraint along specific branches of the evolutionary tree relating the sequences. The output of the program, called DLESS, is available as a track on the University of California Santa Cruz genome browser [http://genome.ucsc.edu/encode].

Evolution also figured prominently in the only talk at the conference given by a non-scientist. In his keynote address, author and journalist Carl Zimmer warned of the perils of ‘genomic myopia’ and challenged computational molecular biologists to create a model of life’s evolution that was consistent with the wealth of knowledge from paleontology and the fossil record. Given the rapid advance of bioinformatics apparent at RECOMB 2006, we have no doubt that our community is up to the challenge.