BIOINFORMATICS

Vol. 17 no. 10 2001
Pages 859–861

Editorial

THE THIRD GEORGIA TECH–EMORY INTERNATIONAL CONFERENCE ON BIOINFORMATICS: IN SILICO BIOLOGY; BIOINFORMATICS AFTER HUMAN GENOME (NOVEMBER 15–18, 2001, ATLANTA, GEORGIA, USA)

Steering & Program Committee: Mark Borodovsky and Eugene Koonin, Co-chairs, Chris Burge, Jim Fickett, John Logsdon, Andrej Sali, Gary Stormo, Igor Zhulin.

Two previous bioinformatics meetings organized by Georgia Tech were held in Atlanta in 1997 and 1999, attracting outstanding researchers from fourteen countries around the globe and establishing this conference as a major academic forum for intensive and open exchange of new ideas. This year we have invited many rising junior scientists to deliver plenary lectures on their discoveries made in silico, with many of these advances driven by the explosion of genomic and proteomic data. The focus of the 2001 conference, in comparison with the two previous ones, has shifted to comparative and evolutionary genomics and proteomics and to new methods and tools which promise further progress in these areas (http://exon.biology.gatech.edu/conference). Other changes include Emory University joining Georgia Tech as a co-organizer of the meeting and a move to a new location. The Sheraton Midtown hotel not only offers a magnificent bird's-eye view of the cosmopolitan capital of the American Southeast, but it is also located near the Atlanta Arts Center and the city's historic area.

Several papers presented at the conference are included into this current Special Issue of *Bioinformatics*, with their publication serving a parallel avenue of presentation for conference participants. Fifteen papers out of twenty-six submitted manuscripts successfully passed peer review and were accepted for publication. The Special Issue starts with the papers devoted to DNA sequencing and DNA sequence analysis, followed by papers on protein sequence analysis and protein structure prediction and ending on analysis of gene expression data and gene networks.

Phan and Skiena provide an algorithm for sequencing by hybridization (SBH) that incorporates realistic error models. In theory, SBH is a very powerful and efficient method of determining sequences from DNA segments. However, most previous algorithms have assumed that the hybridization signals can be obtained without errors. In reality, hybridization between slightly mismatched pairs contribute significantly to the total signal and therefore seriously limit the utility of those algorithms. Phan and Skiena use realistic models for the types of errors that are observed in hybridization experiments and describe an algorithm that can be used to efficiently determine the unknown sequence in the presence of those errors. Theoretical bounds on the performance are provided and are compared with results generated by simulated experiments.

Conservation of genomic regions between closely related species may indicate their important functional properties. The paper by Levy, Hannenhalli and Workman addresses the use of sequence conservation between mouse shotgun reads and a set of transcripts from 502 human genes to identify transcription regulatory regions. They demonstrate that transcription factor binding sites are more likely to appear in conserved regions than in non-conserved regions upstream of genes. Such an enrichment of binding sites was not observed in the conserved protein-coding sequence. Another finding was the observation of co-localization of certain known pairs of transcription factor binding sites inside the conserved regions. Altogether the paper provides a methodology how to detect new regulatory signals and their co-localized pairs, an important step in construction of regulatory models for human genes.

Mechanisms of transcription control in eukaryotes remain to be explained in detail, including the rules of formation of a mosaic of transcription regulation signals encoded in DNA. With part of the picture known and collections of transcription factor binding sites (TFBS) now compiled, the problem remains of how to detect TFBS in clusters near the transcription start. Frith, Hansen and Weng propose a method that searches for clusters of several different types of TFBS'. Their algorithm uses Hidden Markov model with novel architecture and initial tests on muscle specific regulatory regions to show performance that compares favorably with previously known results.

The work by Rogozin, Kochetov, Kondrashov, Koonin and Milanesi contributes to further understanding the sequence code used for fine tuning translation initiation in eukaryotes. The precise mechanism of translation initiation is still not known in detail and the abundant genomic DNA data are, in most cases, lacking experimental evidence for starts of translation and transcription. By mining databases for the most reliable data, the authors' analyses provide an intriguing observation: that a large fraction of 5' UTR regions contain upstream ATG codons. This result is rather unexpected from the standpoint of a sliding mechanism of eukaryotic translation initiation. Another striking observation is that the information content value of the translation start signal (located around a start codon) and the number of upstream ATGs are negatively correlated. These findings lead the authors to several interesting hypotheses on the mechanism of regulation of translation efficiency.

Kraemer, Wang, Guo, Hopkins and Arnold present a refreshing investigation into the degree in which statistical gene-finding programs can be trusted when applied to organisms for which they have not been developed and trained. The organism-specificity of current gene-finding programs and their failure to accurately predict the location of genes in the genomes of 'non-training organisms' is well known. Yet, there has not been a systematic study of the degree of that organism-specificity. The paper fills this gap by considering a case study of five different gene-finding programs (GenScan, HMMGene, GeneMark, Pombe, and FFG) applied to five genomic sequences of *Neurospora crassa*. Four of the five programs were not developed for (and trained on) *N. crassa*. The low accuracy figures, although obtained for a test set of rather limited size, are eye-openers that indicate the necessity of developing and training species-specific gene finders.

Hatzigeorgiou and Reczko report the development and characterization of a new program called DNA Intelligent Analysis for ESTs (DIANA-EST). The program combines neural networks and coding region statistics to predict coding capability of EST data. From an anonymous test set, DIANA-EST correctly predicted 90% of nucleotides as either coding or non-coding, providing a moderate improvement over existing methods.

Kretschmann, Fleischmann and Apweiler attack the notorious problem of automated protein annotation. By applying a data-mining algorithm to the analysis of SWISS-PROT keywords, they have created a huge database of rules and show by cross-validation that these can be used to annotate uncharacterized protein sequences almost as well as is currently done by SWISS-PROT curators.

Bejerano, Seldin, Margalit and Tishby describe an ambitious attempt to identify domain boundaries in protein sequences without constructing a multiple sequence alignment, with this done on the basis of statistical properties of sequence segments. They report encouraging results of automatic identification of known domains. Much more testing is needed, but the method seems to have some real potential for genome-scale applications.

Bolten, Schliep, Schneckener and Schomburg, address the problem of detecting remotely related protein sequences. The approach they use is based on intermediate sequences, where two proteins of undetectable similarity to each other are, nevertheless, both related to the common (intermediate) third sequence. The authors constructed a large directed graph of the sequences in the SwissProt database, based on all local pairwise similarities obtained with local dynamic programming comparisons. They then developed a novel graph-based clustering algorithm and efficient computer routines for performing this large calculation efficiently. An evaluation of the sensitivity based on the SCOP database shows a 24% increase in detected homologs compared to pairwise methods alone.

Mallios addresses the problem of predicting the strength of binding of a given peptide sequence to a class II MHC antibody. The strength of binding was classified into three classes: high, moderate, and none. The binding was modeled as being dependent on the residue types at each of the positions in a peptide. An iterative stepwise discriminant analysis algorithm was trained with peptides of known binding propensity to the HLA-DR1 antibody. This algorithm correctly classified over 90% of the peptides in the test set of several hundred sequences. This new method for prediction of binding propensity may be useful in vaccine design.

Accurate methods for assigning volumes to atoms in protein structures are critical for computation of shapes of protein packing and contact areas involved in protein-protein interaction. Tsai, Voss and Gerstein make an interesting contribution to this field by showing that certain types of atoms behave differently than expected. For example, aromatic carbons with single hydrogen appear to occupy a similar volume as tetrahedral carbons with two hydrogens. A modified atom typing scheme that uses certain numerical criteria produces a more accurate fit to experimental data on atom volumes than the standard chemical typing system.

The work of Fariselli and Casadio is devoted to the problem of predicting the disulfide topology of proteins from amino-acid sequence alone. A protein sequence is represented by a graph with the vertices corresponding to the cysteine residues forming disulfide bridges, and the weights of the edges to the contact potentials. The prediction consisted of finding the graph with the maximum weight. Different residue contact potentials were tested for the accuracy of disulfide bridge prediction. For proteins with four disulfide bridges, the highest accuracy achieved was seventeen times higher than that of a random predictor. The new method may be useful in de novo protein structure prediction.

Jordan, Bishop and Gonzalez use position-specific variability profiles between two paralogous groups of α-mannosidase proteins to provide evidence for functional differences between them. In particular, these methods pinpoint a number of specific residues as likely candidates for functional divergence. Since these residues map to the surface of α-mannosidase, the authors suggest that the functional differences accrued gradually through small, subtle changes in protein sequence.

Yeung, Fraley, Murua, Raftery and Ruzzo describe a method for clustering expression data based on probability models of the data, rather than more commonly-used heuristic methods. This method provides a principled way of evaluating the number of clusters that best fits the data and a means of comparing different models. Results are presented on both simulated data and real data. On simulated data the method is shown to reliably determine

the appropriate model that generated the data. On real data the method is shown to compare favorably with leading heuristic methods and provides the advantage of comparing different numbers of clusters and different models directly.

In a pioneering contribution, Rzhetsky and Gomez develop a mathematical model for the generation and growth of macromolecular networks, particularly for those comprising interactions between different proteins and between proteins and nucleic acids. Even though the model networks have no requirement to adopt pre-conceived topologies, they exhibit the same 'scale-free' statistical properties as *bona fide* molecular networks: small parts of the networks are the same as any larger part. Using their model, the authors determine that the frequencies of molecular domains do indeed follow a power-law distribution. Using this result, Rzhetsky and Gomez derive a simple equation that predicts the total number of distinct interacting domains in a given genome.

Publication of the Special Issue was made possible due to the constant support of the editorial board of *Bioinformatics*.

Editors of the Special Issue: John Logsdon, Jim Fickett, Eugene Koonin, Andrej Sali, Gary Stormo and Mark Borodovsky.