BIOINFORMATICS

Vol. 15 no. 11 1999
Pages 865–866

Editorial

THE SECOND GEORGIA TECH INTERNATIONAL CONFERENCE ON BIOINFORMATICS: SEQUENCE, STRUCTURE AND FUNCTION (NOVEMBER 11–14, 1999, ATLANTA, GEORGIA, USA)

Steering & Program Committee: Pierre Baldi, Mark Borodovsky, Soren Brunak, Chris Burge, Jim Fickett, Steven Henikoff, Eugene Koonin, Andrej Sali, Chris Sander, Gary Stormo

This issue of *Bioinformatics* contains reports on selected papers presented at the international conference in Atlanta. The conference was held at one of the midtown hotels offering a magnificent bird's-eye view to the cosmopolitan capital of the Southeast of the USA. The conference agenda included keynote lectures by Russell Doolittle, Walter Fitch and Walter Gilbert, 19 plenary and contributed talks and more than 50 poster presentations. The focus of the conference was on the most intriguing issues of comparative and structural genomics, functional interpretation of nucleic acid and protein sequences, inferring 3D structure from sequence data and deciphering the history of biomolecular evolution (http://exon.biology.gatech.edu/conference). Submission of manuscripts to the *Bioinformatics* special issue has been a parallel avenue of presentation for the conference participants. Eighteen papers were submitted to the peer review process and nine were accepted for publication. The issue starts with the papers devoted to DNA sequence analysis and gene finding, followed by the papers on protein sequence analysis and protein structure prediction.

Single pass sequencing of the 5′ and 3′ ends of cloned cDNA sequences represents an efficient approach to large scale characterization of the set of genes expressed by an organism. In the past few years huge numbers of such sequences, termed 'expressed sequence tags' or ESTs, have been determined from human and a variety of other organisms, and deposited in public databases. Such sequences represent a powerful resource for identifying genes, but suffer from the relatively high rates of sequencing errors typical of single pass data. Frameshift errors, in particular, can obscure the results of similarity searches based on translations of putative EST ORFs. The paper entitled 'FramePlus: Aligning DNA to protein sequences' by E. Halperin, S. Faigler and R. Gill-More describes a new algorithm specifically tailored to allow for moderate to high rates of frameshift errors. The new techniques resulted in improvements in the ability to detect homologies between ESTs and known proteins in peptide database searches. A.M. Shmatkov, A.A. Melikyan, F.L. Chernousko and M. Borodovsky's paper, 'Finding prokaryotic genes by the "frame-by-frame" algorithm: targeting gene starts and overlapping genes', addresses the important practical problem of gene recognition in prokaryotic genomes, of which over a dozen have already been completely sequenced. The authors suggest an approach to one of the few remaining open problems in prokaryotic gene finding: accurate prediction of gene starts allowing for the possibility of overlapping protein-coding regions, a relatively common occurrence in prokaryotic genomes which seems to be rare in eukaryotes. Their algorithm involves application of a hidden Markov model of gene structure to each of the six global reading frames (three on each strand) of a genome separately, followed by a simple post-processing step to remove completely overlapping genes which rarely occur in nature. Promising results are obtained in identifying gene starts, and the possibility of systematic biases in the annotation of several bacterial genomes is raised.

Benchmarking of the computer tools developed for finding eukaryotic genes is the subject of the paper 'Evaluation of gene prediction software using a genomic dataset: application to *Arabidopsis thaliana* sequences' by N. Pavy, S. Rombauts, P. Dehais, C. Mathe, D.V.V. Ramana, P. Leroy and P. Rouze. The authors have developed the AraSet, a data set of authentic contigs of validated genes of the *Arabidopsis* genome enabling the evaluation of multi-gene models. For this purpose they also introduced new measures of gene finding accuracy reflecting the prediction errors at the protein sequence level. These new measures of accuracy evaluation, along with conventional criteria defined at the site and the exon levels, can be applied to any gene prediction software and to any eukaryotic genome for which a similar data set is built. For the *Arabidopsis* genome the authors have shown that while the accuracy of splice site and exon prediction is quite high, gene modeling accuracy remains very low. The publicly available gene finding programs were shown to differ significantly in their performance. However, it was demonstrated that gene modeling could be further improved by specific combination of gene finding tools.

A. Bansal, in the paper 'An automated comparative analysis of seventeen complete microbial genomes', presents a new improvement in the automated pairwise genome comparison technique used for identifying orthologous genes. This technique was used to compare 17 already sequenced microbial genomes and derive orthologs, orthologous gene-groups, duplications, gene-fusions, genes with conserved functionality and genes specific to groups of genomes.

D. Sankoff's paper, 'Genome rearrangement with gene families', addresses the problem of reconstructing the

evolutionary history of genomes by focusing on gene order and the history of rearrangements rather than measures of sequence similarity. In this work, a generalization of the standard genome rearrangement problem is introduced which allows for the presence of gene families, multiple copies of an ancestral gene, which are recognizable on the basis of sequence similarity but which may no longer be adjacent in contemporary genome sequences. The problem is formulated as a search for the 'true exemplars' of the genome: the members of each of the gene families best representing the original position of the ancestral gene in the last common ancestor of the genomes being compared. A branch and bound algorithm is proposed for efficiently finding the likely set of exemplars and calculating the 'exemplar distance' between a pair of genomes, and the performance is analyzed using primarily simulated data. Because the occurrence of gene families containing many members is widespread in the biological world, it will be interesting to see how this approach fares when applied to more challenging real world examples.

The paper 'Structural basis for triplet repeat disorders: a computational analysis' by P. Baldi, S. Brunak, Y. Chauvin and A.G. Pedersen contributes to the understanding of an important and truly interdisciplinary problem. What is the mechanism acting on the cellular level and triggering a number of degenerative disorders that were shown to be related to the presence of the regions of triplet repeats in human genomic DNA? The authors explore structural characteristics of triplet repeat regions, such as bendability, by means of computer modeling and show that 'pathological' repeats fall into extreme classes. They indicate that mutational expansions of the repeat regions become unstable and likely pathological upon reaching the crucial length of $\sim$150 nt, that correlates with the length of DNA wrapped up in a single nucleosome core particle. This analysis provides new arguments on the involvement of the chromatin structure into the repeat expansion mechanism as well as suggests that expansion of a particular dodecamer repeat, having very high flexibility, may play a role in the pathogenesis of the neurodegenerative disorder multiple system atrophy.

S. Uliel, A. Fliess, A. Amir and R. Unger present 'A Simple Algorithm for Detecting Circular Permutations in Proteins'. The method is based on dynamic programming. While it does not guarantee to find the global minimum, it runs in time proportional to the square of the number of residues in the protein. This is significantly more rapid than other existing methods. Thus, the method will allow identification of circular permutations in a large database of known protein sequences. Such a survey will facilitate studying the importance and role of circular permutations in the evolution of proteins.

P. Baldi, S. Brunak, P. Frasconi, G. Soda and G. Pollastri, in 'Exploiting the past and the future in protein secondary structure prediction', describe a novel method for the prediction of secondary structure in protein sequences. While the new method, a hybrid of a neural network and a hidden Markov model, appears to be at least as accurate as most existing approaches, the main emphasis of the paper is on the development of new algorithmic ideas. The method benefits from long-range dependencies of secondary structure on upstream and downstream residues as well as from multiple sequence information provided by an input sequence alignment. The methodological ideas might also be applied to other problems in bioinformatics, such as the prediction of DNA exon/intron boundaries and promoter regions, RNA secondary structure, and protein functional patterns.

One of the most exciting advances described in this issue is the work of A.V. Lukashin and J.J. Rosa, 'Local multiple sequence alignment using dead-end elimination'. This paper addresses the difficult (likely NP-hard) problem of determining the rigorously optimal local alignment of a set of DNA or protein sequences under the standard 'sum of pairs' scoring system. Local multiple sequence alignment, finding short conserved blocks or motifs common to a set of sequences, has many important applications ranging from finding the sequence signatures of particular protein families to characterizing the binding sites of DNA- or RNA-binding proteins.

The algorithm introduced by Lukashin and Rosa makes use of an algorithmic trick called 'dead-end elimination', a strategy for rapidly eliminating portions of the search space of possible multiple alignments which are provably incompatible with the optimal solution. Using this technique, previously applied with some success to problems of protein folding and design, the alignment algorithm is able in most cases to find the globally optimal multiple alignment of a set of sequences in effectively polynomial time.

Publication of the Special Issue was made possible due to the constant support of the editorial board of *Bioinformatics*. We experienced the synergy of this joint project intended for benefits of both for conference participants and readers of the journal. We wish the *Bioinformatics* journal and the Georgia Tech conference to succeed further in the new Millennium.

Editors of the Special Issue: Chris Burge, Andrej Sali and Mark Borodovsky