Genome-wide Analysis and Comparative Genomics: Session Introduction

I. Dubchak, V. Solovyev, and L. Wei

Pacific Symposium on Biocomputing 8:276-278(2003)

GENOME-WIDE ANALYSIS AND COMPARATIVE GENOMICS

LIPING WEI
Nexus Genomics, Inc.
229 Polaris Avenue, Suite 6
Mountain View, CA 94043
wei@nexusgenomics.com

INNA DUBCHAK
Lawrence Berkeley National Laboratory
MS 84-171, Berkeley, CA 94720
ildubchak@lbl.gov

VICTOR SOLOVYEV
Softberry, Inc.
victor@softberry.com

This is the second year at the Pacific Symposium on Biocomputing (PSB) that a session is devoted to genome-wide analysis and comparative genomics. In the past year, we have witnessed even greater growth of the amount of genomic sequence data. For example, the number of complete eucaryote genome sequences available at NCBI has nearly doubled in the past year, to a total number of eight as of September, 2002. The draft sequences of several major eucaryotes have been published, including mouse and rice. In addition, numerous procaryote genomes have been completely sequenced. The explosion in the amount of genomic sequence data has resulted in unprecedented opportunities for discoveries from computational genome analyses. At the same time, the large amount of data has proven challenging for computational scientists to develop accurate and efficient algorithms. The papers in this section represent excellent examples of new analysis and novel algorithms that are contributing greatly to our understanding of genome biology.

One of the most important and challenging steps in genome annotation is gene prediction. In eukaryotic genomes, which have long noncoding regions and large introns, ab initio gene finders such as Fgenesh and Genscan identify about 90% of the genes, but generate false positive predictions and even more often predict partially incorrect gene structures. Thus, incorporation of as much available evidence as possible for gene prediction is necessary. The paper by Yada at al. presents the DIGIT algorithm that predicts genes by combining the results from several existing gene finders. It was able to successfully discard

many false positive exons predicted by the individual programs and showed remarkable improvements in sensitivity and specificity. Another approach to improve quality of gene prediction is to implement more accurate models of splice site recognition, which is highly nontrivial. An interesting attempt in this direction is presented in the paper by Ott et al. The paper suggests that the splicing of long introns might be facilitated by splicing inner parts of the intron prior to the splicing of the long intron itself.

One of the challenges of the postgenomic era is to understand the regulation of gene transcription. The combinatorial nature of regulation and the practically unlimited number of cellular conditions significantly complicate the experimental identification of transcription factor (TF) binding sites on a large scale. Therefore, computational approaches to reveal potential regulatory elements become very important. Cheremushkin and Kel described a new technique, a variant of phylogenetic footprinting implementation, to mine conserved noncoding regions between human and mouse, and compiled a database of ~60,000 predicted potential TF binding sites. Another paper by Olman at al. introduced a beautiful minimum spanning tree algorithmic solution to reveal regulatory binding sites in a set of similarly regulated genes. Such sets of genes are presumed to share common TF binding sites. They could be extracted from the increasing volume of microarray gene expression data. The paper by Phang et al. outlined a novel non-parametric method called trajectory clustering that was able to cluster groups of genes with known related function better than alternative approaches.

Completion of the sequencing of many genomes and the exponential growth of biological databases present new challenges to sensitive homology searches. The size of the sequences is perhaps the biggest hurdle, since many alignment algorithms were designed for comparing single proteins and are extremely inefficient when processing large genomic intervals. The paper by Kahveci and Singh addresses one of those problems. They proposed an efficient technique to align long genomic strings up to ~100 times faster than the BLAST algorithm.

Finally, three articles deal with different aspects of phylogenetic analysis of genomic data. Distance-based methods are widely used for inferring phylogenies. Recently introduced reversal distance based on the orders of genes is defined as the minimum number of signed/unsigned reversals needed to account for the difference in gene order between two genomes. Wu and Gu presented an effective algorithm to reconstruct optimal distance-based

phylogenetic trees for genomes. Recent findings reinforced the view that while considering evolutionary relationships, we need to account for such genomic events as gene duplication, loss, convergence and lateral (horizontal) gene transfer. The paper by Addrario-Berry et. al. is concerned with evaluating the performance of the model and algorithm for detecting lateral gene transfer events. Such biological processes as hybridization of horizontal gene transfer require network rather than tree structure of relationships. The paper by Nakleh with co-authors reports the development of computational tools for evaluating phylogenetic network reconstruction methods.

The session co-chairs are grateful to all the authors who had submitted their work to this session, and to all the reviewers for their help in the difficult task of choosing the best contributions from a large number of excellent submissions.