Pacific Symposium on Biocomputing 7:112-114 (2002)

GENOME-WIDE ANALYSIS AND COMPARATIVE GENOMICS

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

LIOR PACHTER
Department of Mathematics
UC Berkeley
Berkeley, CA 94720
lpachter@math.berkeley.edu

LIPING WEI
Nexus Genomics, Inc.
390 O'Connor St.
Menlo Park, CA 94025
wei@nexusgenomics.com

One of the key developments in biology during the past decade has been the completion of the sequencing of a number of key organisms, ranging from organisms with short genomes, such as various bacteria, to important model organisms such as *Drosophila Melanogaster*, and of course the human genome. As of February 2001, the complete genome sequences have been available for 4 eukaryotes, 9 archaea, 32 bacteria, over 600 viruses, and over 200 organelles (NCBI, 2001). The “draft” human sequence publicly available now spans over 90% of the genome, and finishing of the genome should be completed shortly. The large amount of available genomic sequence presents unprecedented opportunities for biological discovery and at the same time new challenges for the computational sciences. In particular, it has become apparent that comparative analyses of the genomes will play a central role in the development of computational methods for biological discovery. The principle of comparative analysis has long been a leitmotif in experimental biology, but the related computational challenge of large-scale comparative analysis leads to many interesting algorithmic challenges that remain at the forefront of computational biology research.

The papers in this track represent a cross-section of the ongoing efforts to bridge the gap between comparative computational genomics and biology, and have been selected both for their algorithmic content, and the relevance of their application.

We have also included papers on whole genome analyses, that describe exciting new discoveries that have been made possible only recently with the availability of so much genomic sequence. The breadth of applications that have been addressed is testimony to the explosion of activity in the field of comparative genomics (more generally computational genomics), and the success that it is already heralding in biology. It is clear that there is a lot of low hanging fruit.

The computational perspective of biology has traditionally been that one ought to be able to sequence a genome, then annotate the genes and regulatory elements thereby obtaining the proteins and some understanding of their regulation. Subsequently, solution of the protein folding problem, it was hoped, would elucidate the structure and hence function of the proteins, and then cures for diseases would follow. QED.

Of course every step along the way has proven to be nontrivial, in a highly nontrivial way! Even the sequencing of genomes has proven to be difficult, and the paper by Mulyukov and Pevzner addresses one of the core problems in the new field of assembly, namely the resolution of repeats. Rather than simply suggesting different heuristics and hacks for dealing with repeats in the context of an algorithm, they introduce a beautiful twist to the problem by providing an algorithmic solution for developing experimental assays for resolution of repeats. Hopefully computational biologists will take note that computational methods can interact with experimental biology in a very direct way.

Given that one has genomic sequence at hand, and as we have mentioned there is already plenty of it, the next step is to annotate the sequence. The paper by Kel-Margoulis et al. looks at the important (and difficult) problem of detecting regulatory elements in sequences. They have developed a method of looking for clusters of regulatory elements in genes that are functionally related. Also related to regulatory site detection is the paper by Sze, Gelfand and Pevzner on finding weak motifs in DNA sequences. The annotation of splice sites, while easier than that of regulatory elements, remains an unsolved problem, and the intriguing paper by Patterson, Yasuhara and Ruzzo discusses the possibility of a relationship between pre-mRNA structure and splice sites that might help in splice site detection. Biologists, lacking much hard experimental evidence, have debated the connection between structure of RNA and splicing for some time, so a fresh computational analysis is welcome and perhaps even overdue. Finally, the paper by Holmes and Rubin on the detection of RNA genes in sequences is a beautiful and natural generalization of single organism stochastic context free grammar approaches to two organisms.

Taking a more global view, and looking at protein sequences, the paper by Cline et al. compares protein families in humans, worms, flies and yeast. Also looking at biology from a whole genome perspective is the paper by Imoto, Goto and Miyano which addresses the problem of constructing genetic networks. Such whole genome studies aimed at mapping out the functional and regulatory relationships between proteins are an exciting development that has been heretofore impossible because of the lack of sequence. There are three papers dealing with evolutionary biology: the paper by Goldberg, McCouch and Kleinberg addresses the important problem of generating comparative genomic maps, and the technical paper by Wu and Gu on computing reversal distance is a nice example of some of the algorithms which can lead to estimates of evolutionary distance from such maps. The immediate application of computing reversal distance is the accurate construction of phylogenetic trees (although of course there are other measures of distance which may be even more useful), and the paper by Nakhleh et al. discusses an effective approach to rapidly and accurately constructing phylogenetic trees.

Alignment algorithms lie at the heart of comparative genomics and we have included two important papers that directly address technical issues that are critical in obtaining alignments. The paper by Chiaromonte, Yap and Miller is focused on the problem of obtaining accurate alignments. In particular, they look at the question of how to score nucleotide substitutions in alignments, and how to generate scoring matrices. The work has immediate application to the alignment methods underlying PIPMaker, which is a widely used alignment tool. The paper by Yamaguchi and Maruyama looks at how to generate alignments quickly, which is just as important as generating them accurately. In this interesting paper, they suggest the application of specialized hardware and discuss the associated algorithms.

Finally, the paper by Volkmuth and Alexandrov is an exciting addition in that it proposes a novel way for utilizing comparative genomic information for learning about folding. We believe that such creative ways of finding predictive power in comparative information, coupled with visualization tools such as the new one described by Gherbi and Herisson, hold the promis e of exciting and important biological discoveries based on genomic analysis.

The session co chairs are grateful to the reviewers for their help in choosing the best contributions from a large number of excellent submissions.