Pacific Symposium on Biocomputing 14:111-113 (2009)

COMPUTATIONAL CHALLENGES IN COMPARATIVE GENOMICS: INTRODUCTION TO THE SESSION

B. M. E. MORET*, W. MILLER, P. A. PEVZNER, AND D. SANKOFF

*Laboratory for Computational Biology and Bioinformatics
EPFL (Swiss Federal Institute of Technology)
EPFL-IC-LCBB INJ 230, Station 14
CH-1015 Lausanne, Switzerland
E-mail: bernard.moret@epfl.ch*

Comparative methods have long been a mainstay of biology, particularly evolutionary biology; they are also at the core of medical research based on animal models of human physiology. They find their most challenging and most fitting application, however, in the study of whole genomes, as they are the main tools through which we can make sense of the billions of base-pairs forming the sequence of animal and other genomes. Comparing whole genomes, which is necessarily done through computational methods due to the size of the genomes, has given rise to the research area known as comparative genomics.

Comparative genomics is the tool of choice for identifying genes in both well studied and newly sequenced genomes; for studying the acquisition of virulence or drug resistance in pathogens; for tracking down gene complexes responsible for inheritable diseases or susceptibilities; and for engineering desirable new traits in crops; and for studying many forms of cancers. More generally, comparative genomics is the tool of choice to elucidate how the genetic blueprint translates into specific functions and how that blueprint evolves in populations and into various species.

Comparative genomics uses not just whole-genome sequences, but also dense single-nucleotide polymorphism (SNP) maps, genetic maps, and sequences of individual genes, but it is characterized by its emphasis on a whole-genome approach. Its computational methods include combinatorial optimization, machine learning, and data mining, while much work has also been devoted to visualization of its findings—witness, for example, the many spectacular full-color figures illustrating the correspondences between

Pacific Symposium on Biocomputing 14:111-113 (2009)

the human and mouse genomes.

The focus of our session is on computational models and algorithms. The five papers included in our session (selected from a total of 17 submissions) all exemplify the genome-wide approach of the area. Most use optimization approaches to infer evolutionary events or, equivalently, to annotate regions of the genome according to their evolutionary history.

In their paper on "Identifying parent-daughter relationships among duplicated genes," Han and Hahn tackle the evolution of gene families. When two organisms share a gene family, sorting out orthologs and various paralogs is a prerequisite for a comparative analysis. Many orthologous relationships may exist, yet perhaps the most crucial is that between the "parents" of the families. The authors present a model in which the length of shared syntenic regions is a hidden variable in an HMM model and use an EM algorithm to estimate the parameters and identify parent-daughter relationships among gene family members. They then apply their method to a collection of gene families from six mammalian genomes and use the inferred parental relationships to determine the direction of gene duplication events.

In "A parsimony approach to analysis of human segmental duplications," Kahn and Raphael tackle the two-step model of segmental duplication. These duplications are widespread in the human genome and quite complex, with many appearing to consist of a mosaic of nested duplications. The two-step model allows for such nesting, albeit to just one level. The authors give an integer linear programming formulation of the problem of explaining the data at hand with a minimum number of duplications across the two steps and apply it to the human genome. The nested formulation provides a natural tree relationship and thus a first cut at reconstructing the history of the duplication events.

In "Simultaneous history reconstruction for complex gene clusters in multiple species," Zhang, Song, Hsu, and Miller combine the two previous topics and examine the gene clusters that arise from segmental duplications. These genes form multiple families, yet their history is much more difficult to tease out than those of genes duplicated in simple tandem duplications or through retrotransposition; moreover, the presence of these large and complex duplications creates severe difficulties for sequencing. Using comparative genomics provides some leverage for the problem, leverage that the authors demonstrate how to use.

In "Inferencing genome-wide mosaic structure," Zhang, Wang, McMillan, Villena, and Threadgill take up the crucial problem of inferring the

Pacific Symposium on Biocomputing 14:111-113 (2009)

recombination history of a population of genomes. Repeated meiotic recombination gives rise to a mosaic structure in the genome, with each haplotype block (mosaic piece) traceable to a distinct ancestor from the neighboring blocks. The authors propose a graph model for the formation of these mosaics and provide a dynamic programming algorithm to infer a mosaic with the smallest number of pieces given a population of genomes, showing the results of its application to genome-wide SNP data on mice.

In "An exact solver for the DCJ median problem," Zhang, Arndt, and Tang address the issue of genomic rearrangements from the point of view of phylogenetic reconstruction. While the median of three genomes is an abstract concept, it has become a mainstay of ancestral genome reconstruction; unfortunately, under almost any rearrangement model, computing a median is NP-hard. In this paper, the authors show how to speed up such computations so as to enable application of the technique to nontrivial genomes.

We are very pleased to feature such work at this PSB'09 session and want to take this opportunity to thank attendees, presenters, all submitting authors, and the referees, who together made it possible.