Pacific Symposium on Biocomputing 15:1-2(2010)

COMPUTATIONAL CHALLENGES IN COMPARATIVE GENOMICS SESSION INTRODUCTION

BERNARD M.E. MORET

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

WEBB C. MILLER

Department of Biology
Pennsylvania State University
University Park, PA 16823, USA
E-mail: wcm2@psu.edu

PAVEL A. PEVZNER

Department of Computer Science & Engineering
University of California, San Diego
9500 Gilman Drive, Mail Code 0404, La Jolla, CA 92093-0404, USA
E-mail: ppevzner@cs.ucsd.edu

DAVID SANKOFF

Department of Mathematics and Statistics
University of Ottawa
585 King Edward, Ottawa, ON K1N 6N5, Canada
E-mail: sankoff@uottawa.ca

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 study 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 the human and mouse genomes.

The focus of our session is on computational models and algorithms; this session at PSB'10 follows a previous session on the same theme at PSB'09, which also featured five papers. The five papers included in our session all exemplify the genome-wide approach of the area.

Two of the papers focus on ancestral reconstruction, a topic that has recently attracted much interest, but where assessing the validity of results is obviously very difficult. Hickey and Blanchette, in "A practical algorithm for estimation of the maximum likelihood ancestral reconstruction expected error," provide the first

Pacific Symposium on Biocomputing 15:1-2(2010)

systematic approach to such an assessment, using a direct analog of the phylogenetic bootstrapping process. Gavranovic and Tannier, in "Guided genome halving: provably optimal solutions provide good insights into the preduplication ancestral genome of S. cerevisiae," discuss a specialized application of ancestral inference in which, given a contemporary genome whose lineage is known to have seen a whole-genome duplication, and a closely related genome whose lineage diverged before that duplication and serves as a guide, the preduplication ancestor is inferred.

Two other papers are concerned with metagenomics, where samples are taken of an entire biota (seawater, soil, animal gut, etc.), the samples sequenced, and the resulting sequences (mostly unassembled) placed within a phylogeny of related organisms—a process that has already enabled us to discover very large numbers of new species. Clemente, Jansson, and Valiente, in "Accurate taxonomic assignment of short pyrosequencing reads," puth forth the proposition that the common strategy of assigning metagenomic sequences to the root of an entire clade can be advantageously replaced by a strategy of assigning these sequences to internal nodes that optimize some ROC characteristics. and demonstrate of the use of their strategy on marine and gut data. Essinger and Rosen, in "Benchmarking BLAST accuracy of genus/phylum classification of metagenomic reads," also address the question of proper assignment of metagenomic reads to a phylogeny, but examine the even more common strategy of identifying subsequences with high similarity by using BLAST on an entire database.

Finally, the fifth paper showcases a fascinating application of comparative genomics to what might at first be viewed as a problem in population genetics: how to optimize the choice of a founder population to repopulate a species through captive breeding. Miller, Wright, Zhang, Schuster, and Hayes, in "Optimization methods for selecting founder populations for the captive breeding of endangered species," present formulations and algorithms for selecting a founder population from an existing wild population. Such problems deal with very small populations and target specific collections of alleles, many of which will be represented in only a few individuals—thus genomic methods, which deal with individual genomes, are better suited than standard population genetics methods, which tend to deal with distributions over sizeable populations.

We are very pleased to feature such work at this PSB'10 session and want to take this opportunity to thank attendees, presenters, all submitting authors, and the referees who together made it possible.