

BEYOND GAP MODELS: RECONSTRUCTING ALIGNMENTS AND PHYLOGENIES UNDER GENOMIC-SCALE EVENTS

MICHAEL BRUDNO

University of Toronto

BERNARD MORET

EPFL, Switzerland

RANDY LINDER

The University of Texas at Austin

TANDY WARNOW

The University of Texas at Austin

Multiple sequence alignment (MSA) has long been a mainstay of bioinformatics, particularly in the alignment of well conserved protein and DNA sequences and in phylogenetic reconstruction for such data. Sequence datasets with low percentage identity, on the other hand, typically yield poor alignments. Now that researchers want to produce alignments among widely divergent genomes, including both coding and noncoding sequences it is necessary to revisit sequence alignment and phylogenetic reconstruction under more ambitious models of sequence evolution that take into account the plethora of genomic events that have been observed.

Most current methods postulate only two types of events: substitutions (modeled with a transition matrix, such as PAM or BLOSUM matrices for protein data) and insertions/deletions or *indels* (rarely modelled beyond a simple affine cost function of the size of the gap). While these two events can indeed transform any sequence into any other, this model of genomic events is far too simplistic: substitutions are not location- or neighbor-independent, and indels can be caused by a variety of complex events, such as uneven recombination, insertion of transposable elements, gene duplication/loss, lateral transfer, etc. Moreover, genomic rearrangement events can completely mislead procedures based on most current models, resulting in a total loss of alignment when a homologous element has undergone an inversion or a duplication.

The aim of our session is to bring together researchers in multiple sequence alignment, phylogenetic reconstruction, comparative genomics, DNA sequence analysis, and genetics to examine the state of the art in multiple sequence alignment, discuss how methods can be improved, and whether current projects will suffice for the emerging applications in various biological fields. The four papers in our session, while centering around the topic of sequence comparison, rep-

represent the breadth of interests of scientists in the field: algorithms to generate and analyze alignments, the estimation of phylogenetic trees and how these phylogenies affect alignment algorithms, and analyzing the frequencies of genome rearrangements in various locations of the genome.

Our session has four papers addressing different aspects of the general problem. The paper by Dalca and Brudno presents a unifying view of many sequence alignment algorithms. The authors propose the rectangular scoring scheme framework and demonstrate algorithms to speed up comparison of sequences with arbitrary rectangular scoring. While the resulting program is too slow for whole-genome applications, it can allow for easy prototyping of complex scoring schemes for alignments.

The paper by Landan and Graur addresses the problem of finding the regions of high reliability within a multiple alignment. The authors present an elegant algorithm that determines if the alignments are changed when the sequences are reversed – an indication of a region where the alignment is less reliable. This work has implications for phylogeny estimation, since low-confidence regions within the alignment can then be down-weighted (or even eliminated) during a phylogeny estimation, and thus potentially lead to more accurate phylogenetic estimates.

The paper by Nelesen and colleagues addresses the impact of the choice of guide tree on multiple alignment methods, and on the phylogenetic estimations obtained using the resultant multiple alignments. Their simulation study shows that some methods (for example, ProbCons) are highly responsive to the particular guide tree, while others (for example, Muscle) are less responsive. In addition, they provide a particular technique for producing the guide tree that results in much better estimates of phylogenies than the current gold standard.

The fourth paper in the session by Sinha and Meller addresses the use of genome rearrangements in the estimation of evolutionary relationships between genomes. The potential for genome rearrangements to reveal evolutionary histories is great, but accurate reconstructions require better understandings of the frequencies of the various events, such as inversions, transpositions, and duplications. Sinha and Meller make important inroads on this problem by analyzing how varying definitions of a synteny block affect the observed inversion and breakpoint rates. One of the most interesting conclusions is that the definition of a synteny block has little effect on the estimation of the reuse of breakpoints, shedding additional light on an ongoing academic controversy in the field.

We are excited by the breadth of research taking place in the fields of MSA and phylogeny estimation, and are hopeful that our session will help bring together researchers in these areas. The four papers presented at our session were selected with the help of several reviewers, whose help we gratefully acknowledge.