Phylogenetic Analysis of Whole Genomes
(Keynote Talk)

Bernard M.E. Moret

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

1 Introduction

The rapidly increasing number of sequenced genomes offers the chance to resolve long-standing questions about the evolutionary history of certain groups of organisms, to develop a better understanding of evolution, to make substantial advances in functional genomics, and to start bridging genomics and genetics. Comparative genomics is the term used today for much of the work carried out in whole-genome analysis, correctly emphasizing that the "guilt-by-association" approach used in the analysis of gene and regulatory sequences remains the fundamental tool in the analysis of whole genomes. However, the limitations of pairwise comparisons are even more severe in whole-genome analysis than in sequence analysis and, of course, pairwise comparisons have little to tell us about evolution. Thus we are witnessing a significant increase in phylogenetic research based, not on sequence data, but on larger-scale features of the genome, such as genomic rearrangements, duplications and losses of genomic regions, regulatory modules and networks, chromatin structure, etc.

However, phylogenetic analysis for whole genomes remains very primitive when compared to the analysis of, e.g., coding sequences. In part, this is due to immaturity: while such an analysis was in fact conducted as far back as the 1930s (by Sturtevant and Dobzhansky, using chromosomal banding data and hand computations), the first serious computational attempts are less than 20 years old. The major reason, however, is simply the very complex nature of the data, which makes it very difficult to design good simple models and which causes most questions framed in even the simplest of models to be computationally intractable. A few examples will suffice to illustrate this point. Sequence analysis typically uses character positions as its basic units and assumes some form of independence among the positions, but we lack even a good definition of the basic unit (the syntenic block) most commonly used in comparative genomics. While the phylogenetic community frequently deplores the lack of good tools for the multiple alignment of sequence data, we simply have no tool capable of aligning multiple whole genomes unless they are all very closely related. Whereas computing a parsimony score on any given phylogenetic tree is solvable in linear time, computing a parsimony score on a tree of three leaves under most extant models of genomic rearrangements or duplications and losses is intractable (NP-hard). And while no systematics journal would publish an inferred phylogeny without some form of statistical

J. Chen, J. Wang, and A. Zelikovsky (Eds.): ISBRA 2011, LNBI 6674, pp. 4–7, 2011.

support (bootstrapping values or log-likelihood scores), we have as yet no way of bootstrapping phylogenies built from whole-genome data nor sufficiently good stochastic models to derive likelihood scores.

Fortunately, more and more research groups are working on phylogenetic analysis of whole genomes, so that rapid progress is being made. In this presentation, I will briefly survey the main computational problems, summarize the state of the art for each, and present some recent results from my group that take us closer to a solution to some of these problems.

2 Some Extant Problems

A comparative analysis of complete genomes starts by the identification of syntenic blocks, that is, contiguous regions that are shared, to within some tolerance factor, across the genomes. Ideally, syntenic blocks should be defined in an evolutionary setting, but, as in the case of gene orthology, practical implementations so far have used a variety of heuristics—based on the identification of shared anchors such as genes or other markers and on guidelines about the desired size of such blocks and the amount of dissimilarity tolerable within the blocks. The most recent and ambitious package for the identification of synteny blocks is DRIMM-Synteny which takes into account both duplications and rearrangements. Still missing from the literature is a formal evolutionary definition of synteny, in the spirit of definitions of homology and orthology, and accompanying criteria for selection of appropriate amounts of internal dissimilarity.

Genomic alignment needs much more work. Miller *et al.* developped a pipeline for the alignment of vertebrate genomes in the UCSC Genome Browser. The approach used (an initial star alignment against the human genome, followed by a progressive alignment to place all genomes on the same reference indexing) precludes its extension to more distantly related organisms. The package progressiveMauve is, like DRIMM-Synteny, an improved version of an earlier package, designed to take into account duplications in addition to rearrangements; it computes a multiple alignment, at the sequence level, of several genomes, not relying on any particular reference genome. Its target is clearly the smaller genomes of, e.g., bacteria. Aligning multiple genomes that are only distantly related may require a tree alignment rather than the conventional common indexing of character positions. Events such as chromosomal fusion, fission, or linearization remain to be taken into account.

Constructing a phylogenetic tree based on whole-genome data has seen significant progress. The first published packages, BPAnalysis and GRAPPA, worked only with unichromosomal genomes and were limited in the number of taxa as well as the size of the genomes (the number of syntenic blocks); MGR, which could handle multichromosomal data, scaled poorly, as did the Bayesian package Badger. With the DCJ model of rearrangements, new work was started on pairwise distance estimation and phylogenetic reconstruction, the latter using both parsimony-based methods and distance-based methods. In addition, the use of distance methods led to the first reliable method for confidence assessment. Still missing are robust and scalable methods for phylogenetic reconstruction in the presence of duplicate syntenic blocks, maximum-likelihood methods, and better bootstrapping. All of these methods rely on the prior identification of syntenic blocks; yet, in the case of distantly related taxa, these blocks may have to be

defined in a phylogenetic setting. Simultaneous tree inference and sequence alignment is still in its infancy, so it is no surprise that there has been very little work so far on simultaneous tree inference and syntenic block identification.

Extending the analysis of whole genomes from genomic structure to function starts with regulatory networks and chromatin structure. The former have mostly been studied in single species, but recent work at the Broad (Arboretum) and in my group (ProphyC) have shown that an evolutionary approach can significantly improve the quality of inference, both for entire networks and for modules. Projects for phylogenetic analyses of epigenomic data (such as histone modifications) are starting up everywhere. But models for connecting chromatin structure and regulation remain to be devised and current models for the evolution of regulatory networks leave much to be desired.

3 Some Encouraging Results from My Group

The DCJ (double-cut-and-join) model has considerably simplified the handling of rearrangements and been used in attempts to reconstruct parts of ancestral genomes for yeasts, among other organisms. My group developed a very accurate statistical estimator that takes as input the edit (shortest) distance between two arrangements and returns a maximum-likelihood estimate of the true distance. Later, we gave an ML estimator based on a slight variant of the DCJ model that takes into account duplication of blocks; on simulated data under deliberately mismatched models, the estimator stays within 10% of the true distance in almost all cases and under 5% in most cases. Using this estimator with the FastME program for distance-based reconstruction produces very accurate reconstructions on instances with up to 500 taxa and genomes of up to 20'000 syntenic blocks.

We have also used the DCJ model and work on so-called adequate subgraphs (substructures of the graph representation of rearrangements) to improve the computation of rearrangement medians, the basic step in computing parsimony scores for phylogenetic trees based on rearrangement data. Here the assumption of unique syntenic blocks remains necessary, but with this assumption we demonstrated fast and accurate computations on high-resolution genomic data (10'000 to 20'000 syntenic blocks) as well as very accurate scoring under simulations.

Our incursion into distance methods for rearrangement data suggested introducing perturbations into the distance matrices themselves, yielding the first usable method for evaluating the robustness of a phylogenetic reconstruction. In recent work, we have resampled the adjacencies and obtain discrimination comparable to that demonstrated by conventional phylogenetic bootstrapping for sequence data. While we designed this bootstrapping approach for distance-based methods, extending them to parsimony-based methods is straightforward, although computationally intensive. Thus a serious and longstanding impediment to the use of rearrangement data in phylogenetic inference is nearly overcome.

Functional inferences from large-scale genomic data often involves the inference of regulatory networks for various genes. The difficulty in obtaining comparable data across various species has long restricted such studies to single species, although comparisons were made across various tissues from the same host. Whole-genome RNA-Seq inventories are now providing a richer and more easily comparable source of

expression data across many organisms, thus motivating the development of inference methods based on phylogenetic approaches. My group developed the ProPhyC software to refine inferred networks through the use of phylogenetic relationships and of a simple evolutionary model for regulatory networks. Using this software on inferred networks for closely related species (or tissues) produces significantly improved networks in terms of topological accuracy and thus demonstrates the power of a phylogenetic approach to the analysis of these systems.

4 Conclusions

Phylogenetic inference and, more generally, phylogenetic methods are assuming a greater role in the analysis of whole-genome data. A logical extension of the pairwise comparative approach, phylogenetic methods, while often complex, provide important advantages: