Extending the Reach of Phylogenetic Inference

Bernard M.E. Moret

Laboratory for Computational Biology and Bioinformatics, EPFL, Lausanne, Switzerland
bernard.moret@epfl.ch

One of the most cited articles in biology is a 1973 piece by Theodosius Dobzhansky in the *The American Biology Teacher* entitled "Nothing in biology makes sense except in the light of evolution." It was also around that time that the development of computational approaches for the inference of phylogenies (evolutionary histories) started. Since then, phylogenetic inference has grown to become one of the standard research tools throughout biological and biomedical research. Today, phylogenetic tools receive over 10,000 citations every year. Concurrently, many groups are engaged in fundamental research in phylogenetic methods and in the design and study of computationally oriented models of evolution for systems ranging from simple genetic sequences through entire genomes to interaction networks. Yet, in spite of the fame of Dobzhansky's article and the spread of phylogenetic methods beyond the original applications to systematics, the use of methods grounded in evolutionary biology is not as pervasive as it could be.

In this talk, we illustrate some of the algorithmic problems raised by current research and some of the potential new applications of phylogenetic approaches through several projects carried out in our laboratory. The problems arise from combinatorial and algorithmic questions about models of evolution and approaches to the analysis of whole genomes. The new approaches include an extension of the time-tested and universally used comparative method, as well as applications of phylogenetic approaches to genomic transcripts and cell types, objects not typically studied through the lens of evolution.

Comparing the complete genomes of vertebrates is a daunting problem. Not only does each genome have billions of nucleotides, but almost nothing is known for 90% of even the best studied of these genomes. The standard approach today partitions the genomes into syntenic blocks, contiguous intervals along the genome that are viewed as homologous—as descending from the same contiguous interval in the genome of the last common ancestor (LCA). Since mutations, rearrangements, insertions, and other evolutionary events have transformed the LCA genome in different ways along each evolutionary path, one cannot expect to find high levels of similarity between the sequences defined by these intervals. Instead, one looks for markers, nearly perfectly conserved short sequences that are nevertheless long enough to make accidental conservation highly improbable. Homologous blocks should share (most of) their markers and have few, if any, shared markers with non-homologous blocks. Under most reasonable formulations, this problem is NP-hard and solutions to date are mostly *ad hoc*.

The issue of genomic evolution "in the large," that is, at the scale of markers, genes, or blocks and through rearrangements, duplications, and losses, has been intensely studied for nearly 20 years now, with a number of remarkable algorithmic results. Every new algorithmic result, however, has served mostly to raise interest in more complete or more sophisticated models, or to motivate new and harder problems. Combining

A. Darling and J. Stoye (Eds.): WABI 2013, LNBI 8126, pp. 1–2, 2013.

large-scale duplication events with rearrangements events, for instance, appears crucial to the understanding of genomic evolution, but remains poorly solved to date, even though researchers even now attempt to recreate "ancestral" genomes for families of organisms. Some recent observations on the bench suggest that results that appeared to be artifacts of mathematical models (such as small circular chromosomes created in the process of cutting and regluing chromosomes) may in fact arise in nature, confirming that even abstract research in models and algorithms may lead to scientific breakthroughs in biology.

The comparative method (also known as "guilt by association") attempts to transfer knowledge from one well studied system to another by establishing correspondences. Perhaps the best known example is the transfer of gene annotation from one organism to another using gene homologies. Naturally, such an approach requires a high degree of similarity for the transfer to be successful. We developed a new approach, phylogenetic transfer of knowledge, which leverages known phylogenetic relationships to improve the inference of data about modern systems. We have successfully applied our ProPhyC tool to the refinement of regulatory networks and are currently using it for the refinement and prediction of protein contact networks in protein complexes.

Phylogenetic inference assumes that the systems under study are the product of evolution and share a common ancestor. A phylogeny is simply a tree, with the (unknown) common ancestor at the root and data about the modern systems at the leaves. Evolution, in the form of various events that affect the data used to represent the systems (e.g., sequence data for genomes or directed graphs for regulatory networks), is responsible for the divergence from the common ancestor to the modern forms. Such a model does not apply directly to structures that are influenced by evolution in a less direct, or more complicated manner. Two such are transcripts in genomes exhibiting alternative splicing and cell differentiation in a single organism (or a collection of closely related ones). Transcript evolution takes place at two distinct levels—changes in the underlying gene sequence and changes in the splicing variants. We have developed a two-level framework for inference and used our TrEvoR tool on the entire ASPIC database of alternative transcripts, resulting in much enhanced accuracy in the classification of transcripts.

Cell differentiation is a direct product of development, not evolution, in cells belong to a lineage, not to a clade (group). However, cell types are consistent across individuals of the same species and their characteristic adaptations (e.g., a blood cell, a motor neuron, a red muscle cell, etc.) are the product of long-term evolution, even though in one individual all of these cells are descendants of the same undifferentiated cell in the fertilized egg. Using epigenomic data, it is possible to reconstruct a phylogeny of cell types, that is, a tree in which the children of a node represent various refinements of the single type at the parent. We have carried out such an analysis on human cells of many different types using ChIP-Seq data from the ENCODE project, with robust results that match findings and predictions from developmental biologists.

Our contention is that these are but a few of the numerous further applications of phylogenetic methods in the life sciences, applications that will require both modelling and algorithmic research.