Special Section: Phylogenetics

Daniel H. Huson, Vincent Moulton, and Mike Steel

1 Introduction

Phylogenetics is the reconstruction and analysis of trees and networks to describe and understand the evolution of species, populations, and individuals. It is fundamental to evolutionary biology and finds applications in other areas of classification, such as linguistics. Although the foundations of phylogenetics were laid down many decades ago, it is currently experiencing an exciting renaissance due to the wealth and types of biological data that are now becoming available.

In the months of September to December 2007, key researchers from around the globe working in phylogenetics and related areas gathered together within the "Phylogenetics" program at the Isaac Newton Institute for Mathematical Sciences, in Cambridge, United Kingdom, in order to push the boundaries forward in this important area of mathematical and computational biology. Solutions to problems and new directions of research instigated in this program are already starting to provide new insights to questions that are central to contemporary evolutionary biology. This special section, and five accompanying regular papers, highlights some of the progress achieved. It coincides with the 200th birthday of Charles Darwin, who imagined the history of species as being "represented by a great tree" (*Origin of Species*, Chapter 4).

The four-month program attracted around 200 researchers, with 65 program participants staying in Cambridge for prolonged periods. The program hosted three workshops, along with some shorter meetings, and it was focused on the following main themes: new data types and algorithms in phylogenetics, reticulate evolution, constructing large trees, and mathematical modeling of evolution. These themes provide a rich source of mathematical and computational problems in diverse areas such as combinatorics, algorithmic complexity, graph theory, probability theory, topology, and algebraic geometry. This special section, together with the accompanying five regular papers, provides two or three papers from each of the four themes.

Phylogenetics is a particularly interdisciplinary field, engaging biologists, mathematicians, computer scientists, and statisticians. Not only does biology benefit from the development of new mathematical, statistical, and computational techniques, but the biological problems also enrich these fields and have led to the recent emergence of areas such as "phylogenetic combinatorics" and "phylogenetic algebraic geometry." During the Phylogenetics program, the development of methodology, as well as the underlying theory, were pursued with equal vigor. To stimulate this creative process, we established a website on the PLG program website early in the program entitled "Challenges and conjectures" (http://www.newton.cam.ac.uk/programs/PLG/index.html) and it is remarkable that five of the problems listed there were either solved or had significant progress made on them during the program.

2 Overview of the Contents

We now provide a brief overview of the papers in this special section, before providing some tentative suggestions as to where the future of the field may lie.

2.1 New Data Types and Algorithms in Phylogenetics

The textbook picture of molecular systematics has traditionally viewed biological data as consisting of nicely aligned DNA sequence data from a single gene, from which a phylogenetic tree is then constructed. However, sequence data is more complex. First, sequences arrive unaligned, with insertions, deletions, and sequencing errors, yet most sequence alignment methods are based on first constructing a "guide tree," thereby leading to an annoying circularity in phylogenetic inference. Serrita et al. investigate the performance of a method that aims to solve the multiple alignment problem and phylogenetic reconstruction simultaneously. Grunewald and Moulton's regular paper investigates what happens to a simple phylogenetic approach (maximum parsimony) when different genes—perhaps with different underlying trees —are combined and then analyzed. In the paper by Minh et al., the problem of selecting a subset of taxa of maximal phylogenetic diversity for a given budget is addressed—this problem arises both in conservation biology and molecular genetics, and their solution provides a generalization of two recent algorithmic approaches.

2.2 Reticulate Evolution

It is widely assumed that the evolution of species can be depicted as a tree. For many domains of life, this simple model may be appropriate. Yet, networks are also being regarded as providing more relevant descriptions of molecular evolution. First, it is clear that some parts of the "tree of life" (particularly within prokaryotes) have involved extensive horizontal gene transfer, which makes the very concept of an underlying "tree" problematic. And, even within eukaryotes, reticulate evolution, such as the formation of hybrid species, can occur, for example, in certain plant and fish species. Networks can provide an explicit way to represent this complex history. A second

- *D.H. Huson is with the University of Tübingen, Sand 14, 72076 Tübingen, Germany. E-mail: huson@informatik.uni-tuebingen.de.*
- *V. Moulton is with the School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK. E-mail: vincent.moulton@cmp.uea.ac.uk.*
- *M. Steel is with the Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. E-mail: m.steel@math.canterbury.ac.nz.*

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org.

1545-5963/09/$25.00 © 2009 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

reason for using networks is that different genes often provide conflicting phylogenetic trees caused by processes such as lineage sorting or model misspecification and networks can provide a useful way of representing this conflict. Linz and Semple consider the question of calculating the minimal number of reticulation events required to explain two conflicting trees. Although this problem is hard, they show that it is nevertheless fixed parameter tractible. Cardona et al. investigate metrics to compare different classes of phylogenetic networks and Huson's regular paper shows how techniques that were developed to draw rooted phylogenetic trees can be modified so as to display rooted phylogenetic networks.

2.3 Constructing Large Trees

In the early days of molecular phylogenetics, data would typically be available for just a handful of taxa. This made searching through "tree-space" easy—one could simply check all possible trees or use a branch-and-bound approach. Now it is common to build trees on hundreds or even thousands of taxa and phylogenetic techniques have been developed that combine trees togther into larger "supertrees." Willson investigates some of the mathematical properties of existing supertree methods and shows that, although they can violate a very simple and desirable property, it is possible to design methods that provably satisfy this property. The regular paper by Bordewich et al. investigates the question of whether methods that attempt to search tree space for an optimal tree using popular tree rearrangement operations will necessarily find the "true" tree if the input data fits that tree perfectly. The regular paper by Wu et al. considers the complexity of optimally refining a large but partially-resolved tree given some additional character data, under the maximum parsimony criterion.

2.4 Mathematical Modeling of Evolution

Markov models form the basis of statistical approaches to tree reconstruction. They can be used to study speciation and extinction, as well as to investigate how DNA evolves. The latter is central to statistical methods for inferring phylogeny from genetic sequence data. It has also led to some deep applications of algebra to study the properties of such models and the analysis of polynomial identities, known as "phylogenetic invariants." These approaches can help address fundamental questions about how much one can know about the model from the data it generates. Allman and Rhodes address this "identifiability" question in their paper. Matsen extends the analysis of simple group based models to show that, as well as the equations that phylogenetic invariants provide, polynomial inequalities on site pattern frequencies also convey phylogenetic signal. The regular paper by Mossel et al. formally establishes a result that has been suggested without proof for some time—namely, that tree reconstruction based on ancestral maximum likelihood (a hybrid between parsimony and maximum likelihood) can be statistically inconsistent.

3 Where to from Here? A Perspective on the Future of Phylogenetics

Future research directions in phylogenetics are likely to be strongly influenced by three interacting factors: 1) new types of genomic (and metagenomic) data, which are becoming widely available given new sequencing and resequencing technologies (454, Solexa), 2) the availability of population-level data, requiring the integration of population genetic tools into phylogenetic analysis, and 3) the analysis of large numbers of taxa requiring fast but accurate algorithms to reconstruct and visualize evolutionary histories. We highlight two specific challenges.

Results from metagenomics suggest that communities of microbes do not consist of discrete sets of species in which different organisms have identical or highly similar genomes, but, rather, that for a given species there may be a whole spectrum of organisms displaying many different levels of sequence identity. A central question is how current phylogenetic methods apply in this setting.

Recent and continuing advances in sequencing technology have made it feasible to resequence thousands of genomes per year and a number of projects are underway to resequence the genomes of 1,000 humans and of additional strains and species of model organisms such as *Drosophila* and *Arabidopsis*. What are the appropriate models for infering evolutionary history incorperating both phylogenetic and population data? Phylogenetic methods will have to deal with ever larger data sets, bringing together trees that are inferred at many different locations along the genomic axes of organisms.

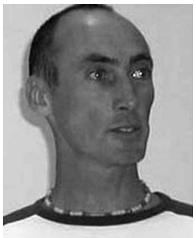
At the theoretical end, research in phylogenetics is likely to be dominated in the near future by questions concerning the mixing rate of Bayesian MCMC approaches, the efficiency of maximum likelihood and distance-based methods (such as Balanced Minimum Evolution and Neighbor-Joining), the development of better supertree and supernetwork methods, and improving existing approaches to current challenging problems. These include developing more sophisticated network-based techniques for representing and modeling reticulate evolution, improving methodology for understanding how phylogenetic information is related to and influenced by environmental factors such as geographical features and climate change, and the use of phylogenies in studying speciation and extinction.

Acknowledgments

The guest editors would like to thank Sir David Wallace and his team at the Isaac Newton Institute for ensuring the smooth running of the Phylogenetics program and Dan Gusfield and the staff of this journal for making this special section possible.

Daniel H. Huson studied mathematics at Bielefeld University and received the PhD degree in 1990. From 1990 to 1999, he held a variety of different research positions at Bielefeld University and was supported during this time by a two-year DFG research scholarship. He then spent two years, 1997 to 1999, as a postdoctoral researcher working with Tandy Warnow at Princeton University. He then joined Celera Genomics as a senior research scientist working in Gene Myers' group. Since 2002, he has been a professor of algorithms in bioinformatics at Tübingen University in Germany.

Vincent Moulton received the PhD degree in mathematics from Duke University in 1994 and did postdoctoral research at the University of Bielefeld, the University of Canterbury, and Massey University. He was a senior lecturer in discrete mathematics at Mid Sweden University (1999-2002) and a professor in bioinformatics at Uppsala University (2002-2004). In 2004, he moved to the University of East Anglia, where he is a professor in computational biology. His research interests are in phylogenetics, computational biology of RNA, metabolic modeling, algorithms in bioinformatics, and the study of discrete structures such as graphs and finite metric spaces.



Mike Steel studied mathematics at Canterbury and Massey Universities (New Zealand) and received the PhD degree in 1989. From 1990-1993, he held various postdoctoral positions in Germany and New Zealand and was appointed to a tenured position at the University of Canterbury in 1994. He is currently a professor and director of the Biomathematics Research Centre at the University of Canterbury and is a principal investigator in the Allan Wilson Centre for Molecular Ecology and Evolution.