Final Report for the ‘Phylogenetics’ Programme

Daniel Huson, Vincent Moulton, and Mike Steel

I. Introduction

PHYLOGENETICS is the reconstruction and analysis of trees and networks to describe and understand the evolution of species, populations and individuals. It is widely used in molecular biology and other areas of classification (such as linguistics), and has both led to and benefited from the development of new mathematical, statistical and computational techniques. 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 September–December 2007, key researchers from around the globe working in phylogenetics and related areas gathered together within the ‘Phylogenetics’ programme at the Isaac Newton Institute for Mathematical Sciences, UK, in order to push forward the boundaries of this important area of mathematical and computational biology. Solutions to problems and new directions of research instigated in this programme are already starting to provide new insights to questions that are central to contemporary evolutionary biology.

II. The main programme themes

The programme aimed to develop our knowledge on the following main themes: new data types in phylogenetics; reticulate evolution; constructing large trees; and mathematical modelling of evolution. These themes, which we shall now describe in more detail, 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.

A. New data types in phylogenetics

Until quite recently most modern methods for constructing phylogenetic trees have been designed with sequence data in mind, usually constructing evolutionary trees from genes as an approximation to species phylogenies. However, the abundance of new types of molecular data (such as whole genomes, expression data, metabolic networks) is creating interesting new challenges for phylogenetics. Not only do we have to reconsider previous estimates of phylogeny in view of new data, but new methods need to be established that allow the incorporation of subtle phylogenetic signals in the data. Moreover, the incorporation of phylogenetic information into bioinformatics methods for tackling problems such as motif discovery in genomes/ biochemical networks (or phylogenetic footprinting as it is sometimes called), can significantly improve sensitivity, although often at the price of introducing hard mathematical and computational variants of well-studied problems.

- *Daniel Huson is with University of Tübingen*
- *Vincent Moulton is with University of East Anglia*
- *Mike Steel is with University of Canterbury*

B. Reticulate evolution

How can we best model reticulate evolution? For example, from genomic data can we determine how much gene transfer occurred early in the Tree-of-Life by comparing the genomes of extant species? Various techniques for building networks have been proposed. For example, since their introduction in the early 1990’s, a rich mathematical theory has started to emerge for representing phylogenetic relationships using so-called split networks (see e.g. Figure 1). These networks, which include median networks and NeighborNets as special examples, provide a snap-shot of data which can indicate the presence of incompatibilities that are often the consequence of non tree-like evolutionary processes. Even so, there is currently great interest in the development of new theories and constructions for phylogenetic networks that provide a more concrete representation of reticulate evolution.

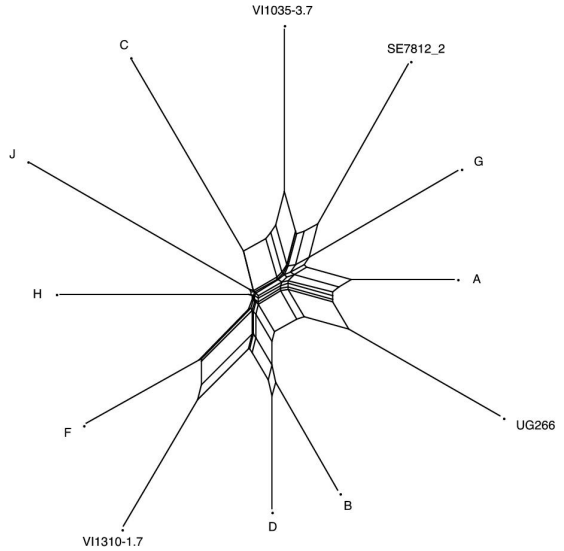

Fig. 1. A split network computed from HIV sequences detailed in *The Phylogenetic Handbook, Cambridge University Press, 2003*. Letters A–J denote HIV-1 subtypes, and the remaining labels denote recombinant viruses.

C. Constructing large trees

Biologists wish to build large trees across thousands of species leading to substantial combinatorial and statistical problems. These trees not only deepen our understanding of the Tree-of-Life, but also provide useful information for the understanding of global biodiversity, a matter of growing public concern. However, popular methods for tree reconstruction (such as maximum parsimony and maximum likelihood) can sometimes be far too computationally expensive for deriving large trees. Moreover, biologists commonly wish to combine several trees from overlapping data sets to obtain overall estimates of phylogeny. Development of methods to provide solutions to these challenges are key to

the success of projects such as the US-based CIPRES initiative to reconstruct the Tree-of-Life that is aimed at developing an infrastructure for computing large trees.

D. Mathematical modelling of evolution

Stochastic models have long played an important role in phylogenetics. Indeed, in early, pioneering work, Yule in 1924 showed how simple branching-type processes could model the distribution of species numbers by genera. More recently statisticians (beginning with Harding in 1971) began to study how the 'shape' of phylogenetic trees could be predicted from simple speciation models. Further investigations by probability theorists and biologists have allowed for features of published trees to be studied, with the goal of learning more about the process of speciation, and testing specific hypotheses. Other processes in phylogenetics where models are of interest include the study of models of character evolution - for example, how does DNA evolve, and how can we use these models to refine methods for tree reconstruction? Another is the use of species-level phylogenetic techniques to study population-level processes through the coalescent process. This process (introduced by John Kingman) has become central to many statistical approaches to studying the evolution of sequences within populations, particularly subject to processes such as recombination, mutation, selection and migration. The study of these models lead to interesting mathematical and computational problems, in areas such as probability theory, algebraic geometry and combinatorics, which are of interest in their own right.

III. Structure of Programme

The programme lasted for 4 months and included 3 workshops together with a half-day meeting aimed at new comers to phylogenetics. It attracted in the order of 200 researchers from all over the world, and over 65 programme participants that stayed in Cambridge for prolonged periods. In addition to the workshop talks, several seminars were delivered during the programme both in Cambridge and across the UK, and a weekly discussion group took place in which new directions were discussed by participants. We now present a brief summary of the 4 main events that took place during the programme.

A. EMBO Workshop on Current Challenges and Problems in Phylogenetics

The workshop took place at the Isaac Newton Institute, September 3-7, and provided a showcase for some recent achievements, challenges and new problems that arise in using mathematical approaches to understand molecular evolution. In line with the main themes of the programme, key topics covered within the workshop included (1) the challenges involved in constructing very large-scale phylogenies, especially in relationship to reconstructing the Tree-of-Life, (2) development of methodologies to reconstruct phylogenetic networks so as to uncover the evolutionary histories resulting from reticulate evolution, (3) extending the construction of gene trees to whole genome phylogenies, and understanding the associated mathematical challenges such as tree mixing and model averaging, and (4) development of methods based on phylogenetic diversity to understand and conserve biodiversity.

Some specific highlights included an invited talk by M. Kucera on the use of stratophenetic tracing in fossil records leading to a comparison between fossil phylogeny of one monophylum of planktonic foraminifera with corresponding SSU rDNA phylogenies. This talk described empirical evidence for long-branch artefacts, large differences in substitution rates and incongruent tree topologies. Another highlight was the talk by I. Ebersberger on mapping human genetic ancestry, which concluded that about 1/3 of our genes evolved as human-specific lineages before the differentiation of human, chimps and gorillas took place.

The workshop was a great success, both in terms of its outstanding scientific standards and in terms of the liveliness, participation and friendliness of the participants. It attracted an impressive group of participants including established scientists, talented young researchers and promising graduate students. In addition to ten world class invited talks of about 1 hour in length each, participants were treated to over 25 outstanding short 20-minute contributed talks, high-lighting the great interest that phylogenetics is generating across various scientific disciplines.

B. Phyloinformatics Workshop

This workshop took place 22-24 October at the e-Science Institute, Edinburgh, UK. Phyloinformatics can be broadly described as the field concerned with the new informatics challenges arising from acquiring, storing and manipulating the phylogenetic data associated with large-scale projects such as constructing and the Tree-of-Life and cataloguing Earth's biodiversity. In the workshop a variety of questions were explored, such as how to compute large phylogenetic trees and visualise/navigate then efficiently?, What is the most efficient way to mine large databases for phylogenetic analysis?, and How should phylogenies be integrated with other data from genomics, geography, stratigraphy, ecology, and development? As a result, and through the 3 discussion sessions, the following points were identified as key for the future of phyloinformatics:

- Coordination of megaprojects (e.g. Global Biodiversity Information Facility (GBIF), Encyclopedia of Life (EoL), CIPRES) to allow for more interoperability.
- Development of new methodologies for phylogenetic tree construction, storage and querying.
- Improved models for tool development to reduce redundancy and allow for different platforms.
- Identifying, prioritizing and filling the gaps in current sequence data.
- Improving outreach to allow other communities the easy use of phyloinformatics tools.

The workshop consisted of 15 talks, 11 from invited experts, plus 3 half hour group discussion sessions on each day (chaired by Mike Sanderson, Mark Westneat, and Olaf Bininda-Emonds, respectively). As with the first workshop, all talks were of extremely high quality, with highlights including Mike Sanderson's talk concerning how to construct the Tree-of-Life from the thousands of phylogenetic trees available in current online data bases, and Mark Westneat's description of the forthcoming web-based EoL.

C. Yggdrasil: Reconstructing the Tree of Life

A Spitalfields Day "Yggdrasil: Reconstructing the Tree of Life", took place at the Isaac Newton Institute on 6 December. Yggdrasil, the 'World Tree' in Norse mythology, provided a vivid image for the field of phylogenetics. The meeting consisted of 4 expository lectures directed at final year undergraduate/beginning

postgraduate students in biology, mathematics, and computer science, highlighting different aspects of phylogenetics.

The first talk, by biologist Peter Lockhart of Massey University (NZ), introduced the problem of inferring phylogenetic trees for chloroplasts, emphasizing the mathematical and biological difficulties of modelling the process of evolution in such organelles. This was followed by computer scientist Tandy Warnow's (University of Texas, USA) talk on computational issues in phylogenetics, which highlighted connections between graph theory and combinatorics and methods of phylogenetic tree reconstruction. After lively conversation over tea, University of Alaska (USA) mathematician John Rhodes spoke on the use of algebraic geometry for theoretical analysis of phylogenetic models. The final speaker of the day, Andreas Dress, Director of the CAS-MPG Partner Institute for Computational Biology in Shanghai, and pictured in Figure 2, discussed the role of models in phylogenetics, illustrating his points with some memorable analogies (such as sphere-shaped cows). He further drew attention to some of the combinatorial aspects of current research projects in this area, including the tight span of metric spaces.

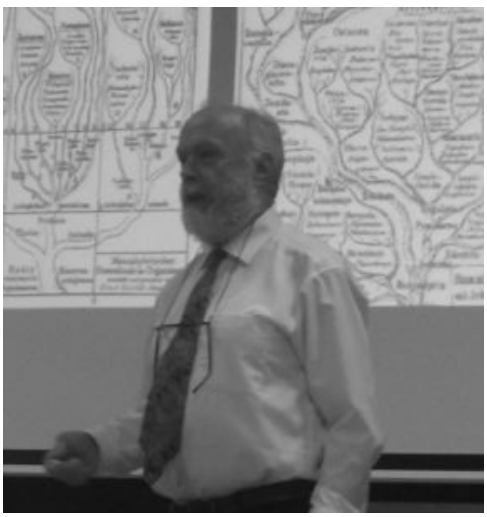

Fig. 2. The Rothschild Visiting Professor, Andreas Dress, discussing Haeckel's Tree-of-Life at the Spitalfields Day.

D. Future Directions in Phylogenetic Methods and Models

The final workshop was held during December 17–21 and attracted 72 participants from 16 countries. The meeting provided both the culmination of the 4-month programme and a glimpse into the future, with reports on results obtained and questions to explore. Each day was based on a different theme, with a keynote speaker setting the scene with a 1-hour seminar. These themes (and speakers) comprised the following: (i) The tree of life – algorithmic and software challenges (Tandy Warnow); (ii) Phylogenetic combinatorics and algebra (Andreas Dress); (iii) Speciation, extinction and tree shape (David Aldous); (iv) The complexities of molecular evolution (Andrew Roger); and (v) Population genetics in phylogeny (Noah Rosenberg). There were 41 talks, with much time spent in informal discussion. Feedback from participants in the exit survey suggested they were very pleased with this last meeting.

IV. Outcomes and achievements

A. Scientific outcomes

1) New data types in phylogenetics: There are now literally thousands of whole genome sequences available, allowing us to dig deeper and deeper into the evolutionary history of present day organisms. For example, phylogenetic trees are now being built for HIV viruses based on whole genome sequences and, using trees such as these, programme participants (e.g. Lemey, Pybus, Rambaut) worked on developing new tools to understand virus evolution, in order to understand problems such as how HIV populations migrate. In related work, other participants (e.g. Gascuel, Spencer, Székely, Vision) grappled with problems in bacterial genome evolution such as how to deal with subsets of genes having different behaviours, and how to compare multiple genomes.

Within the programme it also became clear that there is still quite some debate on how tree reconstruction methods originally designed to deal with single genes should be extended to whole genomes. As part of the process of obtaining a deeper mathematical understanding of how to do this, several of the participants worked on developing a more unified theory for mixture models (e.g. Allman, Kim, Matsen, Rhodes, Steel). These models have been proposed as a way for biologists to analyse data in which certain DNA sequence sites evolve quite differently to other sites, due to structural or functional constraints. Such models can seriously mislead existing phylogenetic approaches, and it is a challenging problem to determine whether methods can be developed that will unambiguously extract phylogenetic signal from data that has evolved according to a mixture model. Although much progress has been made over the last year, and particularly during the programme, further work is needed to fully settle the 'identifiability' question for mixture models. A further insight into the problem of tree reconstruction from non-homogeneous data was a theoretical result concerning the complexity of computing a most parsimonious tree for two genes was obtained (Gruenewald, Moulton).

An exciting new direction for research was also presented by the new generation of sequencing technologies (such as 454 and Selexa sequencing). These technologies deliver large numbers (10^9) of short sequences (40-250nt), and present the possibilities of sequencing short genomes in hours or gathering large numbers of markers from larger genomes. One topical application of such sequencing techniques is metagenomics, the study of genetic material recovered from environmental samples (e.g. the DNA sequences contained in a handful of soil). In this context, participants (e.g. Huson, Rodrigo, Spencer) developed new methods to deal with issues such as How to separate mixtures of genomes? and How to statistically decide the abundance of sequences coming from each genome in the sample?

2) Reticulate evolution: Much interest was generated in the programme concerning the further development of the theory of phylogenetic networks based on acyclic digraphs. These networks can provide an intuitive representation of the evolutionary relationships between species, although surprisingly little is known concerning combinatorial properties of such networks and general methods for their construction. Several lively discussions and talks on this topic led to new insights as to how current network construction methods are related and how to construct and draw such networks. For example, methods were developed for constructing networks from combinatorial data such as triplets

and clusters, as well as new software for their computation (e.g. Dress, Huson, Kelk, Huber, Rupp, Stougie, Willson). Applications of such networks to the modelling of recombination and the reconstruction of whole genome phylogenies (through e.g. combining trees into networks) were also pushed forward (e.g. Holland, Lockhart, Huson, Gusfield, Willson). In related work, new results were developed concerning tanglegrams (Gusfield, St.John), and also concerning the reticulate evolution of languages (Warnow).

3) Constructing large trees: Some approaches developed by graph theorists are already being applied by biologists in the construction of large trees in the form of 'supertrees' (combining trees that classify overlapping sets of species). Within supertree construction, algorithmic approaches were further developed to satisfactorily handle constraints such as edge-lengths, divergence dates, and ancestral taxa (e.g. Semple, Willson), and to build such trees (and networks) from dense data (e.g. Kelk, Huber, Willson). In addition, alternative methods for efficiently constructing large trees based on distance measures and likelihood scores were developed, together with a theoretical analysis of issues in constructing such trees when ancestral data is involved (e.g. Holland, Roch, Whelen, Warnow). Applications of large trees to the understanding of diversity generated a great deal of interest and new results. For example, a conjecture concerning phylogenetic diversity for two trees was solved (see Figure 3), and new methods were developed for improving the applicability of phylogenetic diversity (e.g. Bordewich, Hartmann, Klaere, Rodrigo, Semple, Spillner, von Haeseler).

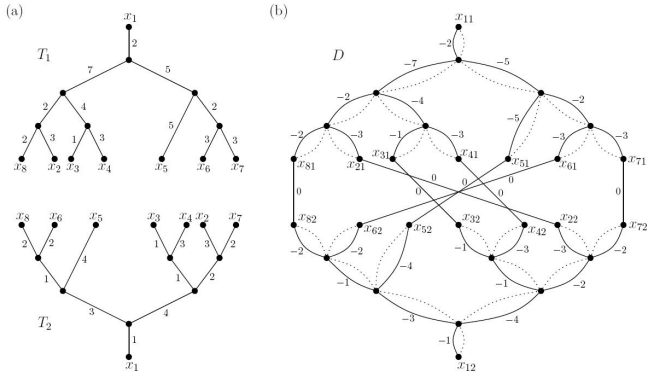

Fig. 3. To efficiently find subsets of the set $\{x_1, x_2, \ldots, x_8\}$ with optimal phylogenetic diversity score relative to the two edge-weighted phylogenetic trees T_1, T_2 pictured in (a), Bordewich, Semple, and Spillner showed that a network flow problem can be solved as illustrated in (b).

4) Mathematical modelling of evolution: One of the main tools in understanding how DNA evolves is the study of Markov models of sequence evolution (on a tree or network), and it is the basis of widely-used likelihood-based and Bayesian approaches to phylogenetics (as well as 'corrected distance' approaches). Participants worked on improving the accuracy of such models through, for example, estimating empirical substitution matrices from huge alignment data bases (e.g. Gascuel, Goldman, Holder). In addition, research was done on the consequences of model mis-specification (e.g. Howe, Lockhart, Naylor, Steel), and on methods for accelerating Bayesian MCMC inference (Nicholls, Rodrigo).

Many stochastic models lead to interesting mathematical problems which are of interest in their own right. For example, Markov models for sequence evolution give rise to polynomial ideals ('phylogenetic invariants') that have a rich algebraic and geometric structure. The theory of such invariants was intensively studied by several participants (e.g. Allman, Matsen, Kim, Rhodes), leading to new results concerning model identifiability and the geometry of phylogenetic models. Related probabilistic questions were also studied (e.g. Mossel, Roch, Steel, Székely) yielding solutions to two problems: (i) How can we efficiently reconstruct species trees from gene trees that conflict due to lineage sorting? and (ii) Is the amount of data required to 'test' whether or not a given phylogenetic tree is 'true' fundamentally less than the amount of data required to reconstruct the phylogenetic tree from scratch?

Participants (e.g. Dress, Gruenewald, Koolen, Moulton, Huber, Spillner, Steel) also pushed forward the new mathematical theory of phylogenetic combinatorics. This subject is concerned with the combinatorial problems involved in modelling evolution and constructing trees. Results were obtained on decompositions of genetic distances based on the tight-span construction of a metric space, and on optimal network realizations of metric spaces. New insights were also gained concerning SPR/TBR combinatorial tree moves and their relation to tree-space (e.g. Bordewich, Gascuel, Huber, Erdös, Steel, Székely).

B. Collaborations

Many important collaborations were started or developed during the programme. In general, participants found it helpful to be able to meet regularly and talk with experts over an extended period. UK participants particularly commented that the programme gave them an excellent opportunity to make new contacts at both a national and international level. In addition, many of the participant's commented on the fact that the programme greatly benefited from both formal and informal discussions. To facilitate these discussions, weekly Monday morning informal discussions were held, generating a great deal of new ideas, and also weekly social events outside work hours, both of which were very well received.

In general, as organisers we were very pleased with the core group of long-term visitors, even though it was slightly smaller than planned due to some late cancellations. Both the Rothschild Visiting Professor (Dress) and the Microsoft Fellow (Lockhart) made valuable contributions to the programme. The flow of short term visitors also provided enriching stream of new ideas to work on. In terms of collaborations, one highlight of the programme was a joint session with the SIS programme participants, in which organisers presented an overview of their respective programmes, followed by discussions.

In general, many of the participants commented on how much they appreciated the working environment in the Institute (in terms of e.g. the open central area, library, and facilities). This undoubtedly acted as a great encouragement for collaboration. A special mention should also be made concerning the staff at the INI. They were unswervingly helpful in contributing to smooth running of programme, ensuring all the participants' needs were met.

C. Publications

One of the main outcomes will be a special issue of the journal *IEEE/ACM Transactions in computational biology and*

bioinformatics, to which we expect about 8-10 papers from the programme participants (all of which will be refereed to usual high standards of TCBB). Publication is planned in early 2009. In addition, one book was started (Phylogenetic networks, D. Huson, R. Rupp), another significantly progressed (Phylogenetic combinatorics, A. Dress, K. Huber, J. Koolen, V. Moulton, A. Spillner), whilst another was almost completed (Reconstructing phylogenies, C. Howe, P Lockhart, D. Morrison).

Of course, many other outputs (including several papers either submitted during the programme or in progress; roughly 30 reported by participants) will be published or presented in other outlets. To stimulate this creative process, we established a website on the PLG programme website early in the programme entitled 'Challenges and conjectures' *http://www.newton.cam.ac.uk/programmes/PLG/index.html*, and it is impressive that five of the problems listed there were either solved, or had significant progress made on them during the programme (in most cases the outcomes will be published).

V. Concluding remarks

As organisers, we were particularly pleased with how the programme progressed. We were encouraged that several leading experts in the field (e.g. Allman, Dress, Mossel, Huber, Lockhart, Semple, Rhodes, Warnow, Willson and others) were able to spend prolonged periods at the institute. The mix of participants ranged across many categories (geography, seniority, gender) as well as disciplines (mathematics, statistics, computer science, biology). It was very pleasing to witness a wide array of mathematical fields in interaction with evolutionary biology — from algebraic geometry, topology and category theory, through to discrete mathematics and probability theory. We expect that the programme will lead to other meetings over coming years involving mathematicians and evolutionary biologists, starting with one in France in June this year, and we have already begun discussions with the director concerning a follow-up meeting to be held at the institute as part of this process.