Random graph models in phylogenetics

Mike Steel (University of Canterbury, Christchurch, New Zealand),
Tanja Stadler (ETH Basel, Switzerland).

September 9, 2016

1 Overview of the Field

Phylogenetics is the reconstruction and analysis of evolutionary trees (and networks) from data, and is widely used in biology and other branches of classification (e.g. linguistics). The leaves of these 'species trees' typically correspond to extant taxa (and, occasionally, earlier sampled fossils), while the interior vertices of the tree correspond to hypothesised ancestral species; branching events correspond to speciation events. The inference of species trees is generally carried out from 'gene trees'. These describe how a gene sampled at the present from individuals of the given species traces back in time to a common ancestor. Phylogenetics is an area that has benefitted greatly from the application of mathematical techniques, particularly combinatorics and probability theory, and a mathematical foundation for this field has developed over the last 50 years [13].

Given a species tree, with branch lengths (the time between speciation events, combined with ancestral population sizes), the gene tree is a random process that is described by the *multispecies coalescent model*, an extension of Kingman's celebrated *coalescent model* from populations to phylogenies [6]. This interplay between gene trees and species trees has been, and continues to be, a hot topic in phylogenetics, and has been

explored in various discussions by our BIRS research group. Gene flow in hybrid evolution can also lead to statistical inconsistency in the estimation of the dominant underlying tree phylogeny [11].

However, random processes also play a key role in several other areas of phylogenetics. For example, the evolution of the species tree is itself a random process (governed by speciation and extinction events). Since one generally does not detect the extinct lineages and because some non-extinct lineages may not have been sampled, we obtain a randomly 'pruned' species tree that is called the 'reconstructed tree'. The stochastic analysis of reconstructed trees has greatly benefitted from recent 'coalescent point process' analysis by Lambert, Stadler and colleagues [8, 12]. The relationship between the reconstructed tree on higher–level taxa (e.g. genera) and that on species is also of particular interest.

A further area where random processes play a key role is in modelling the evolution of discrete characters (e.g. DNA sequences), which are the basis of standard methods for inferring phylogenetic (gene) trees.

Our meeting consisted of each participant presenting some of his/her recent work and open challenges on the blackboard. These informal presentations opened the floor for very lively discussions. In the next section, we describe the main presentation and discussion points

2 Recent Developments and Open Problems

A longstanding problem in biology is to provide a satisfying answer to the question "What is a species?". Biologists have devised numerous concepts of 'species'; however, most have shortcomings of being ill-defined, or restricted to only part of the 'tree of life'. The notion of species for prokaryotes (bacteria and archea) is particularly problematic [7]. Building on recent work of Amaury Lambert [9] and earlier work [10], we had a wide-ranging discussion of various ways to define 'species' based on phylogenies, in which higher–level taxa (eg. genera) can give rise to monophyletic species in various ways (related to a similar approach by [2]. In ecology, there is a well-developed theory for estimating species abundance, based, for example on the neutral theory of biodiversity and related stochastic models.

Within phylogenetics, the notion of a species tree, in which a species comprises a population (or populations), allows for abundances to be included in phylogenetic models. One outstanding question is whether or not there is a non-parametric explanation for the shape of empirical trees fitting the β–splitting model statistic $\beta \sim -1$ (observed in [1] and later studies). Many neutral models predict $\beta = 0$; however, empirical trees tend to be less balanced than this. Models that allow diversification rates to depend on age can be adjusted to give $\beta \sim -1$ but require parameter tuning.

Another area that has benefitted enormously from the input of mathematicians (particularly probability theorists) is the development of statistically consistent methods for inferring species trees from gene trees (under incomplete lineage sorting, and also under lateral gene transfer). In some settings, the notion of a 'tree of life' is often too restrictive, since biological evolution also involves the formation of hybrid species, and processes whereby genes are exchanged between certain taxa (e.g. bacteria). Thus the development of methods for reconstructing, comparing and counting phylogenetic networks is particularly timely.

The notion of a continuous 'tree space' is also important for comparing different phylogenies. Recently distances between trees (and tree averages) in the Billera-Holmes-Vogtmann (BHV) space of phylogenies (a CAT(0) space introduced in [3]) have been shown to be computable. BHV space is quite different from the so-called 'phylogenetic orange' space, in which trees are regarded as 'close together' if they are hard to distinguish from data that has evolved on them.

We also explored better ways to estimate the posterior probability of a tree, based on a given collection of characters. For continuous characters (e.g. those relating to fossil morphology), the approach is often to 'discretise' these; however, other approaches (based on Ornstein–Uhlenbeck or Brownian motion models, and allowing correlations between characters) could also be applied. Finally, the expected loss of biodiversity, as measured by phylogenetic diversity, through random extinction at the present due to climate change and other anthropogenic activity is an important random process (the so-called 'field of bullets' model).

3 Presentation Highlights

Techniques for describing and comparing networks in terms of multi-labelled phylogenetic trees were described by Cecile Ané, leading to lively discussion, and the analysis of hominid fossil skeletal data was dis-

cussed in presentations by Arne Mooers and Tanja Stadler. A recent 'fossilised birth–death process' model has been developed by Standler, and further refinements of this are likely to play a key role in analyses by others (e.g. Alexei Drummond who co-developed the widely-used phylogenetic software package BEAST). Novel methods for estimating posterior probabilities of trees more efficiently (based on Hamiltonian dynamics and on 'regularised' phylogeny reconstruction [14]) were presented by Vu Dinh, leading to fruitful discussions.

Quantifying the extent to which phylogenetic diversity mirrors trait diversity was explored by Arne Mooers, along with ways of relating phylogenetic diversity loss to other species-specific indices of diversity (e.g. 'fair proportion', which is equivalent on rooted phylogenies to the Shapley index from cooperative game theory). A wide-ranging mathematical generalisation of phylogenetic diversity was outlined by Bryant, who (with another mathematician, Paul Tupper) has developed 'diversity theory' with connections to geometry, combinatorial optimisation, and metric space theory [4].

Mathematically, we can analyze B-cell data from immunology in a similar way to species data. Generation of a B-cell clone corresponds to speciation; the death of a B-cell clone correpsonds to species extinction. Thus phylogenetic tools for species evolution may be transferred to immunology. Drawing on his expertise in cancer research, Erick Matsen provided a detailed discussion of the real-time evolutionary processes that apply to B-cells (an important part of the immune system), leading to a number of phylogenetic questions connected with birth–death and branching process models [5]. These topics are currently being explored further in ongoing work between Matsen, Lambert and Stadler.

4 Scientific Progress Made

Significant progress was made during the meeting on a number of phylogenetic issues and problems. Particular advances included exploring the relevance of stochastic models to B-cell evolution, explaining symmetries in birth–death models, unravelling properties of phylogenetic consensus methods.We also obtained new insights into the space of phylogenetic networks, possible metrics for semi-directed networks and local measures of similarity that will be used in future work to summarise bootstrap and Bayesian posterior samples of networks.

5 Outcome of the Meeting

A number of new project collaborations were established among subgroups of the participants. Numerous ideas on how to attack certain key open questions in phylogenetics were put forward. Some of the ideas generated at the meeting will ultimately lead to modifications of mainstream phylogenetic software (e.g. new algorithms for phylogenetic diversity will be encoded into the next version of SplitsTree (by Bryant)). New collaborations have emerged among participants, particularly in relation to extended types of birth–death models (and their application, for example, to analysing B-cell processes, and hominid fossil evolution), as well as new ways to investigate biodiversity loss due to extinction at the present. A further outcome of the meeting was a prediction of the way that phylogenetics will develop over the next 5 years (this is the latest in a series of 5-year predictions hosted on Steel's webpage).

References

[1] D. J. Aldous (2001), Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Stat. Sci.* **16**, 23–34.

[2] D. Aldous, M. Krikun and L. Popovic (2008), Stochastic models for phylogenetic trees on higher-order taxa, *J. Math. Biol.* **56**, 525–557.

[3] L. J. Billera, S. P. Holmes, and K. Vogtmann (2001), Geometry of the space of phylogenetic trees, *Adv. Appl. Math.*, **27**, 733–767.

[4] D. Bryant and P. F. Tupper (2012), Hyperconvexity and tight-span theory for diversities, *Adv. Math.*, **231**, 3172–3198.

[5] R. Duncan and F. A. Matsen (2016), Likelihood-based inference of B-cell clonal families, *PLoS Comput. Biol.*, in press.

[6] L. L. Knowles and L. S. Kubatko (2010), *Estimating Species Trees: Practical and Theoretical Aspects*, John Wiley and Sons, Hoboken.

[7] E. V. Koonin (2015), The turbulent network dynamics of microbial evolution and the statistical tree of life, *J. Mol. Evol.*, **80**, 244–250.

[8] A. Lambert and T. Stadler (2013), Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies, *Theor. Pop. Biol.*, **90**, 113–128.

[9] M. Manceau and A. Lambert (2016), The species problem from the modeler's point of view (manuscript).

[10] R. W. Scotland and M. J. Sanderson (2004), The significance of few versus many in the tree of life, *Science*, **303**, 643.

[11] C. Solis-Lemus, M. Yand and C. Ané (2016), Inconsistency of species-tree methods under gene flow, *Syst. Biol.*, in press.

[12] T. Stadler and F. Bokma (2013), Estimating speciation and extinction rates for phylogenies of higher taxa, *Syst. Biol.* **62**, 220–230.

[13] M. Steel, (2016), *Phylogeny: Discrete and Random Processes in Evolution*, CBMS-NSF Regional conference series in Applied Mathematics (SIAM).

[14] D. Vu, H, Lam, S. Tung, M. A. Suchard and F.A. Matsen (2016), Consistency and convergence rate of phylogenetic inference via regularization, arXiv:1606.03059.