Phylomania 2010

The UTas theoretical phylogenetics meeting

University of Tasmania
School of Maths and Physics
4-5 Nov 2010

Program

Thursday, 4 November

8.15am-9:00am	Registration and coffee
9:00am-9:10am	Welcome
9:10am-9:50am	**John Rhodes**, University of Alaska (Fairbanks) *Semi-algebraic descriptions of the general Markov model*
9:50am-10:30am	**Jeremy Sumner**, University of Tasmania *A Lie algebraic classification of continuous-time Markov models*
10:30am-11:00am	Morning tea
11:00am-11:40am	**Mike Hendy**, Massey University *Radical phylogenetic inversion*
11:40am-12:20pm	**Michael Charleston**, University of Sydney *The current status of cophylogenetic analysis*
12:20pm-1:40pm	Lunch
1:40pm-2:20pm	**Elizabeth Allman**, University of Alaska (Fairbanks) *On the identifiability of 2-tree mixtures*
2:20pm-3:00pm	**Steffen Klaere**, University of Auckland *The link between segregation and phylogenetic diversity*
3:00pm-3:30pm	Afternoon Tea
3:30pm-4:10pm	**David Bryant**, University of Otago *The many wonders of diversity*
4:10pm-4:50pm	**Charles Semple**, University of Canterbury *Submodular functions and biodiversity conservation*
7:30pm	**Phylomania 2010 pub dinner** (venue TBA)

Friday, 5 November

9:00am-9:30am	Coffee
9:30am-9:50am	**Jessica Leigh**, University of Otago *Simulation-based testing in an approximate Bayesian framework*
9:50am-10:10am	**Denise Kühnert**, University of Auckland *Population dynamics in phylogeographic analyse*
10:10am-10:30am	**Alethea Rea**, University of Auckland *Statistical phylogenetics with Nets and a Lasso*
10:30am-11:00am	Morning tea
11:00am-11:40am	**James Degnan**, University of Auckland *Identifying species trees from clade probabilities*
11:40am-12:20pm	**Bjarki Jónsson Eldon**, Oxford University *Concordance between species trees and gene genealogies with multiple mergers*
12:20pm-1:40pm	Lunch
1:40pm-2:20pm	**Matt Phillips**, Australian National University *Marsupial herbivore evolution and the failure of algorithmic morphological phylogenetics*
2:20pm-3:00pm	**Xavier Goldie**, Australian National University *Wattles and Wombats: Molecular rates and diversification in plants and mammals*
3:00pm-3:30pm	Afternoon Tea
3:30pm-4:10pm	**Simon Ho**, University of Sydney *Time-dependent rates of molecular evolution: Evidence and causes*
4:10pm-4:50pm	**Rob Lanfear**, Australian National University *Using molecular phylogenies to estimate the timing of species addition to communities*

Saturday, 6 November

12pm-	**Phylomania 2010 Mount Wellington walk** (details TBA)

Abstracts

Elizabeth Allman

Institute: University of Alaska, Fairbanks, USA.

On the identifiability of 2-tree mixtures

Joint work with S. Petrovic, J. Rhodes, and S. Sullivant.

Phylogenetic data arising on two possibly different tree topologies might be mixed through several biological mechanisms, including incomplete lineage sorting in the case of different topologies, or simply different substitution processes on characters in the case of the same topology. In this talk, we discuss the identifiability of 2-tree mixtures for group-based models. That is, we investigate when a probability distribution of pattern frequencies arising from a mixture model on two trees contains enough information to recover the model parameters. Since this work may be familiar to some, we take time to describe the methods from algebraic geometry that were employed in the proofs.

David Bryant

Institute: University of Otago, New Zealand.

The many wonders of diversity

Joint work with Paul Tupper.

A 'diversity' is a function δ defined on subsets of a set that satisfies the axioms

1. $\delta(A) \geq 0$ for all A;
2. $\delta(A) = 0$ if and only if $|A| \leq 1$;
3. $\delta(A \cup B) + \delta(B \cup C) \geq \delta(A \cup C)$ whenever B is non-empty.

These functions pop up in diverse domains; one topical example is the phylogenetic diversity function that has recently been banned from Systematic Biology. The first ever talk about these abstract diversities was at Phylomania 2009, where I showed how tight span theory extended elegantly to diversity functions. At Phylomania 2010 I'll talk about progress and regression over the past year, about the frustrations of infinity, and about how a fairly simple idea is promising to link disparate domains of pure and applied mathematics.

Michael Charleston

Institute: University of Sydney, Australia.

The current status of cophylogenetic analysis

James Degnan
Institute: University of Canterbury, New Zealand.

Identifying species trees from clade probabilities
Joint work with Elizabeth Allman and John Rhodes.

One method for estimating a species tree from a collection of gene trees is to estimate probabilities of clades (in the rooted setting) or splits (unrooted) from the gene trees, and then to construct the species tree from the weights on the clades or splits. This can be done straightforwardly using an algorithm such as greedy consensus, which accepts the most probable clades one at a time compatible with previously accepted clades until a fully resolved tree is returned. When clade probabilities are generated by the multispecies coalescent, however, this method is known to be statistically inconsistent in the sense that for some species trees and species tree branch lengths, greedy consensus on the clade frequencies will be less likely to return the species tree as the number of input trees increases. This raises the question of whether it is possible to correctly reconstruct the species tree from known probabilities of clades. Many properties of clade probabilities under the coalescent suggest that recovering the species tree from clade probabilities is difficult. For example, the most probable 2-clade (cherry) in a random gene tree is not necessarily a clade in the species tree. Also, clade probabilities on smaller trees (e.g., rooted triples), cannot be written as linear combinations of clade probabilities on larger trees independently of the species tree topology. Despite these challenges, we use linear invariants of clade probabilities to show that clade probabilities do indeed identify the species tree topology under the multispecies coalescent.

Bjarki Jónsson Eldon
Institute: Oxford University, United Kingdom.

Concordance between species trees and gene genealogies with multiple mergers
Joint work with James Degnan.

Concordance between species trees and gene genealogies are considered for gene genealogies allowing multiple mergers of ancestral lineages. Donnelly and Kurtz, Pitman, and Sagitov independently introduced coalescent processes (Lambda coalescent) that allow asynchronous multiple mergers of ancestral lineages. Kingman's coalescent allows only two ancestral lineages to merge at a time.

Lambda coalescents may be more appropriate than Kingman's coalescent for some marine organisms with high fecundity and broadcast spawning. Multiple mergers, ie. allowing any number of ancestral lineages present each time to coalesce, complicates computations.

Two methods will be presented to compute the concordance, one relies on enumeration of all possible sequences of mergers, and the other is the spectral decomposition of the rate matrix. Results will be presented for two and three species and small sample sizes from each, for different Lambda coalescents.

Xavier Goldie
Institute: Australian National University (ANU), Australia.

Wattles and Wombats: Molecular rates and diversification in plants and mammals

Why do some groups of organisms contain many more species than other groups? Understanding the factors that drive and temper the processes of speciation and extinction, including those intrinsic and external to a species' biology, are central to understanding the evolutionary processes that shape diversity patterns. Of particular interest is the covariation of biodiversity with climate,

a ubiquitous pattern in nature. The evolutionary speed hypothesis posits direct mechanistic links between ambient temperature, the tempo of micro-evolution and, ultimately, species richness. In terrestrial systems, species richness increases with both temperature and water availability and the interaction of those terms: productivity. However, the influence of water availability as an independent variable on micro-evolutionary processes has not been examined previously.

Using methods that limit the potentially confounding role of cladogenetic and demographic processes we report evidence that woody plants living in the arid Australian Outback are evolving more slowly than related species growing at similar latitudes in moist habitats on the mesic continental margins. The evolutionary speed hypothesis is prefaced on a causal, directional link between molecular rates and speciation rates. We test this link using mitochondrial and nuclear genes in mammals, and fail to report an association for this group. These results, at odds with patterns reported for other taxonomic groups, indicate that the processes driving speciation may vary considerably in their influence between taxa, and further have implications for the universality of the evolutionary speed hypothesis, as it pertains to global mammalian diversity patterns

Mike Hendy
Institute: Massy University, New Zealand.

Radical phylogenetic inversion

Given a stochastic model of nucleotide substitution on a rooted phylogeny T on n taxa, and a nucleotide at the root, the probabilities P_A of each of the 4^n possible combination of nucleotides at the leaves can be calculated. These probabilities can be expressed as polynomial functions of the probability parameters of the stochastic model on each edge of T. In the case of the Jukes Cantor model, and the (generalised) Kimura 2ST and 3ST models, these probabilities can be expressed using Hadamard conjugations.

A rooted phylogeny T on n taxa has $2n-2$ edges. In the case of the gK3ST model, each edge e has 3 independent probabilities $p_\alpha(e), p_\beta(e), p_\gamma(e)$. Hence given A, any of the 4^n combinations of nucleotides at the leaves, there is a probability P_A which is a polynomial of these $6n-6$ variables. These polynomials are solvable by radicals in the sense that each of the edge parameters $p_\theta(e)$ can be explicitly expressed as a rational polynomial with radicals of the 4^n probabilities P_A.

This talk will show how these polynomials are solved, and why we cannot expect a similar relationship for other stochastic models of nucleotide substitution.

Simon Ho
Institute: University of Sydney, Australia.

Time-dependent rates of molecular evolution: Evidence and causes

More than half a century ago, Kurtén (1959) observed an inverse correlation between the rate of morphological evolution and the time interval over which the rate was measured. Specifically, the rate of morphological change between successive generations was found to exceed macroevolutionary rates by several orders of magnitude. A similar phenomenon has recently emerged in analyses of molecular sequence data, most noticeably when spontaneous mutation rates, measured across a small number of generations, are compared with the much lower rates of evolutionary change measured over geological timeframes.

The striking disparity between short- and long-term rates has important consequences because molecular data are often analysed using molecular-clock methods to estimate evolutionary timescales. If rates depend on the timescale over which they are estimated, it means that the molecular clock cannot be assumed to be constant across timescales. To address this problem, it

is necessary to have a thorough appreciation of the causes of elevated short-term rates. However, a comprehensive characterisation of time-dependence remains elusive because it is difficult to obtain reliable estimates of rates on different timescales.

There are many factors that can lead to elevated short-term rates, but the relative importance of these will vary among taxa. In this talk I will provide an overview of the biological and methodological factors that cause the time-dependence of molecular evolutionary rates. I also survey the published evidence for time-dependent rates from studies of Metazoa and viruses.

Steffen Klaere
Institute: University of Auckland, New Zealand.

The link between segregation and phylogenetic diversity
Joint work with David Bryant.

We derive an invertible transform linking two widely used measures of species diversity: phylogenetic diversity and the expected proportions of segregating (non-constant) sites. We assume a bi-allelic, symmetric, finite site model of substitution. Like the Hadamard transform of Hendy and Penny, the transform can be expressed completely independent of the underlying phylogeny. Our results bridge work on diversity from two quite distinct scientific communities.

Rob Lanfear
Institute: Australian National University (ANU), Australia.

Using molecular phylogenies to estimate the timing of species addition to communities

Communities of species are formed by a combination of immigration, extinction, and diversification. There is a great deal of interest in describing the relative importance of these processes in community formation, and some information about this can be gleaned from molecular phylogenies. We are developing methods which use molecular phylogenies to describe the mode and timing of arrival of new lineages into a community. These methods describe the rates of in-situ diversification and colonisation of lineages into a given community, while accounting for uncertainty in the topology of the tree, the dates of the nodes, and the ancestral biogeographical reconstructions. These methods take account of the fact that molecular phylogenies are dealing with only extant taxa, and will allow the comparison of the arrival times of the extant lineages with a range of null models of community formation.

Jessica Leigh
Institute: University of Otago, New Zealand.

Simulation-based testing in an approximate Bayesian framework

Statistical methods applied to many areas of the sciences are often assessed and marketed using simulation-based performance evaluation. This sort of framework can involve repeated simulation over a large number of combinations of values for relevant parameters, or the selection of a few "pet" parameter values. While the former approach to parameter selection can be inefficient and unwieldy, the second is far from objective and potentially dishonest. We have developed a Markov chain Monte Carlo sampling method to identify regions of parameter space where methods per-

form either well or poorly. Our method is similar to Approximate Bayesian Computation in that it does not involve the calculation of likelihoods, but samples from the probability distribution of interest, rather than an approximation thereof. In addition to describing our method, I will present results from its application to such diverse subject areas as population genetics and public health.

Denise Kühnert
Institute: University of Auckland, New Zealand.

Population dynamics in phylogeographic analyse

Phylogenetic methods for the analysis of infectious diseases advance quickly. Recent improvements include phylogeographic reconstruction of pathogens not only based on sampling dates but also sampling locations. However, apart from the growing amount of genetic data challenging the capabilities of phylogenetic methods there is plenty of epidemiological data available which should also be taken into account when transmission pathways are reconstructed. Recently, the first steps have been taken to reconcile phylogenetic methods with epidemiological approaches. For example, coalescent-based methods have been extended to allow for dynamics within the underlying population, epidemiological analyses have been endorsed by parameter estimates from genetic data.

We simulate viral epidemics based on a stochastic SIR model equipped with empirical estimates of Influenza viruses and reconstruct their phylogeographic history with a popular phylogeographic method. Thereby showing the abilities and boundaries of current methods, this guides us towards new joint methods for the analysis of genetic and case data.

Matt Phillips
Institute: Australian National University (ANU), Australia.

Marsupial herbivore evolution and the failure of algorithmic morphological phylogenetics

Molecular data has attained primacy in phylogenetic inference for extant organisms. However, the fossil record (hence morphology) usually provides the only direct evidence for inferring macroevolution and for calibrating molecular timescales.

Morphological phylogenetic analysis falls broadly into two sets of approaches. One is "informal", emphasizing functional and developmental influences on morphology and informally weighting similarities between taxa; the other is "algorithmic", applying explicit selection criteria, such as maximum parsimony or maximum likelihood for comparing phylogenetic hypotheses. I will show that informal morphology and molecular methods have provided comparable phylogenetic estimates for the most ecologically diverse mammalian order, Diprotodontia (kangaroos, koala, wombats, possums). In contrast, algorithmic approaches have been widely inaccurate. Despite this inaccuracy, algorithmic approaches now dominate morphological phylogeny, largely due to their transparency and potential for statistical rigor. Hence, it is important to understand their present failure. Here I show that much of the inaccuracy can be explained by ecological correlations over-riding phylogenetic signals.

Alethea Rea
Institute: University of Auckland, New Zealand.

Statistical phylogenetics with Nets and a Lasso

John Rhodes
Institute: University of Alaska, Fairbanks, USA.

Semi-algebraic descriptions of the general Markov model
Joint work with E. Allman and A. Taylor.

Phylogenetic invariants for a particular tree and model of sequence evolution are multivariate polynomials satisfied by all probability distributions produced by that model and tree, regardless of parameter values such as edge lengths or mutation rates. However, the zero set of all such invariants is typically larger than the set of distributions arising from the model. There can be probability distributions satisfying all invariants which arise either only from biologically-meaningless complex values of parameters, or from no parameters at all.

An exact characterization of the probability distributions arising from a model and tree require a semi-algebraic description, using polynomial inequalities in addition to invariants. In our work we show how many inequalities can be constructed for a k-state general Markov model through Sylvester's criterion for quadratic forms. In the case of $k = 2$, by also using the hyperdeterminant (a.k.a the tangle) we obtain an exact semi-algebraic description of the model. For larger k, parts of the description are still missing. Although recent work of Zwiernik and Smith in part motivated this study, our approach is different, and overcomes their restriction to 2-state models.

Charles Semple
Institute: University of Canterbury, New Zealand.

Submodular functions and biodiversity conservation
Joint work with Magnus Bordewich.

One of the fascinations of mathematics is that common notions frequently arise in an unexpected settings. Recognising such occurrences and then exploiting the theory of the common notion, often leads to elegant arguments and results. In this talk, we describe such a fascination in the context of biodiversity conservation, where the common notion arising is submodular functions.

Jeremy Sumner
Institute: University of Tasmania, Australia.

A Lie algebraic classification of continuous-time Markov models
Joint work with Jésus Fernández-Sánchez and Peter Jarvis.

In recent work we have discussed the importance of *multiplicative closure* for the Markov models used in phylogenetics. For continuous-time Markov chains, a sufficient condition for multiplicative closure of a model class is ensured by demanding that the set of rate matrices belonging to the model class form a *Lie algebra*. It is the case that some Markov models *do* form Lie algebras (eg.

JC, F81, K3ST and GMM), and we refer to these models as "Lie Markov models". Importantly, other Markov models unequivocally *do not* form Lie algebras (the most conspicuous example being GTR).

The question then naturally arises: How do we generate a full list of Lie Markov models? To answer this question in full generality seems quite difficult as the interaction between the operations of a Lie algebra and the stochastic requirements of rate matrices are somewhat at odds, and it seems that this question has not been addressed in the algebraic or stochastic mathematics literature.

In this talk, we will discuss how we have made signficant progress in generating Lie Markov models by demanding that the models have certain *symmetries* under nucleotide permutations. From a theoretical prespective, we show that the Lie Markov models include, and hence provide a unifying concept for, "group-based" and "equivariant" models, whilst also generating many other interesting examples. We also argue that our scheme is very pleasing in the context of applied phylogenetics, as, for a given symmetry of nucelotide substitution, it provides a natural hierarchy of models with increasing number of parameters.

List of participants

Elizabeth Allman
Institute: University of Alaska (Fairbanks), USA
e.allman@uaf.edu

Anna Brüniche-Olsen
Institute: University of Tasmania, Australia
annabo@utas.edu.au

David Bryant
Institute: University of Otago, New Zealand
david.bryant@otago.ac.nz

Chris Burridge
Institute: University of Tasmania, Australia
Chris.Burridge@utas.edu.au

Michael Charleston
Institute: University of Sydney, Australia
m.charleston@usyd.edu.au

Josh Collins
Institute: University of Canterbury, New Zealand
j.collins@math.canterbury.ac.nz

James Degnan
Institute: University of Canterbury, New Zealand
J.Degnan@math.canterbury.ac.nz

Bjarki Jónsson Eldon
Institute: Oxford University, United Kingdom
eldon@stats.ox.ac.uk

Lynette Forster
Institute: CSIRO CRC Tasmania, Australia
Lynette.Forster@csiro.au

Xavier Goldie
Institute: Australian National University (ANU), Australia
xavier.goldie@anu.edu.au

Mike Hendy
Institute: Massey University, New Zealand
b.r.holland@massey.ac.nz

Simon Ho
Institute: University Of Sydney, Australia
simon.ho@sydney.edu.au

Barbara Holland
Institute: University of Tasmania, Australia
b.r.holland@massey.ac.nz

Peter Jarvis
Institute: University of Tasmania, Australia
Peter.Jarvis@utas.edu.au

Bodie Kaine
Institute: University of Tasmania, Australia
btkaine@postoffice.utas.edu.au

Steffen Klaere
Institute: University of Auckland, New Zealand
steffen.klaere@gmail.com

Denise Kuhnert
Institute: University of Auckland, New Zealand
dkuh004@aucklanduni.ac.nz

Rob Lanfear
Institute: Australian National University (ANU), Australia
rob.lanfear@gmail.com

Jessica Leigh
Institute: University of Otago, New Zealand
jessica.w.leigh@gmail.comnewline

Jonathan Mitchell
Institute: University of Tasmania, Australia
jm06@utas.edu.au

Malgorzata O'Reilly
Institute: University of Tasmania, Australia
marzenao@postoffice.utas.edu.au

Michael Ott
Institute: CSIRO Tasmania, Australia
michael.ott@csiro.au

Matt Phillips
Institute: Australian National University (ANU), Australia
matt.phillips@anu.edu.au

Alethea Rea
Institute: University of Auckland, New Zealand
alethea.rea@gmail.com

John Rhodes
Institute: University of Alaska (Fairbanks), USA
j.rhodes@uaf.edu

James Rowbottom
Institute: University of Tasmania, Australia
james.f.rowbottom@gmail.com

Charles Semple
Institute: University of Canterbury, New Zealand
charles.semple@canterbury.ac.nz

Jeremy Sumner
Institute: University of Tasmania
jsumner@utas.edu.au

Thanks for coming and watch out for Phylomania 2011!