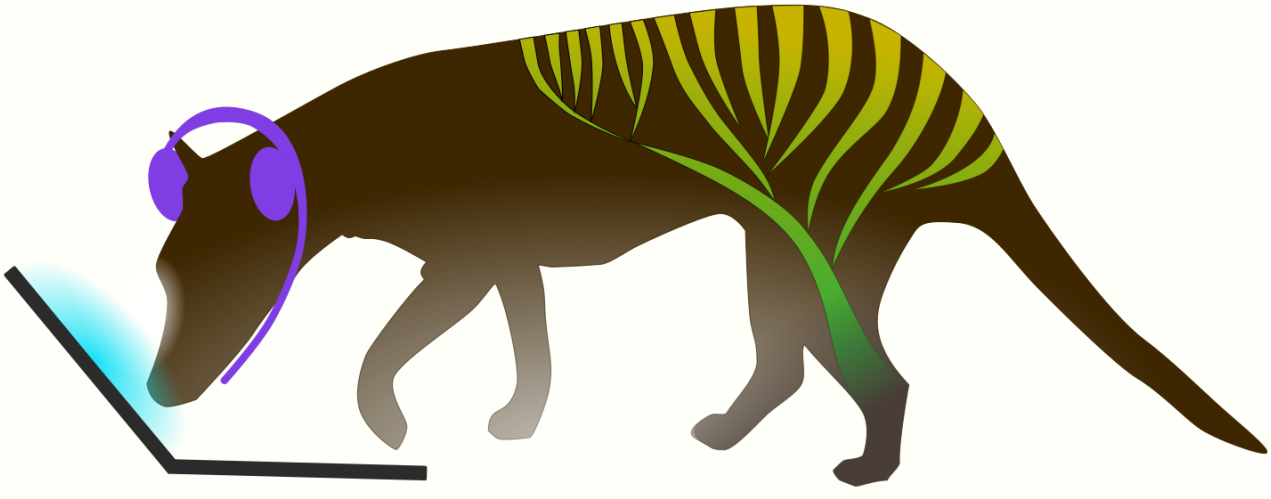
Phylomania 2020
Undeterred and on the web!

All over the world
November 25-27
2020



Venue

For those of you who are attending in person, the room is Social Sciences 209 (SB.AX17.L02.209 - Video Conf (SB.SocSci209))

The video links will be on the Phylomania web page at
https://www.maths.utas.edu.au/phylomania/phylomania2020.html.

Session Guide

Session	Time in Hobart	Theme
Session 1	09:00-10:30 Wed	Discrete methods
Session 2	11:00-13:00 Wed	Models and Inference
Session 3	19:00-21:00 Wed	Methods and Applications I
Session 4	08:00-09:45 Thurs	Networks
Session 5	10:15-12:00 Thurs	Computational methods
Session 6	12:15-12:45 Thurs	POSTER session
Session 7	19:00-20:45 Thurs	Methods and Applications II
Session 8	08:00-09:45 Fri	Algebraic methods+
Session 9	10:15-12:00 Fri	Stochastic models of evolution

Programme				
Session	**Time (AED)**	**Presenter**[a]	**Title**	**Details**[b]
1	9:15	Yoshida	Tropical Principal Component Analysis	L30
1	9:45	Hendriksen	Incompatibility and Universal Tree Sets	L15
1	10:00	Charleston	Landscape gardening in split space	L15
1	10:15	Wicke	Formal Links between Feature Diversity and Phylogenetic Diversity	L15
2	11:00	Felsenstein	Morphometrics on phylogenies: a linear model alternative to the Morphometric Consensus	L45
2	11:45	Dinnage	Autodecoding Evolution: Exploring phylogenetic deep learning for ancestral reconstruction of traits with arbitrarily high dimensionality and complexity	L30
2	12:15	Chan	Inference under the coalescent with recombination	L15
2	12:30	Burden	Feller diffusions for critical and subcritical multi-type branching processes: work in progress	L30
3	19:00	Hoffmann	Bayesian phylodynamics reveal the diversification process of Vanuatu's languages	L30
3	19:30	Nahata∗	Bayesian model comparison of molecular clock models - a phylogenetic simulation study	L15
3	19:45	Liu∗	Is AIC An Appropriate Metric For Model Selection In Phylogenetics?	L15
3	20:00	Crotty	Heterotachy Mapping	L15
3	20:15	Buckley∗	Long-term climatic stability drives accumulation and maintenance of divergent lineages in a temperate biodiversity hotspot	L15
3	20:30	Mitchell	To BIC or not to BIC? Model selection in phylogenomics	P30
3	20:30	Fountain-Jones	Phylodynamics and COVID19	P30
4	8:00	Banos	Identifiability of species networks topologies from genomic sequences using the logDet distance	L30
4	8:45	Francis	The case for normal phylogenetic networks	L30
4	8:30	Bryant	Preconditioning NeighborNet	L15
4	9:15	Huber	Seeing trees in networks	P15
4	9:15	Moulton	Reconstructibility of unrooted level-k phylogenetic networks from distances	P15
4	9:15	Murakami∗	On Cherry-Picking and Network Containment	P30
5	10:15	Matsen	Variational Bayesian phylogenetic inference: where we've been and where we'd like to go	L30
5	10:45	Karcher	Variational Bayesian supertree methods	L30
5	11:15	Beeton	The evolution of PASSÉ	L15
5	11:30	Fourment	Turbocharging variational inference in phylogenetics	P15
5	11:30	Jun	Generalized Pruning	P30
6		Soewongsono∗	The Shape of Phylogenies Under Phase-Type Distributed Times to Speciation and Extinction	poster
6		da Silva∗		poster
6		Jaya∗	Evaluation of recombination detection methods for viral sequence analysis	poster
6		Fisk∗	Phylogenetic Experimental Design via Signal-Noise Framework	poster
6		Jun	Generalized Pruning	poster
7	19:00	von Haeseler	The dynamics of cell division and differentiation in cerebral organoids	L30
7	19:30	Lueg∗	Information geometry for phylogenetic trees	L30
7	20:00	Serra Silva∗	The Effects of Tree Islands on Consensus	L15
7	20:15	Fischer	Refinement stable consensus methods	P30
7	20:15	Smith	Beyond Robinson-Foulds: information-theoretic tree distance metrics	P15
7	20:15	Chernomor	Generating and sampling trees from a phylogenetic terrace	P15
8	8:00	Ardiyansyah∗	Model embeddability for symmetric group-based model	L15
8	8:15	Manuel∗	Analysis of the matrix exponential can guide maximum likelihood-based phylogenetic inference	L30
8	8:45	Stevenson∗	The Hyperoctahedral Group	L30
8	9:15	Shore∗	Phylo-symmetric algebras: mathematical properties of a new tool in phylogenetics	L15
8	9:30	Sumner	Uniformization-stable Markov models	P30
8	9:30	Terauds	Irreducible semigroup-based Markov models	P15
9	10:15	O'Reilly	Matrix analytic methods for the gene-tree species-tree reconciliation problem	L30
9	10:45	Burridge	Does migration promote or inhibit diversification? A case study involving the dominant radiation of temperate Southern Hemisphere freshwater fishes	L15
9	11:00	Soewongsono∗	The Shape of Phylogenies Under Phase-Type Distributed Times to Speciation and Extinction	L15
9	11:15	Diao∗	A subfunctionalization model of gene family evolution predicts balanced tree shapes	L15
9	11:30	Stark	Detecting selection acting on recently duplicated genes.	L15

[a] $*$ $\Rightarrow$ Student; [b] P $\Rightarrow$Pre-recorded; L $\Rightarrow$Live; Talks are 15, 30 or 45 minutes.

Abstracts

Tropical Principal Component Analysis

Ruriko Yoshida, Naval Postgraduate School
ryoshida@nps.edu

Session 1: 9:15AED

Principal component analysis is a widely-used method for the dimensionality reduction of a given data set in a high-dimensional Euclidean space. Here we define tropical principal component analysis (PCA) using the tropical polytope with a fixed number of vertices closest to the data points. We here apply tropical PCA to dimension reduction and visualization of data sampled from the space of phylogenetic trees. Our main results are twofold: the existence of a tropical cell decomposition into regions of fixed tree topology and the development of a stochastic optimization method to estimate the tropical PCA using a Markov Chain Monte Carlo (MCMC) approach. This method performs well with simulation studies, and it is applied to three empirical datasets: Apicomplexa and African coelacanth genomes as well as sequences of hemagglutinin for influenza from New York.

Incompatibility and Universal Tree Sets

Michael Hendriksen, Heinrich-Heine University
michael.hendriksen@hhu.de

Session 1: 9:45AED

(Joint work with Nils Kapust)

We consider the problem of how many phylogenetic trees it would take to display all splits in a given set, a problem related to k-compatibility. A set of trees that display every single possible split of a given set of taxa is termed a universal tree set. In this talk we will find the universal incompatibility function $U(n)$, the minimal size of a universal tree set for n taxa. We will also summarise the proof of this theorem, which appeals to several classical theorems on posets.

Landscape gardening in split space

Michael Charleston, University of Tasmania
michael.charleston@utas.edu.au

Session 1: 10:00AED

Phylogenetic *split space* is a kind of graph whose vertices are splits — potential branches in an unrooted phylogenetic tree, and whose edges join splits when they are adjacent by some rule.
For n taxa there are $2^{(n-1)} - n - 1$ non-trivial splits; exponentially large, but still much less than the number of possible unrooted binary trees ($(2n-5)!! = (2n-5)(2n-7)...(3)(1)$). The splits can also be *weighted*, e.g., by some measure of how well they are supported in a phylogenetic data set. A sophisticated method called *sub-flattening* (Sumner 2017, Bull. Math. Biol.) provides a score for splits in terms of an error estimation via singular value decomposition, and there are some obvious simplistic measures too. If good splits can quickly be found then we might be able to divide-and-conquer, and thus provide yet another new way to try to address the computational problem of finding the best tree(s) in a reasonable amount of time. In order to determine whether splits found in this way are "good", and can be found quickly, requires we understand the landscape of split space: how large are domains of attraction; do "good" splits correspond to those in the true tree, etc.
I will present new experimental findings that may be of interest to the community, describing the utility of the split scores for phylogenetic estimation and some of the landscape characteristics of split space, under three different scoring schemes: sub-flattenings, a very naïve parsimony-like method, and base frequencies. It may even become clear whether this is a good idea or not.

Formal Links between Feature Diversity and Phylogenetic Diversity

Kristina Wicke, The Ohio State University, USA
kristina.wicke@gmail.com

Session 1: 10:15AED

(Joint work with Arne Mooers (Simon Fraser University, Canada), Mike Steel (University of Canterbury, New Zealand))

The extent to which phylogenetic diversity (PD) captures feature diversity (FD) is a topical and controversial question in biodiversity conservation. In this talk, we formalize this question and establish a precise mathematical condition for FD (based on discrete characters) to coincide with PD. In this way, we make explicit the two main reasons why the two diversity measures might disagree for given data; namely, the presence of certain patterns of feature evolution and loss, and using temporal branch lengths for PD in settings that may not be appropriate. We also explore the relationship between the 'Fair Proportion' index of PD and a simple index of FD. We show that the two indices can take identical values for any phylogenetic tree, provided the branch lengths in the tree are chosen appropriately.

Morphometrics on phylogenies: a linear model alternative to the Morphometric Consensus

Joe Felsenstein, Department of Genome Sciences and Department of Biology, University of Washington, Seattle
joe@gs.washington.edu

Session 2: 11:00AED

(Joint work with Fred Bookstein, Department of Statistics, University of Washington, Seattle)

A linear model for morphometrics is briefly presented, which is an alternative to the Generalized Procrustes superpositions used in the standard Morphometric Consensus. In simple cases it makes a superposition of forms which is equivalent to a least-squares Boas superposition, one which is like a Procrustes superposition but does not change size. It can also accommodate more sophisticated models in which the rotation of specimens is treated as uncertain. The computations are a relatively straightforward extension to inference of within- and between-species covariation when multiple specimens of species evolve on a known phylogeny. The framework infers character covariation of both kinds, and can accommodate size variation and allometry. The machinery of the Morphometric Consensus is not needed, so the user is not comforted by the thought that something mysterious, incomprehensible, and transcendent has been done.

Autodecoding Evolution: Exploring phylogenetic deep learning for ancestral reconstruction of traits with arbitrarily high dimensionality and complexity

Russell Dinnage, Institute for Applied Ecology, University of Canberra
russell.dinnage@canberra.edu.au

Session 2: 11:45AED

Deep learning methods have great potential as tools in the biological sciences, including phylogenetics. In this presentation I explore the world of deep generative models, techniques that attempt to learn how to generate complex data from some simpler underlying latent representation (a manifold). I present a method based on regularised evolutionary rate parameters that allows deep generative models to incorporate phylogenetic information into their latent space, which lets them, once trained on large real-world datasets, to generate putative reconstructions of complex biological data at internal nodes of the phylogeny. This framework can be applied to any type of complex biological data, as long as there exists a deep learning method to generate that type of data from a latent vector (generally known as a decoder). I demonstrate the ideas by developing a phylogenetic variational auto-decoder model, and apply it to several different complex datasets that have up to tens of thousands of species, and that also have associated phylogenies. This includes a set of images of animal silhouettes, a set of 3d scans of bird beaks, and a set of species range boundaries for Australia reptiles. These methods allow biologists to make predictions about ancestral characters based on complex statistical structure extracted by these models on large datasets. I discuss whether these predictions are likely to be reliable, and what we might be able to do with them (and if nothing else, what we can learn from them), in the context of machine-human collaboration. In the context of computer science this work provides a way to do "phylogenetic interpolation" within a model's latent space, which, for biological data, is an improvement over simple linear interpolation, the current standard for exploring the output of deep generative models.

Inference under the coalescent with recombination

Yao-ban Chan, School of Mathematics and Statistics / Melbourne Integrative Genomics, University of Melbourne
yaoban@unimelb.edu.au

Session 2: 12:15AED

(Joint work with Ali Mahmoudi, School of Mathematics and Statistics / Melbourne Integrative Genomics, University of Melbourne; David Balding, School of Mathematics and Statistics / School of BioSciences / Melbourne Integrative Genomics, University of Melbourne)

Inferring the genealogical history, known as the ancestral recombination graph (ARG), of a sample of DNA sequences is a long-standing problem in population genetics. Existing methods, such as the state-of-the-art ARGweaver, are limited to a simplified approximation of the underlying model of the coalescent with recombination. In this talk, we utilise a recently developed data structure for recording evolutionary history, the tree sequence, which offers orders-of-magnitude efficiency gains in storing and simulating DNA sequences. This enables us to develop a novel Markov Chain Monte Carlo algorithm to perform probabilistic inference under the full coalescent with recombination for the first time. Our results show improved accuracy over ARGWeaver in terms of inferring recombination rates and other ARG features.

Feller diffusions for critical and subcritical multi-type branching processes: work in progress

Conrad Burden, Mathematical Sciences Institute, Australian National University
conrad.burden@anu.edu.au

Session 2: 12:30AED

(Joint work with Robert Griffiths, Monash University)

Branching processes in which each individual in a population gives birth to a random number of offspring in each generation have been studied extensively since they were first introduced first by Bienaymé and independently by Galton and Watson in the 19th century. They have the property that sub-critical or critical processes, i.e. those for which the expected number of offspring is < 1 or $= 1$ respectively, the population will eventually go extinct with probability 1. Nevertheless, one can still study the distributional properties of such populations conditional on survival: If a population does survive, what does it look like? In particular, we consider subcritical multitype branching processes in which individuals within the population can mutate between different allele types. For this problem we are seeking to categorise the quasi-stationary alllele distribution allele distribution in a declining but surviving population.

Bayesian phylodynamics reveal the diversification process of Vanuatu's languages

Konstantin Hoffmann, Max Planck Institute for the Science of Human History
hoffmann@shh.mpg.de

Session 3: 19:00AED

(Joint work with Mary Walworth, Simon J. Greenhill, Aviva Shimelman, Russell D. Gray, Denise Kühnert)

With more than 120 languages[1] Vanuatu has more languages per capita than anywhere else in the world. This diversity may have been shaped by both language-internal and sociological factors[2], as well as external factors such as disease outbreaks or volcanic eruptions[3] and multiple waves of colonization[4]. These factors suggest that, since the arrival of its first inhabitants, there have been varying periods of population booms as well as periods of substantial decreases.

The Austronesian Basic Vocabulary Database provides 210-item wordlists for over 1500 languages spoken throughout the Pacific region, and includes data for 330 linguistic varieties from Vanuatu. We have recently revised the cognate coding for all the Oceanic languages following the comparative method in consultation with regional experts and published sources[5,6].

We use Bayesian phylodynamic methods to get insights into the diversification history of Vanuatu's languages. Focusing on birth-death models, this raises two methodological challenges: First, the dense language-dialect continuum makes it hard to establish a clear border between languages and thus obscures the otherwise well known sampling proportion. Second, the lack of knowledge about languages that are spoken in the past makes it impossible to jointly infer birth and death rates from the tree consisting solely of extant languages[7].

However, a combined approach using a multi-state birth death model[8] and estimates of pulled diversification rates[8], enables us to detect local variation estimating lineage-specific rates, with confidence about the stability of the overall rates assuming a diversified sampling scenario.

We come to the surprising conclusion that neither of the drastic events that affected Vanuatu had a strong macroevolutionary effect on the island nation's language evolution. Contrarily, they appear to have caused a linguistic resilience through which Vanuatu emerged as the "Galapagos of language evolution".

[1] Harald Hammarström, Robert Forkel, and Martin Haspelmath. *Glottolog 3.3.* (Available online at http://glottolog.org, Accessed on 2020-10-20.), 2018.

[2] Alexandre François. *The dynamics of linguistic diversity: Egalitarian multilingualism and power imbalance among northern Vanuatu languages*. International Journal of the Sociology of Language, (214):85-110, January 2012.

[3] Andrew Hoffmann. *Looking to Epi: Further consequences of the Kuwae eruption, Central Vanuatu, ad 1452*. Indo-Pacific Prehistory Association Bulletin, 26:62–71, 2006.

[4] Cosimo Posth, Kathrin Nägele, Heidi Colleran, Frédérique Valentin, Stuart Bedford, Kaitip W. Kami, Richard Shing, Hallie Buckley, Rebecca Kinaston, Mary Walworth, Geoffrey R. Clark, Christian Reepmeyer, James Flexner, Tamara Maric, Johannes Moser, Julia Gresky, Lawrence Kiko, Kathryn J. Robson, Kathryn Auckland, Stephen J. Oppenheimer, Adrian V. S. Hill, Alexander J. Mentzer, Jana Zech, Fiona Petchey, Patrick Roberts, Choong-won Jeong, Russell D. Gray, Johannes Krause, and Adam Powell. *Language continuity despite population replacement in remote oceania*. Nature Ecology & Evolution, 2(4):731-740, 2018.

[5] Ross Clark. **Leo Tuai: A Comparative Lexical Study of North and Central Vanuatu Languages*. Pacific Linguistics, Canberra, 2009.

[6] John Lynch. *Malakula internal subgrouping: Phonological evidence*. Oceanic Linguistics, 55:399–431, 2016.

[7] Stilianos Louca and Matthew W. Pennell. *Extant timetrees are consistent with a myriad of diversification histories*. Nature, 580(7804):502–505, apr 2020.

[8] Joëlle Barido-Sottani, Timothy G Vaughan, and Tanja Stadler. *A multi-type birth-death model for Bayesian inference of lineage-specific birth and death rates*. Systematic Biology, feb 2020.

Bayesian model comparison of molecular clock models - a phylogenetic simulation study (Student presentation)

Kanika Nahata, Evolutionary and Computational Virology, Rega Institute, KU Leuven
knahata15@gmail.com

Session 3: 19:30AED

(Joint work with Guy Baele (KU Leuven), Mandev Gill (KU Leuven))

In the 1960s, several groups of scientists - including Emile Zuckerkandl, Linus Pauling and Allan Wilson - had noted that proteins experience amino acid replacements at a surprisingly consistent rate across very different species. Since the proposal of such a (strict) clock model, a wide range of different clock model parameterizations have emerged which now take up a prominent place in the field of phylogenetic inference as well as in many other areas of evolutionary biology. In studying pathogen evolution, molecular clocks allow combining the genetic differences between samples and their collection times to estimate time-calibrated phylogenies. Along with the development of increasingly complex clock models comes the need to accurately determine which model is best suited to analyse a particular data set. For this purpose, different marginal likelihood estimators have been developed in recent years to compare relative model fit in a Bayesian framework. These estimators have shown considerable improvements in accuracy, but often at the expense of an increased computational cost. In our simulation study, we examine the performance of these estimators in identifying the correct underlying molecular clock model. The performance of these estimators is also tested in a scenario where they are faced with overparameterization using specific local clock models. Finally, we evaluate clock model performance on various empirical data sets, including pathogen and yeast examples.

Is AIC An Appropriate Metric For Model Selection In Phylogenetics? (Student presentation)

Qin Liu, School of Natrual Sciences, University of Tasmania
qin.liu@utas.edu.au

Session 3: 19:45AED

Akaike Information Criterion (AIC) is a popular tool for model selection in molecular phylogenetics. Under some conditions, AIC is an asymptotically unbiased estimator of the relative expected Kullback-Leibler Divergence (KLD). The values of AIC for a candidate model is calculated as two times the difference of the number of free parameters and the maximum likelihood of the model given the data. In phylogenetics, a couple of recent papers have questioned the effectiveness of AIC in phylogenetics, but there has not been any investigation of mixture or partition models. We are interested in whether AIC is an appropriate tool to compare models, in particular, to compare a partition model and a mixture model in phylogenetics. In "Ghost: recovering historical signal from heterotachously evolved sequence alignments", the authors showed that by comparison with a partition model, the mixture model recovered the correct phylogeny. However, the partition model seemed to have AIC values that were either as low as, or lower than, the AIC values of the mixture model.

One area of our research focuses on evaluating AIC between a partition and a mixture model. In the simulations, we fitted a partition model and a mixture model into the simulated data sets. We evaluated the performance of AIC by using the relative expected KLD, and we compared the best model chosen by AIC with the ones chosen by other criterion, including the Robinson-Foulds distance and the Branch Score. The results show that the best models chosen by these criteria are not always the same. I am going to show some of the preliminary results in my talk.

Heterotachy Mapping

Stephen Crotty, School of Mathematical Sciences, University of Adelaide
stephen.crotty@adelaide.edu.au

Session 3: 20:00AED

Topological inference is often seen as the primary goal in phylogenetics. However, there is much more insight to be gained from large phylogenomic alignments. Such alignments are guaranteed to contain substantial heterotachy, which can be addressed with the use of a mixed branch length model, such as GHOST. We show that phylogenetic signals of biological interest can be mapped back onto the alignment, highlighting sites/genes/regions of potential functional interest.

Long-term climatic stability drives accumulation and maintenance of divergent lineages in a temperate biodiversity hotspot (Student presentation)

Sean Buckley, Molecular Ecology Laboratory, College of Science and Engineering, Flinders University
sean.buckley@flinders.edu.au

Session 3: 20:15AED

(Joint work with Peter Unmack, Institute for Applied Ecology, University of Canberra Michael Hammer, Natural Sciences, Museum and Art Gallery of the Northern Territory; Jonathon Sandoval-Castillo, Molecular Ecology Laboratory, College of Science and Engineering, Flinders University; Luciano Beheregaray, Molecular Ecology Laboratory, College of Science and Engineering, Flinders University)

Anthropogenic climate change is likely to drive regional climate disruption and instability across the globe, and is likely to be exacerbated within biodiversity hotspots. Here, we assessed the role of climatic stability on the accumulation and persistence of phylogenetic lineages in a freshwater fish group (genus Nannoperca) endemic to the southwest Western Australia (SWWA) biodiversity hotspot. Using a combination of genomic and environmental approaches, we investigated population divergence, phylogenetic relationships, species delimitation and ecological niche modelling in SWWA pygmy perches. We identified divergent lineages, including at least two cryptic species, that showed long-term patterns of isolation and persistence. We also developed a novel approach that estimates functional enrichment of diagnostic loci between putative species. This approach pointed to reproductive isolation between putative species potentially linked to genes associated with chromosomal segregation, cytokinetic functions and metabolic processes. Together, these findings suggest a revision of taxonomic and conservation status within N. vittata is required to adequately maintain unique evolutionary lineages into the future. Our results demonstrated strong divergence within the group and highly stable modelled distributions, highlighting the role of climatic stability in allowing the persistence of isolated lineages in SWWA. This biodiversity hotspot is under compounding threat from recent and ongoing climate change, and habitat modification, which may further threaten previously undetected cryptic diversity across the region.

To BIC or not to BIC? Model selection in phylogenomics

Jonathan Mitchell, University of Alaska Fairbanks
jonathanmitchell88@gmail.com

Session 3: 20:30AED

(Joint work with Elizabeth Allman - University of Alaska Fairbanks, John Rhodes - University of Alaska Fairbanks)

The AIC and BIC are commonly used model selection criteria in phylogenetics and phylogenomics. Both criteria can be interpreted as the $(-2\log)$ maximum likelihood with a penalty to avoid overfitting. The penalty terms depend on the geometry of the model. Standard derivations of these criteria assume regularity conditions that are often not met, including in phylogenetic and phylogenomic models, for example, near points where species relationships are hard to determine. Examples include models with non-negative parameters, such as edge length or substitution rate parameters. We generalize the AIC and BIC to models with singularities and boundaries in their parameter spaces. We illustrate the use of these generalized criteria on models involving the multispecies coalescent. Future work is to incorporate these model selection criteria into the MSCquartets R package, which infers species trees and networks under the multispecies coalescent model and tests for model fit.

Phylodynamics and COVID19

Nick Fountain-Jones,
nick.fountainjones@utas.edu.au

Session 3: 20:30AED

(Joint work with Raima Appaw, Scott Carver, Xavier Didelot, Erik M Volz, and Michael Charleston)

Since spilling over into humans, SARS-CoV-2 has rapidly spread across the globe, accumulating significant genetic diversity. The structure of this genetic diversity, and whether it reveals epidemiological insights, are fundamental questions for understanding the evolutionary trajectory of this virus. Here we use a recently developed phylodynamic approach to uncover phylogenetic structures underlying the SARS-CoV-2 pandemic. We find support for three SARS-CoV-2 lineages co-circulating, each with significantly different demographic dynamics concordant with known epidemiological factors. For example, Lineage C emerged in Europe with a high growth rate in late February, just prior to the exponential increase in cases in several European countries. Non-synonymous mutations that characterize Lineage C occur in functionally important gene regions responsible for viral replication and cell entry. Even though Lineages A and B had distinct demographic patterns, they were much more difficult to distinguish. Continuous application of phylogenetic approaches to track the evolutionary epidemiology of SARS-CoV-2 lineages will be increasingly important to validate the efficacy of control efforts and monitor significant evolutionary events in the future.

Identifiability of species networks topologies from genomic sequences using the logDet distance

Hector Banos, Georgia Tech
hbassnos@gmail.com

Session 4: 8:00AED

(Joint work with John Rhodes (University of Alaska Fairbanks), Elizabeth Allman (University of Alaska Fairbanks))

It is known that hybridization plays an important role during the evolutionary process of some species. Therefore phylogenetic trees are sometimes insufficient to describe species-level relationships. We show that most topological features of a level-1 species network are identifiable under the network multi-species coalescent model (NMSC) using the log-det distance between aligned DNA sequences of concatenated genes.

The case for normal phylogenetic networks

Andrew Francis, Centre for Research in Mathematics and Data Science, Western Sydney University
a.francis@westernsydney.edu.au

Session 4: 8:45AED

Numerous classes of phylogenetic network have been defined and studied, with the goal of making the inference of complex evolutionary relationships a reality. In this talk I will present a case for why the class of "normal" networks has emerged as the most promising of these.

Preconditioning NeighborNet

David Bryant, University of Otago
david.bryant@otago.ac.nz

Session 4: 8:30AED

(Joint work with Daniel Huson)

NeighborNet is a method for constructing a network representation for data, developed twenty years ago but still fairly widely used. One stage of the method requires solution of a large, dense, quadratic programming problem, and it is this stage which has limited the application of NeighborNet to a few hundred taxa at best, despite a pretty concerted algorithmic effort. In this talk I'll report on success we have finally had with the use of preconditioned iterative methods, and what those are.

Seeing trees in networks

Katharina Huber, School of Computing Sciences, University of East Anglia, UK
k.huber@uea.ac.uk

Session 4: 9:15AED

A basic problem in phylogenetics is to find ways to combine potentially conflicting gene trees on a set X of species into a graph so that the trees can still be seen in that graph. Phylogenetic networks on X (i.e., certain rooted directed acyclic graphs with leaf set X which represent conflict between gene trees in terms of so-called reticulation vertices, that is, vertices of indegree two or more) have proven the framework of choice to address this problem as they allow one to formalize it in terms of understanding when a phylogenetic tree T is displayed by a phylogenetic network N, that is, T can be obtained from N by deleting for each reticulation vertex of N all but one incoming arcs so that the resulting graph is isomorphic to a subdivision of T. As it turns out, this formalization is too restrictive for some evolutionary processes. In this talk we outline recent progress that we have made to address this problem.

Reconstructibility of unrooted level-k phylogenetic networks from distances

Vincent Moulton, School of Computing Sciences, University of East Anglia
v.moulton@uea.ac.uk

Session 4: 9:15AED

(Joint work with Leo van Iersel, Yuki Murakami, TU Delft)

A phylogenetic network is a graph-theoretical tool that is used by biologists to represent the evolutionary history of a collection of species. One potential way of constructing such networks is via a distance-based approach, where one is asked to find a phylogenetic network that in some way represents a given distance matrix, which gives information on the evolutionary distances between present-day taxa. In this talk, we consider the following question. For which k are unrooted level-k networks uniquely determined by their distance matrices? We consider this question for shortest distances as well as for the case that the multisets of all distances is given.

On Cherry-Picking and Network Containment (Student presentation)

Yukihiro Murakami, TU Delft
yukimurakami07201994@gmail.com

Session 4: 9:15AED

(Joint work with Remie Janssen, TU Delft)

Phylogenetic networks are used to represent evolutionary scenarios in biology and linguistics. To find the most probable scenario, it may be necessary to compare candidate networks. In particular, one needs to distinguish different networks and determine whether one network is contained in another.
In this paper, we introduce cherry-picking networks, a class of networks that can be reduced by a so-called cherry-picking sequence. We then show how to compare such networks using their sequences.
We characterize reconstructible cherry-picking networks, which are the networks that are uniquely determined by the sequences that reduce them, making them distinguishable. Furthermore, we show that a cherry-picking network is contained in another cherry picking network if a sequence for the latter network reduces the former network, provided both networks can be reconstructed from their sequences in a similar way (i.e., they are in the same reconstructible class).
Lastly, we show that the converse of the above statement holds for tree-child networks, thereby showing that Network Containment, the problem of checking whether a network is contained in another, can be solved by computing cherry picking sequences in linear time for tree-child networks.

Variational Bayesian phylogenetic inference: where we've been and where we'd like to go

Frederick "Erick" Matsen, Computational Biology, Fred Hutchinson Cancer Research Center
ematsen@gmail.com

Session 5: 10:15AED

Variational inference (VI) enables Bayesian inference without sampling. It works by fitting an approximation to the posterior distribution, typically minimizing (via gradient descent) a divergence from the true posterior to the approximation. Although VI is a core technique in Bayesian machine learning, its application to Bayesian phylogenetic inference has until recently been hampered by the complexities of tree-valued distributions.
In this talk, I will give an intuitive introduction to VI, and give the foundations required to understand the work that is currently happening on variational Bayesian phylogenetic inference.
This includes several talks submitted to this conference:

- a 'generalized pruning' algorithm that marginalizes out subtree distributions for local gradient updates'
- a supertree method that works to minimize a divergence between a tree distribution on the entire taxon set and given distributions on taxon subsets;
- integration of phylogenetic tree distributions into TensorFlow Probability and PyTorch for flexible phylogenetic modeling and incorporation of covariates.

Variational Bayesian supertree methods

Michael D. Karcher, Fred Hutchinson Cancer Research Center
michael.d.karcher@gmail.com

Session 5: 10:45AED

Given overlapping sets of species, and posterior distributions on phylogenetic tree topologies for each of these taxon sets, how can we infer a posterior distribution on tree topologies for the combined taxon set? Although the equivalent problem for the non-Bayesian case has attracted substantial research, the Bayesian case has not received the attention it deserves. Here, we present a variational Bayesian approach to this problem and demonstrate its effectiveness.

The evolution of PASSÉ

Nick Beeton, CSIRO
nick.beeton@csiro.au

Session 5: 11:15AED

(Joint work with Barbara Holland (UTAS), Larry Forbes (UTAS), Stephen Walters (UTAS), Greg Jordan (UTAS))

Analysing phylogenetic trees using a Binary-State Speciation and Extinction (BiSSE) model is a useful way to test for the presence of trait-based selection in the evolutionary process. Recent advances (Louca and Pennell 2020) exploit the flow of the differential equations describing the likelihoods of such trees, making inference for large trees feasible, but are still slow to find maximum likelihood estimates (MLEs) of parameters. We introduce the Perturbation Approximation of State-based Speciation and Extinction (PASSÉ) model, using an analytic approximation to BiSSE to speed up numerical integration at the cost of introducing some error. Using a large published tree of the diversification of angiosperms (31,749 tips), we show that our method is at least 10x faster at calculating likelihoods and 100x faster at calculating MLEs, while remaining indistinguishable in quality from the exact method.

Turbocharging variational inference in phylogenetics

Mathieu Fourment, University of Technology Sydney
mathieu.fourment@uts.edu.au

Session 5: 11:30AED

Markov chain Monte Carlo algorithms have been the workhorse of Bayesian inference in phylogenetics for almost two decades. Although these algorithms have been successfully used in a wide range of applications they do not scale well to large numbers of sequences. Recent advances in statistical machine learning techniques have led to the creation of probabilistic programming frameworks. These frameworks enable probabilistic models to be rapidly prototyped and fit to data using scalable approximation methods such as variational inference. In this talk I will present some work on phylogenetic variational inference using phylostan, a Stan-based tool. Because trees are unusual statistical objects, phylostan tends to lack the flexibility required for accurately modeling molecular evolution. We can meet this challenge by using more general modeling frameworks such as Pytorch. This framework brings in an efficient machine learning toolbox in a popular programming language that can be easily extended. I will emphasize the current challenges in the new field of phylogenetic variational inference.

Generalized Pruning

Seong-Hwan Jun, Fred Hutchinson Cancer Research Center
sjun2@fredhutch.org

Session 5: 11:30AED

(Joint work with Hasan Nasif, Erick Matsen)

Bayesian phylogenetics is a computationally challenging inferential problem. Classical methods are based on random-walk Markov chain Monte Carlo (MCMC), where random proposals are made on the tree parameter and the continuous parameters simultaneously. The samples from MCMC allows one to quantify uncertainty around the split probabilities of a phylogeny as well as the branch length parameters.
In this paper, we describe a new algorithm that directly generalizes the Felsenstein pruning algorithm by marginalizing out ancestral states and subtrees simultaneously. Utilizing this algorithm, we propose an efficient approach based on composite likelihood method to estimate the parameters governing the phylogenetic tree.
Our experiments demonstrate the estimated parameters compare favorably to the ones obtained via MCMC at a fraction of the runtime.

[Poster] **The Shape of Phylogenies Under Phase-Type Distributed Times to Speciation and Extinction**
(Student presentation)

Albert Christian Soewongsono, School of Natural Sciences, University of Tasmania
albert.soewongsono@utas.edu.au

Session 6: AED

(Joint work with Barbara Holland, Małgorzata O'Reilly)

We consider a macroevolutionary model for phylogenetic trees where times to speciation or extinction events are drawn from a Coxian phase-type (PH) distribution. In this poster, we show that different choices of parameters in our model lead to a range of tree shapes as measured by Aldous' β statistic. Additionally, we show that tree balance is mainly controlled by speciation process while branch lengths are mostly controlled by extinction process.
Key words: Macro-evolutionary model; diversification; stochastic model; tree shape; phase-type distribution.

[Poster] (Student presentation)

Alexandre Bonfim Pinheiro da Silva, Federal University of Rio de Janeiro
alexandrebonfimpinheiro@gmail.com

Session 6: AED

[Poster] **Evaluation of recombination detection methods for viral sequence analysis** (Student presentation)
Frederick Jaya, University of Technology Sydney
fredjaya1@gmail.com
Session 6: AED

(Joint work with Aaron Darling (UTS), Barbara Brito-Rodriguez (UTS))

To accurately infer the evolutionary history of viral genomes, the process of recombination needs to be accounted for and addressed appropriately. A vast choice of recombination detection methods have been developed over the past 20 years, but their ability to address the needs presented by high-throughput sequencing of viral data is unclear.

Here, we present the key considerations for selecting a suitable method for viral analyses. We assess five published methods used to detect recombination in nucleotide sequences - PhiPack (Profile), 3SEQ, GENECONV, UCHIME and gmos. The performance of methods were evaluated with analysis of within-host hepatitis C virus populations, simulated across a wide range of mutation and recombination rates. Scalability was assessed by recording the CPU time required to analyse datasets with $n = 500$, $n = 1000$ and $n = 5000$ sequences per alignment (1680 nt). In addition, empirical datasets of two bovine RNA viruses were analysed by each method and compared with simulation findings.

We find critical trade-offs between the methods, where the most scalable methods (PhiPack (Profile), UCHIME and gmos) may not be suitable for analysis of high coverage, within-host sequencing. Analysis of highly similar sequences (mean pairwise diversity $< 1\%$) produced a high rate of positive detections in PhiPack (Profile), whereas 3SEQ and GENECONV are unable to process these. Overall, the five evaluated methods are inadequate for a rapid and reliable analysis of recombination in large viral datasets, presenting a severe unmet need for the development of scalable and accurate viral recombination detection methods.

[Poster] **Phylogenetic Experimental Design via Signal-Noise Framework** (Student presentation)
J. Nick Fisk, Computational Biology and Bioinformatics, Yale University
jeffrey.fisk@yale.edu
Session 6: AED

(Joint work with Alexander Dornburg, UNC Charlotte; Jeffrey Townsend, Yale University)

In the emergent big-data world of phylogenomics, it is clear that big data results bulwarked by the traditional hallmarks of strong support are sometimes in conflict with one another, and that the resolution of this conflict requires rigorous thought about the sources of conflict and consequently the relative power of data to address phylogenetic hypotheses. Theoretical tools have been derived to address long-standing controversies in experimental design that have occasionally engendered contentious academic debate, such as i) the power of different genes and phylogenetic characters, ii) the relative utility of increased taxonomic versus character sampling iii.) the potential to design taxonomically dense phylogenetic studies optimized by taxonomically sparse genome-scale data. Here, we present an implementation of these theoretical tools to guide phylogenetic experimental design using advances to the phylogenetic signal framework to iteratively rank-order gene-taxa sampling schema and to ensure proposed sampling schema reach the desired power to answer specific phylogenetic hypotheses.

[Poster] **Generalized Pruning**
Seong-Hwan Jun, Fred Hutchinson Cancer Research Center
sjun2@fredhutch.org
Session 6: AED

Bayesian phylogenetics is a computationally challenging inferential problem. Classical methods are based on random-walk Markov chain Monte Carlo (MCMC), where random proposals are made on the tree parameter and the continuous parameters simultaneously. The samples from MCMC allows one to quantify uncertainty around the split probabilities of a phylogeny as well as the branch length parameters. In this paper, we describe a new algorithm that directly generalizes the Felsenstein pruning algorithm by marginalizing out ancestral states and subtrees simultaneously. Utilizing this algorithm, we propose an efficient approach based on composite likelihood method to estimate the parameters governing the phylogenetic tree. Our experiments demonstrate the estimated parameters compare favorably to the ones obtained via MCMC at a fraction of the runtime.

The dynamics of cell division and differentiation in cerebral organoids

Arndt von Haeseler, CIBIV University of Vienna and Medical University of Vienna
arndt.von.haeseler@univie.ac.at

Session 7: 19:00AED

(Joint work with Simon Haendeler, Florian Pflug, Christopher Esk, Jürgen Knoblich, Dominik Lindenhofer)

Many aspects of the development of complex tissues and organs from undifferentiated stem cells are open biological problems. In recent years, a new "3-dimensional" model system, called "organoids", has been developed, which allows studying the differentiation processes of human organs in vitro. At the same time, modern molecular biological methods make it possible to trace the individual contribution (number of offspring aka lineage size distribution) of the undifferentiated stem cell to the developed organoid. We will discuss the offspring size distribution of stem cells grown into cerebral organoids, which partially follows a Zipfian Law. Subsequently, we will present a mathematical model of organoid growth which reproduces and simultaneously explains essential parameters (cell count, lineage size distribution) observed in these complex biological experiments, and show how and under which conditions the empirically observed Zipfian law emerges from this model. In the future, this model is intended to serve as a zero model to study deviations from organoid differentiation and thus possibly to better understand disease patterns.

Information geometry for phylogenetic trees (Student presentation)

Jonas Lueg, University of Göttingen, Faculty for Mathematics and Informatics, Institute for Mathematical Stochastics
jonas.lueg@uni-goettingen.de

Session 7: 19:30AED

(Joint work with Tom Nye, University of Newcastle, Maryam Garba, University of Newcastle, Stephan Huckemann, University of Göttingen)

We propose a new space to model phylogenetic trees. It is based on a biologically motivated Markov model for genetic sequence evolution. As a point set, this space comprises the previously developed Billera-Holmes-Vogtmann (BHV) tree space while its geometry is motivated from the edge-product space. As the latter, our new wald space also involves disconnected forests, it does not contain certain singularities of the latter, though. The geometry of wald space is that of the Fisher information metric of character distributions, either from a discrete Bernoulli or from a continuous Gaussian model. The latter can be viewed as the trace metric of the affine-invariant metric for covariance matrices, the former is that of the Hellinger divergence, or, as we show, equivalent to any metric obtained from an f-divergence, such as the Jensen-Shannon metric. For the latter (continuous) we derive a gradient descent algorithm to project from the ambient space of covariance matrices to wald space and for both we derive computational methods to compute geodesics in polynomial time and show numerically that the two information geometries (discrete and continuous) are very similar. In particular geodesics are approximated extrinsically. Comparison with the BHV geometry shows that our canonical and biologically motived space is substantially different.

The Effects of Tree Islands on Consensus (Student presentation)

Ana Serra Silva, Department of Life Sciences, The Natural History Museum, London and School of Earth Sciences, University of Bristol, UK
a.da-silva@nhm.ac.uk

Session 7: 20:00AED

(Joint work with Mark Wilkinson, Department of Life Sciences, The Natural History Museum, London, UK)

Phylogenetic analyses often yield multiple trees that can be interpreted as comprising distinct subsets based on the notion of adjacency in tree space, itself based on some notion/measure of the distance between two trees. Every pair of trees in the same "island" are connected by a series of trees differing from each other by a single branch rearrangement and every tree in the series is present in the island. In a parsimony setting, the presence of tree islands with very disparate sizes can profoundly affect the commonly used majority-rule consensus (MRC), which can be dominated by groups in large islands. Using variations of the MRC, we show that it is possible to minimise island-size bias, while summarising the topological differences between all islands. We also explore how changing the branch rearrangement threshold can lead to multiple island subsets of the same tree distribution.

Refinement stable consensus methods

Mareike Fischer, Greifswald University, Germany
email@mareikefischer.de

Session 7: 20:15AED

(Joint work with Michael Hendriksen, University of Düsseldorf, Germany)

Consensus methods play a crucial role in mathematical phylogenetics, as they allow for summarizing various gene trees, e.g. resulting from multiple tree reconstruction methods, into one consensus tree. However, it was long unknown if any of the known consensus methods is "future-proof", i.e., robust concerning the introduction of new knowledge; of if even hypothetically a "future-proof" consensus method can exist. In a recent study, Francis and Steel investigated the concept of "future-proofing" with consensus methods. They considered the introduction of additional data and analyzed consensus methods in this regard. In particular, they investigated associative stability – robustness against the introduction of additional trees, and extension stability – robustness against the introduction of additional species. Unfortunately, they found that such future-proofing is not possible, as there exist no regular, extension stable consensus methods. In my talk, I will investigate a related question – can a consensus method be robust against refinement of the input trees? In particular, if you have input trees that are not fully resolved and compare their consensus tree with a second set of input trees that contains resolved versions of the original set, in how far can the second consensus tree contradict the first? We will answer this question and present both negative and some positive results.

Beyond Robinson-Foulds: information-theoretic tree distance metrics

Martin R. Smith, Department of Earth Sciences, Durham University
martin.smith@durham.ac.uk

Session 7: 20:15AED

The Robinson-Foulds (RF) distance is widely used to quantify similarity between phylogenetic trees. The measure tallies the number of bipartition splits that occur in both trees — but this conservative approach ignores potential similarities between almost-identical splits, with undesirable consequences.
"Generalized" RF metrics address this shortcoming by pairing splits in one tree with similar splits in the other. Each pair is assigned a similarity score, the sum of which enumerates the similarity between two trees. The challenge lies in quantifying split similarity: existing definitions lack a principled statistical underpinning, resulting in misleading tree distances that are difficult to interpret. Here, I present new probabilistic measures of split similarity, which allow tree similarity to be measured in natural units (bits). The corresponding tree distance metrics outperform alternative measures against a broad suite of criteria, even without accounting for the non-independence of splits within a single tree.

Generating and sampling trees from a phylogenetic terrace

Olga Chernomor, University of Vienna, CIBIV
o.chernomor@gmail.com

Session 7: 20:15AED

(Joint work with Arndt von Haeseler (Max Perutz Labs, University of Vienna, Medical University of Vienna))

Phylogenetic terraces are collections of trees with an identical score (likelihood as well as parsimony). Terraces occur in the analysis of multi-gene partitioned alignments with missing genes. For sparse alignments (many missing genes per partition) the number of trees on a terrace can be excessively large. Therefore, understanding the influence of these sets of equally scoring trees on algorithms for phylogenetic tree inference could provide valuable information for the improvement of search routines.
In previous work we provided rules to quickly detect whether two neighbouring trees (a pair of trees, which can be transformed one into another by one topological rearrangement) belong to the same phylogenetic terrace or not. We implemented these rules to improve computational efficiency of the state-of-the-art phylogenetic inference program IQ-TREE and exemplified the speedup of terrace-aware implementation on empirical alignments.
Further exploration of terraces and the effects they impose on the tree search routines is hampered by the complexity of associated counting problems and the absence of tools to generate trees from one terrace. Here, we report our approaches to study and characterize terraces. Among others, we will discuss possibilities to generate all trees from a terrace for feasible terrace sizes and to sample trees from a terrace for excessively large terrace sizes, as well as potential applications for terrace summaries and terrace support.

Model embeddability for symmetric group-based model (Student presentation)
Muhammad Ardiyansyah, Aalto University
muhammad.ardiyansyah@aalto.fi
Session 8: 8:00AED

(Joint work with Kaie Kubjas (Aalto University) and Dimitra Kosta (University of Glasgow))

We study model embeddability, which is a variation of the famous embedding problem in probability theory, when apart from the requirement that the Markov matrix is the matrix exponential of a rate matrix, we additionally ask that the rate matrix follows the model structure. We provide a characterisation of model embeddable Markov matrices corresponding to symmetric group-based phylogenetic models. In particular, we provide necessary and sufficient conditions in terms of the eigenvalues of symmetric group-based matrices. To showcase our main result on model embeddability, we provide an application to hachimoji models, which are eight-state models for synthetic DNA. Moreover, our main result on model embeddability, enables us to compute the volume of the set of model embeddable Markov matrices relative to the volume of other relevant sets of Markov matrices within the model.

Analysis of the matrix exponential can guide maximum likelihood-based phylogenetic inference (Student presentation)
Cassius Manuel, Center for Integrative Bioinformatics Vienna, Max Perutz Labs; University of Vienna and Medical University of Vienna
cassiusmanuelperez@gmail.com
Session 8: 8:15AED

(Joint work with Arndt von Haeseler; Center for Integrative Bioinformatics Vienna, Max Perutz Labs; University of Vienna, Medical University of Vienna; Faculty of Computer Science, University of Vienna)

Markov models of DNA evolution are diverse, but they share some analytical properties that can be exploited for inferring phylogenies using maximum likelihood. In particular, I will introduce a closed formula for the matrix exponential of a rate matrix with four states. Then I will present the toy problem of estimating the evolutionary distance t between two sequences. Remarkably, the dominant coefficients of the exponential matrix can be used to decide whether $t = \infty$ is a local maximum point of the likelihood function. An analogous strategy can be employed to discard long branches while optimizing the branch lengths of a given species tree. As a potential application, I will explain how to construct alignments that induce multiple local maxima as a function of one branch length.

The Hyperoctahedral Group (Student presentation)
Joshua Stevenson, School of Natural Sciences, University of Tasmania
joshua.stevenson@utas.edu.au
Session 8: 8:45AED

In the context of estimating genome rearrangement distances, genomes are often represented by signed permutations, which form a group under composition — the hyperoctahedral group. As is often the case with algebraic structures, this group pops up in a number of other places, even just within Phylogenetics. I'll be talking about the hyperoctahedral group and how it can be used to gain a clearer understanding of a few different ideas.

Phylo-symmetric algebras: mathematical properties of a new tool in phylogenetics (Student presentation)
Julia Shore, University of Tasmania, School of Natural Sciences
julia.shore@utas.edu.au
Session 8: 9:15AED

(Joint work with Michael Hendriksen)

In phylogenetics it is of interest for rate matrix sets to satisfy closure under matrix multiplication as this makes finding the set of corresponding transition matrices possible without having to compute matrix exponentials. It is also advantageous to have a small number of free parameters as this, in applications, will result in a reduction of computation time. We explore a method of building a rate matrix set from a rooted tree structure by assigning rates to internal tree nodes and states to the leaves, then defining the rate of change between two states as the rate assigned to the most recent common ancestor of those two states. We investigate the properties of these matrix sets from both a linear algebra and a graph theory perspective and show that any rate matrix set generated this way is closed under matrix multiplication. The consequences of setting two rates assigned to internal tree nodes to be equal are then considered. This methodology could be used to develop parameterised models of amino acid substitution which have a small number of parameters but convey biological meaning.

Uniformization-stable Markov models

Jeremy Sumner, University of Tasmania
jsumner@utas.edu.au

Session 8: 9:30AED

(Joint work with Luke Cooper)

I will discuss the algebraic structure underlying uniformization of continuous-time Markov chains. In particular, I will present recent work which establishes that linear Markov models are "uniformization-stable" if and only if their associated rate matrices occur precisely as the intersection with a Jordan algebra. I will show that this algebraic perspective gives a unified view of this phenomenon spanning disparate model families, including, in particular, the time-reversible hierarchy.

Irreducible semigroup-based Markov models

Venta Terauds, Discipline of Mathematics, University of Tasmania
venta.terauds@utas.edu.au

Session 8: 9:30AED

(Joint work with Jeremy Sumner)

We present a complete characterisation of irreducible semigroup-based Markov models. Semigroup-based Markov models, introduced by Sumner and Woodhams in 2019, are a generalisation of group-based models. They possess many of the pleasing properties of group-based models, with notable examples including the well-known Felsenstein 81 model.

Matrix analytic methods for the gene-tree species-tree reconciliation problem

Małgorzata O'Reilly, University of Tasmania
malgorzata.oreilly@utas.edu.au

Session 9: 10:15AED

(Joint work with Barbara Holland (co-presenting) (School of Natural Sciences, University of Tasmania))

We consider the gene-tree species-tree reconciliation problem, in which the task is to find a suitable gene-tree that fits the given species-tree so as to maximize the likelihood of such fitting, given the available data. We describe a model for the species-tree, a model for the gene-tree, and derive theoretical and algorithmic results for the analysis using matrix-analytic methods. This work is a collaborative effort of the Stochastic Modelling Meets Phylogenetics Group. All code will be made publicly available.

Does migration promote or inhibit diversification? A case study involving the dominant radiation of temperate Southern Hemisphere freshwater fishes

Chris Burridge, University of Tasmania
chris.burridge@utas.edu.au

Session 9: 10:45AED

(Joint work with Jonathan Waters)

Although theory predicts that dispersal is a pivotal influence on speciation and extinction rates, it can have contradictory effects on each, such that empirical quantification of its role is required. In many studies, dispersal reduction appears to promote diversification, although some comparisons of migratory and non-migratory species suggest otherwise.We test for a relationship between migratory status and diversification rate within the dominant radiation of temperate Southern Hemisphere freshwater fishes, the Galaxiidae. We reconstructed a molecular phylogeny comprising >95% of extant lineages, and applied State-dependent Speciation Extinction models to estimate speciation, extinction, and diversification rates. In contrast to some previous studies, we revealed higher diversification rates in non-migratory lineages. The reduced gene flow experienced by non-migratory galaxiids appears to have increased diversification under conditions of allopatry or local adaptation. Migratory galaxiid lineages, by contrast, are genetically homogeneous within landmasses, but may also be rarely able to diversify by colonizing other landmasses in the temperate Southern Hemisphere. Apparent contradictions among studies of dispersal-diversification relationships may be explained by the spatial context of study systems relative to species dispersal abilities, by means of the "intermediate dispersal" model; the accurate quantification of dispersal abilities will aid in the understanding of these proposed interactions

The Shape of Phylogenies Under Phase-Type Distributed Times to Speciation and Extinction (Student presentation)

Albert Christian Soewongsono, School of Natural Sciences, University of Tasmania
albert.soewongsono@utas.edu.au

Session 9: 11:00AED

(Joint work with Barbara Holland, Małgorzata O'Reilly (School of Natural Sciences, University of Tasmania))

We consider a macroevolutionary model for phylogenetic trees where times to speciation or extinction events are drawn from a Coxian phase-type (PH) distribution. We show that different choices of parameters in our model lead to a range of tree shapes as measured by Aldous' β statistic. In particular, it is possible to find parameters that correspond well to empirical tree shapes. Lastly, we derive a likelihood expression for the probability of observing any edge-weighted tree under a model with speciation but no extinction. We perform goodness-of-fit tests for two large empirical phylogenies (squamates and angiosperms) that compare models with Coxian PH distributed times to speciation with models that assume exponential or Weibull distributed waiting times. We found that, in many cases, models assuming a Coxian PH distribution provided the best fit.

Key words: Macro-evolutionary model; diversification; stochastic model; tree shape; phase-type distribution.

A subfunctionalization model of gene family evolution predicts balanced tree shapes (Student presentation)

Jiahao Diao, University of Tasmania, School of Natural Sciences
jiahao.diao@utas.edu.au

Session 9: 11:15AED

(Joint work with Prof Barbara R. Holland, University of Tasmania. Assoc Prof Małgorzata O'Reilly,)

We consider a subfunctionalization model of gene family evolution. A family of n genes that perform z functions is represented by an $n \times z$ binary matrix Y_t where a 1 in the ijth position indicates that gene i can perform function j. Y_t evolves according to a continuous time Markov chain (CTMC) that represents the processes of gene duplication, coding region loss and regulatory region loss with the restriction that each function is protected by selection, meaning that each column in the matrix must contain at least one 1.

We generate gene trees based on the CTMC $\{Y_t, t \geq 0\}$. We analyse the long-run behaviour of the model and specify the conditions where we expect gene trees to continue to grow and where we expect them to have a stable number of genes. We show that different choices of rate parameters for the processes of duplication and loss lead to different tree shapes as measured by two common tree-shape statistics: the β-statistic for measuring tree balance and the γ-statistic for assessing diversification rate. We provide an extension of β to sets of trees. This extension is less biased compared to using the average β value. We find that when the process is stable, gene trees are predicted to have positive values of β indicating balanced trees and negative values of γ indicating that diversification occurs more towards the tips of the tree. When the process is unstable gene duplication dominates and the process is close to following the uniform ranked tree shape (URT) distribution. The results of our analysis suggest that comparing the tree-shape statistics of empirical gene-trees to the predictions presented here will provide a test of the subfunctionalization model.

Detecting selection acting on recently duplicated genes.

Tristan Stark, University of Tasmania, School of Natural Sciences
tlstark@utas.edu.au

Session 9: 11:30AED

(Joint work with Rebecca Kauffman, Maria Maltepes, Ryan Houser, David Liberles - Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia)

We formulate a test for selection acting on duplicated genes shortly after the duplication event occurs. We present a population genetic model describing the evolution of a haploid or diploid population, and calculate prediction bands based on the model for the proportion of the population carrying the duplicated gene over time. We use the prediction band associated with the case in which the duplicated gene is neutral to attempt to reject the hypothesis of neutrality.

Index