Phylomania 2016

Maths & Physics Building,
University of Tasmania, Hobart
November 14-16

Welcome

Phylomania 2016, hosted by the University of Tasmania's Theoretical Phylogenetics Group

Phylogenetics is concerned with one of the most important problems in biology: reconstructing the evolutionary history of present-day organisms from molecular data, such as DNA, or morphological characters. Hidden from view, in phylogenetics software packages used by biologists, are algorithms implementing stochastic and combinatorial methods on binary trees, as well as more general network structures. The mathematics involved represent a unique confluence of probability theory, discrete mathematics, stochastic methods, and statistical inference, as well as algebraic methods such as group theory. There are many important theoretical and practical problems that arise, such as statistical identifiability of models, consistency and convergence of methods. These problems can only be solved using a multi-disciplinary approach. *Phylomania* brings together internationally recognised experts, with the aim of discussing the pressing research problems in phylogenetics.

Phylomania is held within the School of Physical Sciences at the University of Tasmania, Hobart. The meeting brings together mathematicians, statisticians, biologists and computer scientists to discuss current research in phylogenetics. The conference is both informal and leading-edge: some of the very latest ideas, applications, and unsolved challenges, to do with phylogenetic inference and evolutionary biology are voiced here.

Important Information

Location: If you've attended before, you'll already know where the conference is held: Maths and Physics Building (on Google Maps), Clark Road, Sandy Bay Campus.

The lectures will all be held in Physics Lecture Theatre 2, also known as Room 313. Morning and afternoon teas and lunches will all be in Room 333 in the same building, close by.

If you have a poster to show, you may mount it in Room 333.

Wireless Internet Access: If you have `eduroam` set up on your laptop or other mobile computer you can log in to UTAS wireless network with that.

Sadly due to unforeseen administrativitis, we can no longer offer general access wireless internet, so we hope to have temporary accounts set up for you on our system. Look in your conference documents for your specific username and password.

If you have any problems with IT while you're here, you can call the UTAS Service Desk on 03 6226 1818, or just extension 1818 from any office phone.

Excursion: We usually go on an excursion at the end of the conference. There is lots of great walking around Hobart, and we usually take a moderate, low-intensity walk around, before adjourning to a local

hostelry for a reward. Make sure you have sensible shoes for this (sneakers are ok).

If you're planning to come along, bring a sun hat and sun screen!

The UV in Tasmania is very strong and you can easily get bad sunburn if you're not well prepared.

Discussions: At the end of each day we often go to a local pub and this year (again) we have chosen Preachers Bar, which is about a 30 minute walk from campus (see map on the back page).

Disclaimer: I've attempted to not make any mistakes in the production of this programme: please forgive me if I have made some anyway. *Mike Charleston*

TIMETABLE FOR PHYLOMANIA 2016

		Wednesday 16th	Thursday 17th	Friday 18th
		Official Welcome		
	08:50	*Brian Yates, Dean of Faculty of Science, Engineering and Technology*		
		Lectures are all in Lecture Theatre 2, Maths & Physics Building		
Session 1		***Phylogenetic Inference I***	***Models***	***Genome Evolution II***
	09:00 09:20	*Jeremy Sumner*	*Ben Kaehler*	*Venta Terauds*
	09:40	*Michael Charleston*	*Michael Woodhams*	*Paul Slade*
	10:00	*Mathieu Fourment*	*Barbara Holland*	*Malgorzata O'Reilly*
	10:20	*flexible*	*Julia Shore*	*flexible*
		REGISTRATION in Room 333		
10:30 or	10:40	**MORNING TEA & COFFEE in Room 333**		
Session 2		***Genome Evolution I***	***Phylogenetic Inference II***	***Phylogenetic Inference III***
	11:00	*Gavin Huttley*	*John Rhodes*	*Markus Fleischauer*
	11:20	*Jonathan Mitchell*		
	11:40	*Bennet McComish*	*Conrad Burden*	*Peter Jarvis*
	12:00	*Tristan Stark*		
	12:20	*flexible*	*flexible*	*flexible*
	12:30	**LUNCH in Room 333**		
Session 3		***Applications I***	***Coevolution***	***High Throughput Sequencing 1***
	14:00	*Mingxin Liu*	*Michael Hendriksen*	*Allen Rodrigo*
	14:20	*Dorothy Steane*	*Yao-ban Chan*	
	14:40	*Matt Larcombe*	*Nick Beeton*	*Louis Ranjard*
	15:00	*flexible*	*Greg Jordan*	*flexible*
	15:10	**AFTERNOON TEA & COFFEE in Room 333**		
Session 4		***Applications II***	***Poster Session***	***High Throughput Sequencing 2***
	15:30	*Rebecca Jones*	*Gavin Huttley*	*Thomas Wong*
	15:50	*Jorden Stevenson*	*Markus Fleischauer*	*Arndt von Haeseler*
	16:10	*Jakob Butler*	*Matt Larcombe*	**Awards & Closing**
			Michael Charleston	
			Michael Woodhams	
			Scott Whitemore	**Don't forget the traditional outing on Saturday though!**
			and any more are welcome	
		DISCUSSION / PUB		
			Conference Dinner *(not included in registration) 7pm Hope and Anchor Tavern, est. 1807	

Times will be fairly flexible

Presentations

Factoring phylogeny into using fossils to test for correlations between traits and environment

Coevolution

Nicholas J Beeton[1,2]-, Gregory J Jordan[1], Raymond J Carpenter[1], Barbara R Holland[2], Timothy J Brodribb[1]

[1] School of Biological Sciences, University of Tasmania, Private Bag 55, Hobart, Tas. 7001, Australia

[2] School of Physical Sciences, University of Tasmania, Private Bag 37, Hobart, Tas. 7001, Australia

With the advent of modern phylogenetic techniques, an opportunity exists to make better inferences regarding evolutionary processes based on information gained from fossilised taxa. In particular, the family *Proteaceae* (the family that contains *Banksia*, *Protea* and *Grevillea* species) has well-preserved fossils and relatively well-understood phylogenetics, and is an ideal candidate for determining adaptive effects based on changing habitat over the past 70 million years. Stomatal size in this family is known to be related to habitat type in modern species, and there is an apparently clear relationship between preserved stomata size in the fossil record and the estimated temperature at that time. However, basic regression techniques are insufficient to prove that this relationship is inconsistent with simple genetic drift.

To address this problem, we randomly generated matched phylogenetic trees of Proteaceae with a Brownian motion process determining stomatal size, and sampled "virtual fossils" in the same way as the actual data. We find that the fossil record is significantly inconsistent with genetic drift, demonstrating that environmental change drove selection on stomatal size, likely involving habitat effects. Our method has the additional benefit of providing information about the ratio of extinction to speciation $\epsilon = \mu/\lambda$ in the family history.

Inferring rate matrices from the stationary distribution of the multi-allelic neutral Wright-Fisher model for low mutation rates

Phylogenetic Inference II

Conrad Burden[1], Yurong Tang[2]
[1] Research School of Biology, Australian National University
[2] Mathematical Sciences Institute, Australian National University

We address the problem of determining the stationary distribution of the multi-allelic, neutral-evolution Wright-Fisher model in the diffusion limit. A full solution to this problem for an arbitrary $K \times K$ mutation rate matrix involves solving for the stationary solution of a forward Kolmogorov equation over a $(K-1)$)-dimensional simplex, and remains intractable. In most practical situations mutations rates are slow on the scale of the diffusion limit and the solution is concentrated on the corners and edges of the simplex. In this talk I present a practical approximate solution for slow mutation rates in the form of a set of line densities along the edges of the simplex. The solution relies on parametrising the rate matrix as the sum of a reversible part and a set of $(K-1)(K-2)/2$ independent terms corresponding to fluxes of probability along closed paths around faces of the simplex.
The solution is then used to develop a procedure for estimating evolutionary rate matrices from observed site frequency data. The procedure assumes (1) that the data are obtained from a constant size population evolving according to a stationary Wright-Fisher model; (2) that the data consist of a multiple alignment of a moderate number of sequenced genomes drawn randomly from the population; and (3) that within the genome a large number of independent, neutral sites evolving with a common mutation rate matrix can be identified. No restrictions are imposed on the scaled rate matrix other than that the off-diagonal elements are positive and $\ll 1$, and that the rows sum to zero. In particular the rate matrix is not assumed to be reversible.

Comparing the terpene synthase gene family in two assembled eucalypt genomes

Applications II

Jakob Butler[1], Jules S. Freeman[1], Brad M. Potts[2], René E. Vaillancourt[2], Dario Grattapaglia[3], Orzenil B. da Silva Junior[3], Blake Simmons[4], Jeremy Schmutz[5], Kerrie W. Barry[6], David J. Lee[7], Adam Healey[8], Agnelo Furtado[9], Robert J. Henry[9], Abdul Baten[10], Graham J. King[10] and Mervyn Shepherd[10]
[1]School of Biological Sciences, University of Tasmania, Hobart, Australia; [2]School of Biological Sciences and ARC Training Centre for Forest Value, University of Tasmania, Hobart, Australia; [3]EMBRAPA Genetic Resources and Biotechnology, Brasília, Brazil; [4]Joint Bioenergy Instutute, San Francisco, CA; [5]Hudson Alpha, Huntsville, AL; [6]DOE Joint Genome Institute, Walnut Creek, CA; [7]Forest Industries Research Centre, University of the Sunshine Coast, Sippy Downs, Australia; [8]University of Queensland, Brisbane, Australia; [9]University of Queensland/QAAFI, Brisbane, Australia; [10]Southern Cross Plant Science, Southern Cross University, Lismore NSW, Australia

Plant terpenes are diverse compounds which mediate numerous ecological interactions and are of commercial value in many taxa. Eucalypts are characterised by especially high concentrations of foliar terpenes, with the recent assembly of the *Eucalyptus grandis* genome revealing the largest number of terpene synthase genes (TPS) of any sequenced plant genome. While similar representation of TPS was found in the closely related *Eucalyptus globulus*, the extent to which this gene family is conserved more broadly in eucalypts, including in *Corymbia*, a sister eucalypt genus of growing economic importance, is unknown. I will discuss our annotation of the TPS gene family in the recently assembled *Corymbia citriodora* subsp. *variegata* genome and the comparison to *E. grandis* through a phylogenetic analysis incorporating loci positional information from the two assemblies.

Inference of co-evolution from reconciliations

Coevolution

Yao-Ban Chan
University of Melbourne

Phylogenies of species and their constituent genes are linked by mappings called reconciliations, which explain discrepancies by inferring the presence and location of genetic events such as duplications, losses and transfers. Because reconciliations explicitly reconstruct these events, we can analyse them to detect certain biological phenomena. In this talk, we present a statistical method we developed to detect "co-evolving'' genes, which share related evolutionary histories. This phenomenon can arise if the genes are located in the same chromosome region or if they are involved in the same protein complex or biological process. Thus, detecting co-evolution can shed light on proximity and functional genetic relationships.

Experimenting with subflattenings to find trees

Phylogenetic Inference I

Michael Charleston, Jeremy Sumner
School of Physical Sciences, University of Tasmania

A novel method of dimensional reduction for phylogenetic tree models was recently developed by Sumner [1], as detailed in the previous talk. In that method, the patterns at each site in a multiple sequence alignment are counted, and compact sub-matrices of signed counts under a Hadamard (or similar) transformation are constructed. The choice of which transformation to apply is a matter for future investigation: at present we only require that it be orthogonal.
This extends the "flattening" concept of Allman, Kubatko & Rhodes,[2] such that these sub-matrices can be constructed for each possible bipartition, or split, of the taxa in the alignment. We refer to the sub-matrices as *sub-flattenings*. They are significantly smaller — of size that is quadratic *vs* exponential in the number of taxa involved. They also have analogous properties to the flattenings, in particular that, in the absence of stochastic (sampling) error, the rank of such matrices is minimal for splits that correspond to the underlying tree, whereas for splits that are not in the underlying tree, the rank is generically higher.
The mathematical elegance of the sub-flattenings is appealing of itself, but of great practical interest is also how the rank estimate varies in the presence of sampling error, caused by having finite sequences. Since stochastic error generally makes the sub-flattening matrices have full rank, we use singular value decomposition (after Erikson[3]) to obtain the singular values of each sub-flattening matrix, and then measure the sum of squares of all the excess singular values — that is, avoiding the first k terms. This error is smallest for sub-flattenings of minimal *generic* rank, so we attempt to use it as a proxy for the *expected* rank.
We have some preliminary results from a user-friendly C++ program that we have developed to investigate this idea further, showing that the rank error terms do indeed segregate well for splits that are part of the underlying tree. The speed of the implementation even at this early stage is promising, and we look forward to developing it further.

1. Sumner, J. Dimensional Reduction for Phylogenetic Tree Models *arXiv preprint*, 2016
2. Allman, E. S.; Kubatko, L. S. & Rhodes, J. A. Split scores: a tool to quantify phylogenetic signal in genome-scale data. *Systematic Biology*, 2016
3. Eriksson, N. Tree construction using singular value decomposition. In Pachter, L. & Sturmfels, B. (Eds.) *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005

Aligning Biological Networks

Poster

Michael Charleston[1,2], Alexandru Radu[2], Martin McGrane[2]
[1]School of Physical Sciences, University of Tasmania; [2]School of Information Technologies, University of Sydney

Systems biology is full of networks: gene regulatory networks, protein-protein interaction networks, food webs, social and contact networks. These networks change over time, between species, and across ecosystems, but they often retain common features. We leverage common topological features among networks to compare them, e.g., feed-forward loops, highly connected hubs, degree distribution of nodes. We have developed *Node Fingerprinting* (NF) and *Node Handprinting*: state of the art tools to align even large biological networks, with high accuracy and speed, and minimal memory. We have extended NF to the *Biological Network Edit Distance*, which can estimate possible paths between biological networks.

Bad Character Deletion Supertrees

Phylogenetic Inference II

Markus Fleischauer and Sebastian Böcker
Friedrich Schiller University, Jena

Supertree methods combine a set of phylogenetic trees into a single supertree. Similar to supermatrix methods, these methods provide a way to reconstruct larger parts of the Tree of Life. Matrix Representation with Parsimony (MRP) is still the most widely used supertree method today, as the constructed supertrees are of comparatively high quality. Recently, the meta-method SuperFine was introduced, which combines the Strict Consensus Merger (SCM) as preprocessing with MRP and outperforms all other methods. Another recent supertree method is **FlipCut** which tries to resolve incompatibilities in the source trees by flipping '0/1'-entries in their matrix representation. **FlipCut** has guaranteed polynomial running time, and outperformed other polynomial-time supertree methods.
Different from supermatrix methods, supertree methods allow us to analyze large datasets without constructing and analyzing a multiple sequence alignment for the complete dataset. Therefore, supertree methods can be used as part of divide-and-conquer meta techniques like DACTAL, potentially evading the computational complexity of phylogenetic inference methods such as maximum likelihood. Currently the combining step of such divide-and-conquer techniques relies on supertree methods that have to solve NP-hard optimization problems (MRP). Here, we introduce the Bad Character Deletion (BCD) supertree method. We remove the minimum number of clades from the source trees so that the resulting set of trees is compatible. We adapt the **FlipCut** heuristic for the new objective function. Furthermore, we integrate the SCM supertree into our calculations, and show how to use bootstrap values when removing bad clades. On a simulated dataset, BCD outperforms the state-of-the-art algorithms SuperFine and Matrix Representation with Parsimony. BCD supertrees simultaneously produces high-quality supertrees *and* has guaranteed polynomial running time.

Phylogenetic inference with streaming data using sequential Monte Carlo

Phylogenetic Inference I

Mathieu Fourment[1], Aaron Darling[1], Frederick Matsen[2]
[1]University of Technology Sydney; [2]Fred Hutchinson Cancer Research Center

A major shortcoming of current phylogenetic methods is their inability to quickly incorporate new data as it becomes available. Adding new data to an analysis usually requires re-computing the entire analysis. Each analysis run can take decades of CPU time or weeks on a supercomputer facility, making re-analysis of large data sets impractical. This limitation could be addressed by applying a class of Bayesian statistical inference algorithms called sequential Monte Carlo (SMC) to conduct online inference: continuously amending new data and updating the estimate of the posterior probability distribution.
After reviewing the SMC framework, I will describe how SMCs can be adapted to the streaming phylogeny problem by adding new taxa to the backbone tree. Similar to Bayesian MCMC algorithms, designing a good transition kernel is the most challenging aspect of implementing an SMC algorithm and I will present and compare some kernels that vary in complexity and efficiency. Preliminary results show that the streaming phylogeny algorithm can outperform the widely-used MCMC-based algorithm implemented in MrBayes in terms of speed without incurring a significant loss in accuracy.

Behind Every Great Tree is a Great (Phylogenetic) Network

Coevolution

Michael Hendriksen
University of Western Sydney

In a 2015 paper, Francis and Steel showed that there exist non-trivial rooted binary phylogenetic networks (in the class of horizontal gene transfer networks) upon which the distance metric affords a metric on a tree. In this talk we outline how this result can be extended to show that for any tree T there exists a family of these horizontal gene transfer networks such that for any given network N in the family the distance metric d_N affords a metric on T. The family of networks are all 'floating' networks, a subclass of a novel family of networks introduced in this paper, and referred to as 'versatile' networks. We will also briefly characterise versatile networks.

Exploring the consequences of lack of closure for the Goldman-Yang model

Models

Michael Woodhams[1], **Barbara Holland**[1], David Liberles[2], Michael Charleston[1], Jeremy Sumner[1]
[1]University of Tasmania; [2]University of Wyoming

The Goldman-Yang (GY) codon model of evolution is commonly used to try and identify selection. Selection, by its nature, is a heterogeneous process, i.e. it acts on some branches of the evolutionary tree and not others. Our previous work (*Is the GTR model bad for phylogenetics?*) has shown that when evolution occurs under a heterogeneous process it is important to consider the closure properties of models, as non-closed models will give biased estimates of evolutionary distance. It is relatively easy to show that the GY model is not closed, this is a consequence of the fact that it is not linear (meaning that the sum of two GY rate matrices is not a GY rate matrix). This raises the concern that a single GY model fit to a heterogeneous GY process might be biased into either over or underestimating the effect of selection and likewise over or underestimating branch lengths. In this talk we will demonstrate the consequences of lack of closure for estimation of both omega (the selection parameter) and branch lengths. The standard GY model is constructed from an underlying HKY model that acts independently at each codon position, and is then influenced by the genetic code via a parameter omega which modifies the rate of transitions between codons that code for different amino-acids. We consider using different DNA models within a similar structure. However, a DNA model being closed does not imply that the codon model constructed from it is closed (the genetic code prevents this). We were curious to see if using closed (Lie Markov) DNA models helps to reduce the bias that arises from lack of closure in the codon model. The short answer is that it doesn't.

How big are sequence neighbourhoods that affect mutation?

Genome Evolution I

Gavin Huttley[1], Yicheng Zhu[1], Teresa Neeman[1], Von Bing Yap[2]
[1]Australian National University; [2]National University of Singapore

Mutation processes differ between types of point mutation, genomic locations, cells, and biological species. For some point mutations, specific neighbouring bases are known to be mechanistically influential. Beyond these cases, numerous questions remain unresolved including: what are the sequence motifs that affect point mutations? how large are the motifs? are they strand symmetric? and, do they vary between samples? We present new log-linear models that allow explicit examination of these questions along with sequence logo style visualisation to enable identifying specific motifs. We demonstrate the performance of these methods by analysing mutation processes in human germline and malignant melanoma. We recapitulate the known CpG effect and identify novel motifs, including a highly significant motif associated with A$\rightarrow$G mutations. We show that major effects of neighbourhood on germline mutation lie within ± 2 of the mutating base. The implications of our results for molecular phylogenetics will be discussed.

Software tools for analyses of genetic variation: PyCogent3, EnsemblDb3, HomologSampler and MutationMotif

Poster

Gavin Huttley
Research School of Biology, Australian National University

The poster will present brief descriptions of each software tool. If you bring your laptop, I'll help you install them!

- **PyCogent3**: A library that serves as the basis for all evolutionary models developed by the Huttley lab. In addition to the standard time-reversible models, it is capable of specifying general (discrete or continuous-time) Markov processes for nucleotide, tuple (e.g. dinucleotide, codon) or amino acids. It also provides numerous capabilities for data handling and manipulation.
- **EnsemblDb3**: A Python3 library for querying Ensembl's MySQL databases. Also provides tools for downloading and installing these on your own servers.
- **HomologSampler**: A command line tool for batch sampling of homologous sequences from Ensembl MySQL databases. For example, it provides sampling one-to-one orthologs.
- **MutationMotif**: A command line tool for analysing point mutations using log-linear models. Suitable for analysing the influence of sequence neighbourhoods or for comparing the influence of neighbourhoods between samples.

A robust phylogeny of a globally significant Eucalyptus lineage

Applications II

Rebecca Jones[1], Dean Nicolle[2]; Dorothy A. Steane[1], René Vaillancourt; Brad Potts[1]
[1]School of Biological Sciences, University of Tasmania; [2]Currency Creek Arboretum

Robust phylogenies are increasingly needed for the ecologically and economically important Australian tree genus *Eucalyptus*. Applications of such phylogenies include: biodiversity measures; prediction of species responses to climate change and susceptibility to pests and pathogens; assessing the risk and impact of hybridisation between planted and native species; and guiding species selection for hybrid breeding programs. However, most phylogenies to date have been based on one or few genomic regions and/or single sample representation of species. As this iconic Australian genus is known to have been impacted by hybridisation, recent speciation and morphological convergence, phylogenies built on single-exemplar sampling and limited sampling of the genome are likely to result in unresolved and/or misleading phylogenetic relationships. We used multiple samples per taxon (540 samples covering 185 terminal taxa) and over 3000 genome-wide DArT markers to construct a robust phylogeny for a globally significant *Eucalyptus* lineage (sections *Maidenaria*, *Exsertaria*, *Latoangulatae* and related smaller sections), which includes all of the commercially important species, a number of rare and endangered species and unusual taxa with controversial taxonomic treatments. At the higher level, our phylogeny largely matched the morphological treatment of taxonomic sections. At lower levels, however, there were numerous inconsistencies between the morphological treatment and the molecular phylogeny, and taxa within the three main taxonomic sections were generally not monophyletic. Some of the discrepancies appear to be the result of morphological convergence or misclassification, requiring taxonomic reassessment, but many inconsistencies appear to be the products of incomplete speciation and / or hybridisation. Our analysis represents a significant advance on previous phylogenies of this important lineage, providing a robust phylogenetic framework for future evolutionary / ecological studies and practical applications.

Congruence in ecological attributes among old, rare clades. Ancestral states?

Coevolution

Greg Jordan
School of Biological Sciences, University of Tasmania

Reconstructing ancestral ecologies (such as climate and habitat) from phylogenies is very difficult because of extensive convergence – there are too many ways to skin the ecological cat. I argue that clades with high levels of evolutionary niche conservatism should provide better evidence of ancestral ecologies than dynamic clades. Palaeoendemics are clades that are rarer than would be expected for their age. In many cases, this will be because of high levels of niche conservatism. In this talk I will show that both globally and within Tasmania there is strong congruence in some key ecological attributes of palaeoendemic seed plants (notably climates that are both wet and have moderate temperatures). I argue that these attributes are therefore likely to represent ancestral ecological features, and that this approach avoids problems associated with ancestral state reconstruction.

Full Reconstruction of Non-Stationary Strand-Symmetric Models on Rooted Phylogenies

Models

Ben Kaehler
Australian National University

Understanding the evolutionary relationship between species is of fundamental importance to the biological sciences. The location of the root in any phylogenetic tree is critical as it gives an order to evolutionary events. None of the popular models of nucleotide evolution used in likelihood or Bayesian methods are able to infer the location of the root without exogenous information. It is known that the most general Markov models of nucleotide substitution can also not identify the location of the root nor be fitted to multiple sequence alignments with less than three sequences. We prove that the location of the root and the full model can be identified and statistically consistently estimated for a non-stationary, strand-symmetric substitution model given a multiple sequence alignment with two or more sequences. We also generalise earlier work to provide a practical means of overcoming the computationally intractable problem of labelling hidden states in a phylogenetic model.

Standard codon substitution models overestimate purifying selection for non-stationary data

Poster

Ben Kaehler
Australian National University

Estimation of natural selection on protein-coding sequences is a key comparative genomics approach for de novo prediction of lineage specific adaptations. Selective pressure is measured on a per-gene basis by comparing the rate of non-synonymous substitutions to the rate of neutral evolution, typically assumed to be the rate of synonymous substitutions. All published codon substitution models have been time-reversible and thus assume that sequence composition does not change over time. We previously demonstrated that if time-reversible DNA substitution models are applied blindly in the presence of changing sequence composition, the number of substitutions is systematically biased towards overestimation. We extend these findings to the case of codon substitution models and further demonstrate that the ratio of non-synonymous to synonymous rates of substitution tends to be underestimated over three data sets of insects, mammals, and vertebrates. Our basis for comparison is a non-stationary codon substitution model that allows sequence composition to change. Model selection and model fit results demonstrate that our new model tends to fit the data better. Direct measurement of non-stationarity shows that bias in estimates of natural selection and genetic distance increases with the degree of violation of the stationarity assumption. Additionally, inferences drawn under time-reversible models are systematically affected by compositional divergence. As genomic sequences accumulate at an accelerating rate, the importance of accurate de novo estimation of natural selection increases. Our results establish that our new model provides a more robust perspective on this fundamental quantity.

Evidence that diversity dependent and independent process both drive conifer diversification

Applications II

Matt Larcombe[1], 2, Greg J Jordan[2], David Bryant[3], Steven I Higgins[1]
[1] Department of Botany, University of Otago; [2] School of Biological Sciences, University of Tasmania;[3] Department of Mathematics and Statistics, University of Otago

Despite ongoing disagreement about weather species diversity is bounded (diversity-dependent) or unbounded (diversity-independent), there is increasing evidence that both these process act in controlling regional and phylogenetic diversity. This view enables researchers to move beyond polarized studies, to allow quantification of both diversity-dependent and -independent processes during radiations. Here we show that species diversity in conifers is significantly influenced by both diversity-dependent processes (competition) and diversity independent process (niche partitioning and expansion, and clade age). We used a phylogeny of 457 conifer species and processed based species distribution models to describe key aspects of clade niche geometry in this ancient ecologically important group. We then simultaneously estimate the relative effect of different factors using comparative modeling. Our results suggest that while niche expansion and partitioning are significant, competition appears to have the strongest overall effect in shaping clade level species richness in conifers.

What we can learn from the water spider mitochondrial genome?

Applications I

Mingxin Liu, Zhisheng Zhang, Zuogang Peng
University of Tasmania

The mitochondrial genome of the wholly underwater-living spider, *Argyroneta aquatic*, was sequenced and thereby enhancing the available genomic information for Arachnida. The confirmed sequences contained the complete set of known genes present in other metazoan mitochondrial genomes. However, the mitochondrial gene order of *A. aquatica* was distinctly different from that of the most distant *Chelicerata Limulus polyphemus* (Xiphosura), probably because of a series of gene translocations and/or inversions. Comparison of arachnid mitochondrial gene orders for the purpose of phylogenetic inference is only minimally useful, but provides a strong signal in closely related lineages. To test the basal relationships and the evolutionary pattern of tRNA gene rearrangements among Arachnida, phylogenetic analyses using amino acid sequences of the 13 protein-coding genes were performed. An interesting feature, five copies of 135-bp tandem repeat and two copies of 363-bp tandem repeat, was identified in the putative control region. The tandem repeats identified in the *A. aquatica* may help to increase the number of mitochondria and levels of specific enzymes to assist with survival in the aquatic environment. Although it is likely that this is an adaptive response of a terrestrial animal to the colonization of an aquatic habitat, its molecular mechanisms remain to be elucidated.

Ancient and modern genomes reveal that microsatellites maintain a dynamic equilibrium through deep time

Genome Evolution I

Bennet McComish, Michael A. Charleston, Matthew Parks, Carlo Baroni, Maria Cristina Salvatore, Ruiqiang Li, Guojie Zhang, Craig Millar, Barbara R. Holland, David M. Lambert

Tandem repeats of two to six base pair motifs, known as *microsatellites*, are prevalent in both prokaryotic and eukaryotic genomes. Some microsatellites have been shown to be functionally important, but most are assumed to evolve neutrally, and for this reason, along with their abundance and high variability, they have been used extensively in population genetics studies. However, their evolutionary dynamics remain poorly understood, and it is unclear whether microsatellite loci are at equilibrium or experience directional drift in length.
Here we identify more than 27 million microsatellites in modern and ancient Adélie penguin genomes dating to approximately 46.5 kya. In addition, by aligning these loci with microsatellites from a large set of published chordate genomes and mapping them onto a recent phylogeny, we are able to time their evolutionary origin; our data include microsatellites that date to the diversification of chordates more than 500 Mya.
We show that microsatellite length is at a dynamic equilibrium that has remained stable over hundreds of millions of years. While there is length polymorphism among individuals, the overall length distribution for a given locus does not change appreciably over time. Microsatellites can persist over very long time scales, particularly those in exons and regulatory sequence, and some of these microsatellites retain length variability, which suggests they may play a role in the maintenance of evolutionary plasticity.

Phylogenetic Convergence-Divergence Networks

Genome Evolution I

Jonathan Mitchell, Barbara Holland, Jeremy Sumner
School of Physical Sciences, University of Tasmania

Convergence-divergence networks allow for convergence of previously diverged edges and could be used to model introgression. Some three-taxon and four-taxon convergence-divergence networks on the binary symmetric model have desirable mathematical properties, including identifiability. We argue that retaining the molecular clock assumption and introducing convergence parameters can allow for more flexibility in model selection than removing the molecular clock assumption.

On mechanistic modelling of gene content evolution using Markov processes.

Genome Evolution II

Małgorzata O'Reilly
School of Physical Sciences, University of Tasmania

In this talk, I will discuss the new time-heterogeneous Markov model for gene duplication, with the focus on the modelling aspects and potential for further work.

Haplotype reconstruction from Short Read Sequences using Vector Quantization

High Throughput Sequencing I

Louis Ranjard, Allen Rodrigo
Research School of Biology, Australian National University

Uncovering the genetic diversity in an unlabelled mixed sample of individuals is a challenging computational problem. However, this task can be of primary importance. For example, reconstructing the true set of viral haplotypes in an infected host is correlated to the clinical outcome and pathogenesis. We present a vector quantization approach to reconstruct the set of haplotype from a mixed sample of short read sequences when the number of haplotypes is unknown and the sequencing reads are unlabeled. Our method consists in mapping the high dimensional short read sequence space to a lower dimensional space representing the reconstructed haplotypes. We propose to encode each position in the nucleotide sequences as a vector where each element represents the contribution of each nucleotide base. Preliminary computer simulation results will be presented to describe (i) how the true haplotype sequences can be reconstructed and (ii) how these sequences can be used to infer evolutionary parameters of the population under study.

Split Probabilities and the Coalescent Model

Phylogenetic Inference I

Elizabeth Allman[1], James Degnan[2], **John Rhodes**[1];
[1]University of Alaska; [2]University of New Mexico

The coalescent model on a species tree describes the process of incomplete lineage sorting, by which trees for different genes may differ from the species tree and each other. Standard phylogenetic methods infer gene trees, so if the species tree is the goal of data analysis, another layer of inference is needed. The complexity of the coalescent model, though, makes standard statistical approaches computationally unwieldy, limiting the number of taxa that can be dealt with. An alternative is to use gene tree summary statistics, such as the frequencies of splits on the gene trees, as a basis for species tree inference. This talk will indicate how a previously proposed method of species tree inference, NJ_{st}, implemented in ASTRID, uses precisely this information to determine the unrooted species tree topology in a statistically consistent way. Furthermore, though 'split' is an unrooted notion, split probabilities actually contain enough information to determine the rooted topology. The proof of this depends on understanding linear 'split invariants' (analogs of phylogenetic invariants for split probabilities), and certain inequalities related to them.

Evolutionary analyses of short read sequences: an update and a practical application

High Throughput Sequencing I

Allen Rodrigo, Louis Ranjard, Thomas Wong
Research School of Biology, Australian National University

In this talk, I will talk about our ongoing project to analyse a mixed sample of haplotypes from which short-read sequences have been obtained. I will describe two possible scenarios, and then turn to a practical application of the methods we have developed, that reduces the cost of sequencing by a third.

Finding the Lie closure of the Goldman Yang codon model

Models

Julia Shore, Jeremy Sumner, Barbara Holland
School of Physical Sciences, University of Tasmania

Woodhams *et al.* (2016) have found problems with the Goldman Yang (1994) codon model concerning the misestimation of the synonymous/non-synonymous mutation ratio when averaging across branches. Inspired by the successes of Sumner *et al.* (2012), we thought a possible solution to this problem was to find the Lie closure of the model to use as an alternative to the original model. However, the Lie closure of the Goldman Yang model was found to be ridiculously large, too much so to be of practical use. Further investigations are being conducted to find what properties of the Goldman Yang model make it so mathematically difficult to manipulate and whether or not we can "estimate" the Lie closure with a model of a more manageable size.

References:

- Sumner, J. G., Fernández-Sánchez, J., & Jarvis, P. D. (2012). Lie markov models. *Journal of Theoretical Biology*, **298**, 16-31.
- Woodhams, Liberles, Holland, Charleston, Sumner (2016) *in prep.*
- Yang, Ziheng, Goldman, N., & Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution*, **11**(2), 316-324.

Bayes factor tests for measurably evolving populations

Genome Evolution II

Paul Slade, Allen Rodrigo
Research School of Biology, Australian National University

Advances in phylogenetic methodology now allow estimation of evolutionary substitution rates when sequences are obtained at different times; heterochronous sampling. Sequences obtained with sampling time intervals too close yield estimates of evolutionary rates statistically indistinguishable from zero. De facto tests for whether there is sufficient information to estimate a non-zero rate involves sampling-time randomisation followed by the estimation of rates for each randomisation, as a means of generating the null distribution under the hypothesis that the rate is zero. This computationally intensive procedure may require batching hundreds of entire MCMC reiterations. We propose an alternative method that directly compares two competing models and applies a Bayes factor analysis to determine whether heterochronous samples fit the data better than a single isochronous sample. This approach requires only two MCMC runs, and speeds up the process by several orders of magnitude.

Recent developments of a purity-dependent model for microsatellite evolution

Genome Evolution I

Tristan Stark

Previously, we introduced a purity-dependent model for microsatellite evolution. Theoretically, the model was reasonably tractable, and provided an approach to modelling microsatellite evolution which was motivated by the underlying biological mechanics of the system. However, applying the model to real data from whole-genome analysis proved to be problematic. The nature of microsatellite data obtained from Tandem Repeats Finder (TRF) on whole genome data sets is not in correspondence with the state-space of the model. While the model was posed in terms of repeat units, getting data into this form turned out to require some arbitrary decision making to avoid problems at the edge of the state space; some truncation is required, and counting the level of impurity in the terms of the model was open to interpretation. We found that our approach to data-handling could lead to substantial variation in the results of our analysis using the purity-dependent model. Recently we have revisited this model; by making some small adjustments, we arrive at an essentially-similar model which is arguably more realistic, and which sidesteps the issue of making arbitrary data-handling decisions by choosing a state space which closely matches the output of TRF.

Resolving gene tree incongruences in *Nothofagus*: Accommodating for chloroplast dispersal

Applications II

Jorden Stevenson[1], Christopher Burridge[1] Michael Charleston[2], Rebecca Jones[1], René Vaillancourt[1], Barbara Holland[2], Greg Jordan[1]
[1]School of Biological Sciences and [2]School of Physical Sciences, University of Tasmania

Introgressive hybridisation, i.e., the acquisition of genes from other differentiated populations of species through sexual processes, is thought to be an important evolutionary mechanism, especially in plants. It is commonly inferred as a causative mechanism for incongruences among gene trees. However, incomplete lineage sorting, i.e., the retention of lineages that are older than the speciation event, is another mechanism that can give the same gene tree incongruences. In order to disentangle incomplete lineage sorting from introgressive hybridisation, coalescent simulations are commonly employed. However, these simulations do not take into account the differences in dispersal mechanisms commonly observed between nuclear, mitochondrial and chloroplast DNA, and such differences can vastly affect the occurrence of incomplete lineage sorting in populations.

In this study, I used a novel approach to coalescent simulations that employs multiple sub-populations in order to mimic the limited dispersal of chloroplasts found in most plants because of its maternal inheritance (via seed). I studied the gene tree incongruence observed between *Nothofagus cunninghamii* and *N. moorei*. Previous studies had found that two *N. moorei* chloroplast haplotypes from one geographic region nested within a well-supported *N. cunninghamii* clade. This placement is incongruent with both nuclear data and morphological evidence where the species are clearly distinct. My simulations used a total of six sub-populations to represent four of the major *N. cunninghamii* chloroplast haplotypes and two of the major *N. moorei* geographic regions.

My results indicate that the chloroplast DNA gene tree incongruence involving *N. cunninghamii* and *N. moorei* is unlikely to be due solely to incomplete lineage sorting and therefore is more readily explained by ancient introgressive hybridisation, resulting in chloroplast capture. This and other studies, rigorously establishing that organisms have the ability to ‘capture’ genes from another species or population by repeated hybridisation has revolutionised our understanding of the evolution of populations of plants and animals, including *Homo sapiens*. My study shows that even without any morphological intermediates, or known signature from the nuclear genome, evidence can be found for the hybrid origin of some species.

Using genomic information to increase resilience of forest tree plantings

Applications I

Dorothy Steane
School of Biological Sciences, University of Tasmania

The world's forests are being challenged by environmental change due to global warming; increasing exposure to exotic competitors, pests and diseases; and human population pressures. How we take genomics beyond knowledge discovery to aid the development and implementation of adaptation strategies to such change in the required time frame is a major challenge. We here describe a pathway we are exploring with this objective.

While identification of causal genes underlying adaptation and mass screening of germplasm would be ideal, it is not practical. Instead, we are investigating a broad approach that can be applied in many species. We are using population genomics to identify and weight the key environmental drivers that have shaped local adaptation. First, genome-wide scans are used to identify a suite of markers showing signals of disruptive selection. The environmental variables most closely aligned with the variation in this putative adaptive space are then identified. The major environmental correlates are used to develop spatially explicit and biologically-relevant fitness surfaces for contemporary and future climate projections. While functional trait analyses and field trials will provide validation of these models in the long-run, in the absence of better information the models can be used as transfer functions to guide seed source choice or identify components of native gene pools most at risk of future maladaptation. Such an approach is particularly valuable for forest tree plantings, where long generation times and size of forest trees make measurement of life-time fitness difficult.

We demonstrate this approach in the context of climate-adjusted provenancing for ecological restoration (Prober et al. 2015), using several eucalypt species in Australia. Traditionally, local provenances have been favoured due to the 'local is best' paradigm. However, this is increasingly being challenged due to issues of seed supply and quality (e.g. inbreeding), site modification and climate change. Climate-adjusted provenancing (analogous to assisted migration) advocates supplementation of local provenance with germplasm from along a gradient of change (e.g., increasing aridity), with a bias towards provenances that exist in projected analogous climates. The premise of this strategy is to capitalise on inherent genetic diversity and adaptive capacity across a species' range. Apart from identifying and weighting putative climate drivers of adaptation in the target species, our genomic approach has potential to flag whether climate-adjusted provenancing may be compromised by other environmental drivers of adaptation, such as soils.

Reference:

Prober SM, Byrne M, McLean E, Steane DA, Potts BM, Vaillancourt RE, Stock WD: Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. Frontiers in Ecology and Evolution 2015, 3: 65. 10.3389/fevo.2015.00065

Key Words: climate adaptation, genome-wide scans, seed transfer, reforestation, fitness modelling

Dimensional reduction for phylogenetic tree models

Phylogenetic Inference I

Jeremy Sumner
School of Physical Sciences, University of Tasmania

We present a general method of dimensional reduction for phylogenetic tree models. The method reduces the dimension of the model space from exponential in the number of extant taxa, to quadratic in the number of taxa. A key feature is the identification of an invariant subspace which depends only bilinearly on the model parameters; in contrast to the usual multi-linear dependence in the full model space. We discuss a potential application to computation of split weights on a phylogenetic tree.

Computing evolutionary distance between circular genomes

Genome Evolution II

Venta Terauds
School of Physical Sciences, University of Tasmania

For circular genomes with even a small number of regions, estimating evolutionary distances can be computationally intensive. We present a rearrangement model, describe how techniques from representation theory can be applied to simplify distance calculations, and provide some initial results.

Modelling PCR stochasticity and its effects on quantitative NGS experiments

High Throughput Sequencing II

Florian G. Pflug, **Arndt von Haeseler**
(Medical) University of Vienna

Many protocols in modern-day biology use next-generation sequencing (NGS) as a quantitative method, i.e. to measure the abundance of particular DNA molecules. Then, any molecule that remains unsequenced causes a measurement error, and if this affects molecules non-uniformly, results are systematically biased. A major source of such biases is the Polymerase Chain Reaction (PCR), used to amplify DNA prior to sequencing. If it can be adequately modelled, its biases can be predicted and corrected for. Different models of PCR haven been proposed, but none have yet found their way into standard analysis pipelines, owing to a lack of parameter estimates for specific conditions. We thus focus on describing a model whose parameters can be estimated from actual experimental data, while still capturing the main source of biases. We show that this is achieved by viewing PCR as a branching process which, during each cycle, duplicates each DNA molecule with a certain probability, called the reactions efficiency. We combine this model with a simple model of the sampling behaviour of NGS and apply it to published RNA-Seq data. We demonstrate that the reaction efficiency can be estimated from the data, and that the data matches the models predictions well. In particular, we find that the model explains the main observed stochastic effects. Finally, we explore how well we can correct for unobserved molecules, and how much this improves the accuracy of the measured gene transcript abundances.

Automatic bird vocalisation recognition, Part 1

Poster

Scott Whitemore, Michael Charleston
School of Physical Sciences, University of Tasmania

Conservation efforts around the world are becoming increasingly technical. The results and understanding gleaned are used as a basis for making decisions about how we manage our natural resources. To this end, remote sensing has become increasingly popular. When studying an environment, birds are a popular choice, as they are relatively easy to detect and they are sensitive to small changes in environment at lower trophic levels. Unfortunately, conventional high quality species annotation (or 'tagging') requires the services of an expert and is relatively time intensive. As the time and spatial scales that we are interested in expand, manual annotation becomes prohibitively expensive. Indeed, long-term monitoring quickly produces hundreds of hours of audio recordings. If we want to be able to gain a high resolution picture of what is going on in an ecosystem, then we need a way to automatically annotate recordings efficiently and with little human input.
We present a recognition algorithm that will form the heart of a comprehensive long-term bird monitoring workflow. We briefly talk about how we intend to deploy the workflow via a web application server.

Reconstruction of haplotypes with different frequencies from next-generation sequencing datasets

High Throughput Sequencing II

Thomas K F Wong, Louis Ranjard, Allen Rodrigo
Research School of Biology, Australian National University

Next-generation sequencing (NGS) provides a good opportunity to perform deep resolution on haplotype sequences. However, given a mixture of NGS reads from different haplotypes, reconstruction of each haplotype sequence is still a big challenge, especially when there are many (say > 10) haplotypes. In this talk, we will look at a subset of the problem where all the haplotypes have different frequencies in the pool of sequenced DNA. We will propose a novel algorithm to predict the number of haplotypes, and to reconstruct each of the haplotype sequences. Results obtained from simulated datasets showed that the method is promising when the number of haplotypes is up to 20.

Time reversible Lie-Markov DNA mutation models

Models

Michael Woodhams[1], Barbara Holland[1], David Liberles[2], Michael Charleston[1], Jeremy Sumner[1]
[1]School of Physical Sciences, University of Tasmania; [2]University of Wyoming

Time reversible (TR) DNA mutation models have been popular for decades. A more recent development is the Lie-Markov (LM) models, whose defining property that their Markov matrices are closed under multiplication. Most TR models are not LM, and most LM models we have studied are not TR, but there is some overlap - models which are simultaneously LM and TR. The Lie-Markov property is permissive enough to allow very many models, most of which would strike a biologist as arbitrary and not useful. So far our strategy to bring order to this chaos is to insist on symmetries in our LM models (in particular, we allow transitions to behave differently from transversions, but insist all transitions be on an equal mathematical footing with each other, and similarly all transversions.) In this work, we use time reversibility as a property to select a small but interesting set of LM models. We seek to fully characterize the set of models which are simultaneously LM and TR. LM models are a subset of linear models, and we have now fully characterized models which are both TR and linear. Within these linear TR models we have found many, but not all, LM models.

Lie-Markov DNA models are superior to time reversible models when evolution is heterogeneous

Poster

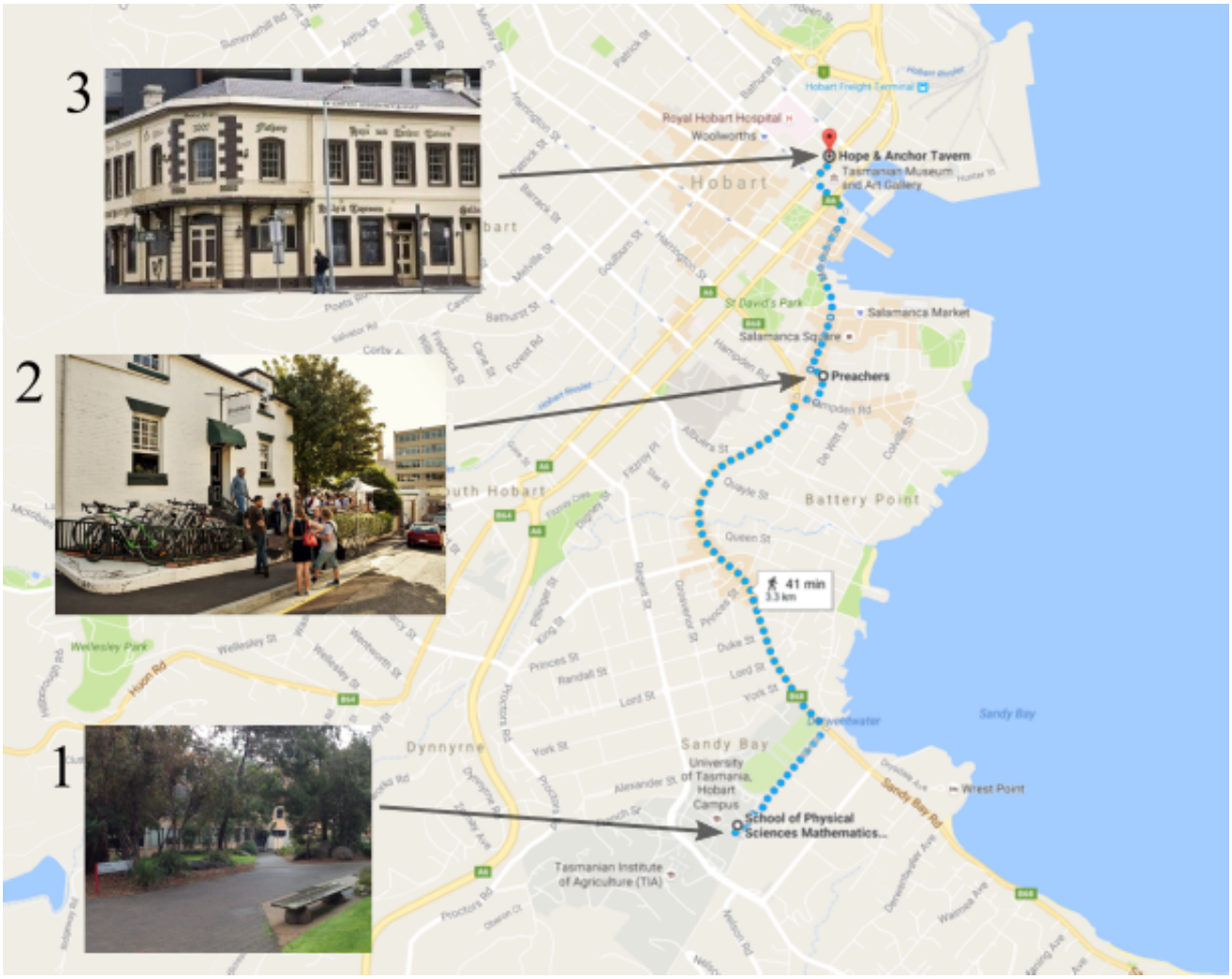
Michael Woodhams, Barbara Holland, Michael Charleston, Jeremy Sumner
School of Physical Sciences, University of Tasmania

The Lie Markov models are DNA mutation models with the property that for any two Markov matrices within the model, their product is also a Markov matrix within the model. This property is useful for analysing non-homogeneous DNA evolution, where model parameters can vary between branches. Without this closure property, non-homogeneous analysis (where we allow our fitted model to vary parameters across the tree) is mathematically inconsistent. DNA models currently in use are almost always time reversble (TR). While some simple models are both TR and Lie Markov, the more complex TR models (e.g. HKY, GTR) are not Lie Markov, and the more complex Lie Markov models are not TR. We present a heirarchy of 37 Lie Markov models which respect the distinction between transitions and transversions. We also present new Monte Carlo results demonstrating the superiority of Lie Markov models over TR models in accurately reconstructing phylogenies in the presences of model inhomogeneity.

Attendees

Nick Beeton	nick.beeton@utas.edu.au
Conrad Burden	conrad.burden@anu.edu.au
Chris Burridge	chris.burridge@utas.edu.au
Jakob Butler	jakob.butler@utas.edu.au
Yao-ban Chan	yaoban@unimelb.edu.au
Michael Charleston	michael.charleston@utas.edu.au
Markus Fleischauer	markus.fleischauer@uni-jena.de
Mathieu Fourment	mathieu.fourment@uts.edu.au
Rhiannon Gray	rhiannon@banach.space
Michael Hendriksen	m.hendriksen91@gmail.com
John Hewson	hewson.john@gmail.com
Steve Higgins	steven.higgins@otago.ac.nz
Barbara Holland	barbara.holland@utas.edu.au
Gavin Huttley	gavin.huttley@anu.edu.au
Peter Jarvis	peter.jarvis@utas.edu.au
Rebecca Jones	rebecca.jones@utas.edu.au
Greg Jordan	greg.jordan@utas.edu.au
Ben Kaehler	benjamin.kaehler@anu.edu.au
Matt Larcombe	matt.larcombe@otago.ac.nz
Mingxin Liu	mingxin.liu@utas.edu.au
Bennet McComish	bennet.mccomish@utas.edu.au
Jonathan Mitchell	jm06@utas.edu.au
Malgorzata O'Reilly	malgorzata.oreilly@utas.edu.au
Louis Ranjard	louis.ranjard@anu.edu.au
John Rhodes	j.rhodes@alaska.edu
Allen Rodrigo	allen.rodrigo@anu.edu.au
Ben Rutherford	benr803@gmail.com
Julia Shore	julia.shore@utas.edu.au
Paul Slade	paul.slade@anu.edu.au
Tristan Stark	tristanstark@gmail.com
Dorothy Steane	dorothy.steane@utas.edu.au
Jorden Stevenson	jordens@utas.edu.au
Jeremy Sumner	jsumner@utas.edu.au
Venta Terauds	venta.terauds@utas.edu.au
Arndt von Haeseler	arndt.von.haeseler@univie.ac.at
Scott Whitemore	scott.whitemore@utas.edu.au
Thomas Wong	thomas.wong@anu.edu.au
Michael Woodhams	michael.woodhams@utas.edu.au

Local Map



Featuring

1. Maths & Physics building, where all the action is happening (view from central concourse);
2. Preachers' Wine Bar, 5 Knopwood Street, where a lot of discussion is going to happen;
3. The Hope and Anchor Tavern (their website), where the Conference Dinner will happen

Index