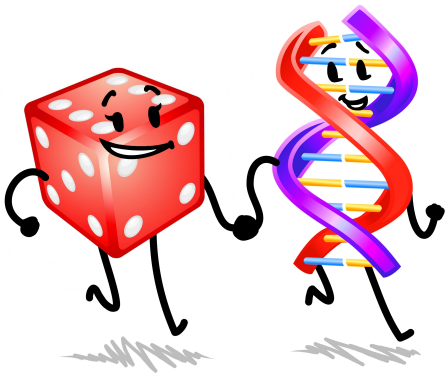
Stochastic Modelling meets Phylogenetics

Collaborative Workshop

University of Tasmania, 16-18 November 2015, Hobart



SPONSORS:

UTAS Career Development Scholarship

ARC Centre of Excellence for Mathematical and Statistical Frontiers

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

The aim of the workshop is to establish a *cross-disciplinary collaboration* in *stochastic modelling* and *phylogenetics*. Stochastic modelling is a key theoretical area of research used in the applications in phylogenetics. Therefore, knowledge of the current advancements in the stochastic models is an advantage for researchers pursuing phylogenetics problems. Similarly, stochastic modelling advancements are stimulated by the real-life problems of significance. Therefore, learning about the needs of the modern phylogenetics is of interest to stochastic modellers. This workshop aims to facilitate training and innovation. Key activities:

1. Overview of the classic stochastic models, with the focus on the recent advancements in the modern literature.
2. Overview of the current big problems and modelling needs of phylogenetics, with the focus on the limitations of the existing models.
3. Discussions and brainstorming of the research ideas, with the focus on the development of a collaborative project encompassing the two research areas:
   (i) stochastic models that need to be constructed and analysed in order to address the modern needs of phylogenetics,
   (ii) phylogenetics methods that could be built on the current unexplored application potential of the existing stochastic models.

Guidelines: As in any collaborative research environment, please respect the origin of research ideas contributed at this workshop.

SMMP Workshop Organizing Committee:

Dr **Małgorzata O'Reilly**

A/Prof. **Barbara Holland**

A/Prof. **Michael Charleston**

Dr **Jeremy Sumner**

A/Prof. **Peter Jarvis**

A/Prof. **Greg Jordan**

Special thanks to:

Ms **Karen Bradford**, *Executive Officer at the School of Physical Sciences*

for fine tuning the administrative details of this workshop.

Prof. **Nigel Bean** *Chair of Applied Mathematics, The University of Adelaide*

Nigel is Professor and Chair of the Applied Mathematics at the University of Adelaide. He is a Chief Investigator of the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). He graduated from the University of Adelaide in 1988 and then went to Cambridge to study for his PhD in stochastic modelling of telecommunication networks under the supervision of Prof Frank Kelly FRS. In 1993 he returned to the University of Adelaide where he has variously been a PostDoc, Lecturer, Director of TRC Mathematical Modelling, and now Professor and Head of Applied Mathematics. Nigel's research has been recognised through the awards of the J.H. Mitchell Medal by ANZIAM and the P.A.P. Moran Medal by the Australian Academy of Science.

Dr **Sophie Hautphenne** *Research Fellow, University of Melbourne*

Sophie is a Research Fellow at the School of Mathematics and Statistics, University of Melbourne, and a Scientist at the Chair of Statistics, Ecole polytechnique fédérale de Lausanne. Since 2015, she is holding an ARC DECRA fellowship at the University of Melbourne. She obtained a PhD in Mathematics from the Université libre de Bruxelles in October 2009. Her fields of research are applied probability and stochastic modelling with a particular focus on branching processes, matrix analytic methods and epidemic models.

Prof. **Mike Steel** *Professor of Maths and Statistics, University of Canterbury*

Mike is a professor of mathematics and statistics and the director of the Biomathematics Research Centre at the University of Canterbury in Christchurch, New Zealand. He is known for his research on modeling and reconstructing evolutionary trees. Mike studied at the University of Canterbury, earning a bachelor's degree in 1982, a masters in 1983, and a degree in journalism in 1985. He then moved to Massey University, where he received his PhD in 1989. He joined the Canterbury faculty in 1994. Mike won the Hamilton Memorial Prize of the Royal Society of New Zealand in 1994; this prize is given annually to a New Zealand mathematician for work done within five years of a PhD. In 1999 he won the research award of the New Zealand Mathematical Society "for his fundamental contributions to the mathematical understanding of phylogeny, demonstrating a capacity for hard creative work in combinatorics and statistics and an excellent understanding of the biological implications of his results."

Prof. **Peter Taylor** *Australian Laureate Fellow, University of Melbourne*

Peter is Professor of Operations Research at the University of Melbourne. He is the Acting Director of the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and a member of Centre for Ultra Broadband Information Networking (CUBIN) in the Department of Electrical and Electronic Engineering. Peter is a former director of the Teletraffic Research Centre and Adstat Solutions at the University of Adelaide. He received a BSc (Hons) in 1980 and a PhD in Applied Mathematics in 1987 from The University of Adelaide. Peter is an internationally-recognised applied mathematician, who specialises in modelling randomly-varying systems. He is a 2013 ARC Australian Laureate Recipient. This ARC fellowship is awarded to outstanding researchers of international repute.

Nigel Bean, University of Adelaide, nigel.bean@adelaide.edu.au

Rhys Bowden, University of Melbourne, rhysbowden@gmail.com

Peter Braunsteins, University of Adelaide, p.braunsteins@student.unimelb.edu.au

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

Jarrad Clark, University of Tasmania, jarradc@utas.edu.au

Axiom Dowling, Honours student, University of Tasmania, axiomd@utas.edu.au

Andrew Francis, University of Western Sydney, a.francis@uws.edu.au

Randall Gray, University of Tasmania, rcgray@utas.edu.au

Rhiannon Gray, University of Tasmania, rgray1@utas.edu.au

Sophie Hautphenne, University of Melbourne, sophiemh@unimelb.edu.au

John Hewson, University of Tasmania, tjhewson@utas.edu.au

Barbara Holland, University of Tasmania, barbara.holland@utas.edu.au

Lucas Hyland, University of Tasmania, Lucas.Hyland@utas.edu.au

Sarah James, University of Adelaide, sarah.james@student.adelaide.edu.au

Peter Jarvis, University of Tasmania, peter.jarvis.edu.au

Greg Jordan, University of Tasmania, greg.jordan@utas.edu.au

Arwin Kahlon, University of Tasmania, kahlon.arwin@gmail.com

Cameron Kirk, University of Tasmania, ckirk@utas.edu.au

Nicholas Matzke, Australian National University, matzke@nimbios.org

Jonathan Mitchell, University of Tasmania, Jonathan.Mitchell@utas.edu.au

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

Damien Palmer, University of Tasmania, damien.palmer@utas.edu.au

Angus Reynolds, University of Tasmania, Angus.Reynolds@utas.edu.au

Sandie Robertson, University of Tasmania, sandier@postoffice.utas.edu.au

Adam Benjamin Rohrlach, University of Adelaide, adam.rohrlach@adelaide.edu.au

Aviva Samuelson, University of Tasmania, aviva.samuelson@utas.edu.au

Barbara Schonfeld, University of Tasmania, barbara.schonfeld@utas.edu.au

Julia Shore, University of Tasmania, jaandrew@utas.edu.au

Tristan Stark, University of Tasmania, tristan.stark@utas.edu.au

Mike Steel, University of Canterbury, mike.steel@canterbury.ac.nz

Jeremy Sumner, University of Tasmania, jeremy.sumner@utas.edu.au

Peter Taylor, University of Melbourne, taylorpg@unimelb.edu.au

Erin Trainer, University of Tasmania, etrainer@utas.edu.au

Scott Whitemore, University of Tasmania, scottw5@postoffice.utas.edu.au

Johanna Wilson, University of Tasmania, J.E.Wilson@utas.edu.au

Michael Woodhams, University of Tasmania, Michael.Woodhams@utas.edu.au

All talks in Lecture Theatre 2. Allow for a discussion at the end of each talk.

8:30–9:00 Early morning tea & coffee, room 328

9:00–9:15 Opening: Małgorzata O'Reilly, Barbara Holland

9:15–10:00 Barbara Holland
Why Phylogenetics?

10:00–10:30 Greg Jordan,
What I wish I knew...

10:30–11:00 Morning tea, room 328

11:00–12:00 Invited speaker: **Mike Steel**
Phylogenetics part I: Evolution on a tree

12:00–12:30 Michael Charleston
Simulation in Phylogenetics

12:30–13:00 Nicholas Matzke
Putting Evolution Into Ecological Niche Modelling

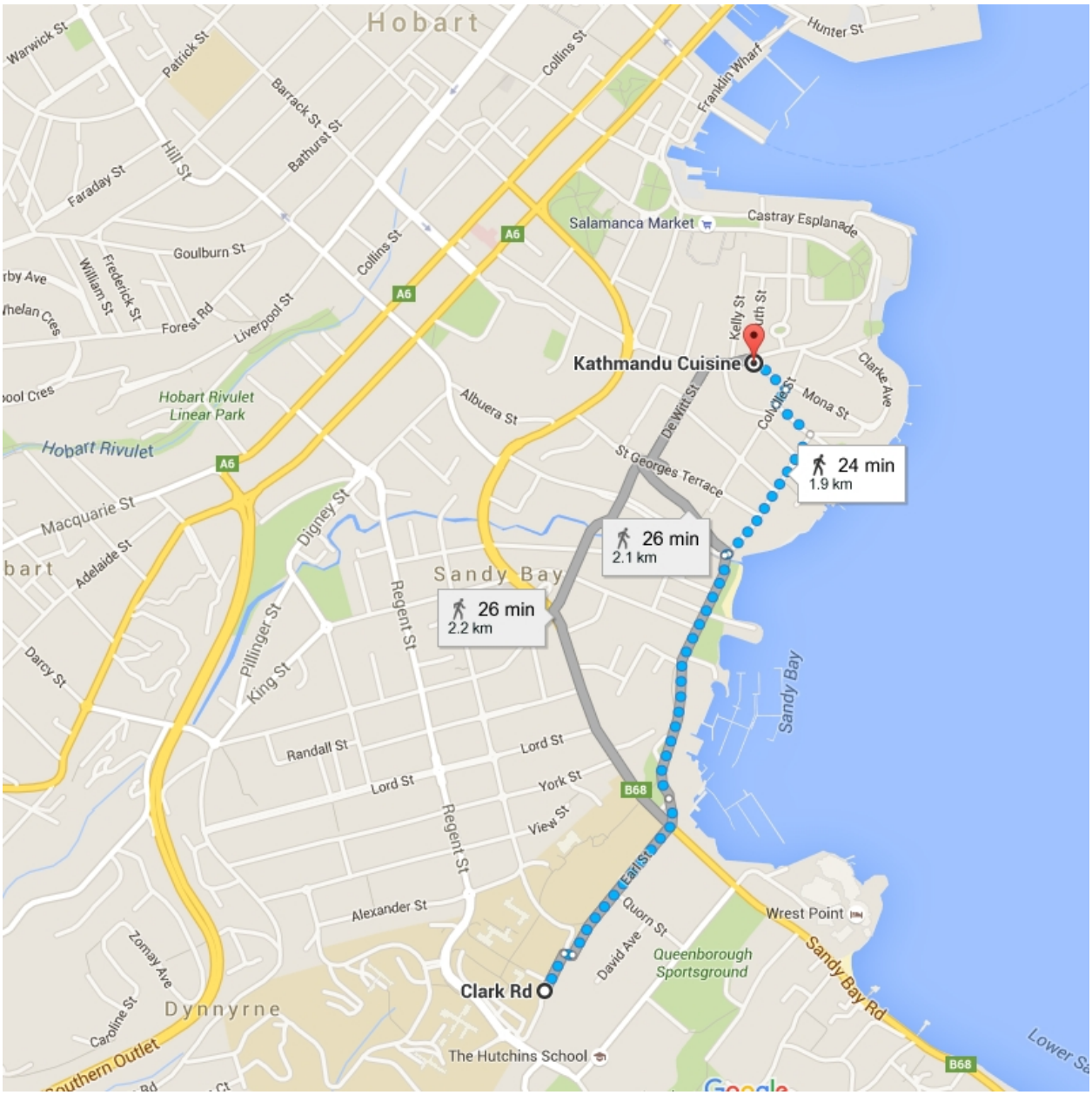
13:00–14:00 Lunch, room 328

14:00–15:00 Invited speaker: **Peter Taylor**
The Role of Physical Understanding in Matrix-Analytic Methods, Part 1

15:00–15:30 Peter Jarvis
Maximum Likelihood and Quantum Random Walks

15:30–16:15 Collaborative discussion:
Limitations of the existing models for Phylogenetics
Chair Barbara Holland
room 333

18:30 Workshop dinner at *Kathmandu Cuisine*
22 Francis St, Battery Point TAS 7004, ph: (03) 6224 8800
`http://www.kathmanducuisine.com.au/`
(see the map attached below)



Directions:

Walk north-east on Clark Rd, 110 m

Turn right towards Earl St, 23 m

Turn left onto Earl St, 500 m

Continue onto Marieville Esplanade, 650 m

Turn right onto Quayle St, 10 m

Continue onto Napoleon St, 350 m

Turn left onto Trumpeter St, 120 m

Turn right onto Colville St, 46 m

Turn left onto Francis St

Destination will be on the left

All talks in Lecture Theatre 2. Allow for a discussion at the end of each talk.

8:30–9:00 Early morning tea & coffee, room 328

9:00–10:00 Invited speaker: **Mike Steel**
Phylogenetics part II: Evolution of trees

10:00–10:30 Michael Charleston
Computational Methods for Tree Search

10:30–11:00 Morning tea, room 328

11:00–12:00 Invited speaker: **Peter Taylor**
The Role of Physical Understanding in Matrix-Analytic Methods, Part 2

12:00–13:00 Invited speaker: **Nigel Bean**
Stochastic Fluid Models (SFMs) part I

13:00–14:00 Lunch, room 328

14:00–15:00 Invited speaker: **Sophie Hautphenne**
An introduction to Branching Processes

15:00–15:30 Peter Braunsteins
Extinction probabilities of branching processes with infinitely many types
PhD Student talk

15:30–16:00 Rhys Bowden
Tree Inference in Computer Network Tomography

16:00–16:30 Afternoon tea, room 328

All talks in Lecture Theatre 2. Allow for a discussion at the end of each talk.

8:30–9:00 Early morning tea & coffee, room 328

9:00–10:00 Małgorzata O'Reilly
Stochastic Fluid Models part II

10:00–10:30 Aviva Samuelson
The importance of choosing the correct model: Stochastic Inspection Model
PhD student talk

10:30–11:00 Morning tea, room 328

11:00–12:00 Invited speaker: **Sophie Hautphenne**
Markovian Binary Trees: definition, computation, and parameter estimation

12:00–13:00 Jeremy Sumner
Lie-Markov Models

13:00–14:00 Lunch, room 328

14:00–15:00 Invited speaker: **Nigel Bean**
Matrix Exponential Distribution

15:00–15:30 Tristan Stark
Modelling the Fate of Gene Duplicates using Phase-Type Distribution
PhD student talk

15:30 Closing: Małgorzata O'Reilly, Barbara Holland

15:45 SMMP Organizing Committee & Collaborators:
Research planning meeting
Chair Małgorzata O'Reilly
room 333

Barbara Holland *School of Physical Sciences, University of Tasmania*

Title: *Why Phylogenetics?*

Phylogenies, descriptions of the evolutionary relationships amongst a set of taxa, underlie almost all aspects of evolutionary biology.

- They can be used to define taxonomies.
- They help us make conservation decisions on the basis of trying to preserve phylogenetic diversity.
- They inform us about the tempo and mode of evolution. For instance, did bird and mammal species really radiate to fill environmental niches left by the dinosaurs, or is this "a Marxist view of evolution that denies a role for competition"? Is New Zealand a "Moa's Ark" that drifted away from Gondwana and spent the next 80 million years evolving unique species or is it rather "The Fly-paper of the Pacific" accepting any old species that blows in from Australia?
- Via ancestral state reconstructions they give us a view of the past that is complementary to the fossil record - for instance they can be used to answer the classic question "What came first - the chicken or the egg?".
- Phylogenies also turn up in less obvious places such as the study of languages, old manuscripts, models of tumour development, or even computer viruses.

Equally startling is the range of mathematics that turns up in phylogenetic studies: stochastic models (of course), combinatorics, probability theory, statistics, algebra, algorithm design and theoretical computer science have all played a role. This talk will give a tour of some phylogenetic applications and different kinds of datasets, and introduce some of the mathematical questions that arise.

Greg Jordan *School of Biological Sciences, University of Tasmania*

Title: *What I wish I knew*

Some important unanswered questions in phylogeny pivot around to a consequence of most applications of phylogeny – that the past is a subset of the present. Current phylogenetic methods only map increasing diversity (i.e. clades can only stay the same or get bigger), but many clades have passed through periods in which they become smaller. Similarly, current methods reconstruct ancestors that are within the range of current samples, but the ancestors can have been well outside the range of the extant clade (think dinosaurs here). These problems are likely to have arisen because current methods of phylogenetic reconstruction and ancestral state reconstruction making unrealistic assumptions about extinction and the speciation process. This seems like a fertile field for modelling – can we better identify the signals of extinction and changes in rates of state change in phylogeny, and can we alter our search methods to take this into account?

I will talk about some examples that have been particularly problematic for me, and plead for some tolerance in my search for answers to the questions that arise. I would also like to argue that we broaden our view of what kind of evidence we should use in reconstructing phylogenies, evolutionary parameters and ancestral states.

Mike Steel *Research Fellow, Australian National University*

Title: *Phylogenetics I: Evolution on a tree*

Stochastic models are now central to the business of inferring phylogenetic trees from genetic data, and for analysing trees (e.g. estimating ancestral states). These models are typically tree-based Markov processes, or mixtures of such processes. In this talk, I will provide an overview of some classical and more recent results concerning the properties of models that are required for accurate tree reconstruction and analysis. Information-theoretic questions, such as how much data is needed, and how far back in time one can reliably infer evolutionary signal will also be addressed. Causes of statistical inconsistency will also be discussed.

Michael Charleston *School of Physical Sciences, University of Tasmania*

Title: *Simulation in Phylogenetics*

Phylogenetic estimation from molecular sequences is made difficult because for almost all real data we have no way of looking back in time to see if we're right. The few exceptions to this are when we have fossil records, from which we might be able to extract phylogenetic information, or when we are lucky enough to have access to molecular data that was extracted during the formation of the tree, such as is commonly done with monitoring the evolutionary dynamics of viruses or bacteria. So, how do we tell how good our methods are?

Computer simulation enables researchers to design experiments and generate data that can be highly simplistic, or very complex, under very tightly controlled conditions. Such data can be as simple as binary characters that flip every fixed time interval, through to much more complex modes of evolution such as considering parameter–rich transition matrices with some sites evolving faster than others, and horizontal gene transfer, and beyond. These data can be provided to phylogenetic estimation methods to see how well, or how badly, they do in recovering the tree that underlay the generation of the data. The speed of computer simulation also means that in principle many different conditions can be exhaustively tested, revealing perhaps areas of strength for some methods and areas of weakness for others.

Simulation has therefore become a central part of the assessment of phylogenetic methods.

This short talk will introduce general approaches to simulating phylogenetic data, such as trees, distances, and sequences. It will help researchers understand the kinds of assumptions that are made, and how the performance of phylogenetic methods can be measured.

Nicholas Matzke *Department of Mathematics and Statistics, University of Canterbury*

Title: *Putting Evolution Into Ecological Niche Modelling*

Traditional Ecological Niche Modelling (ENM) correlates species occurrence data with environmental variables like temperature and precipitation, and uses the ENM to predict potential or actual species distributions (a Species Distribution Model, SDM). ENMs are typically fit one-species-at-a-time, effectively assuming species are independent. This is a peculiar choice if we suspect that species are likely to have phylogenetically autocorrelated environmental responses. I propose Phylogenetic ENM, where the parameters of ENMs for each of a group of related species are jointly estimated along with the parameters of evolutionary models describing the evolution of these niche parameters. I present a preliminary implementation of the phyloENM using JAGS and R, and assess the likelihood that phyloENM will ameliorate some of the issues that beset standard ENM methods, such as overfitting and poor extrapolation ability.

Peter Taylor *School of Mathematics and Statistics, University of Melbourne*

Title: *The Role of Physical Understanding in Matrix-Analytic Methods, Part 1.*

Matrix-analytic methods are used to analyse Markov chain models in which the transition rates of some process of interest, for example a queueing process, are influenced by the dynamics of an underlying 'hidden' Markov chain.

Since Marcel Neuts first showed in the 1970s and 1980s that a particular class of such models, Markov chains of GI/M/1 type, have a matrix-geometric stationary distribution, the interplay between analytic properties and physical interpretation has played a major part in the development of matrix-analytic methods.

Most performance measures of interest in such models can be expressed in terms of the solutions of equations involving matrix power series, which have to be solved numerically. Over the years, various iterative algorithms have been proposed for doing this. In order to establish convergence, and gain information about the speed of convergence, it is often helpful to think about the physical interpretation of the iterates.

I shall discuss the physical interpretation of matrix-analytic algorithms that have been proposed for analysing the simplest class of block-structured Markov additive models, the quasi-birth-and-death processes. *In the first talk, I shall present the necessary background and discuss the physics of the first, linearly-convergent, algorithms.*

Peter Jarvis *School of Physical Sciences, University of Tasmania*

Title: *Maximum Likelihood and Quantum Random Walks*

The last few years have seen the first commercial products exploiting quantum science and its engineering potential, like secure coding systems. As improbable as it seems, there may indeed be prospects ahead for using these technologies for implementing "quantum simulation of stochastic systems". These will go faster, further, better than conventional computation can – a game changer in the making.

Against this backdrop, the talk will investigate a standard tool of probability inference – the likelihood function – in the context of a toy model of a quantum random walk on the line.

Mike Steel *Department of Mathematics and Statistics, University of Canterbury*

Title: *Phylogenetics II: Evolution of trees*

The process of speciation and extinction generates a species tree which biologists attempt to estimate from present-day data (more precisely they estimate the 'reconstructed tree' where extant lineages are usually absent). Since the classical paper of G.U. Yule in 1925, to more recent studies (e.g. Lambert and Stadler 2013) stochastic models for generating birth-death trees have played an important role in statistical phylogenetics. For example, biologists would like to know what the 'shape' of inferred trees might tell us about the processes of speciation and extinction. Also (looking forward in time) one can attempt to predict how much 'evolutionary heritage' might be lost from the tree due to the current high rates of extinction. A further topical issue is that gene trees do not always follow the species tree, due to 'incomplete lineage sorting' and 'lateral gene transfer'. Stochastic models for these two processes of evolving random (gene) trees based on a species tree have led to new insights and techniques for inferring species phylogenies.

Michael Charleston *School of Physical Sciences, University of Tasmania*

Title: *Computational Methods for Tree Search*

The number of trees is very large!! In fact as n, the number of tips, increases, the number of trees increases more than exponentially. (There are $(2n-3)!! = (2n-3)(2n-5)\ldots 3 \cdot 1$ rooted binary trees, for instance.) The huge size of the number of trees alone makes the search for optimal ones very hard; and then if there are more parameters of interest that matter for any tree, such as the lengths of its branches or the proportion of sites that are evolving "fast" or "slow", the problem of finding the best tree or trees becomes frustratingly difficult.

Since we cannot hope to check every tree for optimality, we find a couple of options present themselves. For moderate numbers of species, we can use branch and bound methods to implicitly check sets of trees, ruling out those that cannot be optimal without having to check them explicitly.

When even that is not possible, we can use heuristics to search through the space of trees, and accept that we cannot guarantee that the best tree we find is indeed optimal. Heuristics are generally fast, and usually perform well, but we always have that caveat that we cannot guarantee global optimality.

Finally, we can collect information about the optimal tree in some other way. Markov chain Monte Carlo methods let us search through the tree space in such a way that, if we kept going forever, we would get an accurate picture of the relative "goodness", e.g., in terms of their likelihood, of every tree. This can then be used to calculate the relative likelihoods of particular hypotheses of interest, subject to our prior assumptions, such as when the main radiation of the birds occurred, or how many sites in our sequences are evolving "fast".

This talk introduces the main techniques of finding trees: construction, heuristic search, branch and bound, and MCMC.

Peter Taylor *School of Mathematics and Statistics, University of Melbourne*

Title: *The Role of Physical Understanding in Matrix-Analytic Methods, Part 2.*

Matrix-analytic methods are used to analyse Markov chain models in which the transition rates of some process of interest, for example a queueing process, are influenced by the dynamics of an underlying 'hidden' Markov chain.

Since Marcel Neuts first showed in the 1970s and 1980s that a particular class of such models, Markov chains of GI/M/1 type, have a matrix-geometric stationary distribution, the interplay between analytic properties and physical interpretation has played a major part in the development of matrix-analytic methods.

Most performance measures of interest in such models can be expressed in terms of the solutions of equations involving matrix power series, which have to be solved numerically. Over the years, various iterative algorithms have been proposed for doing this. In order to establish convergence, and gain information about the speed of convergence, it is often helpful to think about the physical interpretation of the iterates.

I shall discuss the physical interpretation of matrix-analytic algorithms that have been proposed for analysing the simplest class of block-structured Markov additive models, the quasi-birth-and-death processes. *In the second talk I will look at the behaviour of the more advanced, quadratically-convergent, algorithms that are currently used. A recurring theme throughout both talks will be the use of different types of censored Markov chains.*

Nigel Bean *School of Mathematical Sciences, University of Adelaide*

Title: *Stochastic Fluid Models (SFMs) part I*

Continuous-time Markov Chains (CTMCs) are the key class of stochastic models used to analyse the evolution of industrial, environmental and biological systems. We use CTMCs to model the transitions between the various states of an underlying physical environment of interest. Stochastic Fluid Models (SFMs) extend this modelling potential by including an extra variable in the model, referred to as the level, which is used to model some continuous performance measure of the system. The literature on the standard SFMs, which are one-dimensional in level (i.e. have one level variable), is well-developed and includes theoretical results and algorithms for the stationary (long-run) as well as transient (time-dependent) analysis.

In this talk I will provide the necessary background to modelling SFMs, focussing on the physical interpretation of the key results and algorithms.

Sophie Hautphenne *School of Mathematics and Statistics, University of Melbourne*

Title: *An introduction to Branching Processes*

Branching processes are powerful stochastic models that describe the evolution of populations of individuals which reproduce and die independently of each other according to specific probability laws. They play an increasingly important role in models of population biology including molecular biology, ecology, epidemiology, and evolutionary theory. Branching processes are also used in other scientific areas, for instance in particle physics, in chemistry, and in computer science. Typical performance measures of these models include the extinction probability of a population, the distribution of the population size at a given time, the total progeny size until extinction, and the asymptotic population composition.

This first tutorial by Sophie will focus on the basic theory of discrete-time Galton-Watson branching processes and continuous-time Markovian branching processes.

Peter Braunsteins *School of Mathematics and Statistics, University of Melbourne*

Title: *Extinction probabilities of branching processes with infinitely many types*

We present some iterative methods for computing the global and partial extinction probability vectors for branching processes with countably infinitely many types. The probabilistic interpretation of these methods involves truncated branching processes with finite sets of types and modified progeny generating functions. Simple probabilistic arguments and coupling methods are used to prove the convergence of the algorithms.

Rhys Bowden *School of Mathematics and Statistics, University of Melbourne*

Title: *Tree Inference in Computer Network Tomography*

Network tomographic topology inference deals with estimating the internal tree structure of a computer network from end-to-end probes through that network. Probes are sent from a single source to multiple destinations at each time point; some probes reach their destination and some are lost, and this information can be used to infer the tree graph. This problem bears many similarities to the inference of phylogenetic trees from DNA sequences: sites in the genome correspond to times when a probe is sent, and changes in the base sequence over time correspond to probe losses.

Previous work has relied upon an assumption that losses on differing links in the tree are independent. I weaken this assumption and show that while it is no longer possible to determine the sequences at internal nodes (even in the limiting case) it is still possible to identify the connectivity structure of the tree.

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

Title: *Stochastic Fluid Models, part II*

Continuous-time Markov Chains (CTMCs) are the key class of stochastic models used to analyse the evolution of industrial, environmental and biological systems. We use CTMCs to model the transitions between the various states of an underlying physical environment of interest. Stochastic Fluid Models (SFMs) extend this modelling potential by including an extra variable in the model, referred to as the level, which is used to model some continuous performance measure of the system. The literature on the standard SFMs, which are one-dimensional in level (i.e. have one level variable), is well-developed and includes theoretical results and algorithms for the stationary (long-run) as well as transient (time-dependent) analysis.

My recent interest has been in extending the modelling potential of the SFMs even further, by considering possibilities such as having more than one level variable, or letting the parameters of the SFMs vary in time. In this talk, I will describe the recent advances in the area of two-dimensional SFMs, and SFMs with cyclic parameters, which offer powerful modelling ability for a wide range of real-life systems of significance.

Aviva Samuelson *School of Physical Sciences, University of Tasmania*

Title: *The importance of choosing the correct model: Stochastic Inspection Model*

Models rely heavily on the assumptions used to create them. If one or more of these assumptions are incorrect, the results obtained may be wildly inaccurate. In this talk, I will illustrate the importance of choosing the correct model by comparing two alternative models for deteriorating systems. Both models belong to a class of stochastic fluid models in which the parameters may change depending on the deterioration level of the system. The key difference between the two models is that the first model assumes that we have perfect information about the deterioration level at all times. The second model on the other hand, which we construct in order to offer a more realistic approach, assumes that we know the deterioration level only at the times when we inspect it, for instance during maintenance.

We construct a numerical example, in which we evaluate two alternative strategies for the maintenance of a machine, using the two models. Strategy 1 is to maintain the system more heavily when it is new. Strategy 2 is to maintain the system more heavily when it is old. The results show that when we incorrectly assumed perfect knowledge and used the first model, the least cost effective strategy was chosen. This example illustrates the importance of choosing a model which is appropriate to the assumptions that are inherent in the real-life problem.

Sophie Hautphenne *School of Mathematics and Statistics, University of Melbourne*

Title: *Markovian Binary Trees: definition, computation, and parameter estimation*

The pressing need to allow for both realism and tractability in the same model of branching process has motivated the development of Markovian binary trees (MBTs). MBTs form a particularly versatile and tractable class of Markovian branching processes, which makes them well-suited for real-world applications. In particular, MBTs have proven to be an excellent modelling tool for applications in population biology and phylogenetics.

This second tutorial by Sophie will focus on the definition and basic properties of Markovian binary trees, as well as some of their applications and parameter estimation methods.

Jeremy Sumner *School of Physical Sciences, University of Tasmania*

Title: *Lie-Markov Models*

Continuous-time Markov models are the work-horse of modern molecular phylogenetics. Underlying these models is a choice of constraints on the free parameters specifying the (unknown) nucleotide substitution rates. As usual, standard statistical arguments regarding parameter estimation come to the fore when choosing the "best" model for a given phylogenetic data set. For example, the standard bias-variance tradeoff dictates that the "best" model depends crucially on the amount and quality of the data under examination. Consequently, the history of phylogenetic model development has ranged from the development of very simple models (all substitution rates equal) to maximally complex models (no constraints at all). Of additional interest is the assumption that the Markov model is homogeneous in time. This is of course a gross simplification as the time-history of molecular substitution pressures is unlikely to have been constant in time. Nonetheless, this assumption is (usually) unavoidable given the difficulty in using a finite amount of present-day data to recover such complicated signals. It is then natural to explore the possibility of phylogenetic models that are at least consistent (in a certain sense) with the possibility of changing substitution rates. *Our recent work on "Lie-Markov models" explores exactly this territory of homogeneous Markov models that are consistent with (the possibility of) a time-heterogeneous substitution process. The key observation is the requirement that the transition matrices arising from the model are algebraically closed under matrix multiplication.* I will review multiple aspects of this work:

(i) The definition and algebraic justification for the Lie-Markov models;

(ii) What can go wrong when a model is not Lie-Markov;

(iii) Examples of well-known models which are Lie-Markov and proof that many oft-used models are not;

(iv) Models which are "almost" Lie-Markov;

(v) Models which are "more than" Lie-Markov;

(vi) Models which are "exactly" Lie-Markov;

(vii) Further algebraic properties of Lie-Markov models;

(viii) Strategies for producing complete classifications of Lie-Markov model.

I am particularly interested in discussing the final aspect (viii). Our work to date has focussed on exploiting discrete permutation symmetries to systematically produce Lie-Markov models. However, these symmetries impose more structure than is necessarily required; often resulting in case (v) where the models are "more than Lie-Markov". It is of on-going interest to me to explore other methods for deriving Lie-Markov models.

Nigel Bean *School of Mathematical Sciences, University of Adelaide*

Title: *Matrix Exponential Distribution*

Markov modulated processes are very common modelling tools and include concepts met earlier in this workshop such as Quasi-Birth-and-Death processes, Markov Binary Trees and Stochastic Fluid Models, as well as further concepts such as Markov modulated Brownian motions. The simplest example is the class of Phase-type distributions (PH), whose study underlies all the concepts listed above. However, when considered as a purely abstract mathematical object, there is a clear generalisation to the class known as the Matrix exponential distributions (ME). In this talk I will give a detailed introduction to the properties of the PH distributions and show the generalisation to the ME distributions. I will then give the physical interpretation of the ME distributions and speculate on how this generalisation could be used to enhance the modelling capability of the above class of Markov modulated models.

Tristan Stark *School of Physical Sciences, University of Tasmania*

Title: *Modelling the Fate of Gene Duplicates Using a Phase-Type Distribution*

After duplication, the fate of a duplicate pair of genes is of great interest in genomics. Gene duplication and subsequent loss is believed to contribute to genome diversification. There has been an increasing focus on developing models to describe these processes. Under the subfunctionalization model functions performed by the original gene can eventually be distributed between the two copies, preserving both copies by selective pressure. Alternatively one copy can be lost entirely with the other performing the full set of functions of the original gene.

Computational biologists use phenomenological models in their empirical analysis that show a sigmoidal hazard rate, however to date mathematical models have not shown this sigmoidal behaviour. In our work, by assuming Poisson rates for null mutations occurring in coding and regulatory parts of genes, we construct a CTMC with a phase-type distribution structure, and two absorbing states corresponding to the fates described above. We derive some measures of interest including hazard rates, mean times to absorption, moments and variances of times to absorption. We show that this model behaves qualitatively similarly to the sigmoidal phenomenological models used in practice. By doing so, we provide mathematically rigorous support for subfunctionalization as an important biological mechanism.

Below is a list of recommended further reading on the topics covered in the workshop.

Books:

1. G. Grimmett and D. Stirzaker, Probability and Random Processes, Oxford University Press, 2001.
2. Theodore E. Harris, The Theory of Branching Processes, Dover Publications, New York, 2002.
3. Haccou, Jagers and Vatutin, Branching Processes: Variation, Growth, and Extinction of Populations, Cambridge University Press, 2005.
4. G. Latouche and V. Ramaswami, Introduction to Matrix Analytic Methods, Piladelphia: American Statistical Association and SIAM, 1999.
5. James Norris, Markov Chains, Cambridge University Press, 2004.
6. S. M. Ross, Introduction to Probability Models, Elsevier, New York, 2007.
7. Mike Steel and Charles Semple, Phylogenetics, Oxford University Press, 2003.

PhD Theses:

1. Ana da Silva Soares, Fluid Queues Building Upon the Analogy with QBD Processes, Université Libre de Bruxelles, 2005.
2. Sophie Hautphenne, An Algorithmic Look at Phase-Controlled Branching Processes, Université Libre de Bruxelles, 2009.
3. Nectarios Kontoleon, The Markovian Binary Tree: A Model of the Macroevolutionary Process, The University of Adelaide, 2006.

Articles:

1. Asmussen S., Bladt M. Point processes with finite-dimensional conditional probabilities (1999). *Stochastic Processes and their Applications*, 82(1):127–142.
2. Bean N.G., O'Reilly M.M., Taylor P.G. Hitting probabilities and hitting times for stochastic fluid flows (2005). *Stochastic Processes and their Applications*, 115(9):1530–1556.
3. Bean N.G., O'Reilly M.M. Spatially-coherent uniformization of a stochastic fluid model to a Quasi-Birth-and-Death process (2013). *Performance Evaluation*, 70(9):578–592.
4. Bean N.G., O'Reilly M.M. A stochastic two-dimensional fluid model (2013). *Stochastic Models*, 29(1):31–63.
5. Bean N.G., Kontoleon N., Taylor P.G. Markovian trees: Properties and algorithms (2008). *Annals of Operations Research*, 160(1):31–50.
6. Bladt M., Neuts M.F. Matrix-exponential distributions: Calculus and interpretations via flows (2003). *Stochastic Models*, 19(1):113–124.
7. Chor B., Hendy M.D., Holland B.R., Penny D. Multiple maxima of likelihood in phylogenetic trees: An analytic approach (2000). *Molecular Biology and Evolution*, 17(10):1529–1541.
8. Chor, B., Steel, M. Do tree split probabilities determine the branch lengths? (2015). *Journal of Theoretical Biology*, 374:54–59.
9. C. Daskalakis and S. Roch (2015). Species trees from gene trees despite a high rate of lateral genetic transfer: A tight bound. ArXiv 1508.01962v1.
10. Hautphenne S., Latouche G., Remiche M.-A. Newton's iteration for the extinction probability of a Markovian binary tree (2008). *Linear Algebra and Its Applications*, 428 (11-12):2791–2804.

11. Holland B.R., Delsuc F., Moulton V. Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to study the origins of placentals and hexapods (2005). *Systematic Biology*, 54(1):66–76.

12. Kempe J. Quantum random walks: An introductory overview (2003) . *Contemporary Physics*, 44(4):307–327.

13. Lambert D.M., Ritchie P.A., Millar C.D., Holland B., Drummond A.J., Baroni C. Rates of evolution in ancient DNA from Adélie penguins (2002). *Science*, 295(5563):2270–2273.

14. Lambert, A., and Stadler, T. (2013). Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies Theoret. *Popul. Biol.*, 90:113–128.

15. Lambert, A. and Steel, M. (2013). Predicting the loss of phylogenetic diversity under non-stationary diversification models. *Journal of Theoretical Biology*, 337:111–124.

16. Lockhart, P. J.; Steel, M. A.; Hendy, M. D.; Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11(4):605–612, PMID 19391266.

17. Mossel E. and Roch S. (2105). Distance-based species tree estimation: information-theoretic trade-off between the number of loci and sequence length under the coalescent. ArXiv:1504.05289v1.

18. Ramaswami V., Taylor P.G. Some properties of the rate operators in level dependent quasi-birth-and-death processes with a countable number of phases (1996). Communications in Statistics. Part C: *Stochastic Models*, 12(1):143–164.

19. Roch S. and Sly A. (2015). Phase transition in the sample complexity of likelihood-based phylogeny inference. ArXiv:1508.01964v1.

20. Semple, Charles; Steel, Mike (2003), Phylogenetics, Oxford lecture series in mathematics and its applications 24, Oxford University Press.

21. Steel, M. (2014). Tracing evolutionary links between species. *American Mathematical Monthly*, 121(9):771–792.

22. Sumner J.G., Fernandez-Sanchez, P D Jarvis P.D. Lie Markov models (2012). *Journal of Theoretical Biology*, 298:16–31.