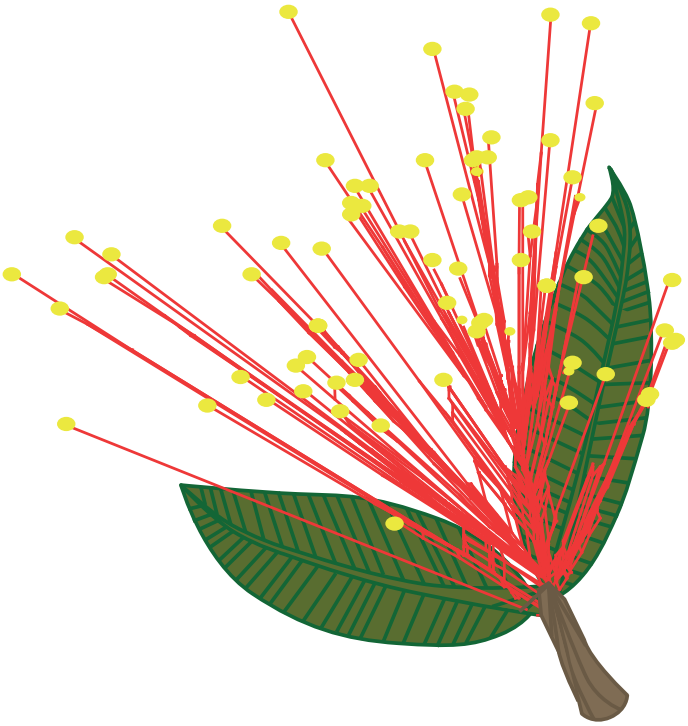

Whitianga '08

NEW ZEALAND PHYLOGENETICS CONFERENCE

10-15 February, 2008
Whitianga, New Zealand

Participants

John Beck	jbeck@truman.edu
David Bryant	d.bryant@auckland.ac.nz
Michael Charleston	mcharleston@it.usyd.edu.au
Alan Cooper	alan.cooper@adelaide.edu.au
Michael Defoin-Platel	michael.defoinplatel@gmail.com
Andreas Dress	andreas@picb.ac.cn
Greg Ewing	gregory.ewing@univie.ac.at
Beata Faller	fallerbeata@yahoo.com
Mareike Fischer	email@mareikefischer.de
Tanja Gernhard	gernhard@ma.tum.de
Gillian Gibb	g.c.gibb@massey.ac.nz
Russell Gray	rd.gray@auckland.ac.nz
Simon Greenhill	s.greenhill@auckland.ac.nz
Klaas Hartmann	kha59@student.canterbury.ac.nz
Joseph Heled	jheled@gmail.com
Mike Hendy	m.hendy@massey.ac.nz
Simon Hills	s.f.hills@massey.ac.nz
Melanie Hingston	mhin027@ec.auckland.ac.nz
Barbara Holland	b.r.holland@massey.ac.nz
Simon Joly	s.joly@massey.ac.nz
David Liberles	liberles@uwyo.edu
Simone Linz	simone_linz@yahoo.de
Peter Lockhart	p.j.lockhart@massey.ac.nz
Sidney Markowitz	sidney@sidney.com
Atheer A Matroud	atheerb@hotmail.com
David Penny	d.penny@massey.ac.nz
Louis Ranjard	l.ranjard@auckland.ac.nz
Alethea Rea	alethea.rea@gmail.com
Howard Ross	h.ross@auckland.ac.nz
Robert Ross	robross45@yahoo.com.au
Raazesh Sainudiin	R.Sainudiin@math.canterbury.ac.nz
Klaus Schliep	k.p.schliep@massey.ac.nz
Charles Semple	c.semple@math.canterbury.ac.nz
Liat Shavit	l.shavit@massey.ac.nz
Andreas Spillner	aspillner@cmp.uea.ac.uk
Mike Steel	m.steel@math.canterbury.ac.nz
Mark Stevens	m.i.stevens@massey.ac.nz
Jeremy Sumner	jsumner@it.usyd.edu.au
Bhalchandra Thatte	bdthatte@gmail.com
Giulia Torricelli	torricelli5@unisi.it
Wai Lok Sibon Li	wli051@ec.auckland.ac.nz

Conference organisers: David Bryant, Alethea Rea, Allen Rodrigo
Graphic: David Bryant, Jamie Kydd

Timetable

Sunday, February 10

19h00	Registration; Wine and Cheese		

Monday, February 11th

9h10	Announcements		
9h15	The joys of being mean.	*Mike Steel*	18
9h40	Maximum likelihood calculations using partial likelihood tensors.	*Jeremy Sumner*	18
10h05	An Equivalence of Maximum Parsimony and Maximum Likelihood revisited	*Mareike Fischer*	8
10h30	*Morning tea*		
11h15	Extreme microevolution in Antarctic micro-arthropods	*Giulia Torricelli*	18
11h40	The Multiscale Nature of Phylogeny: A case study using the New Zealand Pigeon (Kereru)	*Gillian Gibb*	9
12h05	Deciphering the rapid species radiation of the New Zealand alpine rockcress (Pachycladon) species	*Simon Joly*	12
12h30	*Lunch break*		
14h30	Rates curves: causes and solutions	*Alan Cooper*	6
15h00	Comparing models of divergence time estimation	*Michael Defoin-Platel*	7
15h30	*Afternoon tea*		
16h00	On the shape and fabric of human history	*Russell Gray*	10
16h30	Horizontal transmission and cultural phylogenies	*Simon Greenhill*	10
18h00	Fish and chip dinner on the beach.		

Tuesday, February 12th

9h15	Optimizing Diversity with Ecological Constraints	*Beata Faller*	8
9h40	Optimizing Phylogenetic Diversity Across a Forest	*Charles Semple*	17
10h05	Finding subsets of high phylogenetic diversity on special classes of split systems	*Andreas Spillner*	17
10h30	*Morning tea*		
11h15	Snails Through Time: Molecular analysis of New Zealand Marine Gastropods	*Simon Hills*	11
11h40	Detecting and visualizing morphological convergence within the cormorants	*Barbara Holland*	12
12h05	Bird song sequence evolution	*Louis Ranjard*	15
12h30	*Lunch break*		
14h30	An Introduction to Block-Decomposition Theory	*Andreas Dress*	7
15h00	No more biased networks!	*David Bryant*	6
15h30	*Afternoon tea*		
16h00	Lapita - migration throughout near Oceania. Is there evidence for multiple introductions?	*Melanie Hingston*	11
16h30	The Uto-Aztecans: Foragers from the north or farmers from the south?	*Robert Ross*	16

Wednesday, February 13th

Excursion day. Collect packed lunch from the lodge between 8h00 and 9h00.

18h00	*BBQ at the Aotearoa Lodge*

Thursday, February 14th

9h15	Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites	*Peter Lockhart*	13
9h40	Modelling the emergence of genetic coding as a self-organising complex system	*Sidney Markowitz*	13
10h05	Varied accuracy in the prediction of gene function using patterns of co-evolution implicates inconsistency in quality and specificity of Gene Ontology terms	*Wai Lok Sibon Li*	19
10h30	*Morning tea*		
11h15	What can timing information in phylogenetic trees tell us about the speciation process?	*Tanja Gernhard*	9
11h40	Solution to a question of Steel and Hein	*Bhalchandra Thatte*	18
12h05	Reducing clutter in NeighborNet Networks	*Alethea Rea*	15
12h30	*Lunch break*		
14h30	The performance of phylogenetic algorithms in estimating haplotype genealogies	*Greg Ewing*	8
15h00	Kingman's Unlabeled n-Coalescent	*Raazesh Sainudiin*	16
15h30	*Afternoon tea*		
16h00	Sampling trees from evolutionary models	*Klaas Hartmann*	10
16h30	Testing the Reliability of Genetic Methods of Species Identification	*Howard Ross*	15

Friday, February 15th

9h15	Why does my tree look strange? The impact of positive and negative selection on phylogenetic reconstruction	*David Liberles*	12
9h40	The effect on phylogeny of lineage specific structural and functional constraints	*Klaus Schliep*	16
10h05	Lineage Specific Sequence Evolution	*Liat Shavit*	17
10h30	*Morning tea*		
11h00	Bayesian coalescent inference of population history with variable dimensions	*Joseph Heled*	11
11h25	Bias in parametric bootstrapping	*Michael Charleston*	6
11h50	Questions	*David Penny*	14
12h15	*Lunch and departures*		

No more biased networks!

David Bryant

Split decomposition is the 'respected elder' of phylogenetic network methods, but it has problems with crowds and collapses down when there is more than a handful of taxa. The reason is that the criterion used to select splits has a strong negative bias. I'll talk about ways that this bias can be addressed using resampling techniques, providing an (approximately) unbiased and expanded split decomposition.

Bias in parametric bootstrapping

Michael Charleston[†]

Parametric bootstrapping is a widely used method of testing the robustness of phylogenetic inference. It involves simulating data according to a model as estimated from an alignment of molecular sequences, and then re-inferring the best-fit tree for those data; the distribution of such trees enables statistical testing of phylogenies, to answer questions such as "with what probability could this tree be optimal, when the generating tree was that one?" and "are these two trees from different genes on the same taxon set significantly different?"

Even ignoring the axiom that pseudo-samples give pseudo-confidence, sequence simulation and tree inference are thorny issues. Model selection is generally biased, corrections for multiple hits are non-linear, branch lengths are non-independent, and tree space is highly complex.

I will describe some of the effects of this process and the generating tree on branch length estimation, total tree length and tree shape, caused by this process, for both real and simulated data, and discuss some of the ramifications of this bias to phylogenetic inference in real cases.

The evolutionary rates curve: What causes it and can we do anything about it?

Alan Cooper

Recent work has indicated that molecular evolutionary rates appear to exhibit time dependency, in that rate estimates vary according to the time period over which they are measured. This is most noticeable in apparently accelerated short-term rate measurements, and can lead to a considerable over-estimate of divergence times when a molecular clock (relaxed or strict) is used to estimate the timing of recent evolutionary events (eg during the past 1-2 Ma). The estimated rates appear to follow a negative exponential curve, with the most rapid estimates resulting from short-term measurements within families or populations, and the slowest from fossil calibrations.

The issue of rate curves falls at the interface of micro- and macro-evolutionary theory, and is partly explained by changes in terminology and methodological assumptions governing research across this boundary. A curve in rates will always be produced if $d > 0$ at $t = 0$, and possible reasons for this phenomena are explored. One key issue appears to be the essential difference between rate estimates below the species level (which are primarily measuring polymorphism within populations) and the much slower rate recorded above the species level (generally calculated with

a fossil date) which measure the substitution rate, or mutation rate for neutral loci. The latter is the small proportion of polymorphisms that get actually fixed in a species lineage (e.g. by drift, or selective processes) over time. In addition to the effect of polymorphisms, the proportion of saturated sites also appears to play a role in the generation of rate curves.

Several temporally sampled datasets are examined for evidence of rate curves, and different approaches used in an attempt to negate their effect. The results indicate that the problem is widespread, both above and below the species level and that corrections for this behaviour have major impacts on molecular date estimates in the recent past.

Comparing models of divergence time estimation

Michael Defoin-Platel

For inferring divergence time in phylogenetics, it is convenient to assume a constant molecular evolutionary rate over time. This assumption is called the molecular clock hypothesis and provides a means to translate genetic distances into geological times. Deviation from clocklike evolution has been often reported in datasets and the molecular clock hypothesis is therefore violated, particularly when distantly related species are compared, potentially leading to not only an incorrect estimation of species divergence time but also an incorrect inference of phylogenies. In the context of Bayesian phylogenetics reconstruction and divergence date estimation, it is now common to allow every branch to have a different rate of molecular evolution. We propose to review several existing approaches to relax the molecular clock assumption, such as local clock models, auto-correlated and uncorrelated relaxed-clock models. Comparisons are performed using the software Beast for numerical simulations. Finally, the relationship between sequence length and clocklikeness of the data is discussed.

An Introduction to Block-Decomposition Theory

Andreas Dress

Phylogenetic combinatorics deals with the combinatorial aspects of phylogenetic-tree reconstruction. A starting point was the connection between so-called *tight-spans*, trees, splits and additive metrics. In my lecture, I will focus on the rather new developments relating to *block decomposition* of metric spaces reported in (1)...(6) below that allow to canonically decompose any given finite metric space into a sum of pairwise compatible *block metrics* thus providing a far-reaching generalization of the result recalled above.

References

(1) A. Dress, K. Huber, J. Koolen, and V. Moulton, Compatible decompositions and block realizations of finite metric spaces, *Europ. J. Comb.*, in press.

(2) A. Dress, K. Huber, J. Koolen, and V. Moulton, An algorithm for computing virtual cut points in finite metric spaces, Proceedings of COCOA 2007, Lecture Notes in Computing Science.

(3) A. Dress, K. Huber, J. Koolen, and V. Moulton, Cut points in metric spaces, *AML*, in press.

(4) A. Dress, K. Huber, J. Koolen, V. Moulton, and A. Spillner, A note on the metric cut point and

the metric bridge partition problems. submitted.

(5) A. Dress, K. Huber, J. Koolen, and V. Moulton, Block Realizations of Finite Metrics and the Tight-Span Construction I: The Embedding Theorem, submitted.

(6) A. Dress, K. Huber, J. Koolen, and V. Moulton, Block Realizations of Finite Metrics and the Tight-Span Construction II: On Standard Block Realizations of Finite Metrics, in preparation.

The performance of phylogenetic algorithms in estimating haplotype genealogies

Walter Salzburger, Greg Ewing† and Arndt von Haeseler

Genealogies estimated from haplotypic genetic data play a prominent role in various biological disciplines in general and in phylogenetics, population-genetics and phylogeography in particular. Several software packages have specifically been developed for the purpose of reconstructing genealogies from closely related, and, hence, highly similar haplotype sequence data such as mitochondrial DNA. Here, we use simulated datasets with known true topologies to test the performance of traditional phylogenetic algorithms.

Optimizing Diversity with Ecological Constraints

Beáta Faller

Phylogenetic diversity is a measure for describing how much of an evolutionary tree is spanned by a subset of species. Given a taxon set and a corresponding edge-weighted evolutionary tree, a central question in conservation biology is how to choose a fixed number of taxa with maximum diversity. This problem can be solved by the greedy algorithm. However, if we consider a more realistic problem, where we also have an acyclic digraph describing dependencies between taxa, we are only allowed to choose a viable taxon set of given size and the problem becomes more complicated. This talk will discuss the complexity of this latter problem. This is joint work with Magnus Bordewich and Charles Semple.

An Equivalence of Maximum Parsimony and Maximum Likelihood revisited

Mareike Fischer†, Bhalchandra Thatte

The incessantly growing amount of available genetic sequence data requires both stochastic models for nucleotide substition as well as tree reconstruction methods to allow for the inference of phylogenetic trees. Unsurprisingly, such models and methods have therefore been widely discussed in the last decades. Two of the most frequently used tree reconstruction methods are Maximum Parsimony (MP) and Maximum Likelihood (ML), and it is known that these methods sometimes disagree (e.g. in the so-called Felsenstein zone). However, in 1997 Tuffley and Steel carried the analysis of MP and ML an important step further: they proved that under a symmetric multistate model of the evolutionary process, when applied to a single character, MP and ML are equivalent.

In my talk, I will present a short and elementary new proof for this result. Along the way, we combinatorially derive some interesting properties of the likelihood function.

What can timing information in phylogenetic trees tell us about the speciation process?

Tanja Gernhard[†], Erick Matsen, Daniel Ford

In this talk I present a new statistic summarizing the timing of branching events in phylogenetic trees. Our method explicitly considers the relative timing of diversification events between sister clades; as such it is complimentary to existing methods using lineages-through-time plots which consider diversification in aggregate. The method looks for evidence of diversification happening in lineage-specific "bursts", or the opposite, where diversification between two clades happens in an unusually regular fashion. In order to be able to distinguish interesting events from stochasticity, we propose a class of neutral models on trees with timing information and develop a statistical framework for testing these models. Our models substantially generalize both the global-rate speciation-extinction models and the coalescent with ancestral population size variation. I end the talk with an example application: we show that the evolution of the Hepatitis C virus appears to proceed in a lineage-specific bursting fashion.

The Multiscale Nature of Phylogeny: A case study using the New Zealand Pigeon (Kereru)

Gillian Gibb

Evolutionary relationships are studied at a wide range of time scales, though little formal attention is paid to the multiscale nature of the process. In particular, the standard claim of Darwinian evolution that microevolutionary processes are sufficient to explain macroevolution requires consideration of time scales from processes in populations to the origin of major groups. Here we use DNA to integrate studies over the full range from barcoding haplotypes in populations to elucidating the tree of life for birds. We suggest such integration can be achieved relatively painlessly and without acrimony, though some changes in analytical approaches are required. As an example of an integrated study we use the New Zealand pigeon, Hemiphaga novaeseelandiae, and estimate its genetic diversity throughout its range, including the Chatham Islands and Norfolk Island (where it is now extinct). We use longer sequences to identify its nearest relatives within the South Pacific (including the imperial pigeons [Ducula] and fruit doves [Ptilinopus]), and its position within Columbidae (pigeons and doves) generally. Finally, we use its complete mitochondrial genome, together with a sandgrouse (Pterocles namaqua), to study the position of pigeons within the Neoaves radiation. At the deeper levels of phylogeny we wish to reduce the noise in the data and enhance the signal, leading to clearer resolution of the basal nodes of avian phylogeny. Any suggestions of new and interesting ways to integrate multiscale data analyses at would be greatly appreciated!

On the shape and fabric of human history

Russel Gray†

In this talk I will outline two main debates about the nature of human history. The first focuses on the extent to which human history is treelike and the second on the coherence of that history (if every gene has its own history, do words and other cultural traits also chronicle unique pathways?). I will discuss the ability of current methods to address these questions and highlight where further methodological development would be useful.

Horizontal transmission and cultural phylogenies

Simon J. Greenhill†, Tom E. Currie, Russell D. Gray

Phylogenetic tree thinking is beginning to revolutionise studies of linguistic and cultural evolution. However, linguistic and cultural traits are easily transmitted horizontally ("borrowed") between cultures. Indeed, well over 95% of the words in the Oxford English Dictionary aren't English. A loud and persistent debate has centered around the issue of borrowing and whether it invalidates cultural phylogenies or not. Here, we use a natural model of linguistic evolution to simulate borrowing between languages. The results show that tree topologies constructed with Bayesian phylogenetic methods are relatively robust to the effects of realistic levels of borrowing. Inferences about time depth are slightly less robust.

Sampling trees from evolutionary models

Klaas Hartmann

A wide range of evolutionary models have been developed. These are used to test evolutionary hypotheses and provide comparisons with phylogenetic trees constructed from real data. To achieve this it is often necessary to sample – or simulate – trees from these evolutionary models. Sampling trees with a given number of species from these models is more complicated than may first be expected, necessitating some careful mathematical consideration.

In this talk I will show that one seemingly obvious sampling approach is correct for a simple model – the Yule model – but inappropriate for other models. We speculate that the correctness of this approach for the widespread Yule model has resulted in it being applied inappropriately to other models. Finally a simple alternative algorithm that applies to a broad range of evolutionary models is presented.

Multi Loci Bayesian recovery of population dynamics

Joseph Heled

Effective population size is related to genetic variability and is a basic parameter in many population genetics models. Various methods and models have been developed to infer current and past population sizes from genetic data since the introduction of the Coalescent theory in 1982 by JFC Kingman. I will present the Extended Bayesian Skyline Plot, a non parametric method method extending drummond Bayesian Skyline Plot for Multi-Locus data.

Extensive simulations show the accuracy and limitations of recovery and give an indication of the amount of data required for recovering past population dynamics, including bypassing evolutionary bottlenecks. The results show the essential role of multi-locus data and that typical data sets used today are probably too small for obtaining small bounds and providing information past bottlenecks..

Snails Through Time: Molecular analysis of New Zealand Marine Gastropods

Simon Hills†

New Zealand has an excellent marine mollusc fossil record. This fossil record is well documented and has recently been data-based (Crampton et al. 2003). Extant snail lineages are therefore ideal for molecular clock studies.

We have the opportunity to construct phylogenies with a high density of fossil calibrated nodes. For example, the New Zealand Buccinids (Whelks) consist of approximately 50 extant species in 9 genera, supported by 5 genus level and 30 species level fossil calibration dates. This wealth of fossil information enables us to more accurately estimate rates of molecular evolution, and allows comparative studies among lineages.

Additionally, using the New Zealand Volutes, we are analysing the relative phylogenetic information content of several mitochondrial genes, in order to assess the utility of each gene for phylogenetic resolution at different taxonomic levels.

Lapita - migration throughout near Oceania. Is there evidence for multiple introductions?

Melanie Hingston†, Howard Ross, Lisa Matisoo-Smith, Judith Robins

The origin and migration pathways of the Lapita cultural complex have been the subject of discussion between various scientific disciplines. In order to evaluate the different hypothesis, we use the phylogeny of the commensal Pacific rat (Rattus exulans), Kiore, that can be used as proxy for human migration in the Oceanic region. With extended sampling along the Bismarck-Archipelago and the application of Bayesian inference to estimate the relationship and migration rates towards other populations of R.exulans, we aim to clarify whether this area represents one major pathway.

Detecting and visualizing morphological convergence within the cormorants

Barbara Holland†, Martyn Kennedy, Hamish Spencer

We analyse two phylogenetic data sets for the same 33 species of cormorant, one data set is based on osteological features (Seigel-Causey 1988) and one is based on mitochondrial DNA sequences. The two data sets each produce highly supported phylogenies; however, the phylogenies have few points of agreement. We discuss to what extent the patterns in the morphological data can be explained by convergent evolution.

Deciphering the rapid species radiation of the New Zealand alpine rockcress (Pachycladon) species

Simon Joly†, Peter Heenan, and Peter Lockhart

The genus Pachycladon consists of nine species, eight of which are endemic to New Zealand and one that is endemic to Tasmania. These species form a monophyletic lineage that evolved during the last 1-3.5 Myr and that have obtained different morphologies (leaf, inflorescence, growth habit) and occupy a range of habitats (different elevations and soil substrates). For these reasons and because Pachycladon is a close parent of the model plant Arabidopsis thaliana, this genus is seen as a model group to study the genetic of species radiations. Here, we investigate the evolutionary history of the group using several nuclear genes. We first show that the ancestor of the whole group has experienced a singular hybridization event between very distant parents. Then, we tackle the problem of the species phylogeny, which is further complicated in this genus because of the high speciation rate. Both individual and species phylogenetic approaches are used and compared in addressing the problem of species relationships in this genus.

Why does my tree look strange? The impact of positive and negative selection on phylogenetic reconstruction

David Liberles

Standard phylogenetic methods for protein coding genes, from models for maximum likelihood and distance analysis to parsimony, assume that all changes in a sequence are random in a site-independent manner. From the use of such methods, it is assumed that the signal that is generated reflects ancestry. However, both positive directional selection for function and negative selection mediated by protein folding constraints have the potential to generate signals that are not site-independent and that depending upon the accessible mutational paths on the structure/function landscape, may give an alternative systematic signal reflecting convergence. This is in addition to any changes in rate that they impose.

Using minimization of inferred duplication and loss events from a gene tree as an independent signal of phylogenetic fidelity, we subdivide codon positions into bins of average pairwise dN/dS ratios and build phylogenetic trees using each bin. We find, perhaps not surprisingly, that the bins with slightly negative to neutral dN/dS ratios have the truest phylogenetic signals, with deviation found under strong negative and positive selection.

Further, we show in simulation using a lattice model construct that has undergone positive diver-

sifying selection related to a binding function over a phylogenetic tree, that under our simulation conditions, approximately 50% of trees constructed from the resulting sequences show clustering based upon binding capabilities rather than ancestry. However, when the neutral synonymous sites are extracted and used to construct trees, the ancestral signal is recovered.

These results have implications both for systematics and tree building as well as for understanding the coupling of sequence to function and structure on a fitness landscape.

Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites

Nicole Gruenheit, Peter J. Lockhart†, Mike Steel, William Martin

The covarion-like properties of sequences are poorly described and their impact on phylogenetic analyses poorly understood. We demonstrate using simulations that, under an evolutionary model where the proportion of variable sites changes in non adjacent lineages, log likelihood values for rates across site (RAS) and covarion (COV) models become similar, making models difficult to distinguish. Further, although COV and RAS models provide a great improvement in likelihood scores over a homogeneous model with these simulated data, reconstruction accuracy of tree building is low, suggesting caution when it is suspected that proportions of variable sites differ in different evolutionary lineages. We study the performance of a recently developed contingency test that detects the presence of COV-type evolution modified for protein data. We report that if lineagespecific distributions of the proportions of variable sites (pvar) become sufficiently non-overlapping, then the contingency test can incorrectly suggest a homogeneous model. Also of concern is the possibility of different proportions of variable sites between the groups being studied. In a study of chloroplast proteins, interpretation of the test is found to be susceptible to different partitioning of taxon groups, making the test very subjective in its implementation. Extreme intergroup differences in the extent of divergence and difference in proportions of variable sites could be contributing to this effect.

Modelling the emergence of genetic coding as a self-organising complex system

Sydney Markowitz†, Alexei Drummond, Peter Wills

Fundamental questions about how the Genetic Code arose remain a mystery, questions that go to the very heart of understanding the origin of Life. One nearly universal Standard Genetic Code is shared by all living things, indicating the code had its current form very early, before current implementing mechanisms which are built from proteins that depend on the Genetic Code came to be. How did the Genetic Code bootstrap itself? Why is there only one code? Could another code have been possible? How has it maintained stability in the face of mutational and translational errors? Simulating generalised models of genetic coding and protein synthesis as molecular information processing systems, we explore what is required for such a model to self-organise into a stable autocatalytic coding system. Results show emergence of competing coding systems, then dominance by one genetic code. We characterise parameters of the model for which coding emerges.

Comparing models of divergence time estimation

Michael Defoin Platel[†]

For inferring divergence time in phylogenetics, it is convenient to assume a constant molecular evolutionary rate over time. This assumption is called the molecular clock hypothesis and provides a means to translate genetic distances into geological times. Deviation from clocklike evolution has been often reported in datasets and the molecular clock hypothesis is therefore violated, particularly when distantly related species are compared, potentially leading to not only an incorrect estimation of species divergence time but also an incorrect inference of phylogenies. In the context of Bayesian phylogenetics reconstruction and divergence date estimation, it is now common to allow every branch to have a different rate of molecular evolution. We propose to review several existing approaches to relax the molecular clock assumption, such as local clock models, auto-correlated and uncorrelated relaxed-clock models. Comparisons are performed using the software Beast for numerical simulations. Finally, the relationship between sequence length and clocklikeness of the data is discussed.

Questions

David Penny

I must be the most ignorant person I know; consider the following questions (which are more or less randomly selected from a file with 40 such questions). Can we get away from point estimates of times of divergence. It is often more fundamental to know the time period over which a new group evolved, rather then the time of divergences. In other words, the length of the edge (and its variance) is more interesting than the time of the node itself. What are the effects of molecular clock being over-dispersed? Can we combine likelihoods from parsimony (on data where it is the ML estimator) with ML on primary sequence data? We need to know more about the transition from close populational data (where sequences are just one step apart [1-connected], and parsimony is a ML estimator) to having subgroups that are 1-connected, but then more changes between the subgroups. How often do local optima occur on a single tree with real data? Can we get more realistic coalescent trees where multifurcations are possible perhaps a direct calculation from the tree (like ML), not from the data to the tree (like parsimony and distance methods). If there is no information in the data to break a multifurcations, why pretend there is? Now that we have genomic-scale datasets, she we re-look at invariant methods, where long sequences are necessary to reduce the variance. Can we get better estimates of signal/noise ratios, and make use of them? We need to know more about alphabet reduction, when to lump character states together. How can we use more of the information in he data, information that we now discard because we use primarily primary sequence data. I could go on forever, I am just too ignorant.

Bird song sequence evolution

Louis Ranjard†, Howard Ross

Bird songs are an example of cultural evolution and thus the songs of different populations share similarities. Studying the evolution of these cultural traits can reflect relationships between translocated populations of New Zealand Saddleback (*Philesturnus carunculatus*). We mainly focus on the distribution of syllables in songs of birds from different islands. These syllables seem to have a greater stability through evolutionary time than the song itself, therefore we analyse songs encoded as sequences of syllables.

Specific neural networks have been developed to classify syllables, which are used to find evidence of shared vocalizations between songs. Then, phylogenetic analysis can be performed using encoded songs. For example, bayesian simulations performed with software such as BEAST, allow us to get a better insight into the mechanisms of song evolution. Some current work will be presented.

Reducing clutter in NeighborNet Networks

Alethea Rea† and David Bryant

NeighborNet is a method that creates networks based on pairwise distances and supported splits. These networks are created using standard mathematical optimisation tools which generate networks with many parallel splits that make them difficult to interpret. We have taken a statistical approach to reducing the number of splits. Backwards regression removes splits which do not contain a significant amount of information and reoptimises the branches lengths in the resulting network. This technique will be demonstrated on a data set of UN voting patterns.

Testing the Reliability of Genetic Methods of Species Identification

Howard A. Ross†, Sumathi Murugan and Wai Lok Sibon Li

High throughput methods of species identification are based on sequence comparison or clustering techniques. These techniques have received strong criticism on theoretical and practical grounds and, although they have been widely applied, there has been little testing of their reliability. In this study, sequence evolution was simulated in a wide range of evolutionary situations. The reliability of the methods was estimated by determining the frequency with which each method returned the correct species identity in each situation. Sequence similarity and simple clustering methods showed nearly identical performance. A stricter clustering criterion imposed an uncertainty cost with little gain in reliability when species are fully sampled but provided protection against false identification when taxon sampling is incomplete. The rate of speciation relative to the rate of genetic divergence is critical in determining the potential reliability of these methods. The significance of paraphyly is less important than previously asserted.

The Uto-Aztecans: Foragers from the north or farmers from the south?

Robert Ross, Russell Gray and Lyle Campbell

Considerable controversy surrounds the claim that early agricultural dispersals are the main factor that has shaped human linguistic, cultural and genetic diversity. One of the most contentious agricultural dispersal scenarios pertains to the Uto-Aztecan language family. According to the Southern Origin Hypothesis the earliest speakers of Uto-Aztecan languages were maize farmers in central Mexico who spread north into the American Southwest. By contrast, under the Northern Origin Hypothesis speakers of Uto-Aztecan languages were originally foragers in the American Southwest who spread south into Mesoamerica where they borrowed agricultural technology from contiguous Otomanguean and Mayan cultures. I will discuss how we are using lexical data and a Bayesian phylogenetic framework to test these competing hypotheses.

Kingman's Unlabeled n-Coalescent

Raazesh Sainudiin

We derive the transition structure of a Markovian lumping of Kingman's n-Coalescent. The lumped process, referred as the unlabeled n-coalescent here, is a continuous-time Markov chain on the set of all integer partitions of the sample size n. We derive the forward, backward and stationary probabilities of this chain and show that the probability of any given site-frequency spectrum, a commonly used statistics in genomic scans today, can be directly prescribed by integrations over realizations of the unlabeled n-coalescent under the infinitely-many-sites model of mutation. We develop an importance sampler that relies on an augmented unlabeled n-Coalescent forward in time to conduct inference at the empirical resolution of the site-frequency spectrum.

The effect on phylogeny of lineage specific structural and functional constraints

Klaus Schliep[†] Peter Lockhart, Ellen Nisbet

Multigene data sets are becoming the norm for phylogenetic studies; so called 'phylogenomic datasets' may involve hundreds of genes for many species. The size and complexity of these data sets creates a challenge for phylogenetic analysis. As we demonstrate, even linked genes can exhibit different properties of lineage specific evolution.

On the one hand we can improve the estimation of phylogenies, if we incorporate functional information that is phylogenetically informative. On the other hand, we might expect that similar functional constraints will impact on phylogenetic signals in the data. We present some results of a dataset of chloroplast genomes.

Optimizing Phylogenetic Diversity Across a Forest

Charles Semple
A central task in conservation biology is measuring, predicting, and preserving biological diversity as species face extinction. Dating back to 1992, phylogenetic diversity is a prominent notion for measuring the biodiversity of a collection of species. It is now well-known that optimizing phylogenetic diversity on a single tree can easily be done in polynomial time. In this talk, we discuss some recent results describing what can and can't be done easily for optimizing this measure across two or more trees. This is joint work with Magnus Bordewich (Durham University) and Andreas Spillner (University of East Anglia).

Lineage Specific Sequence Evolution

Liat Shavit Grievink[†], Barbara Holland, Mike Hendy and David Penny

Most commonly used phylogenetic models assume a constant process through time; they do not account for lineage specific properties. However it is known that these models are too simplistic, and with time the processes of evolution can change. In particular, it is now widely recognized that as the constraints on the sequences change the proportion of variable sites can vary between lineages. This is expected to affect the ability of phylogenetic methods to correctly estimate phylogenetic trees, especially for long timescales. To date there is no phylogenetic model that allows for change in proportion of variable sites, and the degree to which this affects phylogenetic reconstruction is still unknown. We will present a modification to the program Seqgen-cov that allows the generation of sequences that contain lineage- specific effects, along with a simulation study as a mean of understanding the significance of lineage specific effects on tree reconstruction.

Finding subsets of high phylogenetic diversity on special classes of split systems

Andreas Spillner

A popular way to measure the diversity of a collection of taxa is in terms of their relationships on a phylogenetic tree, as proposed by Faith in 1992. This measure, called phylogenetic diversity, has been used to identify collections of taxa with high diversity. Recently, the definition of phylogenetic diversity has been extended to split systems to take into account situations like non-treelike evolution or to measure diversity by averaging over a set of gene trees. In the talk we are interested in efficient algorithms for finding collections of high phylogenetic diversity on split systems. As the resulting optimization problem is NP-hard in general, we will focus on split systems with a special structure that can be exploited to design efficient algorithms.

The joys of being mean

Mike Steel†

I describe some ways in which one can obtain useful insights into the properties of phylogenetic models – both old and new – that at first look very complex, by simply exploiting standard properties of expectation ('the mean' of a random variable).

Maximum likelihood calculations using partial likelihood tensors

Jeremy Sumner‡and Mike Charleston

I will present a novel approach to maximum likelihood calculation in molecular phylogenetics. By computing "partial likelihood tensors", as opposed to "vectors", it is possible to systematically take advantage of the similarity between many of the character patterns present in aligned sequence data. It will be argued that this approach provides a speed up over Felsenstein's algorithm when the number of data points is of intermediate size; that is, somewhere between a single data point and every possible character pattern present. Comparison will be made to other recently published advances including "sub-tree equality vectors" and "column sorting".

Solution to a question of Steel and Hein

Bhalchandra Thatte

A pedigree is a directed acyclic graph that depicts ancestral relationships between individuals in a population. A somewhat futuristic question in population biology is: what could we say about the pedigree of a large population if we had complete genomes of all individuals in the population? At the Kaikoura meeting in 2006, Mike Steel presented a few combinatorial questions on the reconstruction of population pedigrees. One of the questions demanded reconstruction of a pedigree of a population from the collection of pedigrees on proper subsets of the population. In this talk, I will demonstrate that such a construction is not possible in general.

Extreme microevolution in Antarctic micro-arthropods

Torricelli G.,† Carapelli A., Frati F., Stevens M.I.

Mitochondrial DNA (mtDNA) has been widely used to investigate closely related taxa and phylogeographic relationships among arthropods. The extreme Antarctic environment is characterised by patchily distributed suitable living conditions, and thus soil invertebrate communities are highly fragmented with limited interpopulation gene flow. This scenario provides an ideal opportunity to investigate the possible promotion of microspeciation events by geographical isolation. Springtails are one of the most representative arthropod groups in Antarctica and are highly endemic.

Friesea grisea is the only collembolan species living both in the Transantarctic Mountains (eastern continental Antarctica) and in the Antarctic Peninsula, with congeneric species distributed on sub-Antarctic islands. Here, we sequenced and compared the whole mitochondrial genome from F. grisea collected in the Transantarctic Mountains and from the Antarctic Peninsula. These data reveal a remarkably high intraspecific substitution rate (20% sequence divergence). We then compared the cox1 and 28S genes of the two geographically separated F. grisea Antarctic populations to the sub-Antarctic F. tilbrooki (Macquarie Island) and F. bispinosa (Heard Island). On the basis of these data, the genetic divergence from F. grisea to either of the two sub-Antarctic island species was around 17% for the cox1 gene. We suggest that micro-evolutionary processes within the continental Antarctic biota have occurred despite no discernable morphological divergence, and that our data is the first to support the long-term isolation of all ten springtail species in the Transantarctic Mountains.

Varied accuracy in the prediction of gene function using patterns of co-evolution implicates inconsistency in quality and specificity of Gene Ontology terms.

Sibon Li Wai Lok†, Allen G. Rodrigo, Alexei J. Drummond

Recent studies have shown evidence for the co-evolution of functionally-related genes, as a result of constraints to maintain functional relationships between interacting proteins. Variation in the rate of evolution across gene trees can be partitioned into species-specific effects, gene-specific effects and interaction terms. We outline a procedure for predicting the biological function of a gene by utilising the gene-specific patterns of co-evolution through application of common statistical methods for prediction and clustering. We apply our method to a dataset consisting of genes from 10 prokaryotic species. With generalised linear models (GLMs) as a method of regression, we trained models based on the Gene Ontology (GO) terms that the genes are involved in. We found that the degree of accuracy to which we could predict the function of the gene using GLM models varied substantially across different GO terms. In particular, our model could accurately predict genes involved in ribosomal activity with a ROC area of 89predicting some other GO terms yielded results that were no better than random. We suggest that these results may be attributable to a combination of a lack of functional data, noisy/incomplete functional annotations and poor conventions for defining GO terms/categories. The relevance of our findings to phylogenetic analysis and comparative genomics is discussed.