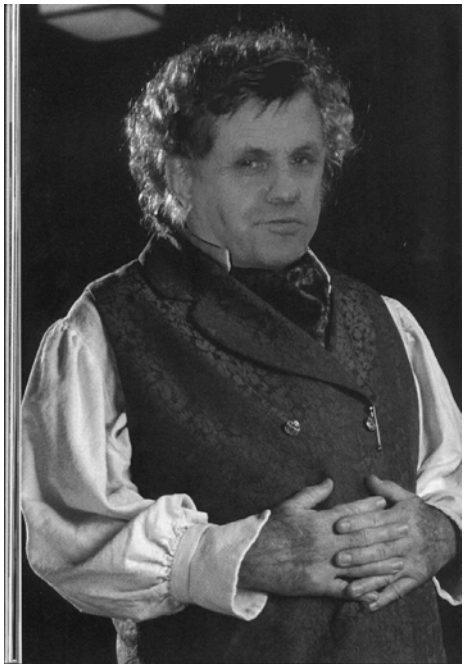
DOOM04



SKOTEL, WHAKAPAPA VILLAGE,
16TH-20TH FEBRUARY, 2004

SPONSORED BY NZIMA PHYLOGENETIC GENOMICS PROGRAMME,
AWCMEE AND HUBBARDS BREAKFAST CEREALS

Sunday 15th

8pm-9pm	**meeting registration at Skotel**

Monday 16th

9.00-9.10am	**welcome**
9.10-9.50am	**Andreas Dress** Quartet methods in phylogenetic analysis
9.50-10.10am	**Stefan Grunewald** A Method to Construct Phylogenetic Networks from Quartet Data
10.10-10.30am	**Mihaela Baroni** Using directed graphs to represent reticulate evolution
10.30-11.00am	**morning tea**
11.00-11.20am	**Mike Steel** The probability of self-sustaining autocatalysis in abstract origin of life models
11.20-12.00am	**Daniel Huson** Computing the Web of Prokaryotic Life
12.00-2.00pm	**lunch**
2.00-2.40pm	**David Penny** Where next in Phylogeny?
2.40-3.00pm	**Bhalchandra Thatte** MANTRA - A Multiple Alignment and Tree Reconstruction Algorithm
3.00-3.20pm	**Mike Hendy** Evolution of tandem repeats
3.30-4.00pm	**afternoon tea**
4.00- 4.20pm	**Russell Gray** How tangled is language evolution? Trees, splits and networks
4.20- 4.40pm	**Quentin Atkinson** How old is the Indo-European language family? A biological solution to a linguistic problem
4.40- 5.00pm	**Simon Greenhill** Language Phylogenies and the Austronesian Migration

Tuesday 17th

9.00-9.10am	**welcome**
9.10-9.30pm	**Elizabeth Allman** Constructing phylogenetic invariants for the general Markov model of sequence mutation
9.30-9.50pm	**John Rhodes** Phylogenetic invariants for the general Markov model of sequence mutation for any number of taxa
9.50-10.10pm	**Matthew Goode** Modeling change in codon substitution using serially sampled sequence data
10.10-10.30pm	**Stephane Guindon** Modelling the site-specific variation of selection patterns along lineages
10.30-11.00am	**morning tea**
11.00-11.20am	**Benny Chor** Maximum likelihood analysis of phylogenetic trees
11.20-11.40am	**Shlomo Moran** Convex Recolorings of Strings and Trees
11.40-12.00am	**Tobias Thierer** A characteristic function approach to perfect and imperfect phylogenies
12.00-12.20pm	**Michael Woodhams** The modified closest Tree
12.30-2.00pm	**lunch**
2.00-2.20pm	**Greg Ewing** Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in two subpopulations
2.20-2.40pm	**James Matheson** Analysis of Historical NZ Respiratory Syncytial Virus (RSV) Data.
2.40-3.00pm	**Howard Ross** Outgroup selection is a critical factor in reconstructions of ancestral sequences of HIV-1 envelope protein
3.00-3.20pm	**David Welch** Reconstructing host genealogy and parasite history - an integrated model
3.30-4.00pm	**afternoon tea**
4.00- 4.20pm	**Philip Daniel** New Supertree methods
4.20- 4.40pm	**Vincent Moulton** Phylogenetic Palantirs
4.40- 5.00pm	**Leonid Rusin** Circumventing limitations of parsimony heuristic search: screening the equal tree space under minimum homoplasy requirement
6.00pm	**barbecue**

Thursday 19th

7.45-9.00am	**breakfast at Chateau**
9.10-9.30pm	**David Bryant** Phylogenetic networks or phylogenetic ink blots?
9.30-9.50pm	**Barbara Holland** Consensus networks and species phylogenies
9.50-10.10pm	**Kim McBreen** Obtaining a species phylogeny for close relatives of *Arabidopsis* in New Zealand
10.10-10.30pm	**Katharina Huber** Modelling the evolutionary past of polyploid species
10.30-11.00am	**morning tea**
11.00-11.20am	**Leon Perrie** New Zealand's Ferns: Rampant Dispersal or Parallel Deceleration of Microevolution?
11.20-11.40am	**Tony Larkum** The whole genome of *Prochlorococcus* and the status of Cyanobacteria
11.40-12.00am	**Christoph Kneip** A nitrogen fixing endosymbiont – the sphaeroid body of *Rhopalodia gibba*
12.00-2.00pm	**lunch**
2.00-2.20pm	**Chris Simon** Progress in the Systematics and Evolution of New Zealand Cicada and their relatives World Wide.
2.20-2.40pm	**Robert Beiko** A Protein-by-Protein Search for Phylogenetic Discovery
2.40-3.00pm	**Tim White** Compressing DNA Sequence Databases with COIL
3.00-3.20pm	**Alex Grossmann** Rank methods in the analysis of expression data
3.30-4.00pm	**afternoon tea**
4.00- 4.20pm	**Lars Jermiin** The Biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated
4.20- 5.00pm	**Tom Britton** Quantifying uncertainty of phylogenetic tree estimates: Bootstrap support vs Bayesian probability

Friday 20th

9.00-10.30am	Informal discussions
11.00-11.30am	**morning tea**

Monday 16th Feb 2004

Andreas Dress

Quartet methods in phylogenetic analysis

Phylogenetic trees can be encoded in terms of (weighted compatible) split systems, (tree) metrics and (weighted Colonius-Schulze) quartet systems. The latter are particularly appropriate for testing the consistency of data. In the lecture, various approaches to quartet-based tree reconstruction will be presented and discussed.

Stefan Grunewald

A Method to Construct Phylogenetic Networks from Quartet Data

In phylogenetic tree reconstruction, it is a common divide and conquer approach to calculate confidence values for all possible quartet trees of the set X of taxa of interest and then to use these confidence values to construct a big tree with leaf set X. For example, this approach is used by the well known tool Tree Puzzle. On the other hand, there are many data sets where the phylogenetic information is better represented by a network than by a tree. I will present a method that uses confidence values of quartet trees to construct planar phylogenetic networks.

Mihaela Baroni

Using directed graphs to represent reticulate evolution

Some evolutionary phenomena, particularly lateral gene transfers and hybridization events, cannot be accurately represented by the classical tree model.

A certain directed graph, called a hybrid phylogeny, that is more general than a rooted phylogenetic tree, could be appropriate for describing reticulate events. In particular, two arcs ending in the same vertex correspond to a hybridization event.

Based on this digraph model, we develop a new framework for representing reticulate evolution, and present some new results.

Mike Steel and Wim Hordijk

The probability of self-sustaining autocatalysis in abstract origin of life models.

The ability of systems of molecular reactions to be simultaneously autocatalytic and sustained by some ambient `food source' of simple molecules may have been an essential step in the origin of life. We describe a polynomial-time algorithm that determines whether any given set of molecules, reactions and catalysations contains a subsystem that is both autocatalytic and able to be sustained from a given subset of the molecules. We then use this algorithms to investigate random catalytic networks - in particular a model of Kauffman. Using simulations and some analytic techniques we investigate the rate of catalysis that is required for the emergence of autocatalytic and sustaining subsystems.

Daniel Huson

Computing the Web of Prokaryotic Life

It has been argued that the "Tree of Life" is perhaps really a "Web of Life", as mechanisms such as hybridization, recombination and swapping of genes probably play a central role in evolution. Consequently, methods that compute phylogenetic networks, rather than trees, are of practical interest. The first such method to be suggested and implemented was "split decomposition", which, in practice, is limited to small numbers of taxa. More recent methods, such as Neighbor-Net (a hybrid of the popular Neighbor-Joining method and split decomposition) and consensus networks, do not exhibit this limitation. Such methods do not produce a network directly, but rather output a set of weighted "splits". We have developed and implemented algorithms that solve the problem of producing networks, so-called splits graphs, from such split systems, efficiently. Previously published algorithms are not suitable for these large data sets. Based on this, we present and discuss several versions of a "Web of Prokaryotic Life".

David Penny

Where next in Phylogeny?

For the past 20 years the phylogenetics community gives the appearance of being obsessed with sampling error. Are the sequences long enough to give us ‘confidence’ in the results of phylogeny? The recent paper by Rokas et al. with 106 genes should start to change our attitude to this ‘fashion’ - with increasingly longer sequences the sampling error is decreasing and that concern is being replaced by actually getting the models right! Inconsistency from model misspecification (incorrect assumptions about evolution) becomes the major problem. Under the instrumentalist (frequentist) approach the paradigm has been to accept the best-fitting model – even if we know the model is wrong – shut up, just accept it. Perhaps we need a more data-centered approach, less belief about our ‘methods’, more concern for the information in the data. Given this approach I will outline some areas that appear ripe for rapid development.

Asymmetric models may be the first to gain here. Tensors give us much more information about change in process. Traditionally we blame the data (our models are fine, if we get the wrong answer then it is data that is wrong). On the scientists’ approach, the data is fine, our models are wrong. **Signals in sequences** need to be evaluated more precisely, including Lento and triangle plots. Yes-no statistical measures on composition biases are less use, need magnitude tests - how much of the signal is explained by the model, there is an analogy with analysis of variance. Does removal of sites under positive selection improve the phylogenetic signal? **Information theory**. How much useful information is there in DNA sequence data, and how much do we actually use? What is the gain in information we actually get out of our trees? **Landscapes.** Huelsenbeck and Bayesian vs Landscape analysis, why does landscape analysis do so very much better. SpectroNet and analysis of sampled trees from MCMC. What is the relationship between Bayesian methods and maximum integrated likelihood?

Appropriate data types will become clearer - for **Long term** evolution, primary sequences may be less important for the tree itself (times may be another issue). We require 'rdc' (rare genomic changes) such as gene duplications, major insertions/deletions, gene order, position of transposable elements, intron number and position, and so on. Increasingly this will require genomic-level amounts of data. In a sense, experimentalists have been outwitting theoreticians by already finding other data types. Fortunately the theoreticians are catching up. For **Short term** evolution, sequences are abundant (not sparse) We do not have to average over missing data, thus parsimony is the ML estimator for the tree (or network) - but what about for rates? If we don't require a Markov model, are there different questions or approaches?

Rates, we need the minimum number of changes in process, and 'rate' will need to be expanded to see if only some nucleotide pairs change rate. Interpolations under rate variation are unbiased, but we know little on how the variance is affected. For change in rates (clock) we need more interest in magnitude tests. **Theoretical** stuff on trees and data. R-Y coding (for example), is it a nested subset of nucleotide coding (and therefore can use Likelihood ratio tests). Similarly, what difference does it make whether a tree is treated as a single 'parameter', or as a compatible subset of splits? Is there a better way for 'likelihood' on morphological data; will it help by using the Stein paradox in statistics (using a large range of averages over different sites). Perhaps that is enough ideas for one day, there are a lot more.

Bhalchandra Thatte

MANTRA - A Multiple Alignment and Tree Reconstruction Algorithm

MANTRA is an ongoing project in which our goal is to simultaneously reconstruct the phylogeny and multiple alignment of a set of n sequences. I will introduce the problem, and a heuristic algorithm that does at most O(n^2) pairwise sequence alignments.

Mike Hendy

Evolution of tandem repeats

Tandem repeats, whether short or long, may be useful phylogenetic information, even within populations. However to exploit this, we need to have an understanding of models of their evolution. For long repeats accumulated nucleotide substitutions give useful information. For short identical subsequences, we must exploit the repeat frequencies.

Russell D. Gray, Robb Rutledge, Flavia Filimon and David Bryant

How tangled is language evolution? Trees, splits and networks

The idea that much of recent human history might reflect pure trees of phylogenetic descent is appealing simple. It has stimulated numerous researchers to investigate the extent to which genes, languages and cultures are bound together in co-diverging trees of evolutionary history. Increasingly studies have used computational phylogenetic methods to make inferences about linguistic history and the evolution of cultural traits (Warnow, 1997; Gray & Jordan, 2000; Pagel, 2000; Holden, 2002; O'Brien et al, 2002; Jordan & Shennan, 2003; Holden & Mace, 2003; Rexova et al. 2003; Gray & Atkinson, 2003). However, a persistent criticism of this approach is that cultural evolution is far from tree-like (Moore 1994, Terrell 1988, Terrell et al 2001). Not only might patterns of genetic, linguistic, and cultural diversity reflect different histories, each of these histories might be strikingly reticulate. As one participant at a recent symposium on phylogenetic methods in archaeology growled, "This is not history. This is history put in nested boxes!"

What is needed to get beyond the impasse of these polarised *a priori* views is an analytic approach that enables us to assess where on the continuum between a pure tree and a totally tangled network any particular case may lie. More specifically, this approach should be able both to identify the particular languages where borrowing has occurred and detail the exact characters that were borrowed. In this talk we will outline one such method - NeighbourNet (Bryant & Moulton, in press) - using Indo-European and Polynesian lexical data to demonstrate the potential and possible pitfalls of this approaches.

Simon Greenhill and Russell Gray

Language Phylogenies and the Austronesian Migration

Recently there has been an increased appreciation of the importance of migration scenarios in human prehistory. Diamond and Bellwood (2003) go so far as to claim that migration is the most important process in human holocene history. However, quantitative tests of these migration hypotheses are rare. One exception is Gray and Jordan's (2000) test of the 'Express Train' scenario of Austronesian expansion. This study used phylogenetic methods to construct a language tree and quantitatively evaluated the extent to which the tree supported the 'Express Train' model. However, there were some flaws in this study, and in this talk we will discuss these problems, and outline the results of the solutions we have recently implemented.

Quentin Atkinson and Russell Gray

How old is the Indo-European language family? A biological solution to a linguistic problem.

Languages, like genes, provide vital clues about human history (Gray & Jordan, 2000). The origin of the Indo-European language family is 'the most intensively studied, yet still most recalcitrant, problem of historical linguistics' (Diamond & Bellwood, 2003). Due to slow rates of genetic change, admixture, and the relatively recent timescales involved, genetic analyses have not conclusively resolved debates about time-depth in Indo-European. Languages, however, change much faster than genes and so contain more historical information at shallower time-depths. Despite this, traditional means of linguistic analysis have also been unable to provide convincing evidence of Indo-European origins. This is primarily due to problems associated with variable rates of language evolution and an inability to quantify the degree of statistical uncertainty in estimated ages. Here, we explore the application of new phylogenetic date estimation techniques to linguistic data. These methods are able to estimate divergence times and the uncertainty associated with each estimate, even under conditions of rate heterogeneity. We combine maximum-likelihood models of language evolution, Bayesian inference of phylogeny and rate smoothing algorithms to test between two theories of Indo-European origin - the 'Kurgan expansion' and 'Anatolian farming' hypotheses. The Kurgan hypothesis centres on possible archaeological evidence for an expansion into Europe and the near-East by Kurgan horsemen beginning in the sixth millennium BP (Gimbutas, 1973). The Anatolian hypothesis claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000 to 9,500BP (Renfrew, 1987). In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800BP and 9,800BP (Gray & Atkinson, 2003). The results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the Bayesian analysis.

Diamond, J. & Bellwood, P. Farmers and Their Languages: The First Expansions. Science 300, 597 (2003).

Gimbutas, M. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500-2500 B.C. Journal of Indo-European Studies 1, 163-214 (1973).

Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426, 435-439 (2003).

Gray, R.D. & Jordan, F.M. Language trees support the express-train sequence of Austronesian expansion. Nature 405, 1052-1055 (2000).

Renfrew, C. Archaeology and Language: the Puzzle of Indo-European origins. London: Jonathan Cape (1987).

Tuesday 17th Feb 2004

Elizabeth S. Allman

Constructing phylogenetic invariants for the general Markov model of sequence mutation.

A phylogenetic invariant for a probabilistic model of biological sequence evolution along a tree is a polynomial that vanishes on the expected frequencies of base patterns at the terminal taxa. Except for small trees and simple models, this has been a difficult problem, with the most progress made for group-based models via Hadamard conjugation.

We explain a construction of invariants for the general Markov model, and thus all sub-models, of k-base sequence evolution on an n-taxon tree, for any k and n. The method depends primarily on the observation that certain matrices defined in terms of expected pattern frequencies must commute, and yields many invariants of degree k+1, regardless of the value of n. While these commutation relations are most easily understood for a 3-taxon tree, they generalize to arbitrary trees.

Although the construction does not give all invariants, it gives enough to identify joint distribution arrays arising from general Markov model parameters provided that the array is in a certain set. Since biological data is likely to lie in that set, these invariants may be sufficient for phylogenetic applications.

John A. Rhodes

Phylogenetic invariants for the general Markov model of sequence mutation for any number of taxa.

Finding all phylogenetic invariants for a model of molecular evolution means finding all polynomial relationships among entries of the joint distribution array of patterns in aligned sequences that the model predicts.

For applications, a smaller set of invariants may be sufficient, as long as we can be sure any array satisfying those invariants will satisfy all unknown invariants as well. In the language of algebraic geometry, such invariants are said to set-theoretically define the phylogenetic variety.

We explain the relationships between phylogenetic invariants and flattenings of the joint distribution table. Moreover, for the general Markov model with k bases, we show that if we understood phylogenetic invariants for a 3-taxon tree, then we would have a set-theoretic understanding of the phylogenetic variety for any number of taxa n. In the k = 2 case, we do understand the 3-taxon tree. In the most interesting case, when k = 4, the 3-taxon phylogenetic ideal is still not fully understood, though commutation relations give non-trivial elements of this ideal.

This work makes connections to Bayesian models with hidden variables, and problems arising in complexity theory.

Matthew Goode

Modeling change in codon substitution using serially sampled sequence data

Previous studies have described the use of nucleotide sequences sampled over time to estimate substitution rates over a series of time intervals. In this talk we discuss extending this approach to codon substitution models. In particular, we examine using the codon model of Neilsen and Yang 1998, and allowing parameters associated with selection and proportion of site classes to vary over time. We discuss how this approach can be used to aid in detecting changes in selection over time in populations where evolution can be measured, e. g., HIV and other rapidly evolving viral populations. We present the results of a study into the effectiveness of the method.

Stephane Guindon

Modelling the site-specific variation of selection patterns along lineages

Patterns of substitution between codons disclose the way Darwinian selection acts at the protein level. The nonsynonymous (amino acid-altering) to synonymous (silent) substitution rate ratio is indeed a straightforward measure to detect a neutral, a positive (diversifying) or a negative (purifying) selection process of molecular evolution. Recent stochastic models of nucleotide substitution between codons assume that the selection pattern may vary across sites but remains constant among lineages. Despite some work that has been done to relax this constraint (see Yang and Nielsen, 2002), no model provides a real statistical framework to deal with switches between selection processes at individual sites during the course of evolution. This paper describes a new approach that allows the site-specific selection process to vary along lineages of a phylogenetic tree. The rates of changes of selection process are explicit parameters that are adjusted under a maximum likelihood framework. The analysis of eight HIV-1 env homologous sequence data sets shows that the fit of our model is significantly better than one that does not take into account switches between selection pattern in the phylogeny.

Benny Chor

Maximum likelihood analysis of phylogenetic trees

Among various methods for constructing phylogenetic trees from sequence data, maximum likelihood (ML) is considered the most reliable, and it is widely used. Yet, there are many issues regarding ML that are not very well understood. These include the number of local maximum points on the likelihood surface, and the computational complexity of ML. I will describe several results and open problems in this area.

Shlomo Moran

Convex Recolorings of Strings and Trees

A coloring of a tree is convex if the vertices that pertain to any color induce a connected subtree; a partial coloring (which assigns colors to some of the vertices) is convex if it can be completed to a convex (total) coloring. Convex coloring of trees arises in areas such as phylogenetics, linguistics, etc. e.g., a perfect phylogenetic tree is one in which the states of each character induce a convex coloring of the tree. Research on perfect phylogeny is usually focused on finding a tree so that few predetermined partial colorings of its vertices are convex.

When a coloring of a tree is not convex, it is desirable to know "how far" it is from a convex one. One common measure for this is based on the parsimony score, which is the number of edges whose endpoints have different colors. In this paper we study another natural measure for this distance: the minimal number of color changes at the vertices needed to make the coloring convex. This can be viewed as minimizing the number of ``exceptional vertices" w.r.t. to a closest convex coloring. We also study a similar measure which aims at minimizing the number of ``exceptional edges" w.r.t. a closest convex coloring. We show that finding each of these distances is NP-hard even for strings. In the positive side we present few algorithms for convex recoloring of strings of trees: First we present algorithms for optimal convex recoloring of strings and trees, which for any fixed number of colors are linear in the input size. Then we present fixed parameter tractable algorithms and approximation algorithms for convex recoloring of strings and trees. We also discuss generalization of our algorithms to weighted trees, and to non-uniform cost model, in which the cost of overwriting one color by another may depend on the specific colors involved. (joint work with Sagi Snir).

Tobias Thierer

A characteristic function approach to perfect and imperfect phylogenies

In a homoplasy-free model of evolution, the labelling of a phylogenetic tree is always convex, i.e. the set of nodes (species) labelled with any specific character state is connected. However, biological data is almost never perfect and also doesn't closely follow the simplified models, yielding no or multiple convex and near-convex possible labellings on a given tree topology. We present a characteristic function approach for generalised binary characters (character states 0, 1 and "unknown") to assess the number of possible labellings for any given number of 0 -> 1 and 1 -> 0 character state transitions. This allows a biologist to quickly overview perfect and near-perfect phylogenies, and thus both to estimate the reliability of a solution and to identify likely errors in the input data. A generalisation of the approach can be used to calculate the likelihood of a leaf-labelling, given a tree and state transition probabilities.

Michael Woodhams

The Modified Closest Tree

Maximum Likelihood (ML) for phylogenetic inference from sequence data remains a method of choice, but has computational limitations. In particular it can not be applied for a global search through all potential trees when the number of taxa is large, and hence a heuristic restriction in the search space is required. We have derived a quadratic approximation to the likelihood function whose maximum is easily determined for a given tree, and for which a branch and bound search is feasible. The derivation depends on Hadamard conjugation, and hence is limited to the simple symmetric models of Kimura and of Jukes and Cantor. Preliminary testing has demonstrated its accuracy is close to that of ML.

Greg Ewing

Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in two subpopulations

We present a Bayesian statistical inference approach for estimating migration rates and effective sizes in two subpopulations, using sequence data collected at different times. By using Markov Chain Monte Carlo (MCMC) integration, we take account of the uncertainty in genealogies and parameters. We recover information about the unknown true ancestral coalescent tree, population size, mutation rate and the migration rate from the temporal and spatial sequence data. We show that the method recovers the truth with simulated data. Finally we illustrate this method on HIV data with asymmetric migration rates between body compartments.

James Matheson

Analysis of Historical NZ Respiratory Syncytial Virus (RSV) Data.

In this talk I will present my results so far in looking at samples of Respiratory Syncytial Virus (RSV) sampled from around New Zealand over the last 20 years. I will discuss some of the challenges of analysing this type of data and look at how various phylogenetic techniques such as Median networks and UPGMA characterize the data and what information can be extracted.

Howard Ross

Outgroup selection is a critical factor in reconstructions of ancestral sequences of HIV-1 envelope protein

Ancestral DNA and protein sequences can be reconstructed on phylogenetic trees, and could be the basis for HIV-1 vaccine design. We investigated the variation in the ancestral sequence arising from the methods used in reconstruction. We reconstructed the ancestral sequence of 118 HIV-1 B-subtype envelope proteins using different (a) methods of rooting the phylogenetic tree, (b) algorithms for estimating the phylogenetic tree, (c) algorithms for reconstructing the ancestral sequence on the phylogenetic tree, and (d) forms of reconstruction. The reconstructed ancestral sequences differed on average at 11% of sites (range 0.5 - 23%) with the method of rooting the phylogenetic tree being the greatest source of variation. Ancestral sequences were progressively more different from the circulating B-subtype sequences or their consensus when reconstructed on trees rooted at a computed centre, or with the M-group consensus, individual D-subtype or D-subtype consensus sequences as outgroups. The computationally predicted structural and immunological properties of the ancestral sequences, reconstructed on trees rooted with either the computed centre-of-tree or the D-subtype sequences outgroup, were compared with the predictions for circulating B-subtype sequences. Sequences reconstructed on trees with centre-of-tree rootings had energetically more favourable predicted 3-D structures, and had a greater number of predicted medium-affinity immunological binding sites than did other reconstructions. Sequences reconstructed on trees rooted with D-subtype sequences had a greater number of predicted high- and low-affinity immunological binding sites. However, direct comparison with the HIV-1 immunology database revealed that almost every reconstructed ancestral sequence contained all of the common epitopes. Within the limitations of our dataset, we conclude that sequences reconstructed on trees with centre-of-tree rootings may more closely resemble circulating sequences, but their immunological potency may not differ from sequences reconstructed by other methods, a result with potential significance for vaccine design.

David Welch

Reconstucting host genealogy and parasite history - an integrated model

In many situations, the genealogy of a parasite closely follows the genealogy of its host species. If the parasite is passed only vertically (from parent to child at birth), the correspondence is perfect. If it is passed only horizontally (between any two members of a population at any time), the two histories diverge immediately. In this talk, I'll present a new stochastic model of a parasite that is transmitted at birth and via contact. The model gives rise to a graph structure similar to Kingman's coalescent. I'll discuss how this graph, which represents the coupled host-virus genealogy, can be reconstructed using a Bayesian approach. I'll motivate the talk with an example of a cat population hosting the Feline Immunodeficiency Virus.

Philip Daniel

New SuperTree Methods

I shall out line the results presented in my recently submitted masters thesis. First, I shall describe a supertree method that determines whether a collection of rooted phylogenetic trees is displayed by a unique minimal tree. Second, I shall describe two notions of compatibility and displays—perfect compatibility and ancestral compatibility—that handle rooted semi-labeled trees. I will also describe two new supertree methods that follow these two notions.

Vincent Moulton

Phylogenetic Palantirs

Recently, much excitement has been generated in the chicken-scratching community about spaces of trees. In this talk we will present recent results on some combinatorial structures that are related to such spaces.

Leonid Rusin

Circumventing limitations of parsimony heuristic search: screening the equal tree space under minimum homoplasy requirement

Among discrete methods of phylogenetic inference, parsimony is the fastest and most widely used but also most susceptible to unequal evolutionary rates. If the amount of homoplasy at the level of individual characters is high enough, it may graft taxa erroneously on the basis of homoplastic characters while searching for most parsimonious trees during additional sequence replicates. Ultimately, heuristic search may stall within a local tree island isolated from the global optimum by branch swappings that can not be implemented given a current combination of taxa and thus will fail to find the true tree. We present an easy approach to circumvent limitations of heuristic search and reveal reliable nodes at different levels of resolution. Experiments were done with a complete SSU rDNA alignment of 36 nematode and 6 other metazoan groups, which contained clear disparities in rates of nucleotide substitution across taxa. To isolate signal/noise ratio available in data at different levels of resolution, a series of subalignments was generated by gradually removing partitions with respect to number of substitutions assigned to explain the most parsimonious tree. A series of equal tree spaces was produced by processing the subalignments with parsimony. To elaborate a measure of reliability of a node, we computed values of homoplasy index (HI) for each putative synapomorphy of a node. Analysis of tree spaces revealed nodes reconstructed on the basis of at least one character with HI values close to or equal zero. Some of them were recurrent, while others were not present in the initial tree and yet most of them corresponded to higher nematode taxa, which monophyly has already been substantiated from independent evidence (morphology or SSU rRNA secondary structure). To study cases when alternative nodes were supported by characters with almost equally high HI

values, we analysed the total amount of homoplasy generated between taxa by a topology through computing pairwise homoplasy distances for trees in tree spaces. Analysis of corresponding homoplasy matrices revealed nodes, which content is in accord with pairwise homoplasy distance distribution, *i.e.* joined taxa generate less homoplasy with respect to each other than to other taxa. Nodes that meet the minimum homoplasy requirement were drawn from the pooled equal tree space and used for constraining a heuristic search with the initial alignment. None of the equal trees found fully coincided with the compiled phylogeny, although it received the highest likelihood score by Shimodaira-Hasegawa one-tailed test using RELL bootstrap.

Thursday 19th Feb 2004

David Bryant

Phylogenetic networks or phylogenetic ink blots?

There are many valid (damning?) criticisms of phylogenetic network methods. Splits graphs are hard to interpret, and easy to miss-interpret. Distance based methods like Neighbor-Net and Splits decomposition are highly prone to noise in the distance data, and its not clear what sense (if any) to make of a spectronet when the data is not tree-like. No-one would deny that one can produce wonderfully intricate graphics with methods like NeighborNet... but how can we assess no way to systematically assess the significance of what we see? I'll talk about the problems, and report on progress towards finding solutions.

Barbara Holland

Consensus networks and species phylogenies

Phylogenetic methods such as bootstrapping and Bayesian MCMC produce collections of trees rather than a single tree. Typically consensus methods are used to summarise this information. However, consensus methods by definition are not capable of displaying conflicting hypotheses. We present a method that generalises the notation of consensus trees to consensus networks. It uses median networks to visualise those splits that are common to some threshold proportion of the input trees.

Multiple trees may also arise through the analysis of independent genes. Here we can apply the same method to visualise species phylogenies.

The method is demonstrated by displaying the output of both a bootstrap and an MCMC analysis of the Murphy et al. (2000) mammal data set, and also to display the level of incongruence in a set of 106 gene trees for 8 species of yeast that appeared in a recent issue of Nature (Rokas et al., 2003).

Kim McBreen and Peter Lockhart

Obtaining a species phylogeny for close relatives of *Arabidopsis* in New Zealand

Consensus networks are currently under investigation as a means for both investigating phylogenetic uncertainty and for visualising species phylogenies. We present results using this technique to study close relatives of *Arabidopsis thaliana* (the dicot plant whose genomes have been completely sequenced). New Zealand *Pachylcadon* complex comprise morphologically and ecologically diverse species that are closely related to *Arabidopsis thaliana.* We are interested in their potential for studying the genetics of morphological and ecological diversification.

Katharina Huber, Vincent Moulton and Bengt Oxelman

Modelling the evolutionary past of polyploid species

Evidence suggests that polyploidization that is, doubling of the genome via duplication or hybridization, played a major role in the evolution of the higher plants. Modelling the evolutionary past of such species is therefore an interesting problem.

In this talk, we will discuss a formalization of this problem and present a construction for a phylogenetic network that is guaranteed to minimize the number of evolutionary events required to explain the evolution of a collection of polyploid species.

Leon Perrie

New Zealand's Ferns: Rampant Dispersal or Parallel Deceleration of Microevolution?

The origins of New Zealand's biota remain controversial: were the ancestors of the extant biota present on the ancestral New Zealand landmass when it separated from Gondwana approximately 80 million years ago (vicariance), or did they arrive only after New Zealand was completely surrounded by ocean (long-distance dispersal)? Penalised Likelihood analysis using the program r8s of a *rbcL* data set encompassing global fern diversity indicates that either (1) the majority of the New Zealand fern groups investigated must have arrived by dispersal, or (2) a deceleration in the rate of microevolution has occurred independently across multiple groups. The likelihood for each of these alternative explanations is discussed, as are the implications for other plant groups.

Tony Larkum, Lars Jermiin and Pete Lockhart

The whole genome of *Prochlorococcus* and the status of Cyanobacteria

Two whole genomes of *Prochlorococcus* were announced last year as well as the whole genome of an oceanic, deep-water *Synechococcus* cyanobacterium. *Prochlorococcus* algae were only discovered in the late 1980s but are now recognised as the perhaps most

important primary producers in the oceans. They are of particular interest because they possess a very different light-harvesting compliment of proteins than typical Cyanobacteria. Instead of possessing phycobiliproteins assembled into phycobilisomes which funnel energy into the two photosystems , *Prochlorococcus* algae possess a chlorophyll-based family of light-harvesting proteins (pcb proteins). Pcb proteins bind both chlorophyll *a* and chlorophyll *b*. This is very unusual because chlorophyll *b* is very rare in Cyanobacteria (two other groups exist with pcb proteins which bind Chl *b* – *Prochloron* and *Prochlorothrix*). The possession of pcb and Chl *b* has caused some people to suggest that these three groups of algae should be placed in a special Phylum – the Prochlorophytes. However 16 S rRNA analysis suggests that these groups fall within the cyanobacterial clade and, in fact, on different branches of the cyanobacterial tree.

It is against this background and driven by a desire to clarify the situation that the whole genomes have been carried out. The results present us with a very interesting situation and one which poses important questions on a number of levels.

The analyses present us with two genomes that are very different from each other. Apart from housekeeping genes and protein machinery genes, the two genomes are very different from each other. However one of the genomes does bear substantial similarity with the deep-water *Synechococcus*. In fact the only real similarity between the two strains of *Prochlorococcus* is the possession of pcb genes. Oddly , and frustratingly, the obvious gene for the biosynthesis of Chl *b*, Chl *a* oxygenase, is not present in either genome (this is the gene used to make Chl *b* in the other prochlorophytes, in green algae and in plants). And no obvious gene for the synthesis of Chl *b* has been identified. So at present we do not know how Chl *b* is formed in *Prochlorococcus*, but we do know that Chl *b* is present and is attached to pcbs.

Thus we can say that the *Prochlorococcus* strains do not resemble each other in anything like the extent that we would normally regard as defining a species or genus – although 16s rRNA analysis does group them. Thus the possession of pcb and Chl *b* and the lack of phycobiliproteins turns out to be a rather poor character to use for higher order classification. On the other hand we do not have any good characters, apart from 16S rRNA, to classify Cyanobacteria at present. These results also force us to ask what is the relevance of the possession of pcb /Chl *b* synthesis genes, which are found in the Cyanobacteria, *Prochloron, Prochlorothrix, Prochlorococcus* and *Acaryochloris* (which posseses the unique Chl, Chl *d*)?

Christoph Kneip

A nitrogen fixing endosymbiont – the sphaeroid body of *Rhopalodia gibba*

The diatom *Rhopalodia gibba* harbours DNA-containing cell inclusions named sphaeroid bodies, which are separated from the cytoplasm of the cell. The ultrastructure of this cell inclusion resembles that of recent cyanobacteria, although the sphaeroid bodies are lacking typical pigmentation. These features lead to the supposition that these inclusions are symbionts.

By isolation of intact sphaeroid bodies, subsequent purification of genomic DNA and analysis of 16S rDNA sequences, we can show that this cell inclusion groups together with free living nitrogen fixing cyanobacteria. Further experiments demonstrate

that the sphaeroid bodies` genome codes for genes of the *nif*-family and that subunits of nitrogenase are present. Thus, we demonstrate that the sphaeroid bodies are endosymbionts which provide nitrogen to its diatom host cell, in contrast to higher plants, where extracellular symbionts fix nitrogen.

We suggest that the intracellular sphaeroid bodies of *Rhopalodia gibba* may represent a vertically transmitted, permanent endosymbiotic stage in the transition from a free living diazotrophic cyanobacterium to a nitrogen fixing eukaryotic organelle.

Chris Simon

Progress in the Systematics and Evolution of New Zealand Cicada and their relatives World Wide

I will present an informal five-minute review of our research progress in the systematics and evolution of New Zealand cicadas and their relatives world wide since last year's meeting in Kaikoura. The mtDNA/ef1a trees for all species and subspecies of the NZ cicada genus Kikihia have now been finalized and songs analyzed for all species. All songs for the Maoricicada species and subspecies have also been recorded and analyzed. The number of taxa of world-wide Cicadettini included in our phylogenetic tree has been increased significantly (adding more mtDNA and ef1a data as well as some morphology). We are currently developing two phylogenetics projects in collaboration with Thomas Buckley (Landcare Mt. Albert) and Max Moulds (Australian Museum): Phylogenetics of Australian Cicadettini and their relatives world wide, and Phylogeography of the New Zealand *Kikihia muta* complex, *Maoricicada campbelli* complex, and *Rhodopsalta curentata/leptomera* complex.

Tim White

Compressing DNA Sequence Databases with COIL

Publicly available DNA sequence databases such as GenBank are large and growing at an exponential rate, presenting serious storage and data communications problems. Currently, this data is usually kept in large “flat files,” which are compressed using standard techniques; this strategy rarely achieves good compression. I investigated a new approach to compression, *edit-tree coding*, which finds sets of similar sequences and construct trees on them: each tree is encoded by choosing one sequence as the root and recording it verbatim, then recording all edge mutations or *edits*; decoding entails traversing trees and applying edits to reconstruct the sequences at the leaves. This scheme, as implemented in the freely available COIL program, reduces the storage space required for GenBank’s Mouse EST sequence database by over 50% compared to traditional techniques, and there is potential for even greater compression ratios. I will discuss some of the theoretical and implementation issues involved.

Alex Grossmann

Rank methods in the analysis if expression data

Given a family of expression data in a genome, a natural "distance" between genes is the Kendall covariance matrix. We use this "distance" on data from *E. coli* and from *B. subtilis*, obtaining groups of "significant" genes for sets of experimental conditions. We are testing the relevance of these groups.

Lars S. Jermiin, Simon Y. W. Ho, Faisal Ababneh, John Robinson, and Tony Larkum

The Biasing Effect of Compositional Heterogeneity on Phylogenetic Estimates may be Underestimated

Compositional heterogeneity among diverging lineages can lead to errors in the phylogenetic estimates of topology and edge lengths, but it remains unclear how large this heterogeneity is required to be in order to cause such errors. We used phylogenetic analysis of sequence data generated by Monte Carlo simulation to assess the effect of compositional heterogeneity on inference of trees with short internal edges. We found that (*i*) compositional heterogeneity in sequence data increases the difficulty with which short internal edge lengths can be inferred using the maximum-parsimony method, the maximum likelihood method with an F81 model of nucleotide substitution, and the neighbor-joining method with distances estimated using the Jukes-Cantor model of nucleotide substitutions; and (*ii*) that the LogDet method, unlike the other methods assessed, had no difficulty in inferring the internal edge under conditions where the other methods failed. The results highlight the importance of assessing prior to a phylogenetic analysis whether the data violate the phylogenetic assumption of stationarity, and call into question recent reports that have implied that compositional heterogeneity is of minor concern to phylogenetic analysis. It is concluded that as the number of sequences in phylogenetic data becomes larger, so does the potential problems caused by compositional heterogeneity and the need to assess whether the assumption of compositional homogeneity is violated by the data intended for phylogenetic analysis. Methods to assess compositional homogeneity in sequence data are briefly reviewed.

Tom Britton

Quantifying uncertainty of phylogenetic tree estimates: Bootstrap support vs Bayesian probability

Abstract: Two commonly used methods for expressing the uncertainty (or conversely the support) for an estimated phylogenetic tree topology are "bootstrap support values" and "Bayesian posterior probabilities". It has been empirically observed that the latter is almost always larger than the former. However, the few available theoretical arguments for this comparison suggest that there is no such systematic difference. In the talk we will present a heuristic argument supporting the empirically observed systematic difference.

Participants

Name	Contact details
Martyn Kennedy	
Kim McBreen	
Alex Grossman	
Barbara Holland	
David Welch	
Leonid Roussine	
Christoph Kneip	
Johan Kahrstrom	
Phil Novis	
Mihaela Baroni	
Leon Perrie	
Bettina Greese	
Michael Woodhams	
Stephane Guindon	
James Matheson	
Matthew Goode	
Quentin Atkinson	
Dietmar Cieslik	
Vincent Moulton	
Katharina Huber	
Prof Anthony Larkum	
Dr L Jermiin	
Adrian Paterson	
Tobias Thierer	
Simon Greenhill	
Daniel Huson	
Tom Britton	
Shlomo Moran	
Stefan Grunewald	
Greg Ewing	
Howard Ross	
Judith Robins	
Russell Gray	
Andreas Dress	
Robert Beiko	
Steve Trewick	
Mary Morgan Richards	
Chris Simon	
John Rhodes	
Elizabeth Allman	
Peter Lockhart	
Mike Hendy	
David Penny	
Mike Steel	
Tim White	
Bhalchandra Thatte	
Philip Daniel	
David Bryant	
Benny Chor	
Dave Phillips	
Susan Wright	
Carl Masak	
Mirko Wojnowski	
Andrew Grimm	