SOUTH 2012 - TALK TITLES AND ABSTRACTS

- UPDATED 23 JANUARY 2012

REMCO BOUCKAERT
BEAST 2

DAVID BRYANT
Where now for phylogenetic networks? A review and some open problems *[no abstract]*

MICHAEL CHARLESTON
Practical approaches to cophylogenetic analysis

JOSH COLLINS
Finding maximal acyclic agreement forests

MAREIKE FISCHER
The impact of 'non-heredity' on tree reconstruction in practice

PAUL GARDNER
Annotating a plethora of RNA motifs

GILLIAN GIBB
Next-generation mitogenomics: new results from modern and fossil xenarthrans

NICOLE GRUENHEIT
Transcriptomics of New Zealand alpine plants

STEFAN GRÜNEWALD
The quartet distance between phylogenetic trees

STEPHANE GUINDON
From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages

JOSEPH HELED
The coalescent and local mixing

SIMON HILLS
Marrying molecules and morphology in marine molluscs

BARBARA HOLLAND
Distance corrections for Dollo data

JENNIFER HOYAL CUTHILL
Cophylogenetics and the study of bacterial lateral gene transfer

LARS JERMIIN
MtArt-2012: A new MtArt model of amino acid substitution

STEFFEN KLAERE
Goodness of fit tests in phylogeny

AARON KLEINMAN
The size of common subtrees of pairs of phylogenetic trees

DENISE KÜHNERT
Inferring epidemiological parameters while reconstructing phylogenetic histories

JESSICA LEIGH
Where now for haplotype networks? A review and some open problems *[no abstract]*

DAVID A. LIBERLES
Towards a general model for duplicate gene retention and loss

SIMONE LINZ
Picking cherries to merge phylogenetic trees into a temporal network

DAVID MARSHALL
Difficulties with tree-length estimation in likelihood-based phylogenetic analyses

BENNET MCCOMISH
Multiple optima of likelihood on trees from real data

JONATHAN MITCHELL
Distinguishing convergence in phylogenetic models

DAVID PENNY
Are the processes of microevolution sufficient for macroevolution.- integrating molecular evolution over different time scales

ANTHONY POOLE
Illuminating the twilight zone of sequence similarity: use of phylogenetic networks of protein structure for classification of superfamilies

STEFAN PROST
Implementing demographic models in the study of human history in the Pacific

LOUIS RANJARD
Estimating dispersal from genealogies

RAAZESH SAINUDIIN
Ancestries of recombining population pedigrees

SCOTT C. SCHMIDLER
Bayesian protein structure alignment and phylogeny

CHARLES SEMPLE
Realizing phylogenetic networks with local information

HELEN SHEARMAN
A comparison of phylogenetic diversity and maximum minimum distance

CHRIS SIMON
A more detailed look at species swarms in the NZ cicada genus *Kikihia*

MIKE STEEL
Peculiar properties of Penny's 'pully principle'

KATHERINE ST. JOHN
Walks of treespace

JEREMY SUMNER
Is the general time-reversible model bad for molecular phylogenetics?

LEO VAN IERSEL
Approximating the minimum number of reticulations needed to explain two conflicting gene trees

MICHAEL WOODHAMS
The triangle inequality and additivity in phylogenetic distances

CHIEH-HSI WU
Bayesian model selection of substitution models and their site assignment

RURIKO YOSHIDA
Nonparametric estimation of phylogenetic tree distributions

BOJIAN ZHONG
Systematic error in seed plant phylogenomics

JOE ZHU
Clade probabilties under neutral evolutionary models

Abstracts

Author: Remco Bouckaert

Title: BEAST 2

Abstract:

BEAST 2 is the latest release and a complete rewrite of the popular BEAST 1 software. In this talk, we highlight the main new features, including check pointing, add-on architecture, and BEAUti support for not only writing but also reading XML. A few new models will be demonstrated that are implemented in BEAST 2, but not BEAST 1. One aim of BEAST 2 is to make it much easier for developers to implement new phylogenetic models and in this talk we point out how this is achieved and show ways to get you started. For more information and pointer to documentation see http://beast2.cs.auckland.ac.nz.

Author: Michael Charleston

Title: Practical approaches to cophylogenetic analysis

Abstract:

Cophylogenetic analysis is the study of historical relationships between groups of ecologically linked taxa, based on the phylogenies of both and known associations between their tips, of the form "this louse infects that pocket-gopher" or "this virus infects these primates". The problem is often approached using cophylogeny mapping techniques, where the "dependent" phylogeny, P, is mapped into the "independent" phylogeny, H, so as to minimise an overall cost function based on the coevolutionary events that are implied by the mapping. This is by far the most intuitive and interpretable method but it is computationally intractable. We present a new integer linear programming (ILP) formulation that can be used to deal at least in part with this intractability. We show that the ILP is correct and that it can solve realistic instances of the problem (up to approximately 40 host tips) in reasonable time.

(Joint work with: Bin Zhou)

Author: Josh Collins

Title: Finding maximal acyclic agreement forests

Abstract:

In some talks I've given recently I've referenced finding an acyclic agreement forest for a set of trees that is minimal under a particular ordering. This talk will go into more depth on the ordering, the algorithm to find one `quickly', and whether we can say anything sensible about the distribution of maximal acyclic agreement forests among acyclic agreement forests.

Author: Mareike Fischer

Title: The impact of 'non-heredity' on tree reconstruction in practice

Abstract:

In a recent study, we investigated a conjecture by Arndt von Haeseler concerning the Maximum Parsimony method for phylogenetic estimation. The conjecture suggested that a Maximum Parsimony tree for a particular (DNA) alignment necessarily has subtrees of all possible sizes which are most parsimonious for the corresponding subalignments. We answered the conjecture affirmatively for some special cases but negatively for the general case. Biologically, the latter case was very counterintuitive as it suggested that even a unique most parsimonious tree on n taxa need not have anything in common with the most parsimonious trees(s) on n-1 taxa – a scenario which we refer to as 'non-heredity'.

However, it remained an open question how relevant this problem is in practice and what impact different evolutionary models have on the occurrence of non-heredity. Therefore, I will present in my talk some new simulation studies that tackle this question and explain for which models non-heredity can or cannot be ignored.

Author: Paul Gardner

Title: Annotating a plethora of RNA motifs

Abstract:

Identifying the function of RNAs turns out to be a ridiculously hard problem. Yet the discovery of novel RNAs of unknown function is accelerating. In this talk I will discuss how I am building tools for addressing this problem. My approach is to annotate known and identify novel RNA motifs that are involved in a range of functions, including catalysis, structure and transcription signaling.

Author: Gillian Gibb

Title: Next-generation mitogenomics: new results from modern and fossil xenarthrans.

Abstract:

Xenarthra is one of the four major clades of placental mammals, and is represented today by just three living groups: armadillos, sloths and anteaters. These unusual species are the relicts of a South American radiation of mammals that diversified during the Tertiary, but had become largely extinct by the end of the Pleistocene.

Despite the phylogenetic distinctiveness of Xenarthrans, mitochondrial genome data has been sorely lacking, with only four genomes reported so far. Using next-generation sequencing technology, we increased the number of complete Xenarthran mitochondrial genome sequences from four to 27.

These new genomes include representatives of almost all species in the 14 extant genera of Xenarthra, including the rare and cryptic fairy armadillos from the genera Chlamyphorus and Calyptophractus.

The sequencing of these distantly related genomes was possible using de-novo assembly of Illumina single-end reads generated from indexed genomic libraries on a single Illumina lane. We show that this strategy allows successful genome assembly from several museum specimens, opening new possibilities for exploring this understudied museo-diversity.

Moreover, using the same strategy we successfully retrieved the complete mitochondrial genome of the extinct ground sloth Mylodon darwinii from a 10,000 year old bone sample at more than 50x coverage. These new mitogenomic data will allow the reconstruction of a comprehensive phylogenetic framework and timescale incorporating the recently extinct diversity of xenarthrans.

(Joint work with: Melanie Kuch, Hendrik Poinar and Frederic Delsuc)

Author: NICOLE GRUENHEIT

Title: Transcriptomics of New Zealand alpine plants

Abstract:

Transcriptome analysis is increasingly being used to study the evolutionary origins and ecology of non-model plants. One issue for both transcriptome assembly and differential gene expression analyses is the recurrent occurrence in plants of whole genome duplication (WGD) and hybridization resulting in higher ploidy levels. The divergence of duplicated genes following WGD creates near identical homeologues that can be problematic for de novo assembly and also reference based assembly protocols that use short reads (35 – 100bp). Here we report a successful strategy for the assembly of two transcriptomes made using 75 bp Illumina reads from *Pachycladon fastigiatum* and *Pachycladon cheesemanii*. Both are allopolyploid plant species (2n=20) that originated in the New Zealand Alps about 0.8 million years ago. In a systematic analysis of 19 different coverage cutoffs and 20 different k-mer sizes we showed that i) none of the genes could be assembled across all of the parameter space ii) assembly of each gene required an optimal set of parameter values and iii) these parameter values could be explained in part by different gene expression levels and different degrees of similarity between genes. To obtain optimal transcriptome assemblies for allopolyploid plants, k-mer length and coverage cutoff need to be considered simultaneously across a broad parameter space. This is important for assembling a maximum number of full length ESTs and for avoiding chimeric assemblies of homeologous and paralogous gene copies.

Author: STEFAN GRÜNEWALD

Title: The quartet distance between phylogenetic trees

Abstract:

The quartet distance is one way to quantify how different two phylogenetic trees on the same taxa set are. It is defined to be the number of subsets of cardinality 4 of the taxa set for which the restrictions of the trees are different. Bandelt and Dress showed in 1986 that the maximum distance between two binary trees, when normalized by the number of all 4-sets, is monotone decreasing with n. They conjectured that the limit of this ratio is 2/3. I will generalize the conjecture to not necessarily binary trees. It turns out that the general conjecture is equivalent to the original one. An

advantage of the new formulation is that we can look at simple trees where the number of splits stays small while the number of taxa goes to infinity. I will give a proof for the simplest cases and speculate how the problem might be solved in the future.

Author: STEPHANE GUINDON

Title: From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages

Abstract:

The accuracy and precision of species divergence date estimation from molecular data strongly depend on the models describing the variations of substitution rates along a phylogeny. These models generally assume that rates randomly fluctuate along branches from one node to the next. However, for mathematical convenience, the stochasticity of such process is ignored when translating rate trajectories into branch lengths. I this talk, I will show that this simplification has a strong impact on the precision with which the autocorrelation of rates along a phylogeny is estimated. I will describe a new approach that explicitly considers the average substitution rates along branches as random quantities, resulting in a more realistic description of the time-dependent variations of evolutionary rates. This new method provides more precise estimates of the rate autocorrelation parameter as well as divergence date estimates. Altogether, the proposed approach is a step forward to designing biologically relevant models of rate evolution that are well-suited to data sets with dense taxon sampling which are likely to present rate autocorrelation.

Author: JOSEPH HELED

Title: The coalescent and local mixing

Abstract:

Under the coalescent, the parent of each individual is chosen at uniform from all possible individual in the previous generation. This is known as perfect mixing, and is, of course, totally unrealistic.

Using simulations I will contrast the classic coalescent with a local mixing model, where parents have to be geography near.

Author: SIMON HILLS

Title: Marrying molecules and morphology in marine molluscs

Abstract:

Proper consideration of the evolution of life on earth must make a suitable phylogenetic evaluation of the 99% of organisms that are now extinct. Due to the lack of preserved DNA for the overwhelming majority of extinct organisms, the only data by which these evolutionary relationships can be assessed is morphological. However, independent phylogenetic analysis of molecular and morphological data usually reveals substantially different signals. Poor resolution, convergence and ecophenotypic variability plague phylogenetic reconstructions where molecular and morphological based analyses are compared.

Using robust molecular and morphometric datasets we explore these issues in the New Zealand marine mollusc genus Alcithoe. In evaluating disagreement between molecular and morphological data in Alcithoe we found that an ecological variable, maximum habitat depth of species, is correlated with a significant conflicting signal in the morphological dataset. We then tested if a phylogeny that is more similar to the molecular based reconstruction can be generated from the morphological data after filtering out characters that correlated with water depth. Furthermore, we examine the rates of morphological evolution of selected characters in order to assess the effect of rate heterogeneity on phylogenetic reconstruction with morphological characters.

Through this examination of molecular and morphological data with extant species we aim to identify an optimal set of morphological characters that will enable more accurate phylogenetic reconstruction and the inclusion of extinct species.

(Joint work with: James Crampton)

Author: BARBARA HOLLAND
Title: Distance corrections for Dollo data

Abstract:
We investigate distances on binary (presence/absence) data in the context of a Dollo process, where a trait can only arise once on a phylogenetic tree but may be lost many times. We introduce a novel distance, the Additive Dollo Distance (ADD), and derive it by multiple independent lines of reasoning including an intriguing link to the LogDet distance. Simulations of Dollo data are used to compare a number of binary distances including ADD, LogDet, Nei Li and some simple, but to our knowledge previously unstudied, variations on common binary distances. We apply the ADD in two different contexts. The first application is to a recent Diversity Arrays Technology (DArT) data set to study the phylogeny of Eucalyptus. The second application is to gene family presence/absence on bacteria from the COG database. This data has previously been used in a conditioned genome reconstruction (CGR) approach. We demonstrate that the LogDet distance performs poorly in the context of a Dollo process, which may have implications for its use in connection with CGR.

(Joint work with: Michael Woodhams, Dorothy Steane, and Vincent Moulton)

Author: JENNIFER HOYAL CUTHILL
Title: Cophylogenetics and the study of bacterial lateral gene transfer

Abstract:
Cophylogenetic analysis is a method for inferring the joint history of two associated phylogenies (the "host" and "associate"). Cophylogenetics can be applied to a number of analogous co-evolutionary systems (parasites and hosts, species and biogeographic areas, genes and species), which can be described using a common set of historical events. These events include codivergence (joint

divergence of host and associate lineages), duplication (divergence of an associate lineage without host divergence), loss (due, for example, to failure to colonize a divergent host), and finally host switch (divergence of an associate lineage followed by a colonization of a new host). This host switch model can represent novel parasite infections, migration of species, or the lateral transfer of genes between species. By explicitly modelling lateral gene transfer in this way, cophylogenetics may aid the study of this importance mechanism of microbial evolution. However, modelling lateral gene transfer presents some ongoing challenges, particularly where transferred genes show low sequence divergence. This talk will describe a sequence similarity analysis of antibiotic resistance genes in pathogenic Gammaproteobacteria (focussed on Vibrio cholera the agent of cholera), which suggests multiple lateral gene transfers. The talk will then explore the use of cophylogenetic analysis in the search for the evolutionary origins of these antibiotic resistance genes.

(Joint work with: Michael Charleston and Paul Greenfield)

Author: LARS JERMIIN

Title: MtArt-2012: A new MtArt model of amino acid substitution

Abstract:

In this seminar, we focus on the role that model misspecification has played in a recent study of molecular evolution. The study led to the development of the MtArt model of amino acid substitution (Mol. Biol. Evol. 24, 1-5 [2007]), which is now often used in phylogenetic studies of polypeptides encoded by arthropod mitochondrial genomes.

An initial survey of the alignment of amino acids used to infer the MtArt model indicates that the data are unlikely to have evolved under globally stationary, reversible, and homogeneous (SRH) conditions, implying that a critical assumption behind the use of the MtArtmodel was violated during its estimation. Using column-and-row permutations of a heat map containing p-values frommatched-pairs tests of symmetry allowed us to identify a subset of sequences that are consistent with evolution under globally SRH conditions. Using these sequences, we inferred a revised MtArt model of amino acid substitution that is consistent with its intended purpose (i.e., to estimate evolutionary patterns and processes under globally SRH conditions). The new MtArt model is the first amino acid substitution model to have been inferred using sequences found to be consistent with evolution under globally SRH conditions.

The effect of using the new MtArt model, instead of the old MtArt model, is illustrated using a phylogenetic study of polypeptides encoded by the mitochondrial genomes of Bactrocera, a dipteran genus.

(Joint work with: Iker Irisarri, Federico Abascal, David Posada and Rafael Zadoya)

Author: Steffen Klaere

Title: Goodness of fit tests in phylogeny

Abstract:

The purpose of a goodness of fit test is to determine whether a reconstructed model is actually supported by the data. There have been a few tests introduced before, with varying success. One major issue regards the high amount of rejections of models. Thus, most users decide to ignore such tests.

Recently, we have introduced an alternative test, MISFITS, which acknowledges the fact of rejection and proposes to decompose the data into model generated and "other". While this approach is more appealing to users, the tools employed are still under debate.

I will discuss here the merits of our method, propose adjustments to the tools, and address some questions posed by users.

Author: Aaron Kleinman

Title: The size of common subtrees of pairs of phylogenetic trees

Abstract:

Let T be a phylogenetic X-tree. Given a subset $Y \subseteq X$ we can consider the Y-tree $T|_Y$ obtained by restricting T to the leaves in Y and contracting edges whose endpoints are vertices are of degree 2. We show there is a constant c such that for any size X and any two X-trees T_1, T_2 there is a $Y \subseteq X$, $|Y| \geq c \log |X|$ such that $T_1|_Y = T_2|_Y$. This answers a question posed by Mike Steel.

Author: Denise Kühnert

Title: Inferring epidemiological parameters while reconstructing phylogenetic histories

Abstract:

Rapidly evolving viruses such as HIV, HCV and Influenza virus are of major interest in phylogenetics. They are special in that their ecological and evolutionary processes act on the same timescale and are therefore entangled. The cross-reaction of the two processes must be accounted for when inferring epidemiological parameters and/or phylogenetic history.

Our aim is a joint epidemiological phylogenetic, i.e. phylodynamic, analysis of genomic data by incorporating the dynamics of an SIR model into Bayesian phylogenetic inference. A new version of the birth-death model (Stadler, 2010) that incorporates sampling-through-time and can also be extended to allow birth, death and sampling rates to change over time can be parametrized to facilitate modeling SIR-like population dynamics while reconstructing phylogenetic history and simultaneously estimating epidemiological parameters.

Author: David A. Liberles

Title: Towards a general model for duplicate gene retention and loss

Abstract:

A mechanistic model is proposed that is consistent with the expected different time-dependent loss/retention profiles for the nonfunctionalization, neofunctionalization (plus nonfunctionalization), subfunctionalization (plus nonfunctionalization), and dosage balance mechanisms. Preliminary validation of this model with simulated data from a network of duplicated interacting protein-encoding genes was performed. This is a step towards the development of a probabilistic gene tree/species tree reconciliation method that enables mechanistic inference of underlying evolutionary mechanisms acting on duplicate genes.

(Joint work with: Ashley I. Teufel)

Author: Simone Linz

Title: Picking cherries to merge phylogenetic trees into a temporal network

Abstract:

Phylogenetic networks are now frequently used to explain the evolutionary history of a set of species for which a collection of gene trees reveals inconsistencies. However, the reconstructed networks are not always temporal. If a phylogenetic network is temporal, then it satisfies the time constraint of instantaneously occurring reticulation events; i.e. all species that are involved in such an event coexist in time. Furthermore, although a collection of phylogenetic trees can often be merged into a temporal phylogenetic network, many algorithms do not necessarily find it since their primary optimization objective is the minimization of the number of reticulation events. In this talk, we present a characterization for when two rooted binary phylogenetic trees admit a temporal phylogenetic network. This characterization is based on the existence of a particular type of sequence with which cherries can be picked from two phylogenies. Furthermore, we explore some algorithmic questions related to calculating the minimum number of reticulation events needed to merge two phylogenies into a temporal network.

(Joint work with: Peter J. Humphries and Charles Semple)

Author: David Marshall

Title: Difficulties with tree-length estimation in likelihood-based phylogenetic analyses

Abstract:

Recent studies have drawn attention to problems with branch- and tree-length estimation in Bayesian phylogenetic analysis. Phylogenetic signal for tree length is often poor compared to that of topology and other model parameters, and Bayesian priors can have undesired effects on tree-length estimation. Even without Bayesian priors involved, likelihood-based methods can suffer unexpectedly severe distortion of tree-length, especially in analyses using a priori data partitioning and complex substitution models. Topological conflict within a data partition can cause distortion of

partition-rate multipliers and large biases in tree length, usually toward longer trees. Tree-length bias is of the greatest significance when the branch length or substitution rate itself is a parameter of interest, as in some divergence-time analyses, but it also may affect topology and node support.

Author: BENNET MCCOMISH

Title: Multiple optima of likelihood on trees from real data

Abstract:

It is known that the maximum likelihood function can have multiple local optima on a given tree. Some simulation studies suggest that this is not likely to affect tree-building, but these simulations used data generated on a single tree. In contrast, it has been shown that simple mixture models can generate data where multiple optima can occur even on the tree with the highest likelihood. Here we present results generated using actual biological sequence data, which tend to confirm the simulation results, providing further reassurance that the value of maximum likelihood as a tree selection is not often compromised by the presence of multiple local optima.

(Joint work with: Klaus Schliep and David Penny)

Author: JONATHAN MITCHELL

Title: Distinguishing convergence in phylogenetic models

Abstract:

Given a set of taxa, can the probability distribution from the model where those taxa evolved independently with no convergence be distinguished from the model where convergence occurred? This is a particularly important question because it involves determining whether the case of genetic cross-over (eg. hybridisation or horizontal gene transfer) is distinguishable from the case of the evolution of independent lineages. The constraints on the probability distributions arising from the models will be analysed by transforming the distributions into their appropriate basis (i.e. the Hadamard basis for the binary symmetric model). By comparing the probability spaces from two models it can be determined whether the models are distinguishable from each other. I will start by looking at the simplest model, the two taxa binary symmetric model, before looking at the more complicated three and four taxa models.

Author: DAVID PENNY

Title: Are the processes of microevolution sufficient for macroevolution.- integrating molecular evolution over different time scales

Abstract:

We need a more formal analysis of the continuity of evolutionary processes over widely different time scales, and this has been done to some extent with RNA viruses. The so-called molecular clock of molecular evolution showed that the rate of neutral mutations per generation also equalled the long-term of change. However, there is now more interest on the apparent acceleration of shortterm rates, or the apparent acceleration of rates when speciation is occurring. We have extended the calculations into the 'twilight zone' before fixation or loss of mutations has occurred. There are interesting, and predictable effects from population structures modelled as different forms of connected graphs. We find a range of effects from those of population size, population structure, and speciation. Similarly, there are popular hypotheses with regard to extinctions at the Cretaceous/Tertiary boundary that make unwarranted assumptions about the insufficiency of microevolutionary processes.

Author: ANTHONY POOLE

Title: Illuminating the twilight zone of sequence similarity: use of phylogenetic networks of protein structure for classification of superfamilies

Abstract:

There are numerous cases where macromolecular structure appears better conserved than sequence. However, without clear models for how structure evolves over time, it can be difficult to make objective assessments of homology versus convergence. As a step towards addressing this problem, we sought to examine whether protein structure carries a detectable evolutionary signal consistent with signal at the sequence level, and whether existing forms of structure-based classification identify this signal. To this end, we present a structure-based phylogenetic network of the ferritin-like superfamily. This superfamily encompasses proteins with a broad range of biological functions, all of which share a common structural core, but amino acid sequence similarity across its members is extremely low. To assess the extent to which structural data may carry evolutionary information, we compared the evolutionary relationships suggested from our structure-based phylogenetic network with available independent evidence, including sequence-based subphylogenies, dimerisation geometries, and classification in structure-based protein databases. Our structure-based phylogenetic network recovers established major families, indicating compelling overlap between these independent forms of evidence, and suggesting that considerable evolutionary signal is preserved in three-dimensional structures. An advantage of structural phylogenies over other classification schemes is that these enable examination of specific evolutionary relationships between structures. While assigning homology beyond the twilight zone of sequence similarity still requires human inference, the approach described here improves the evidence-base from which to draw such conclusions.

Author: STEFAN PROST

Title: Implementing demographic models in the study of human history in the Pacific

Abstract:

The Pacific is an outstanding region to study human prehistory and evolution. Colonization of the Pacific comprises the earliest movement of modern humans out of Africa with the subsequent settlement of Australia and Papua New Guinea (Near Oceania) about 50.000 years ago and the latest movement of people into Remote Oceania about 3.500 years ago. The study of human demographic history in the Pacific is currently on an inflection into a new era. New sequencing technologies, improved techniques to retrieve DNA from subfossil remains and new data analysis approaches will most likely revolutionize our understanding of this complex human achievement. A problem with most statistical analyses applied to study human settlement in the Pacific is that conclusions about the underlying demography and demographic changes are subjective and thus cannot give quantifiable support for or against different underlying scenarios. Here I will present the use of model-based analysis to get a better insight into this complex human achievement and outline pros and cons of this approach.

Author: LOUIS RANJARD

Title: Estimating dispersal from genealogies

Abstract:

Extensive sampling of individuals from populations of interest is becoming common in ecological studies. Phylogenetic reconstructions from such data are used to assess the corresponding genealogies and estimate evolutionary histories. Moreover, one can try to understand the history of the colonisation of the space to identify source populations, migration corridors or the migration behaviour of the species under study. We propose a niche model of colonisation where the probability of migrating from one location to another depends on the geographical distance and whether the destination location has been colonised in the past and is therefore occupied or not. Using a Bayesian framwork, we aim to estimate the parameters of this model of colonisation given the localisation of sampled individuals and their genealogy. We simulate migration histories in a two-dimensional finite space and use our approach to estimate the model parameters. Results about these simulations will be presented.

Author: RAAZESH SAINUDIIN

Title: Ancestries of recombining population pedigrees

Abstract:

We derive the exact one-step transition probabilities of the number of lineages that are ancestral to a random sample from the current generation of a bi-parental population that is evolving under the discrete Wright-Fisher model with n diploid individuals. Our model allows for a per-generation recombination probability of r. When r=1 our model is equivalent to Chang's model for the zygotic pedigree. When r=0 our model is equivalent to Kingman's coalescent model for the cytoplasmic, mitochondrial or sub-karyotic tree defined by a DNA locus that is free of intra-locus recombination, and when 0 < r <1 our model can be thought to track the cytoplasmic pedigree with paternal leakage probability r or to track a sub-karyotic pedigree defined by a haploid DNA locus from an autosomal chromosome that has an intra-locus recombination probability r. Thus, our discrete-time Markov chain model is an r-homotopy that contains Chang's model for the zygotic pedigree, Kingman's discrete coalescent model for the cytoplasmic tree, and the discrete sub-karyotic pedigree model that may be approximated by Hudson's and Griffiths' ancestral recombination graph (ARG). We provide the first explicit transition probabilities of this discrete time Markov chain of the number of ancestors to a random sample from the present time and study its stationary distribution. We study three properties of this r-specific ancestral size Markov chain: time to most recent common ancestor (MRCA) of the population, time at which all individuals are either common ancestors to all present day individuals or ancestral to none of the present day individuals, and the fraction of common ancestors at this time. These results generalize the two main results in Chang's model.

(Joint work with: Bhalchandra Thatte)

Author: SCOTT C. SCHMIDLER

Title: Bayesian protein structure alignment and phylogeny

Abstract:

We present a stochastic process model for the joint evolution of protein primary and tertiary structure, suitable for use in alignment and estimation of phylogeny. Indels arise from a classic Links model and mutations follow a standard substitution matrix, while backbone atoms diffuse in three-dimensional space according to an Ornstein-Uhlenbeck process. The model allows for simultaneous estimation of evolutionary distances, indel rates, structural drift rates, and alignments, while fully accounting for uncertainty. The inclusion of structural information enables phylogenetic inference on time scales not previously attainable with sequence evolution models. The model also provides a tool for testing evolutionary hypotheses and improving our understanding of protein structural evolution.

Author: : CHARES SEMPLE

Title: Realizing phylogenetic networks with local information

Abstract: Results that say local information is enough to guarantee global information provide the theoretical underpinnings of many reconstruction algorithms in evolutionary biology. Such results include Buneman's Splits-Equivalence Theorem and the Tree-Metric Theorem. The first result says that, for a collection $\mathcal C$ of binary characters, pairwise compatibility is enough to guarantee compatibility for $\mathcal C$, that is, there is a phylogenetic (evolutionary) tree that realizes $\mathcal C$. The second result says that, for a distance matrix D, if every 4×4 distance submatrix of D is realizable by an edge-weighted phylogenetic tree, then D itself is realizable by such a tree. In this talk, we investigate results of this type for structures more general than trees.

Author: HELEN SHEARMAN

Title: A comparison of phylogenetic diversity and maximum minimum distance

Abstract:

Phylogenetic diversity and maximum minimum distance each aim to maximise the diversity in a set of species for conservation biology. In this talk I will compare the performance of these two methods. I will compare in what settings maximum minimum distance captures greater diversity than phylogenetic diversity and vice versa. This will include the robustness of these methods to errors in the distance estimates in the phylogeny and distance matrices.

Author: CHRIS SIMON

Title: A more detailed look at species swarms in the NZ cicada genus *Kikihia*

Abstract:

Last year at Leigh I discussed the idea that Pleistocene interglacial contact between recently diverged species in the NZ grass-cicada "*Kikihia muta* complex" has resulted in a complex pattern of gene exchange across species boundaries. We used microsatellite loci to estimate current and past gene flow at secondary contact between lineages defined on the basis of courtship songs, morphological traits, and mitochondrial phylogenies. Our phylogeographic studies of the genus *Kikihia* identified 20 potential hybrid zones between species pairs that vary widely in their times of mtDNA divergence (between 20,000 and 3.5 million years). These well-supported molecular phylogenies, dated using Bayesian molecular relaxed-clock methods, were used as a framework to understand species interactions. The mating song and female response of each species (controlling pre-zygotic isolation), has also been characterized throughout each species' range. That preliminary study examined the introgression of alleles at three contact zones involving four species (*Kikihia "nortwestlandica"*, *K. "southwestlandica"*, *K. muta muta*, and *K. "tuta"*). This year I present additional data representing many more localities for the west coast contact zone between *K. "nortwestlandica"*, *K. "southwestlandica"*. We find that, as often happens with more sampling, the story is more complicated that initially realized. Introgression from a third species, *Kikihia "murihikua"*, has affected allele frequencies in the southernmost populations and that introgression between the two "westlandica" species is more widespread.

(Joint work with: Beth Wade)

Author: KATHERINE ST. JOHN

Title: Walks of treespace

Abstract:

An NNI-walk is a sequence of trees, $T_1,T_2,\ldots,T_k$, where each tree, T_i , differs from T_{i+1} ,by a single NNI move. David Bryant's Combinatorial Challenges ask what is the shortest walk that visits all trees and what is the shortest walk that visits every tree in an SPR neighborhood. We will address recent progress on these intriguing conjectures.

Author: MIKE STEEL

Title: Peculiar properties of a polarity 'pully principle'

Abstract:

Neutral macroevolutionary models, such as the Yule model, provide a prior distribution on rooted trees under certain assumptions on the random process of speciation and extinction. Such models

can provide a strong signal as to the approximate location of the root when only the unrooted phylogenetic tree is known, particularly as the number of leaves grow. On the other hand, some models are completely uninformative with respect to root location since they convey the same probability to any re-rooting of a tree (i.e. they satisfy a type of 'pulley principle'). Here we show that there is just one model that satisfies this property and which is sampling consistent; all other discrete phylogenetic models that are sampling consistent convey some information as to the location of the ancestral root in an unrooted tree.

Author: JEREMY SUMNER

Title: Is the general time-reversible model bad for molecular phylogenetics?

Abstract:

The general time reversible model (GTR) is presently the most popular model used in phylogenetic studies. However, GTR substitution matrices are not closed under matrix multiplication. In this talk I will give examples that demonstrate why this deficit may pose a problem for phylogenetic analysis and interpretation. I will also briefly discuss recent progress on classifying models which are algebraically closed: the so-called "Lie Markov models".

Author: LEO VAN IERSEL

Title: Approximating the minimum number of reticulations needed to explain two conflicting gene trees

Abstract:

There are several reasons why different genes can have different evolutionary histories. One possible cause are reticulate evolutionary events, such as recombination, hybridization and horizontal gene transfer. In the presence of such events, two genes can have different gene trees, and a well-known problem is to compute the minimum number of reticulations necessary to explain the conflicts between these trees. A more constructive variant of this problem is to generate a phylogenetic network that displays both gene trees and has a minimum number of reticulations. This well-studied problem is known to be NP-hard and fixed parameter tractable. However, less is known about the possibility to approximate this problem. Is it possible to find a solution in polynomial time that is provably at most some constant factor away from the optimum?

Author: MICHAEL WOODHAMS

Title: The triangle inequality and additivity in phylogenetic distances

Abstract: A phylogenetic distance is additive if, under some model of evolution, the distance between two sequences is (in the absence of sampling error) equal to the divergence time between the sequences. Informally, a distance formula satisfies the triangle inequality if you can never get from A to B faster via a third point C than you can by travelling directly. I attempt to prove that these two properties are mutually exclusive.

Author: Chieh-Hsi Wu

Title: Bayesian model selection of substitution models and their site assignment

Abstract:

There are a number of models proposed to elucidate across-site variation of substitution pattern or rate, as well as methods that perform model selection. However, some questions still remain, including which substitution models should be used at which sites of the sequence alignment. Here, this is referred to as the substitution model partition problem. We present a Dirichlet process mixture (DPM) model for nucleotide alignment data that accounts for across-site heterogeneity of substitution pattern and rate simultaneously. This model estimates the number categories of substitution model and rate, in addition to the assignment of sites to categories. Moreover, it enables Bayesian selection over a set of standard nucleotide substitution models for each substitution model category. We investigate whether the inference on coalescent parameters produced by analyses using the DPM model are significantly different to those produced by the analyses assuming that substitution pattern is homogeneous across sites.

Author: Ruriko Yoshida

Title: Nonparametric estimation of phylogenetic tree distributions

Abstract:

As population-based models of gene trees such as coalescents have been developed to more accurately model distributions of gene trees across the genome, meanwhile detection of horizontal gene transfer and discordances among gene trees have become important problems in phylogenetics. In this talk we take a nonparametric approach to estimating distributions of phylogenetic trees, and identifying outliers in a sample which may not have been drawn from the same distribution as the majority of the sample. As an application, the main evolutionary process generating gene tree histories can be estimated from observed gene trees, and “outlier” gene trees can be identified which potentially have singular evolutionary events such as unusual neofunctionalization or horizontal gene transfer. Our approach mimics kernel density estimation, using popular distances between trees to define “kernels.” In short, the estimated probability of a tree *T* is influenced by empirical frequencies of nearby trees, with the level of influence determined by the distance to *T*.

In contrast to parametric models such as coalescents, using a nonparametric approach avoids issues such as model mis-specification which might potentially confound detection of outlier trees. We compare our methods to comparable one-class support vector machines from machine learning, and demonstrate superior accuracy in outlier detection on simulated data. We also apply our methods to a well known dataset of yeast genes, identifying the same outlier genes as previous analysis. Our software `GKDEtrees` for estimating tree distributions is implemented in R and c++ and is freely available for download.

(Joint work with: D. Haws, P. Huggins, and G. Weyenberg)

Author: BOJIAN ZHONG

Title: Systematic error in seed plant phylogenomics

Abstract:

Resolving the closest relatives of Gnetales has been an enigmatic problem in seed plant phylogeny. The problem is known to be difficult because of the extent of divergence between this diverse group of gymnosperms and their closest phylogenetic relatives. Here we investigate the evolutionary properties of conifer chloroplast DNA sequences. To improve taxon sampling of Cupressophyta (non-Pinaceae conifers) we report sequences from three new chloroplast (cp) genomes of Southern Hemisphere conifers. We have applied a site pattern sorting criterion to study compositional heterogeneity, heterotachy and the fit of conifer chloroplast genome sequences to a GTR + G substitution model. We show that non-time reversible properties of aligned sequence positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction of these seed plants. When 2250-3000 of the most varied sites in our concatenated alignment are excluded, phylogenetic analyses favour a close evolutionary relationship between the Gnetales and Pinaceae – the Gnepine hypothesis. Our analytical protocol provides a useful approach for evaluating the robustness of phylogenomic inferences. Our findings highlight the importance of goodness of fit between substitution model and data for understanding seed plant phylogeny.

(Joint work with: Oliver Deusch , Vadim V. Goremykin , David Penny, Patrick J. Biggs , Robin A. Atherton and Peter James Lockhart)

Author: JOE ZHU

Title: Clade probabilities under neutral evolutionary models

The Yule model and the coalescent model are two neutral stochastic models for generating trees: a rooted Yule tree describes the speciation from the top of the tree; a coalescent tree models lineages coalesce back in time from bottom of the tree. Although, these two models are quite different, they lead to exactly the same distributions of tree topologies, as well as the probabilities of monophyletic groups (clades). In this project, we extended earlier work, derived exact formulas of computing probabilities of clades in rooted trees, and extended these probabilities for clans in unrooted cases.