Phylomania

The UTas theoretical phylogenetics meeting

University of Tasmania
School of Maths and Physics
29-30 Oct 2009

Program

Thursday, 29 October

9:00am-9:40am	Registration and coffee
9:40am-9:50am	Welcome
9:50am-10:30am	**John Rhodes**, University of Alaska (Fairbanks) *Identifiability of Phylogenetic Models for Morphological Data*
10:30am-11:00am	Morning tea
11:00am-11:40am	**Jeremy Sumner**, University of Tasmania *The algebra of Markov models on phylogenetic split networks*
11:40am-12:20pm	**Stefan Gruenewald** CAS-MPG Partner Institute for Computational Biology, Shanghai *Quartet-Based Methods to Construct Phylogenetic Networks*
12:20pm-1:40pm	Lunch
1:40pm-2:20pm	**Klaas Hartmnan**, University of Tasmania *Challenges in prioritising birds for biodiversity conservation on a global scale*
2:20pm-3:00pm	**Barbara Holland**, Massey University *A powerful low-parameter method for inferring quartets under the General Markov Model*
3:00pm-3:30pm	Afternoon Tea
3:30pm-4:10pm	**Michael Woodhams**, University of Sydney *The Mosaic Model of Nucleotide Substitution*
4:10pm-4:50pm	**Roger Brown**, Australian National University *Distance measures derived by spectral methods applied to categorical sequences*

Friday, 30 October

9:20am-9:50am	Coffee
9:50am-10:30am	**David Bryant**, University of Auckland *How diverse is your tight span?*
10:30am-11:00am	Morning tea
11:00am-11:40am	**Conrad Burden**, Australian National University *Alignment-free sequence similarity measures using k-word matches*
11:40am-12:20pm	**Qiang Li** CAS-MPG Partner Institute for Computational Biology, Shanghai *Estimation of evolutionary distances based on word composition*
12:20pm-1:40pm	Lunch
1:40pm-2:20pm	**Michael Charleston**, University of Sydney *A likelihood method for cophylogenetics*
2:20pm-3:00pm	**James Degnan**, University of Canterbury *Identifiability of rooted species trees from unrooted gene tree distributions*
3:00pm-3:30pm	Afternoon Tea
3:30pm-4:10pm	**Peter Jarvis**, University of Tasmania *Markov invariants for phylogenetic rate matrices derived from embedded submodels*

Abstracts

Roger Brown

Institute: NCI National Facily, Australian National University
bregor@sf.anu.edu.au

Distance measures derived by spectral methods applied to categorical sequences

The spectra of a protein amino-acid sequences, viewed as a categorical sequences, contain information relating to the molecular structure of target proteins as well as features due to the periodic nature of alpha-helices and beta-strands etc. Thus, distance measures based on the comparison of such spectra look further into the characteristics of the molecules being compared than do measures that rely on site by site comparison of aligned sequences. Furthermore, the method when formulated in terms of sequences alone, makes no assumptions regarding substitution probabilities and similar concepts and thus is quite distinct from most other methods.

David Bryant

Institute: Department of Mathematics, University of Auckland, New Zealand
d.bryant@auckland.ac.nz

How diverse is your tight span?
Joint work with Paul Tupper.

The tight span is a simple and elegant construction that takes a metric and returns a cell complex (or a split network). Andreas Dress's epic *Trees, Tight Extensions of Metric Spaces, and the Cohomological Dimension of Certain Groups: A Note on Combinatorial Properties of Metric Spaces* should rightly viewed as the ancestor of all split based phylogenetic networks (spectronet, NeighborNet, split decompositon etc.). The construction has many fascinating and elegant properties, but the attraction for us phylogeneticist mathematicians is that when a metric is treelike, the method returns a tree, and as the metrics get less tree-like, the method returns networks that look less and less treelike.

Nevertheless, distance based methods as so 1980s. For several years we've been looking for ways of defining tight spans on pattern distributions, a statistically responsible tight span. Very very recently, we made a major step forward. We have found what we think is the appropriate tight span definition for *diversity measures*, generalisations of metrics that assign diversity values (like phylogenetic diversity) to subsets of the taxa set. I will introduce the tight span for metrics, then give our definition for diversity measures, talk about its tight span and, all going well, hint how this could provide a way to consistently construct networks under a general Markov model.

Conrad Burden

Institute: Centre for Bioinformation Science, Australian National University
conrad.burden@anu.edu.au

Alignment-free sequence similarity measures using k-word matches

A common problem faced by biologists is obtaining a measure of similarity between sequences related by common ancestry or related function. The most popular, currently available sequence matching algorithms attempt to align long sequences. This may not always be appropriate when related sequences have been rearranged or spliced, or when identifying short regulatory motifs. We

are developing an alignment free method, called k-word matches. The idea is to use as a comparison statistic the number of exact or partial short word matches of a given pre-specified length. We have found accurate representations of the statistical properties of word match counts under suitable null hypotheses, and are developing fast computer algorithms for biological applications.

Michael Charleston
Institute: School of IT, University of Sydney
m.charleston@usyd.edu.au

A likelihood method for cophylogenetics

Using phylogenies of ecologically linked taxon sets to infer historical patterns of coevolution is known to be a very hard problem, yet it is of great importance if we want to uncover the dynamics of coevolving species. Nothing evolves in isolation, after all.

Given the input of host (independent) tree H, dependent (parasite/pathogen) tree P and associations $\varphi : L(P) \mapsto L(H)$, we attempt to recover the best explanation for the differences between P and H. By far the most successful procedure in terms of interpretability is to map P into H, but this is computationally very demanding, even in its simplest forms: it's recently been shown to be NP-complete (in accordance with what one might expect of the "all biologically interesting computational problems are at least NP-complete" rule).

Another approach is to find the most probable reconstruction that could lead to our observed P and φ, given H and under some probabilistic model. This likelihood approach has seen little development, in part due to the mixed nature of discrete events and continuous processes. I present a new treatment of the problem that can be used to estimate the most likely reconstructions using Markov chain Monte Carlo methods, using a discretized model of cophylogenetic evolution.

James Degnan
Institute: Biomathematics Research Centre, University of Canterbury, New Zealand
J.Degnan@math.canterbury.ac.nz

Identifiability of rooted species trees from unrooted gene tree distributions
Joint work with Elizabeth Allman and John Rhodes.

A common distinction in phylogenetics is made between species trees and gene trees. Species trees describe the history of population divergences between related populations, while gene trees describe the evolutionary relationships that best describe the ancestry for a single gene sampled from different species. Using Kingman's coalescent applied to multiple species, the multispecies coalescent model describes probabilities of gene trees given a species tree. Each rooted, binary species tree topology on n species together with edge weights on internal branches of the species tree induces a distribution on the possible rooted, binary gene trees on the same n species. Properties of rooted gene tree distributions have been studied for several years. We consider distributions of unrooted gene trees by considering the probability under the multispecies coalescent model that a random rooted gene tree has a given unrooted topology. We show that for four species, the unrooted gene tree distribution does not identify the rooted species tree topology. However, we use linear invariants on unrooted gene tree probabilities to show that for any binary species tree with five or more species, unrooted gene tree distributions identify both the rooted species tree topology and its internal edge weights.

Stefan Gruenewald
Institute: CAS-MPG Partner Institute for Computational Biology, Shanghai, China
stefan@picb.ac.cn

Quartet-Based Methods to Construct Phylogenetic Networks

Usually, the aim of a phylogenetic analysis is to construct a tree. However, in case of reticulate evolution or ambiguous data, a single tree is not sufficient to display all observed signals. One solution that is increasingly appreciated by biologists is to construct a more general network than a tree. The most commonly used generalization of unrooted trees is a splits graph and the most successful methods to construct them are NeighborNet and Split Decomposition. Both methods use pairwise dissimilarities between the taxa as their input. However, when it is decided which splits are contained in the output networks, weighted quartets (splits of four of the taxa into two pairs) play a crucial role. Therefore, it is a natural approach to compute quartet weights directly from the raw data (e.g. sequences), rather than taking a detour via distances. I will present two methods following this approach, QNet and SuperQ, and some ideas for future work.

Klaas Hartmann
Institute: TAFI, University of Tasmania
Klaas.Hartmann@utas.edu.au

Challenges in prioritising birds for biodiversity conservation on a global scale

The Zoological Society of London runs the Edge of Existence programme. This programme aims to identify those species that are most evolutionarily distinct and globally endangered. I have been working on the mathematical framework underlying the global bird prioritisation that is currently being finalised. In this talk I will describe the approach we have taken and in particular, focus on the challenges resulting from the sparse dataset available (only about half of the world's bird species have useful sequence data).

Barbara Holland
Institute for Fundamental Sciences, Massey University, New Zealand
b.r.holland@massey.ac.nz

A powerful low-parameter method for inferring quartets under the General Markov Model
Joint work with Jeremy Sumner and Peter Jarvis

Sumner et al (2005, 2006, 2008) have developed a theory of Markov invariants that extends the Log-Det distance measure, which has long been used in phylogenetics, to larger sets of taxa. Here we focus on the invariants for quartets (dubbed squangles) and show that they provide a powerful tool for phylogenetic estimation. Simulations reveal that they have similar power to NJ and ML under simple models such as the Jukes-Cantor, but as expected by theory - they remain consistent when base frequencies drift over the tree, whereas other methods can become inconsistent. On the other hand, the invariant-based method can become inconsistent when data is generated under a rates across sites model and/or with proportions of fixed sites, however, parameters have to become quite extreme before this is the case. We discuss how the squangles can be used to derive weights for the different quartet topologies and how these can be used in combination with programs such as QNet and quartet puzzling to do larger phylogenetic analyses.

Peter Jarvis
Institute: School of Maths and Physics, University of Tasmania
Peter.Jarvis@utas.edu.au

Markov invariants for phylogenetic rate matrices derived from embedded submodels
Joint work with Jeremy Sumner.

We consider novel phylogenetic models with rate matrices that arise via the embedding of a progenitor model on a small number of character states, into a target model on a larger number of character states. Adapting representation-theoretic results from recent investigations of Markov invariants for the general rate matrix model, we give a prescription for identifying and counting Markov invariants for such symmetrically embedded models, and we provide enumerations of these for low-dimensional cases. The simplest example is a target model on 3 states, constructed from a general 2 state model the $2 \hookrightarrow 3$ embedding. We show that for 2 taxa, there exist two invariants of quadratic degree, that can be used to directly infer pairwise information from observed sequences. A simple simulation study verifies their theoretical expected values, and suggests that, given the appropriateness of the model class, they have greater statistical power than the standard (log) Det invariant (which is of cubic degree for this case).

Qiang Li
Institute: CAS-MPG Partner Institute for Computational Biology, Shanghai, China
qiang.d.li@gmail.com

Estimation of evolutionary distances based on word composition

Compositional vector approach provides an alignment-free method for phylogenomics based on K-string composition. It has reached a high degree of agreement with taxonomy, but the meaning of branch lengths of the resulted trees was unclear. We propose a heuristic probabilistic model for the evolution of K-string composition to calibrate the compositional vector trees, and estimate the working range of K. It leads to even simpler methods for distance estimation, solely based upon the presence or absence of K-strings, which yield phylogenetic trees with meaningful branch lengths.

John Rhodes
Institute: Department of Mathematics and Statistics, University of Alaska (Fairbanks), USA
j.rhodes@uaf.edu

Identifiability of Phylogenetic Models for Morphological Data
Joint work with Elizabeth Allman and Mark Holder.

Parsimony remains the most common method of phylogenetic inference for morphological data, despite its known inconsistency for data produced under the simplest plausible model of character evolution. Using a more statistically well-founded method of inference, however, requires accounting for an inherent ascertainment bias in morphological data: constant characters, and perhaps non-parsimony informative characters, cannot be reliably recorded.

To overcome parsimony's inconsistency, Lewis (2001) and Nylander-et al. (2004) introduced simple models that condition on the types of observed characters that are more reliably recorded, for use in inference with likelihood or Bayesian methods. However, to establish that inference is consistent under these models requires proving the identifiability of their parameters, a foundational issue they did not address.

Using algebraic methods, and in particular known phylogenetic invariants for a k-state general

Markov model, we prove that a class of models encompassing those mentioned above indeed has identifiable parameters, for sufficiently large trees. This highlights the usefulness of understanding the algebraic structure of phylogenetic models in addressing theoretical issues.

Jeremy Sumner
Institute: School of Maths and Physics, University of Tasmania
jsumner@utas.edu.au

The algebra of Markov models on phylogenetic split networks
Joint work with Peter Jarvis and Barbara Holland.

It is known that so called (abelian) "group-based" Markov models on phylogenetic trees can be extended quite naturally to arbitrary split systems. In this talk, I will show that a similar extension can be achieved for general Markov models. This is achieved by studying the (Lie) algebra of the generators of the continuous-time Markov chain together with the "splitting" operator that generates the branching process on phylogenetic trees. The resulting representation of the general Markov model is such that it can immediately be extended in a natural way to include arbitrary splits. This extenstion contains the standard tree model as a special case, but has the potential to associate an individual weight and rate matrix to any additional incompatible splits that we wish to include. Under restriction of the general model to the parameter space of a group-based model, it is intriguing that our extension is identical to the group-based approach only on trees; as soon as any incompatible splits are introduced the two approaches are inconsistent with each other. Within the context of this inconsistency, I will close with a tentative arguement that our approach to extending to arbitrary split systems has desirable properties that the group-based extension does not share.

Michael Woodhams
Institute: School of IT, University of Sydney
woodhams@it.usyd.edu.au

The Mosaic Model of Nucleotide Substitution
Joint work with Michael Charleston and Jeremy Sumner.

When dealing with deep phylogeny, we cannot reasonably assume that one single evolutionary model applies uniformly over the whole tree. The mosaic model takes a number of submodels, each of which is a traditional evolutionary model (e.g. GTR+I). It then applies these models patchwork-like over a phylogeny, so that each submodel applies over a subset of the full tree. A given branch in the phylogeny may have more than one submodel applying to it, in which case a weighted average of the submodels is used, the weights becoming additional parameters to fit.

The mosaic model is being implemented using the PAL library in Java (Drummond, Strimmer et al.) as a basis. The mosaic model breaks a number of standard evolutionary model assumptions, and we discuss the program code needed to support these changes. Some very preliminary Monte Carlo results will be presented.

List of participants

Roger Brown
Institute: NCI National Facily, Australian National University
bregor@sf.anu.edu.au

David Bryant
Institute: University of Auckland, New Zealand
d.bryant@auckland.ac.nz

Conrad Burden
Institute: Centre for Bioinformation Science, Australian National University
conrad.burden@anu.edu.au

Chris Burridge
Institute: School of Plant Science, University of Tasmania
Chris.Burridge@utas.edu.au

Michael Charleston
Institute: School of IT, University of Sydney
m.charleston@usyd.edu.au

James Degnan
Institute: Biomathematics Research Centre, University of Canterbury, New Zealand
J.Degnan@math.canterbury.ac.nz

Des Fitzgerald
Institute: School of Maths and Physics, University of Tasmania
D.FitzGerald@utas.edu.au

Stefan Gruenewald
Institute: CAS-MPG Partner Institute for Computational Biology, Shanghai, China
stefan@picb.ac.cn

Klaas Hartmann
Institute: TAFI, University of Tasmania
Klaas.Hartmann@utas.edu.au

Barbara Holland
Institute for Fundamental Sciences, Massey University, New Zealand
b.r.holland@massey.ac.nz

Peter Jarvis
Institute: School of Maths and Physics, University of Tasmania
Peter.Jarvis@utas.edu.au

Greg Lee
Institute: TAFI, University of Tasmania
greg.lee@utas.edu.au

Qiang Li
Institute: CAS-MPG Partner Institute for Computational Biology, Shanghai, China
qiang.d.li@gmail.com

Rob Little
Institute: School of IT, University of Sydney
rhtlittle@gmail.com

Jonathan Mitchell
Institute: School of Maths and Physics, University of Tasmania
jm06@utas.edu.au

John Rhodes
Institute: University of Alaska (Fairbanks), USA
j.rhodes@uaf.edu

Jim Stankovich
Institute: Menzies Centre, Tasmania
jsumner@utas.edu.au

Jeremy Sumner
Institute: School of Maths and Physics, University of Tasmania
jsumner@utas.edu.au

Michael Woodhams
Institute: School of IT, University of Sydney
woodhams@it.usyd.edu.au