Spring Phylogenetics at UEA

Organisers: Katharina T. Huber and Sarah Bastkowski

Date: Tuesday, 17 April 2012

Venue: D'Arcy Thompson Room

Programme

13:00 – 13:35: Leo van Iersel, *Minimizing Hybridizations*

13:35 – 14:00: Jo Dicks, *Identification and analysis of rDNA sequence microheterogeneity for fine-scaled phylogenetic estimation*

14:00 – 14:35: Philippe Gambette, *Structure and enumeration of level-k phylogenetic networks.*

14:35 – 15:05: Break

15:05 – 15:30: Rad Suchecki, *Padre2012: Constructing and comparing MUL-trees.*

15:30 – 15:55: Sarah Bastkowski, *SuperQ: A new method to construct weighted super-networks from partial trees*

15:55 – 16:20: Andrei-Alin Popescu, *Extending APE to incomplete distances*

16:20 – 16:45: Katharina Huber, *Lassoing phylogenetic trees*

17:00 - : *Informal discussions in the student union pub*

Abstracts

Minimizing Hybridizations

Leo van Iersel,
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands.

Given a collection of trees describing the evolution of different genes, a well-studied problem is to compute the minimum number of hybridizations necessary to explain the conflicts among the trees. A constructive variant of this problem is to construct a rooted binary network that contains each of the given trees and has a minimum number of vertices with indegree 2. I will review this problem and present several recent results.

Identification and analysis of rDNA sequence microheterogeneity for fine-scaled phylogenetic estimation

Jo Dicks
John Innes Centre

Ribosomal RNA genes, known as ribosomal DNA or rDNA, are frequently used for phylogenetic analysis because their sequences exhibit high levels of conservation across species boundaries. However, many researchers have noted problems in phylogenetic estimation arising from sequence heterogeneity within an organism's genome. The rDNA genes are found in tandem arrays of tens or even hundreds of repeating units, at one or more loci within a genome. The sequences of each unit in an array were once thought to be identical but it is now known that mutations may occur, causing heterogeneity amongst units. Opposing these divergent mutational processes, unit sequences are homogenised through concerted evolutionary processes such as unequal sister chromatid exchange (USCE) and gene conversion (GC).

Whole genome sequence data are now enabling us to quantify the microheterogeneity apparent in the rDNA genes. Furthermore, modelling the concerted evolutionary processes that shape this observed sequence variation both offers the potential to both gain insight into the mode and tempo of these processes and to use them for fine-scaled phylogenetic estimation of closely related organisms.

Here, we will outline our recent progress towards these goals. We will discuss the identification of rDNA sequence variation within two closely-related yet contrasting yeast species, using bespoke Perl software such as TURNIP (http://www.ncyc.co.uk/software/turnip.html), and the use of this variation in phylogenetic estimation. Finally, we will present early work in the modelling of the USCE and GC processes that will ultimately enable us to turn an historic phylogenetic problem into a phylogenetic opportunity in the estimation of accurate within-species evolutionary trees.

Joint work with: Claire West, Steve A James, Robert P Davey, and Ian N Roberts.

Structure and enumeration of level-k phylogenetic networks

Philippe Gambette,
Université Paris-Est Marne-la-Vallée. Paris. France.

Phylogenetic networks generalize the tree concept to model Evolution. In rooted phylogenetic networks, some vertices have two parents: they correspond to species resulting from hybridization between two ancestral species, or having received a gene from another species through horizontal gene transfer. The "level" of a phylogenetic network measure how much its structure differs from a tree structure. Considering phylogenetic networks with bounded level allows to obtain efficient reconstruction or comparison algorithms. I will present some properties of phylogenetic networks with bounded level, showing how they can be obtained from a finite set of generators, and present some results about the enumeration of level-1 and 2 networks.

Padre2012: Constructing and comparing MUL-trees

Rad Suchecki,
University of East Anglia.

Multi-labeled trees (or MUL-trees) are used to reconstruct evolutionary past of polyploid species. The key feature of a MUL-tree is that more than one of its leaves can be labelled by a single species. Consequently the problem of constructing a consensus MUL-tree from a set of multi-labeled gene trees is much more complex than in the case of phylogenetic trees. In addition, for a given input, multiple non-isomoprhic consensus MUL-trees can be constructed. Each of these MUL-trees needs to be scored, so that a most parsimonious one (in a well defined sense) can be identified. Additionally, it may be of interest to compare the constructed MUL-trees with one-another and with the input MUL-trees. In recent work, we have presented the MultiCons heuristic based on an established algorithm that constructs a consensus MUL-tree from a set of multi-labeled gene trees. In related work, we have addressed the problem of comparing MUL-trees by investigating a number of distance measures for them from the theoretical perspective. We have also implemented one of these distance measures, namely the Maximum Agreement Subtree (MAST) distance for MUL-trees. In this talk we provide some implementation details of MultiCons and the MAST distance, which are both implemented in JAVA and form a part of the next release of the Padre software package.

SuperQ: A new method to construct weighted supernetworks from partial trees

Sarah Bastkowski,
University of East Anglia

Presenting evolutionary data from different trees for a set of taxa in a joint network is an important problem in phylogenetics. A tree construction for one single gene does not always correspond to the species phylogeny. So a common approach is to sequence many genes, construct trees for each of them and fuse the trees into a supertree or a network. This becomes even more difficult if the input consists of partial trees, i.e. several taxa are missing in these gene trees. In the SuperQ algorithm, quartets, which are derived from the input trees, are used toand a circular split network. The next challenge is to and good weights for the splits in the resulting network. We describe a new approach to this problem and present some results on the performance of SuperQ on simulation data.

Extending APE to incomplete distances

Andrei-Alin Popescu,
University of East Anglia

APE is a popular R package used for phylogenetic inference and analysis. Among its methods, APE contains implementations of algorithms for inferring phylogenies from a distance matrix on a set of taxa. However in modern phylogenetic analysis, such distance matrices frequently contain missing information, which until recently APE could not handle. In this talk we briefly outline some of the algorithms which can be used to infer phylogenetic trees from such datasets and which we have implemented into APE.

Lassoing phylogenetic trees

Katharina Huber,
University of East Anglia,

A classical result, fundamental to evolutionary biology, states that an edge-weighted tree T with leaf set X, positive edge weights, and no vertices of degree 2 can be uniquely reconstructed from the set of leaf-to-leaf distances between any two elements of X. In biology, X corresponds to a set of taxa (e.g. extant species) and the tree T describes their phylogenetic relationships. Provided one has access to all distances, and these are known to be sufficiently close to the distances induced by some (as yet unknown) tree, then that tree, together with its edge weighting, can be computed with some degree of confidence from those distances in polynomial time using, for example, Neighbor-Joining. However, much of the data being generated even by modern genomic methods have patchy taxon coverage whereby only certain pairs of taxa have a known (or, at least, sufficiently reliable) distance. This raises interesting mathematical questions (besides the obvious statistical and algorithmic ones) concerning tree reconstruction from such incomplete data some of which we will address in this talk.