Midsummer Phylogenetics at UEA

Organisers: Katharina T. Huber and Sven Herrmann

Date: Friday, 17 June 2011

Venue: D'Arcy Thompson Room

(Note the venue change!)

Programme

13:00 – 13:20: Eva Czabarka, *Optimal placement of multiplication events on a species tree.*

13:20 – 13:40: Rad Suchecki, *Constructing consensus polyploid phylogenies.*

13:40 – 14:00: Athena Chu, *Phylogenetic analysis of the avenacin gene cluster.*

14:00 – 14:20: Jo Dicks, *Two practical problems in the phylogenetic analysis of plant datasets.*

14:20 – 14:30: Andrei-Alin Popescu, *Extending APE to handle incomplete distances.*

14:30 – 15:00: Break

15:00 – 15:20: Sarah Bastkowski, *SuperQ: a new method to construct weighted super networks from partial trees.*

15:20 – 15:40: Sven Herrmann, *Phylogenetic trees and k-dissimilarity maps.*

15:40 – 16:00: Josh Collins, *Clustering character data into differing tree topologies.*

16:00 – 16:30: Barbara Holland, *Phylogenies from DArTs: A stochastic Dollo process with censored data.*

The organisers would like to thank the Roberts Fund for its support.

Abstracts

SuperQ: a new method to construct weighted super networks from partial trees

Sarah Bastkowski,
University of East Anglia, Norwich, UK.

One major problem in phylogenetics is presenting evolutionary data from different trees for a set of taxa in a joint network. It gets even more difficult if several taxa are missing in the trees. In the SuperQ algorithm, quartets, which are derived from the input trees are used to find a circular split network. The challenge is then to find good weights for the edges in this network. We describe a new approach to this and present some preliminary results on experiments that we carried out to evaluate the performance of this new approach as compared to other different edge weighting strategies.

Clustering character data into differing tree topologies

Josh Collins,
Massey University, Palmerston North, New Zealand.

It is possible to say individual character sites in a data sequence have definitely evolved on some tree topology. However, in some data sets, such as those arising from hybridisation, it may not always be the same tree. This talk will cover the details of a genetic algorithm that attempts to cluster such data into small sets of trees in a sensible way using various extensions of Maximum Parsimony. At the end will be discussed the inherent shortcomings and possible improvements of the implementation.

Phylogenetic analysis of the avenacin gene cluster

Athena Chu,
John Innes Centre, Norwich, UK.

Avenacin is an anti-microbial secondary metabolic triterpene accumulated in the root tips of oat plants, protecting them from broad-spectrum disease. The avenacin biosynthetic genes are physically linked, co-regulated by identical promoters and collectively by chromatin de-condensation. These genomic features of the avenacin biosynthetic gene cluster are believed to have resulted from stringent selection for adaptation to the environment of invasive soil micro-organisms. The gene cluster structure exhibited by avenacin has long been described in bacteria, fungi, and C. elegans, but has only recently been discovered in plants. To date five genes from the avenacin biosynthetic cluster have been characterized, and all derive from large multigene families that have diversified to perform unique reactions. Sad1 is the first gene in the avenacin pathway, the signature enzyme, leading to the synthesis of β-amyrin. The tailoring enzymes that then modify the β-amyrin backbone, Sad2, Sad7, Sad9, and Sad10, are hypothesized to have undergone highly dynamic evolutionary processes, resulting from both selective pressures on individual genes and the pathway as a whole and the tremendous genomic plasticity of plants, in order to obtain gene innovations and new pathway formation. Detailed evolutionary analyses have recently been performed on Sad1 and Sad2. Here, the gene evolutionary trajectories of Sad7, Sad9, and Sad10 have been elucidated to identify the selective pressures on each gene. Using the available sequenced plant genomes, we can further identify key events in the formation of the avenacin gene cluster by studying the genomic locations of the Sad gene orthologues in the context of gene phylogeny. In parallel, avenacin related metabolites have been identified among wild grasses in the tribe Aveneae to relate the evolution of the avenacin biosynthetic genes to the species phylogeny of Aveneae core and the Aveneae lineages.

Optimal placement of multiplication events on a species tree

Eva Czabarka,
University of South Carolina, Columbia, USA.

The placement of a minimal number of multiplication events on a species tree that explain a set of duplication episodes on gene trees corresponds to placing the minimal number of intervals that cover an interval structure of the tree and satisfy some additional constraints. This leads to a Gallai-type minimax theorem, that we will describe. This is joint work with Todd Vision and Laszlo Szekely.

Two practical problems in the phylogenetic analysis of plant datasets

Jo Dicks,
John Innes Centre, Norwich, UK.

Recent phylogenetic analyses of plant datasets have focused on two types of plant dataset. In the first type, we have been interested in determining whether a phylogenetic tree or a phylogenetic network is most appropriate for describing it. Together with collaborators at the University of East Anglia and the Institute of Food Research, Norwich, we have developed a simple statistical test (the TreeFit test) that can answer this question for a wide variety of biological datasets. In the second type, we have been analysing simultaneously the phylogenetic trees of one or more individual gene families alongside the genomic distributions of the constituent genes, in order to make inferences about the genomic events that have led to the observed gene trees. Here, we will give an outline of these problems, referring to real plant datasets. We will then look forward to new research in these areas, discussing new ideas that could be used to solve these two problems.

Phylogenetic trees and k-dissimilarity maps

Sven Herrmann,
University of East Anglia, Norwich, UK.

Phylogenetic trees are versatile tools in the analysis of sequence data. Several approaches exist to construct such trees from data using metrics or distances, and the Tree Metric Theorem of Dress gives an explicit condition when such a metric defines a tree. More recently, (see, e.g., Levy, Yoshida and Pachter [Mol. Biol. Evol. 23(3):491–498. 2006]), it was suggested to use the phylogenetic diversity not of pairs, but of triplets, quartets or, in general, k-tuples, to construct trees from the given data. Maps assigning values to such k-tuples of taxa as opposed to pairs are called k-dissimilarity maps.

In this talk, we will analyse when such k-dissimilarity maps correspond to trees and give generalisations of the Tree Metric Theorem.

This is joint works with Katharina Huber, Vincent Moulton and Andreas Spillner.

Phylogenies from DArTs: A stochastic Dollo proces with censored data

Barbara Holland,
University of Tasmania, Tasmania, Australia.

Diversity Array Technologies (DArTs) are a relatively new kind of DNA marker system that seem like they could be usefully applied to phylogenetics (a few papers have already explored this). Like marker systems such AFLP and RFLP, the method produces presence abscence data, but unlike these methods it is very unlikely for shared presences to occur by chance.

The basic idea is as follows. One or a small number of genomes are selected to form the genomic representation. Two enzymes are used to cut the DNA from these genomes at certain recognition sites (a rare 6bp recognition site and a more frequent 4bp recognition site). Fragments of DNA whose ends have been cut by two rare recognition sites are amplified. These fragments, which are said to form the genomic representation, are arranged on a microchip. Other genomes can then been checked to see which fragments within the genomic representation they have copies of in their own sequence. For each other genome that is compared to the genomic representation this results in a binary sequence that indicates presence (1) or absence (0) of each of the fragments.

The first obvious advantage of this approach is that it creates a representation of the whole genome rather than just a few genes. This alleviates the problem of picking a small set of genes that may not be representative of the evolutionary history of the species. The second advantage is that in comparison to an individual site, long fragments of DNA are very unlikely to be similar due to chance. So if two species share a fragment it is vastly more likely that they share it due to common ancestry rather than due to a chance similarity.

To use these data for phylogenetics it would be useful to develop a likelihood equivalent of Dollo parsimony (in which characters can be lost multiple times but gained only once), such

models have already been explored in the context of language evolution and gene content evolution. However, another complicating issue is the censoring effect created by only being able to see those fragments that were in the original genomic representation, i.e. fragments that are shared by a group of species but that are not present in the original species used to make the genomic representation are missing from the data.
Joint work with Dorothy Steane, Michael Woodhams and Vincent Moulton.

Extending APE to handle incomplete distances

Andrei-Alin Popescu,
University of East Anglia, Norwich, UK.

tba

Constructing consensus polyploid phylogenies

Rad Suchecki,
University of East Anglia, Norwich, UK

Multi-labeled trees (or MUL-trees) along with phylogenetic networks 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 standard phylogenetic trees. Furthermore, for a given input, multiple distinct consensus MUL-trees can be constructed, and thus each of these MUL-trees needs to be scored, so that a most parsimonious one (in a well defined sense) can be identified. We present an improved implementation of an established algorithm that constructs a consensus MUL-tree from a set of multi-labeled gene trees. To aid the understanding of the approach, we have prepared an experimental pipeline which allowed us to identify certain limitations of the original approach and also to evaluate several speed-ups and improvements. Both the old and the new implementations are tested using randomly generated data as well as previously published biological examples. The reduction in memory consumption as well as run-time makes it feasible to construct consensus MUL-trees from much larger and more complex input datasets.