Abstracts

Applications of Phylogenetic Networks in Yeasts

Teun Boekhout

Yeasts, usually defined as unicellular fungi, are polyphyletic and occur in two main domains of the fungal kingdom, the Ascomycota and Basidiomycota. Our understanding of the evolutionary history and phylogeny of these microorganisms has improved significantly since DNA markers have been introduced (> 20 years ago). Initially this concerned only two regions of the ribosomal DNA (rDNA) locus, the D1D2 domains of the Large SubUnit (LSU) and the ITS 1 + 2 spacers of the rDNA. Meanwhile other genes have been introduced, such as RPB1, RPB2, TEF-1alpha, actin, etc. Moreover, yeasts are among the best studied domains in terms of comparative genomics. Most studies, incl. ours, still rely strongly on the analysis of traditional phylogenetic trees that are be generated using different algorithms. However, in cases of horizontal gene transfer, speciation by hybridization, as well as the analysis of hybrid complexes these methods will fail and network approaches are needed. In the analysis of the evolutionary history of an important human fungal pathogen, Cryptococcus gattii, that is causing epidemics in British Colombia (Canada), the Pacific North West of USA, but also in Brazil, West Australia and that is emerging in Mediterranean Europe, hybridization events played an important role to shape the population structure and the presence of virulence attributes. This was demonstrated and visualized using network analysis. A major challenge for the future will be the analysis of whole genome data that are rapidly emerging for yeasts [and fungi in general]. In my opinion network analysis may be able to detect horizontal gene transfer, hybridization events, and contribute to the analysis of hybrid genomes, and so on. The appealing presentation of the resulting networks also needs attention, as does the regular updates for new genomes that appear.

Quasi-median networks as a tool of exploratory data analysis

Hans-Jürgen Bandelt

Networks are labeled and weighted connected graphs which, in contrast to trees, may contain cycles. A network can accommodate ambiguities concerning evolutionary pathways and thus simultaneously represent alternative estimates of a phylogeny. A data-display network may either directly represent the given character data or certain splits inferred from the data via estimated trees or via decomposition/projection or approximation procedures operating on derived distances. Further technical distinctions are whether or not the interior nodes of networks can be interpreted as character-state sequences, and whether or not a network is rooted.
Every table of aligned deoxyribonucleic acid (DNA) sequences or other character-state sequences can be represented by its accompanying quasi-median network, which is mathematically speaking a certain dual to the data table. If the sequences are binary, that is, all sites have only two states, then "quasi-median" is specified to "median". Quasi-median networks thus do not depend on any kind of approximation, estimation or heuristic and are thus closest to the data. Inasmuch as the data are faithfully represented, quasi-median networks can be very large and too complex for visualization in the plane. Therefore one usually has to resort to some sort of simplification. One option is to use mutation-filtered data, which exclude hypervariable sites. This strategy is routinely applied in forensic genetics in order to detect sequencing and documentation problems in mtDNA data submitted to the forensic database

EMPOP. Another option will be thinning of a quasi-median network by removing highly ambiguous portions of the network in a systematic way.
This can result in a disconnected subnetwork, which could subsequently be reconnected by some heuristic such as median-joining (MJ). This approach modifies the pruning method proposed by Katharina Huber and co-workers more than a decade ago.

From phylogenetic trees to phylogenetic networks, and beyond

Eric Bapteste

The growing recognition that mobile genetic elements (e.g. viruses and plasmids) are prevalent and play a major evolutionary role, even in the evolution of cells, and the recognition of the importance of introgressive processes (e.g. recombination, lateral gene transfer, symbioses), is transforming our representations of evolution. The vast majority (and possibly the quasi-totality) of evolving entities seem to be genetically mosaic, although to variable extents. Therefore, it is becoming increasingly obvious that the model of a universal tree greatly under-estimates the actual diversity of evolutionary objects and phenomena. First, this model is typically centered on one evolutionary process (e.g. vertical descent), and on one type of object: simple lineages belonging to one level of biological organization (i.e. monophyletic groups of cellular beings). Second, trees are acyclic graphs, and phylogenetic methods seeking to reconstruct them adopt constraining assumptions leading to filter out numerous data (however rich in non-tree evolutionary signal) from the evolutionary analyses.

This realization has prompted the development of phylogenetic networks methods that aim to overcome some of the limits of the tree model (i.e. to offer an analytical framework to study mosaic entities whose components can come from distinct levels of biological organization, the emergence of phylogenetically composite lineages, and to allow for the reconstruction of evolutionary graphs that are not necessarily tree-like, etc.). These developments are invaluable advances that enrich our understanding of biological evolution. Nonetheless, in this talk I will argue that it might be desirable that phylogenetic networks of the future are much more than Œaugmented trees¹, in order to avoid (if it can be) that they inherit some conceptual limits from tree-thinking.

In particular, I will insist that a systematic bias in favour of simple lineages and vertical inheritance should be avoided, as could be observed in some treatments of trees with taxonomically patchy distributions, or with weak phylogenetic signals. Such a bias can produce artefactual dichotomies, and thus hypothetical ancestral lineages, that do not always correspond to real biological entities. I will also discuss the need to interpret phylogenetic networks with a specifically designed network-thinking, because the objects and relationships investigated by networks cannot always be considered equivalent to, or simple extensions from, those studied in trees. Before I try to suggest some early leads for a network-thinking, I will get back to the problem of a priori massive data exclusion to promote the future development of inclusive evolutionary networks, in addition to strictly phylogenetic networks. I will then conclude by raising a more modest concern: the possibility to implement tests to avoid eventual issues of circularity in phylogenetic network reconstruction, in particular to avoid the reconstruction of one network when multiple networks would be necessary.

Phylogenetic Networks with Recombination

Dan Gusfield

In response to the scope of the conference, ``The future of phylogenetic networks'', and in response to the request from the organizers to relate recombination networks to other models of phylogenetic networks, this talk will do two things. First explain the role of recombination networks in association mapping, i.e., methods to locate genes that influence some genetic trait (disease or commercial trait). That application illustrates one of the future (as well as present) directions for the use of recombination networks. Second, I will *try* to explore the significant technical issues of invisible and Steiner nodes in both recombination networks and cluster-based phylogenetic networks. Hopefully, examining the commonalities and differences of these technical results will lead to better understanding, or to the transfer of results from one model to the other. This second part of the talk is really an experiment, i.e., talking about things that I don't know much about, rather than about things that I think I know something about.

Using phylogenetic networks to describe hybrid evolution in plants

Barbara Gravendeel

A common speciation mechanism in plants is hybridisation resulting in polyploidy. Due to this process, plants display a high degree of chromosomal rearrangements by gene duplication, elimination, inversion and translocation. This process causes variation in numbers and distribution patterns of DNA loci among related species and makes molecular phylogenetic reconstructions of their evolution challenging.
I will present several cases in which application of a combined approach, i.e. combining DNA sequences, karyotype and chromosome painting data provided the information required to resolve phylogenetic relationships.

Bridging the cultural/philosophical divide between mathematicians and biologists

Katharina Huber

Phylogenetics provides mathematicians (viewed in a broad way) and biologists with many exciting opportunities to work together to explore (and hopefully answer) some of the exciting questions surrounding the problem of how life on earth has evolved. However these collaborations also pose interesting challenges for both fields. Together with the audience, we will explore some of them in this talk.

Expanding the application of networks in plant phylogenetics

Scot A. Kelchner

Use of networks in plant phylogenetics historically has been limited to cases in which a researcher suspects or detects hybridization and gene flow. Networks became a common a posteriori way to visualize such signal in the data. Although the nature of plant genetics makes character conflict prevalent in phylogenetic data sets, there is a reluctance by plant systematists to use networks for purposes other than gene flow description. This might be due less to the availability of tools and more to cultural factors: (i) the dominance of tree thinking in systematics; (ii) the tractability of trees for statistical measures accepted by the community; (iii) the absence

of routine exposure to alternative uses for networks; (iv) insufficient training in available network techniques; and (v) an expectation that character conflict is inevitable but can be overcome sufficiently by adding more data.

In this talk, several potential uses for networks in plant phylogenetics will be proposed. Examples come from ongoing efforts to resolve phylogenetic relationships among bamboos. Most techniques are related to exploratory data analysis (EDA), although I also explore how networks might be useful for screening loci in combined data matrices, for identifying error in data assembly, and for increasing our understanding of unexpected but Œsupported¹ tree resolution. Finally, I will suggest that expanding the use of networks in plant phylogenetics will require several convincing, high profile demonstrations of the useful information that networks can provide to phylogeny estimations, particularly now that we have entered the phylogenomic era of massive data sets.

What mathematical optimization can, and cannot, do for biologists

Steven Kelk

Broadly speaking, mathematical optimization is the science of selecting an “optimal” element from a space of possible solutions. The notion of “optimal”, and the structure of the search space, are both defined mathematically. Classical examples of mathematical optimization within phylogenetics include construction of trees under the Maximum Parsimony (MP) or Maximum Likelihood (ML) criteria. However, all computational methods that construct phylogenetic networks or trees involve (implicitly or explicitly) some form of mathematical optimization.

A recurring problem at the interface between mathematics and biology is that, outside mathematical circles, the strengths and limitations of mathematical optimization (as a tool for hypothesis generation) are often poorly communicated and understood. Consequently, it can be very difficult for a non-mathematician to accurately interpret and contextualize the output of software packages and algorithms.

In this talk I will try and summarize, in non-mathematical language, some of the main strengths and weaknesses of mathematical optimization within phylogenetics, particularly phylogenetic networks. Strengths include the rigorous formalization of the concept “best”, while weaknesses include the over-simplification of complex biological phenomena and computability problems. I will also try and demystify some of the terminology used by mathematicians, such as optimality, objective function, computational (in)tractability, approximations, heuristics and uniqueness. The goal is to set the stage for the more mathematical talks that will be presented at the workshop.

Phylogenetic networks: where are we now, and where do we need to go? A general biological perspective

David A. Morrison

The history of biology has been one of moving slowly from the description of "natural history" to the concept of "biological science". One important component of this transformation was the realization that a scientist must take into account evolutionary history when studying any biological system. Therefore, the most important recent advance in the biological sciences has been the development of explicit methods for constructing phylogenies. Phylogenies have

proliferated into all branches of biology, forming the framework on which to hang all biological knowledge.

However, to date phylogenetic analysis has been based principally on the model of a dichotomous tree, rather than on the original conception (in 1755) of a phylogenetic network. This talk discusses how far we have come in developing suitable network models for phylogenetic analysis, and where we realistically need to get to before these models will become as widespread in biology as their tree counterparts. This talk thus sets the biological scene for the workshop, as well as a possible agenda.

There have been two "relationship" threads running in parallel through the history of biology: affinity relationships and genealogical relationships.These are still in use today, as undirected graphs and directed acyclic graphs, respectively. Affinity networks are used for (among other things): exploratory data analysis, displaying data patterns, displaying data conflicts, and testing phylogenetic hypotheses. Genealogical networks are used for detecting reticulations due to: recombination, hybridization and introgression, horizontal gene transfer, and genome fusion.

Mathematically, we have a fine array of methods for calculating affinity networks, and these are appearing more and more often in the literature. However, there is a dearth of methods for producing genealogical networks. Some of the factors that might be desirable for networks to explicitly model evolution include: adding the concept of "reticulation events" to that of the current substitution / indel models; separating the effects of randomness and rooting from reticulation; having an explicit null model for reticulation; using mixture models; and having methods to quantify robustness of branch / reticulation estimates.

We also need to bear in mind that biologists and mathematicians look at the world in quite different ways, with different objectives and different assumptions. Clear communication between the two groups is sometimes lacking; and this workshop is designed to forge and foster linkages between the two groups.

“But where are the bootstrap values?” Robustness, reliability and confidence issues

Vincent Moulton

Bootstrap values are commonly generated when constructing phylogenetic trees, and also for certain types of phylogenetic networks, such as NeighborNets.

However, many network building methods don't produce such values, and it can be tricky to interpret them even when they are available. We shall explore this issue (with audience participation!) as well as the more general question as to how confident we can be with the networks that we generate, and how we might go about developing appropriate methods to help quantify this.

Quantifying reticulation across many phylogenies

Charles Semple

A problem that continues to attract the interests of mathematicians and computer scientists is the following: Given a collection P of phylogenies in which each phylogeny correctly represents the evolution of some gene, what is the smallest number of reticulation events to simultaneously

explain the phylogenies in P? The problem dates back at least to Hein (1990) and yet, despite the ever increasing number of publications on the problem, satisfactory methods for answering it have only been obtained for when |P|=2 and P consists of binary phylogenies. What makes the problem difficult when |P| >=3 or P consists of non-binary phylogenies? In this talk, we highlight some of these difficulties and discuss recent progress.

Statistical issues in phylogenetic network models

Mike Steel

What do incompatible gene trees tell us about the pattern of species evolution? In particular, is this incompatibility a signal for non-tree-like (reticulate) evolution, or do gene trees simply disagree for stochastic reasons (lineage sorting, systematic error etc)? A related issue is whether a 'species tree' has a well-defined meaning, or whether the history of life can only be represented by a highly tangled web. In the first part of this talk, I describe how different (reticulate and non-reticulate) stochastic processes can lead to different signals in the data, and I will describe some of the ways in which evolution might be described by a 'central tendency' species tree. In the second part of the talk, I will describe some recent joint work on species tree reconstruction when genes have evolved under a model of lateral gene transfer. A typical question is: 'could we reconstruct a species tree on (say) 200 species from lots of gene trees, if each gene has been laterally transferred into other lineages, on average, ten times?'

Validating methods for constructing evolutionary phylogenetic networks with a bank of real biological datasets

James B. Whitfield

A large variety of existing biological problems can be approached using phylogenetic networks. This talk will focus on a few topics that might especially enhance the usage of networks by practicing systematists, and present a few examples. The emphasis will be particularly on split networks, using gene trees as source data.

1) using networks to summarize the congruence and conflict among many genes (gene trees vs. species trees) in genome-scale phylogenetics - can we develop bookkeeping and associated graphical methods and realistic measures of support for supernetworks base on partial trees? (I'll bring and show a large dataset from bees as a starting point)

2) How can we know how effective our methods for detecting horizontal gene transmission really are? This is a big issue especially for organisms that have vitally important endosymbionts - I'll present a few biological examples from a few well-studied systems.

3) Integrating ecology with phylogeny in multi-trophic-level interactions
- can we meaningfully link food webs (networks) with phylogenies (trees or networks)? If so, what questions could we ask/potentially solve with these networks? This a new, but rapidly developing, topic. I'll illustrate with a large example of wasps that parasitize caterpillars that feed on plants.

I will conclude with a discussion of development of a bank of test datasets that could be used for validating some phylogenetic network methods. An obvious issue is how to know what a "correct" (i. e.,validated) answer is with real biological data.