Database research

- Rod Page. Title: An outsider's perspective on TreeBASE.
  Abstract: The current TreeBASE is a black hole – data disappears in and is difficult to extract again. Furthermore, no use is made of the wealth of information that could be linked to data in TreeBASE. The only external links TreeBASE contains are author email addresses. Yet, given a GenBank sequence or a paper title one can go to the Internet and readily extract information on genes, specimen localities, PubMed records, citation links, images, taxonomic authorities, etc. The search interface is limited, and locks users into primitive and often fruitless searches. TreeBASE is a walled garden in a time when the world is discovering data integration, federated searches, and “mashups.” Designing new, improved (read bigger) relational database schema does nothing to address these issues. If the community wants a useful tool that tells us what we know (and what we don’t know) about the tree of life, and enables the kind of integrated research that we systematists so often say is only possible with a phylogenetic underpinning, then I suggest we need something rather different. This talk will sketch some problems with TreeBASE, discuss some ideas relating to globally unique identifiers, metadata, inference, and the Semantic web, and will end with the author running from the room hotly pursued by Bill Piel.
  Links:
  - http://darwin.zoology.gla.ac.uk/ rpage/lsid/examples/lsid.pdf,
  - http://darwin.zoology.gla.ac.uk/ rpage/TreeBASE/,
  - http://ispecies.org,
  - http://ispecies.blogspot.com/, and

Algorithms

- Nick Pattengale, Department of Computer Science, UNM. Title: A Sublinear-Time Randomized Approximation Scheme for the Robinson-Foulds Metric.
  Abstract: The Robinson-Foulds (RF) metric is the measure most widely used in comparing phylogenetic trees; it can also be computed in linear time using Day’s algorithm. When faced with the need to compare large numbers of large trees, however, even linear time becomes prohibitive. In this paper, we present a randomized approximation scheme that provides, with high probability, a \((1 + \varepsilon)\)-approximation of the true RF metric for all pairs of trees in a given collection. Our approach is to use a sublinear-space embedding of the trees, combined with an application of the Johnson-Lindenstrauss lemma to approximate vector norms very rapidly. We discuss the consequences of various parameter choices (in the embedding and in the approximation requirements). We also implemented our algorithm as a Java class that can easily be combined with popular packages such as Mesquite; in consequence, we present experimental results illustrating the precision and running-time tradeoffs as well as demonstrating the speed of our approach.

- Derek Ruths, Department of Computer Science, Rice University. Title: Detecting and Reconstructing Reticulate Evolutionary Events in Microbes: Challenges and Methods.
  Abstract: Large classes of microbial organisms evolve through mechanisms other than lineal descent. Two reticulate processes in particular, interspecific recombination and horizontal gene transfer (HGT), have been implicated as major mechanisms among viruses and bacteria. When these events occur,
the evolutionary history of a set of organisms cannot be represented by a tree; hence, detecting and reconstructing them is a necessary step in building the Tree of Life, particularly the prokaryotic branch. In this talk, I will discuss the challenges involved in identifying and classifying reticulate evolutionary events in prokaryotes. I will also present two methods our group has developed: RECOMP is a method for detecting interspecific recombination and RIATA-HGT is the first polynomial-time heuristic for reconstructing arbitrary HGT events.

  
  Abstract: We study the problem of enumerating substrings that are common amongst genomes that share evolutionary descent. For example, one might want to enumerate all identical (therefore conserved) substrings that are shared between all mammals and not found in non-mammals. Such collection of substrings may be used to identify conserved subsequences or to construct sets of identifying substrings for branches of a phylogenetic tree. For two disjoint sets of genomes on a phylogenetic tree, a substring is called a discriminating substring or a tag if it is found in all of the genomes of one set and none of the genomes of the other set. Given a phylogeny for a set of m species, each with a genome of length at most n, we develop a suffix-tree based algorithm to find all tags in $O(nm \log m)$ time. We also develop a sublinear space algorithm (at the expense of running time) that is more suited for very large data sets. We next consider a stochastic model of evolution to understand how tags arise. We show that in this setting, a simple process of tag generation essentially captures all possible ways of generating tags. We use this insight to develop a faster tag discovery algorithm with a small chance of error. However, tags are not guaranteed to exist in a given data set. We thus generalize the notion of a tag from a single substring to a set of substrings whereby each species in one set contains a large fraction of the substrings while each species in the other set contains only a small fraction of the substrings. We study the complexity of this problem and give a simple linear programming based approach for finding approximate generalized tag sets. Finally, we use our tag enumeration algorithm to analyze a phylogeny containing 57 whole microbial genomes. We find tags for all nodes in the phylogeny except the root for which we find generalized tag sets.

- Sebastien Roch, Department of Statistics, The University of California at Berkeley. Title: Phylogenetic Tree Reconstruction by Maximum Likelihood is NP-Hard.
  
  Abstract: I will sketch a short proof that reconstructing phylogenies by maximum likelihood is NP-hard. The proof uses a connection between likelihood and parsimony observed by Tuffley and Steel.
  
  Links:

New Heuristics for Large-Scale Phylogeny Reconstruction

- Tiffani Williams, Department of Computer Science, Texas A&M University. Title: Cooperative Rec-I-DCM3: A Population-Based Approach for Reconstructing Phylogenies. Co-author: Marc L. Smith.
  
  Abstract: We discuss the use of cooperation as a general technique for designing faster algorithms for reconstructing phylogenetic trees. Currently, the best-performing MP heuristic is Rec-I-DCM3. Each iteration of a Rec-I-DCM3 search is guided by a single tree. We describe a new algorithm called Cooperative Rec-I-DCM3, which maintains a population of Rec-I-DCM3 trees to search the tree landscape. Our results demonstrate that Cooperative Rec-I-DCM3 outperforms its sequential counterpart by at least an order of magnitude in terms of convergence to best-known MP scores.
  
  Links to relevant papers and urls: http://faculty.cs.tamu.edu/tlw/Publications.html.
• Usman Roshan, Department of Computer Science, New Jersey Institute of Technology. Title: Boosting phylogeny reconstruction using Recursive-Iterative-DCM3. Primary co-authors: Bernard Moret, Tandy Warnow, and Tiffani Williams.

Abstract: The Disk Covering Method (DCM) family of algorithms improve an existing phylogeny reconstruction method by deploying it in a divide-and-conquer framework. DCMs divide the input set of species into smaller subsets, compute phylogenies on the subsets using the given base phylogeny reconstruction method, merge the phylogenies using the Strict Consensus Merger into a supertree, and refine the resulting supertree to make it binary. The latest and most effective DCM to date is Recursive-Iterative-DCM3 (Rec-DCM3) which combines the basic divide-and-conquer strategy with iteration. In this talk I will describe the Rec-DCM3 method and present experimental results for improving phylogeny reconstruction using MP heuristics (with TNT), ML heuristics (with RAxML), and simultaneous alignment and phylogeny reconstruction heuristics (using Poy). URL for papers: http://www.cs.njit.edu/usman/papers.html.

• Alexandros Stamatakis, High Performance Computing Bioinformatics, Institute of Computer Science, Foundation for Research and Technology, Heraklion, Crete, Greece.

Abstract: RAxML-VI is a program for Maximum Likelihood-based inference of huge phylogenies up to 25,000 taxa. This talk will cover the usage of RAxML-VI in practice for large-scale phylogenetic inference. In addition, it will address a recent performance comparison with other popular ML programs. Finally, the potential of using the CAT model of rate heterogeneity as a work-around for the significantly more compute- and memory-intensive GAMMA model will be discussed.

URLs:

- Papers about RAxML (PDFs): www.ics.forth.gr/~stamatak (publications frame)
- CAT versus GAMMA comparison (preprint): wwwbode.in.tum.de/~stamatak/gamma.pdf.

Mathematical Phylogenetics.

• Eric Vigoda, Georgia Institute of Technology. Title: Phylogeny of Mixture Models: Maximum Likelihood and Ambiguity. Co-author: Daniel Stefankovic.

Abstract: We address phylogenetic reconstruction when the data is generated from a mixture distribution. In our work we consider data coming from a mixture of trees which share a common topology, but differ in their edge weights (i.e., branch lengths). We first show the pitfalls of popular methods, including maximum likelihood and Markov chain Monte Carlo algorithms. We then determine in which evolutionary models, reconstructing the tree topology, under a mixture distribution, is (im)possible. We prove that every model whose transition matrices can be parameterized by an open set, either has ambiguous mixture distributions, in which case reconstruction is impossible in general, or there exist linear tests which identify the topology. This duality theorem, relies on our notion of linear tests and uses ideas from linear programming duality.


• Radu Mihaescu, Department of Mathematics, the University of California at Berkeley. Title: Better Reliability Criteria for the Neighbor Joining Algorithm Co-authors: Dan Levy, Radu Mihaescu, and Lior Pachter.

Abstract: We show that the neighbor-joining algorithm is a robust quartet method for constructing trees from distances. This leads to a new performance guarantee that contains Atteson’s optimal radius bound as a special case and explains many cases where neighbor-joining is successful even when Atteson’s criterion is not satisfied. We also provide a proof for Atteson’s conjecture on the optimal edge radius of the neighbor joining algorithm. The strong performance guarantees we provide also hold
for the quadratic time fast neighbor-joining algorithm, thus providing a theoretical basis for inferring very large phylogenies with neighbor-joining.

Link to Kevin Atteson's paper: http://www.springerlink.com/media/12t6dn1w3n4tm0nj13/contributions/t/9/

- Constantinos Daskalakis, Department of Computer Science, University of California at Berkeley. Talk title: *Optimal Phylogenetic Reconstruction*. Co-authors: Sebastien Roch and Elchanan Mossel.

Abstract: Reconstructing a tree on \( n \) leaves requires at least \( c \log n \) evolutionary characters, for some constant \( c \). In fact, if the mutation rate is high the required number of characters is polynomial in \( n \) and algorithms that match this lower bound have been found recently. This leaves open the question of whether reconstruction can be done with fewer characters for models with small mutation rates. We prove a conjecture of Mike Steel that \( k \log n \) characters - for a constant \( k \) - suffice as long as the mutation rate is below some specific threshold, known in statistical physics as the reconstruction threshold for spin-glasses. Our algorithm matches the lower bound for phylogenetic reconstruction up to a multiplicative constant. Links to papers


- Cameron Hill, Department of Computer Sciences, the University of California at Berkeley. Title: *Constructing accurate forests from distance matrices*. Co-authors: Constantinos Daskalakis, Alexander Jaffe, Radu Mihaescu, Elchanan Mossel, and Satish Rao.

Abstract: I will discuss our paper *Constructing maximal accurate forests from distance matrices* (to appear in the proceedings of RECOMB 2006). Underlying our work is the following problem, which is inspired by the work of [1]:

**MAXIMAL EDGE-DISJOINT FOREST (MEDF)** Given a distance matrix \( D \), find a pairwise edge-disjoint forest \( F = \{T_1, \ldots, T_r\} \) such that each \( T_r \) is accurate and \( r \) is as small as possible.

MEDF is hard. We provide a solution of the problem after relaxing the edge-disjointness condition, with probabilistic accuracy guarantees. That is, our theoretical algorithm constructs a forest which is accurate with arbitrarily high probability. The correctness guarantee arises from a test for the reliability of each quartet topology the algorithm attempts to use in reconstruction. Our subtree reconstruction algorithm (derived from that in DCM1) improves on that in [1] and yields an \( \tilde{O}(n^3) \) running time. We also show that our method obtains almost all trees with \( poly - \log(n) \) length sequences, improving on [FewLogs1]. The running time of our method permits effective implementation, and experiments have shown that it produces highly accurate forests even with quite short sequences and converges quite rapidly with increasing sequence length. Moreover, our method markedly outperforms Neighbor Joining.


Simulators for Complex Evolution

- Monique Morin, The Department of Computer Science, The University of New Mexico. Title: *NETGEN: Simulator for Creating Phylogenetic Networks*.

Abstract: By comparing generated and reconstructed phylogenetic networks, confidence in reconstruction techniques is achieved. NETGEN is an event driven simulator for generating phylogenetic networks with sequences. The traditional birth-death model used often in biology to create phylogenetic trees is extended to support diploid hybrids and/or variable rate lineages. Hybrid decisions are made according to sequences associated with lineages, which are developed in conjunction with the topology itself. Additional options such as outgroups, non-ultrametricity, hybridization limit, and root sequence specification allow networks to be customized. This talk presents an overview of the NETGEN simulator, key features and options, and results of initial validation efforts. An extension of the Newick
format which includes hybrids is also proposed. The source code is available for download under GNU General Public License. (1)

Relevant url: http://www.phylo.umm.edu/~morin.

• Eric Miller, Section of Integrative Biology, The University of Texas at Austin. Co-author: Lauren Ancel Meyers.

Title: RiboSim: Connecting Microevolutionary and Macroevolutionary Processes and Providing Standardized Sequence Data.

Abstract: The lack of standardized data hinders direct comparison between phylogenetic reconstruction algorithms. Data sets from natural taxa lack a known, true tree to test the accuracy of reconstruction methods, while mathematically simulated data potentially reduce the complexity of microevolutionary and macroevolutionary processes.

We constructed a simulation program that examines the evolution of 16S ribosome sequences. The program tracks the lineage of each sequence and gives the true phylogeny that relates the sequences to each other. As the simulation progresses, the evolutionary histories of the current sequences are recorded, thereby allowing recovery of the complete phylogenetic tree. The terminal RNA sequences, but not internal sequences, are also saved.

The simulation incorporates knowledge about known 16S sequences and RNA folding to determine which ribosomes are more likely to replicate. Point mutations, insertions, and deletions at each generation are sources of variation in the ribosomes. These mutations can cause speciation events if the resulting ribosome is more similar to known sequences of a different species.

The rate of mutation also changes along lineages. Fluctuating population sizes allow genetic drift to act on the population, while a changing folding environment allows traversal across non-optimal sequence space.

This simulation allows us to test how mutations interact with the environment to produce new species. We can ask how population-level phylogenetic trees differ from species-level trees. The output from this simulation can be used to compare the accuracy of phylogenetic reconstruction and alignment algorithms.

Multiple Sequence Alignment

• Kevin Liu, The Department of Computer Sciences, The University of Texas at Austin. Title: Improving the Accuracy of Phylogenetic Tree Reconstruction Methods with Simple Alignment Gap Information, Co-authors: C. Randal Linder, Li-San Wang, and Tandy Warnow.

Abstract: We quantify the impact on topological accuracy when phylogenetic reconstruction methods incorporate affine gap calculations. Current methods either ignore gapped regions completely or treat gaps as a separate state, and emphasize substitution events. However, biologists believe that indel events are also significant and better analyzed using affine gap penalties. Sankoff introduced the Generalized Tree Alignment problem in the context of multiple sequence alignment and tree estimation, and we aim to improve on work done on this problem by better addressing the usage of gap penalties in MSAs.


Abstract: Micro-indels are small insertion or deletion events (indels) that occur during genome evolution. The study of micro-indels is important, both in order to better understand the underlying biological mechanisms, and also for improving the evolutionary models used in sequence alignment and phylogenetic analysis. The inference of micro-indels from multiple sequence alignments of related genomes poses a difficult computational problem, and is far more complicated than the related task of inferring the history of point mutations. We introduce a tree alignment based approach that is suitable
for working with multiple genomes and that emphasizes the concept of indel history. By working with an appropriately restricted alignment model, we are able to propose an algorithm for inferring the optimal indel history of homologous sequences that is efficient for practical problems. Using data from the ENCODE project as well as related sequences from multiple primates, we are able to compare and contrast indel events in both coding and non-coding regions. The ability to work with multiple sequences allows us to refute a previous claim that indel rates are approximately fixed even when the mutation rate changes, and allows us to show that indel events are not neutral. In particular, we identify indel hotspots in the human genome.

- Nicholas Bray, The Department of Mathematics at UC Berkeley. Title: MAVID. Co-author: Lior Pachter.
  
  Abstract: To appear.