

CIPRES All Hands Meeting 2009; Posters:

(1) Hyun Jung (Justin) Park and Luay Nakhleh, Rice University

Title: Inferring Phylogenetic Networks Using Maximum Parsimony

Authors: Hyun Jung Park, Guohua Jin, and Luay Nakhleh

Abstract: Maximum parsimony is one of the most commonly used criteria for reconstructing phylogenetic trees. Recently, this criterion was extended to the domain of phylogenetic networks, and its application to detecting reticulate evolutionary relationships was demonstrated. However, one of the major problems with this extension has been that it favors more complex evolutionary relationships over simpler ones, thus having the potential for overestimating the amount of reticulation in the data. An ad hoc solution to this problem that has been used entails inspecting the improvement in the parsimony length as more reticulation events are added to the model, and stopping when the improvement is below a certain threshold. We address this problem in a more systematic way, providing a statistical framework for addressing two questions. First, how many reticulation events are required to explain the evolution of a sequence data set under the maximum parsimony criterion? Second, what is the support for the hypothetical reticulation events identified by the criterion? For the first question, we provide a  $\$p\$$ -value computation to determine the amount of reticulation required. A particular use of this computation allows for answering the question of how tree-like the evolutionary history of a data set is. For the second question, we employ a bootstrapping procedure to estimate the significance of the inferred reticulation events. We have implemented both methods in our NEPAL software tool (available publicly at <http://bioinfo.cs.rice.edu/>), and studied their performance on both biological and simulated data sets. While our studies show very promising results, they also highlight issues that are inherently challenging when applying the maximum parsimony criterion to detect reticulate evolution.

---

---

(2) Cuong Than and Luay Nakhleh, Rice University

Title: Species Tree Inference by Minimizing Deep Coalescences

In a 1997 seminal paper, W. Maddison proposed minimizing deep coalescences, or MDC, as an optimization criterion for inferring the species tree from a set of incongruent gene trees, assuming the incongruence is exclusively due to lineage sorting. In a subsequent paper, Maddison and Knowles provided and implemented a search heuristic for optimizing the MDC criterion, given a set of gene trees. However, the heuristic is not guaranteed to compute optimal solutions, and its hill-climbing search may make it slow in practice. We provide two exact solutions to the problem of inferring the species tree from a set of gene trees under the MDC criterion. One solution is based on a novel integer linear programming (ILP) formulation, and another is based on a simple dynamic programming (DP) approach. Industrial-strength ILP solvers, such as CPLEX, make the first solution appealing, particularly for very large-scale instances of the problem, whereas the DP-based solution eliminates dependence on proprietary tools, and its simplicity makes it easy to integrate with other genomic events that may cause gene tree incongruence. Using the exact solutions, we analyze a data set of 106 loci from eight

yeast species, a data set of 268 loci from eight Apicomplexan species, and several simulated data sets. We show that the MDC criterion provides very accurate estimates of the species tree topologies, and that our solutions are very fast, thus allowing for the accurate analysis of genome-scale data sets. Further, the efficiency of the solutions allow for quick exploration of sub-optimal solutions, which is important for a parsimony-based criterion such as MDC, as we show. We show that searching for the species tree in the compatibility graph of the clusters induced by the gene trees may be sufficient in practice, a finding that helps ameliorate the computational requirements of optimization solutions. Further, we study the statistical consistency and convergence rate of the MDC criterion, as well as its optimality in inferring the species tree. Finally, we show how our solutions can be used to identify potential horizontal gene transfer events that may have caused some of the incongruence in the data, thus augmenting Maddison's original framework. We have implemented our solutions in the PhyloNet software package, which is freely available at <http://bioinfo.cs.rice.edu/phyloNet>.

---

---

(3) Jeffrey Boore and Susan Fuerstenberg, JGI

Title: Evolutionary Analysis as the Basis for Interpreting, Comparing, and Presenting Genomes: The GATOR and PHRINGE System

Abstract: We anticipate a great acceleration in whole genome sequencing over the next few years. Current tools for interpreting, comparing, and presenting these data cannot handle the expected pace, lack integration, require extensive IT support and computational expertise, and do too little to facilitate biological discovery. In particular, the standard “browser” format is anachronistic, with the genome assembly, rather than the biological information, being the organizing principle. It requires great manual effort to identify any particular gene or biochemical pathway. Fortunately, two new developments are enabling a better approach. First, next generation sequencing technology allows very deep sequence coverage of the set of expressed genes. For example, 200-fold mean sequence coverage can be obtained on a typical transcriptome for about \$20,000. This means that genes can be modeled with much greater accuracy, so even the early stages of analysis can focus on biological discovery instead of manual gene curation. Second, we have developed an effective tool (“PHRINGE”, for Phylogenetic Resources for Interpreting Genomes) for assigning orthologous and paralogous relationships among genes based on phylogenetic analysis of complete gene sets. In the absence of biochemical characterization, the best inference of gene function is by inferring that orthologous genes retain the same function. This is incorporated into the “GATOR” (Genome Analysis Tools and Online Resources) system under development, a “gene-centric”, user-friendly, streamlined approach to genome interpretation, comparison, and presentation. The entry point is the gene catalog itself, sortable by many categories, including domain content, intracellular location, SNP content, biochemical pathway, protein characteristics, number of members in any gene family, and many others. Users can view evolutionary trees, gene colinearity maps, and links to protein structures for all genes in multiple sequenced genomes.

---

---

(4) Susan Fuerstenberg and Jeffrey Boore, JGI

Poster title: Will Your Favorite Genome be Sequenced?

Abstract: The past decade has seen a fantastic acceleration in our ability to sequence complete genomes. We anticipate that this exponential increase will continue over the next few years. Will your favorite genome be among those sequenced? We review the several technologies that are state-of-the-art for genome sequencing, describe their capabilities, provide an interpretation of the pros and cons of each, and argue for the most appropriate mix of techniques for whole genome sequencing. We describe the factors important in the choice of genomes to target, the various mechanisms of funding and coordinating such a large project, and the options for follow-on analyses that will maximize biological discovery and the utility of the work to the broader scientific community.

---

(5) John Harshman, Pacbell

Title: Avian introns have been shrinking for the last hundred million years

Abstract: Analysis of 33 introns in 15 loci from 169 bird species, shows a mean reduction in length of 0.03% per lineage per million years. This reduction is slow but consistent over the tree. Of 336 branches in the tree, only 52 show a net gain in total sequence length. Total loss over all 7.9 billion lineage-years of evolution is 37468 bases. There is no correlation between the age of a branch and its amount of loss; the rate has remained relatively constant over avian evolution. There is, however, a correlation in direction and amount of change among adjacent branches.

Authors: John Harshman, Michael J. Braun, Frederick S. Sheldon, Edward L. Braun, Rauri C. K. Bowie, Jena L. Chojnowski, Shannon Hackett, Kin-Lan Han, Christopher J. Huddleston, Rebecca T. Kimball, Ben D. Marks, Kathleen J. Miglia, William S. Moore, Sushma Reddy, David W. Steadman, Christopher C. Witt, and Tamaki Yuri

---

(6) Chia Shen, Harvard University

Authors: Michael Horn and Chia Shen

SDR Lab, School of Engineering and Applied Sciences, Harvard University

Title: INVOLV - Interactive Exploration of Life on Earth on Multi-Touch Tables

INVOLV ([www.involvweb.org](http://www.involvweb.org)) is a multi-touch tabletop system that uses the Voronoi treemap algorithm, together with multimedia and tangible objects, to create an interactive visualization for the Encyclopedia of Life (EOL). INVOLV is an ongoing project of the Scientists' Discovery Room Lab (SDR) at the Initiative in Innovative Computing at Harvard University. At the present, the visualization includes over 1.2 million named species organized in a nine-level hierarchy. In attempting to visualize this taxonomy of life on earth, we have begun to integrate phylogenetic information about the evolutionary history of organisms from the Tree of Life ([www.tolweb.org](http://www.tolweb.org)) with our existing visualization. We propose an extension of the Voronoi treemap algorithm that employs force-directed graph drawing techniques to overlay supplemental phylogenetic

information on top of the treemap. This technique has the added benefit of guiding the spatial layout of regions in the treemap based on the evolutionary relatedness of the various taxonomic groups. The intended users of INVOLV will range from informal science education to scientists.

---

---

(7) John Heraty, University of California, Riverside

Poster title: **RAPID and Standard Bootstrap and the Phylogeny of Hymenoptera**

John Heraty and James Munro

Department of Entomology, University of California, Riverside, CA 92521

[john.heraty@ucr.edu](mailto:john.heraty@ucr.edu), [james.munro@ucr.edu](mailto:james.munro@ucr.edu)

**Abstract:** Hymenoptera are the ants, bees and wasps and represent one of the most diverse insect lineages originating during the Jurassic Period. One dataset from the Hymenoptera Tree-of-Life project, based on a by-eye alignment of 18S, 28S, COI and Ef1-alpha (7150 bp), was used to compare results from different approaches using RAxML. For one analysis, a more typical method of a priori optimization of estimating initial parameters was employed and followed by independent searches for the Best Known Likelihood (BKL) tree and then a Standard Bootstrap analysis to estimate support values (SBS). A second analysis used CIPRES for a Rapid Bootstrap Analysis (RBS), which estimates a set of best trees and parameters for the subsequent BKL search. BKL trees produced from different RBS search strategies (numbers of replicates) involved convergence on similar trees with slightly different BKL scores. Notably, there was little or no RBS support for a monophyletic Hymenoptera despite a relatively long-branch and consistent strong support (98-100%) from the SBS searches.

---

---

(8) Sheng Guo and Junhyong Kim, University of Pennsylvania

Poster title: Large-Scale Simulating of RNA Evolution by an Energy-Dependent Fitness Model

Poster authors: Sheng Guo, Li-San Wang, Junhyong Kim

Poster abstract: Simulated nucleotide sequences are widely used in theoretical and empirical molecular evolution studies. Conventional simulators generally use substitution matrices, e.g., HKY and GTR, to model the rate of nucleotide change, and frequently carry additional augmentations such as modeling rate variation among sites by a gamma distribution. Such purely sequence-based simulation is relatively straightforward to conceptualize and implement, however, may be inherently inadequate to capture the complex temporal and spatial variation and constraint in molecular evolution. One new strategy is to consider the effect of phenotypic fitness upon the evolution of genes. The secondary structure of an RNA can be viewed as one of its phenotypes. In this work, we use the folding free energy of the secondary structure of an RNA as a proxy for its phenotypic fitness, and simulate RNA evolution by a mutation-selection population genetics model. Basically, a mutant is randomly sampled from the mutant ensemble of

an RNA, and may then fix with a probability determined by its phenotypic fitness and the effective population size. The mutation can be a substitution, an insertion or a deletion. Because the two-step process is conditioned on an RNA and its mutant ensemble, we no longer have a global substitution matrix, nor do we explicitly assume any for this inhomogeneous stochastic process. After introducing the base model of RNA evolution, we outline the heuristic implementation algorithm and several model improvements. We then discuss the calibration of the model parameters and demonstrate that in phylogeny reconstruction with both the parsimony method and the likelihood method, the sequences generated by our simulator, rnasim, have greater statistical complexity than those by two standard simulators, ROSE and Seq-Gen, and are close to empirical sequences.

---

---

(9) Christiane Weirauch & Dimitri Forero, University of California, Riverside

Poster Title: Phylogenetic analyses of Reduviidae using the CIPRES Portal

---

---

(10) Monique Morin, University of New Mexico

Poster title: Reconstructing Phylogenetic Networks with Single-Diploid-Hybrid Events

Eukaryotic species are known to hybridize. When two diploid parent lineages each contribute half of their DNA sequences to a new species, a diploid hybrid is formed. If this new species proves viable and reproduces, a more complex evolutionary history results. The reconstruction of these topologies can be pursued from both biological and algorithmic perspectives. Our previously reported work has focused on developing a unique simulator to co-evolve a source topology and its sequences. This implementation merges biological and computational approaches by allowing the sequences to influence the subsequent reticulate topology. Specifically, this presentation introduces new quantitative descriptors of timing, diversity, and impact, which can be applied to most reticulate events. Furthermore, we propose a novel, biologically-inspired algorithm that uses knowledge of hybrid-descendant extant taxa to reconstruct single-diploid-hybrid networks. We discuss the performance of this technique on topologies where hybrid occurrence varies in location and parental diversity.

---

---

(11) Tiffani Williams, Texas A&M University

Poster title: Effective Techniques for Summarizing Large Collections of Evolutionary Trees

---

---

(12) Luay Nakhleh, Rice University

Title: Evolutionary phylogenetic networks: models and issues

Abstract: Phylogenetic networks are special graphs that generalize phylogenetic trees to allow for modeling of non-treelike evolutionary histories. The ability to sequence multiple genetic markers from a set of organisms and the conflicting evolutionary signals that these markers provide in many cases, have propelled research and interest in phylogenetic networks to the forefront in computational phylogenetics. In this talk, I will discuss the evolutionary phylogenetic networks model, which explicitly represents reticulation events, and discuss issues with the reconstruction of these networks.

---

---

(13) Jijun Tang and William Arndt, University of South Carolina

Title: Genome rearrangement analysis for large genomes

---

---

(14) Serita Nelesen, The University of Texas at Austin  
Ugochukwu C. Anokwuru and Gail N. Maduro, Huston-Tillotson College

Title: Finding the Best MRP Tree

Abstract: The new supertree method, Superfine, was designed to produce improved estimates of supertrees. In this summer internship between Huston-Tillotson University and UT-Austin, we discovered that Superfine also produces better solutions to maximum parsimony, when applied to MRP matrices. We report on these results on real (biological) datasets.

Authors: Ugochukwu C. Anokwuru, Gail N. Maduro, Serita Nelesen, M. Shel Swenson, C. Randal Linder, and Tandy Warnow

---

---

(15) Kevin Liu and Serita Nelesen, The University of Texas at Austin

Title: SATé: Simultaneous Alignment and Tree estimation

Authors: Kevin Liu, Sindhu Raghavan, Serita Nelesen, C. Randal Linder, and Tandy Warnow

Abstract:

Inferring an accurate evolutionary tree of life requires high-quality alignments of molecular sequence data sets from large numbers of species. However, this task is often difficult, slow, and idiosyncratic, especially when the sequences are highly diverged or include high rates of insertions and deletions (collectively known as indels). We present SATé (Simultaneous Alignment and Tree estimation), an automated method to quickly and accurately estimate both DNA alignments and trees with the maximum likelihood criterion. In our study, it improved tree and alignment accuracy compared to the best

two-phase methods currently available for datasets of up to 1000 sequences, showing that coestimation can be both rapid and accurate in phylogenetic studies.

---

---

(16) Kevin Liu, The University of Texas at Austin

Title: POY\*: improved estimations of trees using POY

Abstract: POY is a heuristic for the NP-hard "minimal treelength" problem, which takes as input a set of unaligned sequences and produces a tree for the sequences, along with sequences at the internal nodes, so as to minimize the total cost of all the evolutionary events (substitutions plus indels). While some studies have shown that POY produces trees that are not as topologically accurate as standard two-phase methods, we show that when POY is used with an affine gap penalty it can produce trees that are topologically as accurate as the leading two-phase methods.

Authors: K. Liu, S. Nelesen, S. Raghavan, C.R. Linder, and T. Warnow.

---

---