FINAL REPORT

DOE Award Number: DE-FG02-06ER25734
Date of Report: March 12, 2014
Period Covered: August 15, 2006 — August 14, 2009

Tools for Accurate and Efficient Analysis of Complex Evolutionary Mechanisms in Microbial Genomes

Luay Nakhleh

Department of Computer Science
Rice University
6100 Main Street, MS 132
Houston, TX 77005
Phone: (713) 348 3959
Fax: (713) 348 5930
Email: nakhleh@cs.rice.edu

1 Introduction

Microbes are tiny organisms that constitute a large group of all organisms on Earth. Their relationship to humans, animals, and plants is very delicate, and hence large resources have been dedicated to studying and understanding these organisms. Given the wide array of potential microbial applications, which include toxic-waste cleanup, carbon management, and energy production, the DOE Microbial Genome Program was established to sequence and classify microbial genomes (see http://www.microbialgenome.org).

Discovering the genomic innovations in microbes bears a great significance on understanding these organisms, the risks that some of them pose, and the contributions that others make. Unlike Eukaryotic organisms which evolve mainly through lineal (vertical) descent, microbial evolution is much more complex, and hence mapping and understanding microbial genomes pose a great challenge. One of the major evolutionary mechanisms that shape genomic innovations in microbial organisms is horizontal gene transfer (HGT), which is believed to be ubiquitous in bacteria. Despite the ubiquity of HGT and the main roles it plays in microbial genome diversification, very few tools have been developed for its detection. Further, most of these tools either have poor performance or are *ad hoc* extensions of existing techniques, limiting their applicability to specific datasets.

I proposed to develop computationally efficient tools for accurate detection and reconstruction of these complex evolutionary mechanisms, thus enabling rapid and accurate annotation, analysis and understanding of their genomes. To achieve this goal, I proposed to address three aspects.
(1) Mathematical modeling. A major challenge facing the accurate detection of HGT is that of distinguishing between these two events on the one hand and other events that have similar "effects." I proposed to develop a novel mathematical approach for distinguishing among these events. Further, I proposed to develop a set of novel optimization criteria for the evolutionary analysis of microbial genomes in the presence of these complex evolutionary events.
(2) Algorithm design. In this aspect of the project, I proposed to develop an array of efficient and accurate algorithms for analyzing microbial genomes based on the formulated optimization criteria. Further, I proposed to test the viability of the criteria and the accuracy of the algorithms in an experimental setting using both synthetic as well as biological data.
(3) Software development. I proposed the final outcome to be a suite of software tools which implements the mathematical models as well as the algorithms developed.

2 Technical Accomplishments

2.1 Research Work

Horizontal gene transfer (HGT) is a mechanism by which species/organisms transfer genetic material to each other, typically across species boundaries, but in some cases also within the organisms of the same species. The most commonly used method for detecting HGT is the phylogeny-based one. In this method, gene trees are compared to the species tree, and incongruence is used to indicate HGT events. The objectives of our research project is to identify the challenges associated with this approach, and solving them in order to enable efficient and accurate detection of HGT at the scale of whole genomes.

In this project, we achieved all the goals we set out to work on in the proposal:

1. We have extended the RIATA-HGT method so that it accurately handles non-binary trees and computes multiple solutions. Reconstructed gene trees are often non-binary, meaning that they are not fully resolved (mainly due to lack of phylogenetic signal). Therefore, methods for handling non-binary trees appropriately are required. Further, multiple optimal scenarios of reconciling species and gene trees may exist, and the number may be exponential in the number of HGT events. We have extended RIATA-HGT so that it computes a large portion of the solution space, and represents them in a compact manner.

2. We have established complexity results for a set of combinatorial problems that related phylogenetic networks to their constituent trees and clusters, and provided improved efficient parameterized algorithms.

3. Observing that measures of topological difference between a pair of phylogenetic networks failed to satisfy the properties of a metric, we have developed the first metric for comparing phylogenetic networks topologies.

4. We have continued the study of the maximum parsimony criterion for phylogenetic networks, and established new theoretical results on the criterion, as well as new efficient heuristics for computing the parsimony length of a phylogenetic networks.

5. To assess the confidence of an inferred HGT event by the RIATA-HGT method in particular, and topology-based HGT detection methods in general, we have developed two novel approaches, one based on the bootstrap values of the gene tree branches, and the other based on integrating the maximum parsimony criterion with the topology-based inference. The results are very promising, as we have shown on multiple biological data sets.

6. We have developed a novel two-stage approach for identifying the bacterial strain or species tree. The first phase is an efficient identification of candidate strain/species tree topologies, using a novel computation based on maximal cliques, and the second is a novel estimation of strain/species tree divergence times based on a mixed integer linear programming (MILP) approach. Inferring an accurate strain/species tree has a significant impact on the quality of inferred HGT events, since these are inferred based on comparing the topologies of the strain/species tree against the gene trees.

7. Finally, we have incorporated all the results in the PhyloNet software package, for which we have developed an extensive user's manual, with sample files, and submitted a manuscript the describes the tool).

Some of the developed methods have also been implemented in the NEPAL software tool and is available publicly on our website (the website URL is provided below). Further, all these results have been published in international journals and conference proceedings (listed below), and have been presented at various universities and conference meetings.

2.2 Research Findings

Our research produced several findings:

1. While comparing the topologies of species/gene trees may result in a high percentage of false positive predictions of HGT events, incorporating the bootstrap values computed for the branches of the species and/or gene trees result in significant improving in the predictions, significantly eliminating false positives, while maintaining the true ones.

2. While we established in the previous period of the project that accuracy of the maximum parsimony criterion in a standalone setting, we have demonstrated that a careful integration of the criterion with RIATA-HGT does combined the accuracy of the former with the speed of the latter, thus achieving a new fast and accurate technique for inferring HGT events.

3. We have found that careful resolution of non-binary gene trees, by using the topology of the species tree as a basis for the resolution, significantly improves the quality of HGT detections from non-binary trees. It is important to note that most software tools for estimating HGT events based on species/tree topology comparison do *not* handle non-binary trees, which is a limiting factor, since in real applications, it is often the case that the trees are non fully resolved. Therefore, our work in this area presents a novel advancement.

4. While we had established that the number of optimal HGT scenarios may be exponential in the number of HGT events in an optimal scenario, we have observed that these multiple scenarios have great sharing (intersections) among each other. We have shown how to exploit this sharing to efficiently and compactly compute and represent these solutions.

5. While existing methods for establishing the species tree rely heavily on searches of the huge tree space and an expensive probabilistic evaluation of each tree, we have shown that by using the coalescent times computed for the internal nodes of the gene trees, one can obtain a species tree, along with its divergence times, much more efficiently by combining the concepts of maximal cliques of a compatibility graph with mixed integer linear programming.

6. On the data analysis side, we have analyzed twelve genomes of the *Staphylococcus aureus* bacteria and established, for the first time, the rate of homologous recombination in bacteria.

2.3 Software Development

Our research results have been implemented in two software packages, PhyloNet and NEPAL; the websites of both tools are given in Section 4.

2.3.1 PhyloNet

PhyloNet is a suite of software tools, written in JAVA, for reconciling species/gene trees and evaluating phylogenetic networks. It contains many utilities, including:

- GENCPLEX and GENST: utilities for inferring the strain/species tree (topology and times) from a collection of gene trees, given by their topologies and coalescent times.
- MAST: a utility for computing maximum agreement subtree (MAST) of a pair of trees. Beside its central role in our own heuristic for computing species/gene tree reconciliation scenarios, it is also useful as a standalone utility, since the MAST of two trees has been long used as a metric for comparing tree topologies in estimating the performance of phylogenetic methods.
- RF: a utility for computing the Robinson-Foulds (or, symmetric difference) distance between a pair of trees. This metric is the most commonly used one for comparing phylogenetic trees, and we use it as a building block in one of our own measures for comparing phylogenetic networks.
- RIATAHGT: a utility that implements our heuristic for reconciling species/gene trees and reconstructing phylogenetic networks. The method is extended to handle non-binary trees, multiple solutions, and to compute confidence values of inferred HGT events.
- LCA: a utility for computing the *least common ancestor* (LCA) of a set of taxa in a give tree. This is a central utility in our RIATAHGT heuristic, but is also very useful for other applications, and hence the decision to make it a standalone utility.
- RECOMP: a utility that implements our methods for detecting interspecific recombination in a sequence alignment, based on the parsimony criterion and a sliding window.
- CHARNET: a utility that computes the clusters, trees, and tripartitions defined by a phylogenetic network.
- CMPNETS: a utility that implements our three measures for comparing phylogenetic networks.
- NETPARS: a utility that computes the parsimony score of a set of sequences labeling the leaves of a given phylogenetic network.
- COUNTCOAL: a utility for computing the number of coalescence scenarios that reconcile a pair of species/gene trees.

2.3.2 NEPAL

NEPAL is a suite of software tools, written in C++, for inferring and evaluating phylogenetic networks based on our extensions of the maximum parsimony and maximum likelihood criteria. Further, the NEPAL method has been extended so as to have an implementation of the RIATA-HGT/MP integrated approach.

3 Project Personnel

Two of my graduate students, **Cuong Than** and **Hyun Jung Park** worked extensively on the algorithmic parts of this research, as well as on developing the software tools. Both of them successfully defended their Master's and PhD theses, as follows:

- Cuong Than. *Computer Science*, Rice University.
  Joined my group: January 2006.
  Date of Master's defense: May 2008.
  Master's thesis title: Reconstruction of Phylogenetic Networks and Their Relationships with Trees and Branches.
  Date of PhD defense: October 2009.
  PhD dissertation title: Inference of Parsimonious Species Phylogenies from Multi-locus Data.
  First position: Post-doctoral fellow, University of Michigan, Ann Arbor. (Start date: Nov 1, 2009. Mentor: Prof. Noah Rosenberg.)

- Hyun Jung Park. *Computer Science*, Rice University.
  Joined my group: Aug 2007.
  Master's thesis: waived (prior MSc degree).
  Date of PhD defense: Mar 2012.
  PhD dissertation title: Towards Accurate Reconstruction of Phylogenetic Networks.
  First position: Post-doctoral fellow, Baylor College of Medicine. (Start date: May 1, 2012. Mentor: Prof. Wei Li.)

In addition I have collaborated with **Dr. Guohua Jin**, who was a research scientist at the Department of Computer Science, Rice University. Dr. Jin's expertise is in high-performance computing, a discipline whose integration in this project I valued as essential, if we were going to be able to scale the performance of methods to genome-wide data (which we did).

To develop the stochastic framework that distinguishes among the various factors that cause species/gene tree incongruence, I had argued that an integration of phylogenetics and population genetics methodologies would be essential. To achieve that, I collaborated with **Prof. Hideki Innan** and his group at the Graduate University for Advances Studies in Japan. Their area of expertise is evolutionary population genetics.

To establish theoretical properties of problems related to horizontal gene transfer, I collaborated with **Dr. Iyad Kanj**, from DePaul University, and **Dr. Ge Xia** from Lafayette College on proving NP-hardness of some problems, as well as introducing efficient parameterized algorithms.

Last but not least, **Dr. Sagi Snir** from University of Haifa and **Dr. Tamir Tuller** from Tel Aviv University, and my group have collaborated on establishing computational complexity results for the parsimony and likelihood criteria for phylogenetic network reconstruction and evaluation.

4 Products and Technologies

4.1 Publications

- **‡: Nakhleh's group member.**

1. L. Nakhleh, "Evolutionary phylogenetic networks: models and issues." In *The Problem Solving Handbook for Computational Biology and Bioinformatics*, L. Heath and N. Ramakrishnan (editors). Springer, 125-158, 2010.

2. H.J. Park[‡], G. Jin, and L. Nakhleh, "Bootstrap-based support of HGT inferred by maximum parsimony." BMC Evolutionary Biology, 10: 131, 2010.

3. L. Nakhleh, "A Metric on the Space of Reduced Phylogenetic Networks". IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(2): 218-222, 2010.

4. L. Nakhleh, D. Ruths[‡], and H. Innan, "Gene Trees, Species Trees, and Species Networks." In *Meta-analysis and Combining Information in Genetics*, R. Guerra and D. Goldstein (editors). CRC Press, 275-293, 2009.

5. C. Than[‡] and L. Nakhleh, "Species tree inference by minimizing deep coalescences." PLoS Computational Biology, 5(9): e1000501, 2009.

6. G. Jin, L. Nakhleh, S. Snir, and T. Tuller, "Parsimony Score of Phylogenetic Networks: Hardness Results and a Linear-time Heuristic." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6(3): 495-505, 2009.

7. C. Than[‡], D. Ruths[‡], and L. Nakhleh, "PhyloNet: A Software Package for Analyzing and Reconstructing Reticulate Evolutionary Relationships." BMC Bioinformatics, 9: 322, 2008.

8. C. Than[‡], G. Jin, and L. Nakhleh, "Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer." Proceedings of the Sixth RECOMB Comparative Genomics Satellite Workshop, 2008. Lecture Notes in Bioinformatics (LNBI #5267), pp. 113-127, 2008.

9. I. Kanj, L. Nakhleh, C. Than[‡], and G. Xia, "Seeing the Trees and Their Branches in the Network is Hard." Theoretical Computer Science (TCS), 401: 153-164, 2008.

10. C. Than[‡], R. Sugino, H. Innan, and L. Nakhleh, "Efficient Inference of Bacterial Strain Trees From Genome-scale Multi-locus Data." The 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB). Bioinformatics, 24: i123-i131, 2008.

11. C. Than[‡] and L. Nakhleh, "SPR-based Tree Reconciliation: Non-binary Trees and Multiple Solutions." Proceedings of the Sixth Asia Pacific Bioinformatics Conference (APBC 08), 251-260, 2008.

12. I. Kanj, L. Nakhleh, and G. Xia, "The Compatibility of Binary Characters on Phylogenetic Networks: Complexity and Parameterized Algorithms." Algorithmica, 51: 99-128, 2008.

13. C. Than[‡], D. Ruths[‡], H. Innan, and L. Nakhleh, “Confounding Factors in HGT Detection: Statistical Error, Coalescent Effects, and Multiple Solutions.” Journal of Computational Biology, 14(4): 517-535, 2007.

14. G. Jin, L. Nakhleh, S. Snir, and T. Tuller, “Inferring Phylogenetic Networks by the Maximum Parsimony Criterion: A Case Study.” Molecular Biology and Evolution, 24(1): 324-337, 2007.

15. I. Kanj, L. Nakhleh, C. Than[‡], and G. Xia, “Seeing the Trees and Their Branches in the Network is Hard.” Proceedings of the Tenth Italian Conference on Theoretical Computer Science (ICTCS ’07), 82-93, 2007.

16. G. Jin, L. Nakhleh, S. Snir, and T. Tuller, “A New Linear-time Heuristic Algorithm for Computing the Parsimony Score of Phylogenetic Networks: Theoretical Bounds and Empirical Performance.” Proceedings of the International Symposium on Bioinformatics Research and Applications (ISBRA’07). Lecture Notes in Bioinformatics (LNBI #4463), pp. 61-72, 2007.

17. D. Ruths[‡], J.T. Tseng, L. Nakhleh, and P.T. Ram, “De novo Signaling Pathway Predictions based on Protein-Protein Interaction, Targeted Therapy and Protein Microarray Analysis.” Proceedings of the RECOMB Satellite Workshop on Systems Biology and Proteomics. Lecture Notes in Bioinformatics (LNBI #4532), pp. 109-119, 2007.

18. G. Jin, L. Nakhleh, S. Snir, and T. Tuller, “Efficient Parsimony-based Methods for Phylogenetic Network Reconstruction.” The European Conference on Computational Biology (ECCB), 2006. (published in the journal Bioinformatics)

19. C. Than[‡], D. Ruths[‡], H. Innan, and L. Nakhleh, “Identifiability Issues in Phylogeny-based Detection of Horizontal Gene Transfer.” Proceedings of the Fourth RECOMB Comparative Genomics Satellite Workshop, 2006. Lecture Notes in Bioinformatics (LNBI #4205), pp. 215-229, 2006.

20. G. Jin, L. Nakhleh, S. Snir, and T. Tuller, “Maximum Likelihood of Phylogenetic Networks.” Bioinformatics, 22(21): 2604-2611, 2006.

21. G. Jin, L. Nakhleh, S. Snir, and T. Tuller, “Efficient Parsimony-based Methods for Phylogenetic Network Reconstruction.” Bioinformatics, 23: e123-e128, 2006.

4.2 Talks

1. “From Gene Trees to Phylogenetic Networks: Computational Approaches.” Symposium on Biological Networks from Genes to Populations, The 17th Annual Meeting of the Society for Molecular Biology and Evolution (SMBE 2009), Iowa City, IA, Jun 2009.

2. “From Gene Trees to Species Phylogenies: Computational Techniques.” The Integrative Biology Seminar Series, School of Biological Sciences, The University of Texas at Austin, Feb 2009.

3. “Efficient Search for the Species Tree in the Compatibility Graph of Gene Trees.” The *Estimating Species Trees* Workshop, The University of Michigan, Ann Arbor, Jan 2009.

4. “Novel algorithmic techniques for gene tree reconciliation in bacterial genomes.” The Evolution 2008 Conference, The University of Minnesota, Jun 2008.

5. “Phylogenetic Networks: Reconstruction and Evaluation.” The ITES Networks Cluster Seminar, The University of Houston, Apr 2008.

6. “From Gene Trees to Species Phylogenies.” Computational Aspects of Biological Information, Microsoft Research, Dec 2007.

7. “Efficient Reconstruction of Species Trees from Genome-scale Multi-locus Data Under the Coalescent.” Department of Computational and Applied Mathematics, Rice University, Nov 2007.

8. “Horizontal Gene Transfer Detection: Issues and Algorithms.” The 5th Bertinoro Computational Biology Meeting, Bertinoro, Italy, May 2007.

9. “Computational Methodologies for Inferring and Analyzing Networks: Species, Individuals, and Molecules.” Collaborative Research Symposium, Rice University and the Texas Medical Center, Nov 2006.

10. “Modeling and Reconstructing Non-treelike Evolutionary Relationships in Species and Populations.” Keck Seminar, Gulf Coast Consortia/Keck Center, Rice University, Oct 2006.

11. “Phylogenetic Networks and Reconstruction of Horizontal Gene Transfer.” Corporate Affiliates Meeting, Department of Computer, Rice University, Oct 2006.

12. “Identifiability Issues in Phylogeny-based Detection of Horizontal Gene Transfer.” RECOMB Comparative Genomics, Montreal, Canada, Sep 2006. Co-authors: C. Than, D. Ruths, and H. Innan.

4.3 Websites

1. http://bioinfo.cs.rice.edu/phylonet

2. http://bioinfo.cs.rice.edu/nepal

4.4 Software Tools

We have developed two major software packages that incorporate all developed algorithmic techniques in them.

1. The PhyloNet software package.
   A package of software tools to compare phylogenetic trees, reconstruct phylogenetic networks and metrics for their topological comparisons. A user’s manual and executable are currently available at http://bioinfo.cs.rice.edu/phylonet

2. The NEPAL software package.
   A package of software tools for reconstructing and evaluating phylogenetic networks based on our proposed extension of the maximum parsimony and maximum likelihood criteria. A user’s manual and executable are currently available at http://bioinfo.cs.rice.edu/nepal