Information Technology Research (ITR):

Building the Tree of Life—A National Resource
for Phyloinformatics and Computational Phylogenetics

The CIPRES Project

FINAL Report for NSF Cooperative Agreement
(Abridged)

EF EF-0715370 (UT Austin and subcontracts)

1 Overview of the CIPRES project

This is the final report of one of the grants to the CIPRES (Cyber Infrastructure for Phylogenetic Research) project, which began formally in October 2003, funded by a five-year, $11.6M ITR (Information Technology Research) award from the National Science Foundation, and which was funded by no-cost extensions.

The objective of the project is to design, develop, and implement a hardware and software infrastructure to facilitate work with phylogenies and, in particular, to support an attempt at assembling the Tree of Life, the phylogeny of all organisms on the planet. The original CIPRES team was made of five lead institutions and another eight affiliated ones, comprising a total of 33 faculty researchers. The team has since grown to 16 institutions and gained additional faculty conducting research in algorithm design, database design, software architecture design, and modelling and simulation, developing a central software library with well defined and fully documented APIs, setting up a high-performance central compute and database server, and preparing a wide range of outreach tools.

Initially Bernard Moret was the director of the project, but Tandy Warnow became the director of the project upon Moret's departure for EPFL in Switzerland. She works with the Executive Committee to set directions, evaluate progress, and assign responsibilities.

Overarching Goal of CIPRES:

- To provide the computational infrastructure needed to reconstruct phylogenies for millions of taxa.

Our approach towards this goal has four principal components: (i) algorithmic research and development, (ii) concomitant research in modelling and simulation (to provide benchmark datasets and

to improve the quality of stochastic models), (iii) professional software development to provide a base of code that can run on a large server or be downloaded to a researcher's lab machines, and (iv) community engagement for training, feedback, dissemination, etc..

We believe that, through the development of this infrastructure, biologists will be able to investigate new scientific questions with greater subtlety and far-reaching consequences than before. Because all life is descended from earlier life, the process of speciation necessarily produces groups of organisms that share sets of traits. Therefore, understanding biology in all of its facets, e.g., morphology, biochemistry, biophysics, physiology, genomics, proteomics, etc., requires knowledge of the evolutionary relationships of organisms. The greatest promise of the CIPRES project is to provide the means of reconstructing the Tree of Life, in itself the ultimate resource for understanding biology through the comparative method. In that sense, there is no end to the list of potential benefits to Biology, so the list we now give hits only a few high points, increasing in scope from a narrow phylogenetic scope to a full planetary ecosystem.

- Better models of DNA sequence evolution and gene order evolution and better phylogenetic inference go hand-in-hand. CIPRES will refine both, using better trees to refine models, and better models to infer better trees. With large-scale phylogenies, it will become possible to determine whether different models are more appropriate in different parts of the tree, as well as whether different models pertain to different classes of genes or regions of the genome for different groups of organisms. Such determinations will lead to a better understanding of how evolution has acted to shape organisms as well as of some of the constraints within which they have evolved.
- Biologists have long been interested in how and why organisms have become adapted to both the external environment and the internal environment of their own cells/soma. A fairly complete tree will enable researchers to understand the extent and origin of individual adaptations and then extend that understanding to adaptive evolution in general.
- As biologists begin to understand better the genetics of adaptive evolution, it will become possible to address sophisticated questions about whether, in the process of evolution, life has explored the full parameter space of adaptive solutions. If it turns out that (as seems likely) only a small proportion of possibilities have been tried and new unexplored configurations are highlighted, we will be able to identify alternative solutions, the understanding of which may prove instrumental in maintaining the health of the planet and its residents.

Some benefits will accrue to the scientific community at large: CIPRES is, in a very fundamental way, an interdisciplinary project: it has an interdisciplinary focus, recruited a multidisciplinary team and proceeded to turn it into an interdisciplinary one, and, more importantly, every team member is well aware of the key role of interdisciplinary work within the project, or, to put it more starkly, of the fact that neither biologists alone nor computer scientists alone have any chance of succeeding in this endeavor. Finally, a number of our members are competent across a large range of subjects: for instance, our chief software architects are biologists, while some of our chief modelers are computer scientists; each of these researchers has gained a deep appreciation for the methods, culture, and values of some of the other disciplines represented within CIPRES. All of these factors make CIPRES a project that is really working in an interdisciplinary manner (as opposed to the multidisciplinary approach common to many large projects); we hope that CIPRES can serve as one example of how interdisciplinary research can be successfully conducted.

CIPRES has four specific interacting aims:

- Develop and implement algorithms that can routinely handle datasets with a million taxa.
- Develop advanced stochastic models of evolution for use in phylogenetic estimation and to provide standardized benchmarks for rigorous performance evaluation.
- Develop a robust software architecture that is modular, extensible, and optimized for performance.
- Provide outreach, education and training in phylogenetics to the general public and provide technical leadership to the scientific community.

Underlying these four aims is an effort in databases, which are needed for reconstruction, for curation of datasets and analyses, and for simulation. We thus constituted five overlapping working groups within CIPRES: Algorithms, Databases, Simulation and Modelling, Core Development, and Outreach. Each of these groups has designated leaders who, together with the five lead PIs, constitute the Executive Committee of CIPRES. The current membership of the executive committee is listed below.

Mark Holder (software)
Junhyong Kim (simulations)
Wayne Maddison (software)
Mark Miller (UCSD PI and core team)
Brent Mishler (outreach)
Satish Rao (Berkeley PI)
David Swofford (FSU PI)
Val Tannen (databases)
Tandy Warnow (UT PI and algorithms)

This write-up constitutes the entire report to the National Science Foundation. Due the the large number of institutions, participants, publications, etc., we put together this write-up using the structure of Fastlane annual reports and uploaded it as a single file.

The structure of the report is as follows. We provide the list of participants in Section 2. In Section 3 we discuss our educational activities (beyond direct supervision of students and postdocs, and the formal educational activities organized by the Outreach Focus Group). We then continue with the reports of the four different focus groups on algorithms (Section 4), software and the central resource (Section 5), simulation and modelling (Section 6), outreach activity (Section 7), and databases (Section 8). We summarize our contributions to science, human resources, and broader impact, in Section 9; details of the research contributions are provided in the relevant sections. Section 10 lists publications written by the group during the course of the grant, as well as project artefacts other than those focused on education.

2 Participants

We list all participating institutions, foreign collaborators, faculty, postdocs, graduate and undergraduate students, and staff, who were involved in the project for any significant amount of time, whether or not they drew any funds from the project, and whether or not they are currently active.

2.1 Participating Institutions

The CIPRES project is a community endeavor which now consists of 16 institutions, led by four collaborating institutions (UT-Austin lead, UC Berkeley, UCSD, and Florida State University). The current member institutions of CIPRES are:

- American Museum of Natural History
- Florida State University
- Georgia Institute of Technology
- New Jersey Institute of Technology
- North Carolina State University
- Rice University
- Texas A&M University
- University of Arizona
- University of British Columbia
- University of California, Berkeley
- University of California, San Diego
- University of Connecticut
- University of Pennsylvania
- University of South Carolina
- University of Texas, Austin
- Yale University

2.2 Faculty.

All faculty members listed below worked 160 hours or more on the project in some project year.

- David Bader, Prof. of Computing at Georgia Inst. of Technology. US citizen. Home Page: http://www.cc.gatech.edu/~bader/.
- Francine Berman, Prof. of Computer Science and Director of the San Diego Supercomputing Center, UCSD. Home Page: http://www.sdsc.edu/about/Director.html
- Susan Davidson, Prof. of Computer and Information Sciences, U. Penn. Home Page: http://www.cis.upenn.edu/~susan/.
- Michael Donoghue, Prof. of Ecology and Evolutionary Biology, Yale. Home Page: http://www.yale.edu/eeb/donoghue.
- Steven Evans, Prof. of Statistics (joint appointment in Mathematics), UC Berkeley. Home Page: http://www.stat.berkeley.edu/~evans/
- David Hillis, Prof. of Integrative Biology, UT Austin. Home Page: http://www.zo.utexas.edu/faculty/antisense/index.html
- Mark Holder, Assist. Professor, Department of Ecology and Evolution, University of Kansas. Software lead. Home Page: http://people.ku.edu/~mtholder/
- John Huelsenbeck, Prof. of Integrative Biology, UC Berkeley. Home Page: http://ib.berkeley.edu/people/faculty/person_detail.php?person=319
- Warren Hunt, Prof. of Computer Sciences, UT Austin. Home Page: http://www.cs.utexas.edu/~hunt/
- Robert Jansen, Prof. of Integrative Biology, UT Austin. Home Page: http://www.biosci.utexas.edu/ib/faculty/jansen.htm
- Sampath Kannan, Prof. of Computer and Information Sciences, U. Penn. Home Page: http://www.cis.upenn.edu/~kannan/
- Richard Karp, Prof. of Computer Science, UC Berkeley. Home Page: http://www.cs.berkeley.edu/~karp/
- Junhyong Kim, Prof. of Biology, U. Penn. Home Page: http://www.bio.upenn.edu/faculty/kim/
- Paul Lewis, Prof. of Ecology and Evolutionary Biology, U. Conn. Home Page: http://www.eeb.uconn.edu/people/plewis/
- C. Randal Linder, Associate Prof. of Integrative Biology, UT Austin. Home Page: http://www.biosci.utexas.edu/IB/faculty/linder.htm.
- David Maddison, Prof. of Entomology, U. Arizona. Home Page: http://david.bembidion.org/

- Wayne Maddison, Prof. of Zoology, U. British Columbia. Home Page: http://salticidae.org/wpm/home.html
- Lauren Ancel Meyers, Assoc. Prof. of Integrative Biology, UT Austin. Home Page: http://cluster3.biosci.utexas.edu/research/meyers/LaurenM/Lauren_M.html
- Daniel Miranker, Prof. of Computer Sciences, UT Austin. Home Page: http://www.cs.utexas.edu/users/miranker.
- Brent Mishler, Prof. of Integrative Biology, UC Berkeley. Home Page: http://ucjeps.berkeley.edu/people/mishler.html.
- Bernard Moret, Prof. of Computer Science, EPFL (Swiss Institute of Technology, Lausanne), Switzerland (formerly at UNM Dept. of Computer Science). Home Page: http://people.epfl.ch/bernard.moret
- Elchanan Mossel, Assoc. Prof. of Statistics and Computer Science, UC Berkeley. Home Page: http://www.stat.berkeley.edu/~mossel/
- Spencer Muse, Associate Prof. of Statistics, NCSU. Home Page: http://spensermuse.aas.duke.edu/~spencermuse/.
- Eugene Myers, Prof. of Computer Science, UC Berkeley (now at Janelia Farm Research Campus of the Howard Hughes Medical Instiute). Home Page: http://research.janelia.org/myers/
- Luay Nakhleh, Assoc. Prof. of Computer Science, Rice. US Permanent Resident, white, male, no diabilities. Home Page: http://www.cs.rice.edu/~nakhleh/
- Christos Papadimitriou, Prof. of Computer Science, UC Berkeley. Home Page: http://www.cs.berkeley.edu/~christos/
- William Piel, Assoc. Director of Evolutionary Bioinformatics, Peabody Museum of Natural History, Yale University. Home Page: http:www.treebase.org/~piel.
- Satish Rao, Prof. of Computer Science, UC Berkeley. US citizen, Home Page:http://www.cs.berkeley.edu/~satishr
- Usman Roshan, Associate Prof. of Computer Science, NJIT. US Home Page: http://cs.njit.edu/usman/
- Stuart Russell, Prof. of Computer Science, UC Berkeley. Home Page: http://www.cs.berkeley.edu/~russell.
- David L. Swofford, Senior Research Scientist, Duke Institute for Genome Sciences and Policy. Home Page: http://www.genome.duke.edu/centers/ceg/swofford/.
- Jijun Tang, Associate Prof. of Computer Science and Engineering, U. South Carolina. Home Page: http://www.cse.sc.edu/~jtang/
- Val Tannen, Prof. of Computer and Information Sciences, UPenn. Home Page: http:www.cis.upenn.edu/~val.

- Paul Turner, Assoc. Prof. of Ecology and Evolution, Yale. Home Page: http://www.yale.edu/turner/home/index.htm
- Tandy Warnow, Prof. of Computer Sciences, UT Austin. Home Page: http://www.cs.utexas.edu/users/tandy
- Ward C. Wheeler, Curator of Invertebrates, AMNH. Home Page: http://www.amnh.org/science/divisions/invertzoo/bio.php?scientist=wheeler
- Tiffani Williams, Assistant Prof. of Computer Science, Texas A&M. Home Page: http://faculty.cs.tamu.edu/tlw/

2.3 Primary professional Staff.

All staff members listed worked 160 hours or more on the project for some year, with the exception of some unpaid staff who worked on the AMNH educational outreach (and we indicate these as such).

Staff that are paid by CIPRES

- Alex Borchers, Staff Member, San Diego Supercomputing Center, UCSD.
- Adam Cathers, San Diego Supercomputing Center, UCSD. US citizen, white, male, no disabilities.
- Lucie Chan, Staff Member, San Diego Supercomputing Center, UCSD.
- April Davidson, Project Coordinator, UNM.
- T. Phong Dinh, Staff Member, San Diego Supercomputing Center, UCSD.
- Mark Dominus, Staff Member, U. Penn.
- Kevin Fowler, San Diego Supercomputing Center, UCSD.
- Paul Hoover, Staff Member, San Diego Supercomputing Center, UCSD.
- Dana Jermanis, Senior Software Developer, San Diego Supercomputing Center, UCSD.
- Terri Liebowitz, Staff Member, San Diego Supercomputing Center, UCSD.
- Brian Lucena, Visiting Faculty, UC Berkeley.
- Madhu Madhusudan, Staff Member, San Diego Supercomputing Center, UCSD.
- Tim McPhillips, San Diego Supercomputing Center, UCSD.
- Mark Miller, Team Leader, San Diego Supercomputing Center, UCSD. Home Page: http://www.sdsc.edu/~mmiller.

- Erica Ocegueda, Staff member, UNM.
- Cynthia Perrine, Education Coordinator, UC Berkeley.
- Jin Ruan, Staff Member, San Diego Supercomputing Center, UCSD.
- David Stockwell, Staff Member, San Diego Supercomputing Center, UCSD.
- Rahul Suri, Staff Member, UT-Austin.
- Ashton Taylor, Staff Member, San Diego Supercomputing Center, UCSD. Home page: http://www.digitalmudstudios.com.
- Can Tran, Staff Member, San Diego Supercomputing Center, UCSD.
- Brandan White, San Diego Supercomputing Center, UCSD.
- Tracy Zhao, Staff Member, San Diego Supercomputing Center, UCSD.

Staff that are not paid by CIPRES

- Laurie Alvarez, Staff member, UT-Austin.
- Adriana Aquino, Staff. Worked less than 160 hours in all grant years.
- Daniel Aviv, Staff. Male, worked less than 160 hours in all grant years.
- Joel Cracraft, Staff. Male Worked less than 160 hours in all grant years.
- Louise Crowley, Student, Female, Worked less than 160 hours in all grant years.
- Mohammad Faiz, Staff. Male, Worked less than 160 hours in all grant years.
- Megan Harrison, Postdoc, Female Worked less than 160 hours in all grant years.
- Jay Holmes, Staff. Male Worked less than 160 hours in all grant years.
- Daniel Janies, OSU-Staff. Male Worked less than 160 hours in all grant years.
- Maritza Macdonald, Staff. Female, Hispanic Worked less than 160 hours in all grant years.
- Mordecai MacLow, Staff. Male Worked less than 160 hours in all grant years.
- Mark Norell, Staff. Male Worked less than 160 hours in all grant years.
- Susan Perkins, Staff. Female Worked less than 160 hours in all grant years.
- Paola Predraza, Postdoc, Female, Hispanic Worked less than 160 hours in all grant years.
- Lorenzo Prendini, Staff. Male Worked less than 160 hours in all grant years.
- David Randle, Staff. Male Worked less than 160 hours in all grant years.
- Zobar Ris, Staff. Male Worked less than 160 hours in all grant years.
- Monique Scott, Staff. Female, African American, Worked less than 160 hours in all grant years.
- Cate Starr, Staff. Female, Worked less than 160 hours in all grant years.
- Ellen Trimarco, Staff. Female, Worked less than 160 hours in all grant years.

2.4 Postdoctoral Fellows.

Postdoctoral fellows funded by CIPRES

- Michael Alfaro, UCSD (postdoctoral fellow of John Huelsenbeck). Home Page: http://www.eeb.ucla.edu/indivfaculty.php?FacultyKey=10361
- Mark Holder (postdoctoral fellow of Dave Swofford while at FSU).
- Peter Midford, UBC (postdoc of Wayne Maddison). Home Page: http://mesquiteproject.org/midford/.
- Sagi Snir, UC Berkeley (postdoc of Lior Pachter). Home Page: http://math.berkeley.edu/~ssagi/
- Shel Swenson, UT Austin (postdoc of Tandy Warnow).
- Rutger Vos, Simon Fraser U. (research fellow with Wayne Maddison). Home Page: http://rutgervos.blogspot.com/.

Postdoctoral fellows not funded by CIPRES

- François Barbançon, UT Austin (postdoctoral fellow of Tandy Warnow).
- Tanya Berger-Wolf, UNM (postdoctoral fellow of Bernard Moret).
- Sarah Cohen-Boulakia, U. Penn (postdoctoral fellow of Val Tannen)
- Steve Fisher, U. Penn (postdoctoral fellow of Junhyong Kim).
- Fan Ge, University of Pennsylvania.
- Sergei Kosakovsky Pond, UCSD (research fellow working with Spencer Muse).
- Yelena Shvets, UC Berkeley (postdoc of Steve Evans and Monty Slatkin).
- Li-San Wang, U. Penn (postdoctoral fellow of Junhyong Kim),
- Cam Webb, Yale U. (postdoctoral fellow of Michael Donoghue).
- Tiffani Williams, UNM (postdoctoral fellow of Bernard Moret).

2.5 Graduate Students.

Students paid (mostly partially) through CIPRES:

- François Barbançon, UT Austin (student of Dan Miranker).

- Nicholas Bray, UC Berkeley (student of Lior Pachter).
- Kevin Chen, UC Berkeley (student of Lior Pachter and Satish Rao).
- Shirley Cohen, UPenn (student of Val Tannen and Susan Davidson). Home Page: http://www.seas.upenn.edu/~shirleyc.
- Costis Daskalakis, UC Berkeley (student of Satish Rao).
- Nick Eriksson, UC Berkeley (student of Bernd Sturmfels).
- Yu Fan, U. Conn (student of Paul Lewis).
- Kirsten Fisher, UC Berkeley (student of Brent Mishler).
- Ganesh Ganapathy, UT Austin (student of Tandy Warnow).
- Denise Green, UC Berkeley (worked with Brent Mishler).
- Sheng Guo, U. Penn (student of Junyong Kim).
- Tracy Heath, UT Austin (student of David Hillis).
- Cameron Hill, UC Berkeley (student in the Mathematics Department).
- David Kysela, Yale University (student of Paul Turner).
- Ruth Kirkpatrick, UC Berkeley (worked with Brent Mishler).
- Henry Lin, UC Berkeley (student of Satish Rao).
- Kevin Liu, UT Austin (student of Tandy Warnow).
- Wenguo Liu, UT Austin (student of Dan Miranker).
- Andrew McGregor, U. Penn (student of Sampath Kannan).
- Frank Mannino, NCSU (student of Spencer Muse).
- Rui Mao, UT Austin (student of Dan Miranker).
- Radu Mihaescu, UC Berkeley (student of Lior Pachter and Satish Rao).
- Eric Miller, UT Austin (student of Lauren Meyers).
- Luay Nakhleh, UT Austin (student of Tandy Warnow).
- Manikandan Narayanan, UC Berkeley (student of Dick Karp).
- Serita Nelesen, UT Austin (student of Warren Hunt).
- Smriti Ramakrishnan, UT Austin (student of Dan Miranker).
- Samantha Riesenfeld, UC Berkeley (student of Dick Karp).
- Sébastien Roch, UC Berkeley (student of Elchanan Mossel).
- Usman Roshan, UT Austin (student of Tandy Warnow).

- Ariel Schwartz, UC Berkeley (student of Gene Myers and Lior Pachter).
- Rebecca Shapley, UC Berkeley (worked with Brent Mishler).
- Stephen Smith, Yale (student of Michael Donoghue).
- Errol Strain, NCSU (student of Spencer Muse).
- Jeet Sukumaran, Kansas (student of Mark Holder).
- Shel Swenson, UT Austin (student of Tandy Warnow).
- Kunal Talwar, UC Berkeley (student of Christos Papadimitriou and Satish Rao).
- Andres Varón, CUNY (student of Ward Wheeler).
- Rutger Vos, U. British Columbia (student of Wayne Maddison).
- Yifeng Zheng, U. Penn (student of Susan Davidson and Junhyong Kim). Home Page: http://www.cis.upenn.edu/~yifeng.
- Derrick Zwickl, UT Austin (student of David Hillis).

Graduate students paid through other grants:

- Matt Ackerman; PhD student at Missouri State University, Funded by Google (Summer of Code participant), supervised by Mark Holder.
- Dan Adkins, UC Berkeley (student of Satish Rao).
- Stanislav Angelov, U. Penn (student of Sanjeev Khanna and Sampath Kannan).
- Maud Artaud, UCSD (visiting student from France).
- Jason Caravas; PhD student at Wayne State University. Funded by Google (Summer of Code participant), supervised by Rutger Vos.
- Guojing Cong, UNM (student of David Bader).
- Siobain Duffy, Yale U. (student of Paul Turner).
- Fan Ge, U. Penn (student of Junhyong Kim).
- Eric Gottlieb, M.S., UNM CS, of Bernard Moret.
- Boulos Harb, U. Penn (student of Sampath Kannan).
- Chris Harrelson, UC Berkeley (student of Satish Rao).
- Kris Hildrum, UC Berkeley (student of Satish Rao).
- Bonnie Kirkpatrick, UC Berkeley (student of Steve Evans).
- Mahesh Kulkarni, UNM (student of Bernard Moret).

- Melanie Langlois, UCSD (visiting student from France).
- Richard Liang, UC Berkeley (student of Steve Evans).
- Brian Moore, Yale U. (student of Michael Donoghue).
- Monique Morin, UNM (student of Bernard Moret).
- Anneke Padolina, UT-Austin (student of Randy Linder).
- David Suarez Pascal; PhD student at UNAM. Funded by Google (Summer of Code participant), supervised by Mark Holder.
- Nicholas Pattengale, UNM (student of Bernard Moret).
- Sindhu Raghavan, UT-Austin (student of Tandy Warnow).
- Peter Ralph, UC Berkeley (student of Steve Evans).
- Derek Ruths, Rice (student of Luay Nakhleh).
- Allan Sly, UC Berkeley (student of Steve Evans).
- Krister Swenson, UNM (student of Bernard Moret).
- Jijun Tang, UNM (student of Bernard Moret).
- Ruth Timme, UT Austin (student of Randy Linder).
- Fang Yue, U. South Carolina (student of Jijun Tang).
- David Zhao, UT Austin (student of Tandy Warnow).
- Lijuan Zhao, UNM (student of Bernard Moret).

2.6 Undergraduate Students (partial)

These students were not funded by CIPRES. Most did not work for 160 hours or more in any grant year.

- Abraham Bachrach, UC Berkeley.
- Kevin Bullaughey, U Penn.
- Chris Crutchfield, UC Berkeley.
- Alex Jaffe, UC Berkeley.
- Sun Jin Lee, Yale U..
- Ving Ian Lei, UT Austin (student of Dan Miranker).
- Jenny Liu, UC Berkeley.

- Erik Lewis, UC Berkeley.
- Zack Mahdavi, UT Austin.
- Diana Miachalek, UC Berkeley.
- Kavya Rao, UCSD (intern at SDSC).
- Apurva Shah, UC Berkeley.
- Jennifer Vo, UCSD (intern at SDSC).
- Yul Yang, Yale U.

2.7 Other students

- Jorge Alva, high school student, intern at SDSC.

2.8 International collaborators

We have several international collaborators, including Prof. Olivier Gascuel, U. Montpellier (France), Dr. Pablo Goloboff (Argentina), Prof. Daniel Huson, U. Tübingen (Germany), Prof. Jens Lagergren, Royal Inst. Technology (Sweden), Prof. David Sankoff, U. Ottawa (Canada), Dr. Alexis Stamatakis, Greece. and Prof. Michael Steel, U. Christchurch (New Zealand).

2.9 Organizational Partners.

NESCent. A major domestic partner is the NSF-funded Center for Evolution Synthesis, NESCent; we have maintained contact with NESCENT directors from its initial days. Several of our participants have taught in NESCent workshops and courses, and have mentored students at the Google Summer of Code activities held at NESCent. We have also coordinated funding activities for postdocs and sabbaticals as well. For example, CIPRES wanted to fund Derrick Zwickl for a postdoc year, and NESCent was also interested in funding him, so we came to a mutually agreeable arrangement through which NESCent funded a 2-year postdoc for Zwickl, and CIPRES funded a 1-year sabbatical for Paul Lewis at NESCent. In addition to Zwickl, two of our former doctoral students (Kirsten Fisher and Ganesh Ganapathy) and one of our current doctoral students (Stephen Smith) have received NESCent postdoctoral fellowships. Perhaps most importantly, we have collaborated with NESCent in the development and maintenance of TreeBASE-II (see the Databases section for more about this collaboration).

AToL groups. A second major domestic partner is the collection of AToL-funded groups, and CIPRES members have given presentations on CIPRES at AToL PI meetings. Several CIPRES members are part of AToL teams and keep us in touch with those teams. We have collaborated with a few AToL centers (e.g., the NemATol project) to help them analyze their data while gaining

valuable experience with the behavior of our tools on large biological datasets. We also invited AToL participants to our own group meetings (most notably, to the All-Hands Meeting held in Austin in early 2006). Among other things, we learned of the importance to AToL teams of good algorithms for multiple sequence alignment and of the expectation from these teams that CIPRES would work on this problem—something that was not in our proposal nor in our cooperative agreement, but that we have since then made substantial progress on (and were subsequently awarded an AToL grant to continue).

SEEK. Finally, another major domestic partner is the SEEK project, also funded by a large NSF ITR award, with major components at UNM. SDSC, and Kansas. Details about our developing collaboration with SEEK, much of which centers on the integration of the CIPRES framework and the Kepler workflow tool, are to be found in Section 5.

3 Educational activities

In this section we describe a few of the educational activities provided by CIPRES. This includes new courses taught, but also includes activities organized for its students and postdoctoral trainees.

3.1 Graduate student meetings

CIPRES held All-Hands meetings each of its first three years, which included research talks and posters, and were open to the general public. These meetings included educational activities organized specifically for students and postdocs working for CIPRES:

- The first took place at the Marconi Conference Center in northern California, organized for the Algorithms group students and postdocs.
- The second took place in Taos, NM, in July 2005. This meeting was for explicitly for students and postdocs only (no faculty or senior personnel allowed), but all students were invited.
- The third took place during the second All-Hands Meeting, in Austin in February 2006, and was also for students and postdocs only.

Marconi Meeting, December 2004. The Algorithms group held a workshop at the Marconi Conference Center (California), with the aim of providing a deeper education in biological motivation and applications for the students working on algorithm development. Tutorials were given by CIPRES participants Evans (Statistics), Warnow (Computer Science), and Linder (Biology), and research seminars led by Linder (reticulate evolution), Moret (gene-order phylogeny), and Rao (large-scale optimization). About 45 people attended, most of them students or postdocs from Computer Science and Statistics. Four biologists (Ruth Timme, PhD student at UT Austin, Randy Linder, Wayne Maddison, and Dave Swofford) and two SDSC staff members (Terri Liebowitz and Alex Borchers) also attended.

Taos Meeting, July 2005. About 30 students and postdocs attended the Taos (New Mexico) meeting, roughly equally divided between biologists and non-biologists (computer scientists, mathematicians, and statisticians). Faculty and other senior personnel were not included in this meeting, at the request of the students; this helped participants in asking potentially naive questions and also allowed a freer choice of topics for discussion. A questionnaire was sent out to the participants, with about a 25% response rate; tabulated answers and quotes are provided on the CD.

The schedule for the meeting was as follows:

- Monday the 18th:

  07:00–09:00: Continental breakfast provided
  09:00–10:00: Introductions

10:00–10:15: Snack
10:15–12:00: Group (Phylogenetic Reconstruction—Brian O'Meara)
12:00–14:00: Lunch catered at hotel: New Mexican Buffet
14:00–14:30: Group (Networks—Sagi Snir)
14:30–14:45: Snack
14:45–16:00: Group (Networks—Sagi Snir)
16:15–18:00: Group (Applications of Phylogeny—Lucia Peixoto)

- Tuesday the 19th:

  07:00–09:00: Continental breakfast provided
  09:00–10:00: Group (MSA—Corrie Moreau)
  10:00–10:15: Snack
  10:15–11:00: Group (MSA—Corrie Moreau)
  11:00–13:00: Lunch (on our own)
  13:00–14:45: Group (Databases—Shirley Cohen)
  14:45–16:15: Snack + Free Time
  16:15–18:00: Group (Modeling and Simulation—Tracy Heath)
  19:00: Dinner at Joseph's Table

Austin Meeting, February 2006. The meeting in Austin, TX, followed the second All-Hands Meeting (AHM) for the entire CIPRES project. Many graduate students and postdocs presented talks at the AHM; there was also an actively attended poster session. The Austin Student Meeting described here was held the day after the AHM 2006 meeting concluded; once again, only students and postdocs were invited.

- Saturday evening: Dinner at Dona Emilia's, 19:00.
- Sunday morning: Building/Room: Taylor 3.128

  09:00-10:00 - Coffee and introductions
  10:00-10:50 Two group discussions:
  - Speeding up the ML algorithm (Leaders: Alexis Stamatakis, Derrick Zwickl)
  - Things that confound phylogeny (Leaders: Kris McGary, Ruth Timme)

  Short Break
  11:00-12:00 Two group discussions
  - Multiple sequence alignment (Leaders: Kevin Liu, David Zhao)
  - Trees as input (Leaders: Shel Swenson, Tracy Heath, Serita Nelesen)

3.2 Mini-Symposium and Workshop in Evolutionary Simulations, 2006

The Modelling and Simulations group held a highly successful Mini-Symposium and Workshop in Evolutionary Simulations March 31-April 1, 2006 which brought together various people using

simulation-based approaches to study evolution. Highlights included keynote talks by Christina Burch on viral evolution, Carlo Maley on individual-based simulation of tumor evolution, John Yin on detailed mechanistic simulation of viral reproduction, and Paul Higgs in RNA evolution and simulation. A full list of the participants is as follows:

- Abedon, Stephen T. Ohio State University
- Beckmann, Kevin, Penn State University
- Burch, Christina, University of North Carolina
- Cowperthwaite, Matt, University of Texas, Austin
- Dang, Kristen K., UNC-Chapel Hill
- Draghi, Jeremy, Yale University
- Duffy, Siobain, Yale University
- Strain, Errol, NC State University
- Fisher, Stephen, University of Pennsylvania
- Guisinger, Mary, University of Texas, Austin
- Heath, Tracy, University of Texas, Austin
- Higgs, Paul, McMaster University
- Hillis, David, University of Texas
- Karmarkar, Vidyadhar, Penn State University
- Kim, Sangtae, University of Florida
- Kysela, David, Yale University
- Landweber, Laura, Princeton University
- Leebens-Mack, Jim, Penn State University
- Maley, Carlo C., The Wistar Institute
- Mannino, Frank, NC State University
- Miller, Eric, University of Texas, Austin
- Moody, Michael, Indiana University
- Moore, Michael J., University of Florida
- Muse, Spencer, NC State University
- Peixoto, Lucia, University of Pennsylvania
- Poggio, Andy, UC Berkely

- Turner, Paul, Yale University
- Wang, Li-San, University of Pennsylvania
- Warnow, Tandy, University of Texas, Austin
- Wilke, Claus, University of Texas, Austin
- Yin, John, University of Wisconsin-Madison
- Yue, Feng, University of South Carolina
- Yuri, Tamaki, Smithsonian Institute
- Zheng, Yifeng, University of Pennsylvania

3.3 New courses taught

Several of the CIPRES faculty (e.g., Warnow, Hillis, Linder, etc.) regularly teach evolution and phylogenetics, both to biology students and to computer science students. However, new courses were also created in response to the growing interest in this research area at the collaborating institutions. We briefly describe these new courses.

UC Berkeley Statistics. Steve Evans is a Professor of Statistics, with a joint appointment in Mathematics. In Fall 2004, Evans taught an undergraduate 3 unit "honors seminar" course (Stat 157) on phylogeny. In Fall 2005, Evans did the same on population genetics with a phylogenetic component. Each course had 15 students. In Fall 2005, Evans also gave a lecture in our VIGRE undergraduate lecture series that covered his research on historical linguistics with Warnow. In Spring 2005, Evans ran a graduate reading group on phylogeny for students from Statistics, Integrative Biology and Computer Science with 7 students. In Fall 2005, Evans helped run a joint Statistics - Integrative Biology reading group on fitness landscapes and speciation with about 8 students. In Spring 2006, Evans participated in a joint Integrative Biology - Statistics - Computer Science reading group on gene trees versus species trees with about 30 participants.

UT Austin, Computer Sciences. Warnow created a computer science undergraduate course for non-majors, in which multiple sequence alignment and phylogenetic reconstruction were the scientific applications for the computer science concepts taught in the course.

The UT-Austin group hosted two undergraduate students from Huston-Tillotson, a historically black college in Austin, Texas, each of the 2008 and 2009 summers, and trained them through a computational phylogenetics research project.

3.4 Instructional Materials

Many CIPRES faculty has created educational materials in phylogenetics, specifically designed for computer scientists, mathematicians, and statisticians. Some of these efforts have produced book chapters, tutorials, and webpages.

- Book chapters: see (4; 211; 99; 208).
- Tutorials held at international meetings. In particular, we have created tutorials at the CSB meeting (overview of phylogenetics), at PSB 2004 (reticulate evolution), and PSB 2008 (multiple sequence alignment and phylogeny estimation under complex models of evolution).
- NESCent educational activities. Several of our main participants (e.g., Rutger Vos, Bill Piel, Wayne and David Maddison) were instructors in the NESCent computational phyloinformatics courses. In addition, CIPRES participants have mentored Google Summer of Code students (Caravas, Pascal, and Ackerman).
- Workshops. Many of the CIPRES faculty teach at the annual Woods Hole Meetings, and train new students in modern phylogenetic tools (such as Phycas, for example).

See http://www.cs.utexas.edu/users/tandy/tutorials.html for some of these tutorials and book chapters.

4 Algorithms

4.1 Personnel

Focus leader: Tandy Warnow

All personnel listed worked 160 hours or more in a grant year, unless otherwise indicated (in fact, only certain staff who worked on the AMNH outreach activity worked less than 160 hours in any calendar year for CIPRES).

Senior Personnel:

- David Bader, School of Computing, Georgia Institute of Technology
- Michael Donoghue, Yale University, Ecology and Evolution
- Steve Evans, Mathematics and Statistics, UC Berkeley
- John Huelsenbeck, Integrative Biology, UC Berkeley
- Warren Hunt, Computer Sciences, UT-Austin
- Sampath Kannan, Computer and Information Sciences, The University of Pennsylvania
- Richard Karp, Computer Science, UC Berkeley
- C. Randal Linder, Integrative Biology, UT-Austin
- Bernard Moret, EPFL (Switzerland); formerly of Computer Sciences, Univ. of New Mexico
- Elchanan Mossel, Statistics and Computer Sciences, UC Berkeley
- Luay Nakhleh, Computer Science, Rice University
- Christos Papadimitriou, Computer Science, UC Berkeley
- Satish Rao, Computer Science, UC Berkeley
- Usman Roshan, Computer Science, New Jersey Institute of Technology
- Stuart Russell, Computer Science, UC Berkeley
- Alexandros Stamatakis, EPFL, Switzerland (foreign collaborator)
- Jijun Tang, Computer Science, The University of South Carolina
- Li-San Wang, Biology, The University of Pennsylvania
- Tandy Warnow, Computer Sciences, UT-Austin
- Ward Wheeler, American Museum Natural History
- Tiffani Williams, Computer Science, Texas A&M

Students and postdocs funded by CIPRES

- Michael Alfaro. Postdoctoral fellow, UCSD Biology, of John Huelsenbeck.

- Nicholas Bray. PhD student, Berkeley Math, of Lior Pachter.
- Kevin Chen. PhD student, Berkeley Math, of Lior Pachter and Satish Rao.
- Costis Daskalakis. PhD student, Berkeley CS, of Christos Papadimitriou.
- Nick Eriksson. PhD student, Berkeley Math, of Bernd Sturmfels.
- Ganesh Ganapathy. PhD student, UT-Austin CS, of Tandy Warnow and Vijaya Ramachandran.
- Cameron Hill. PhD student, Berkeley Math, of Satish Rao.
- Alex Jaffe. Undergraduate student, Berkeley CS, of Satish Rao.
- Henry Lin. PhD student, Berkeley CS, of Satish Rao.
- Kevin Liu. PhD student, UT-Austin Computer Sciences, of Tandy Warnow.
- Andrew McGregor. PhD student, Penn CIS, of Sampath Kannan.
- Radu Mihaescu. PhD student, Berkeley Math, of Lior Pachter and Satish Rao (Berkeley CS).
- Luay Nakhleh. PhD student, UT-Austin CS, of Tandy Warnow.
- Manikandan Narayanan. PhD student, Berkeley CS, of Dick Karp.
- Samantha Riesenfeld. PhD student, Berkeley CS, of Dick Karp.
- Sébastien Roch. PhD student, Berkeley Statistics, of Elchanan Mossel.
- Usman Roshan. PhD student, UT-Austin CS, of Tandy Warnow.
- Ariel Schwartz. PhD student, Berkeley CS, of Gene Myers and Lior Pachter.
- Sagi Snir, postdoctoral fellow of Lior Pachter and Satish Rao at UC Berkeley, Mathematics Department.
- Michelle (Shel) Swenson, PhD student, UT-Austin Computer Sciences, of Tandy Warnow and C. Randal Linder
- Kunal Talwar. PhD student, Berkeley CS, of Christos Papadimitriou and Satish Rao.
- Andres Varón, PhD student, AMNH, of Ward Wheeler.
- Derrick Zwickl. PhD student, UT-Austin Biology, of David Hillis.

Students and postdocs not funded by CIPRES

- Dan Adkins. PhD student, Berkeley CS, of Satish Rao.
- François Barbançon. Postdoctoral fellow, UT-Austin CS, of Tandy Warnow. (Funded by CIPRES when he was a PhD student of Dan Miranker.)
- Tanya Berger-Wolf. Postdoctoral fellow, UNM CS, of Bernard Moret.
- Eric Gottlieb, M.S., UNM CS, of Bernard Moret.
- Chris Harrelson, PhD. student of Satish Rao, Berkeley CS.
- Boulos Herb, PhD. student of Sampath Kannan, University of Pennsylvania Computer and Information Sciences.
- Kris Hildrum, PhD student of Satish Rao, Berkeley CS.

- Bonnie Kirkpatrick, PhD student, Berkeley
- Jenny Liu. Undergraduate student of Satish Rao, Berkeley CS.
- Zack Mahdavi. Undergraduate student, UT-Austin CS, of Tandy Warnow. Honors thesis on maximum likelihood and multiple sequence alignment.
- Diana Miachalek. Undergraduate student, Berkeley CS, of Satish Rao.
- Nicholas Pattengale, M.S., UNM CS, of Bernard Moret.
- Sindhu Raghavan, UT-Austin, Computer Science, PhD student
- Stephen Smith, Yale University, PhD student of Michael Donoghue.
- Jijun Tang. PhD student, UNM CS, of Bernard Moret.
- Li-San Wang. Postdoctoral fellow, Univ. of Pennsylvania Biology, of Junhyong Kim.
- TIffani Williams. Postdoctoral fellow, UNM CS, of Bernard Moret.
- David Zhao. PhD student, UT-Austin, CS, of Tandy Warnow.

4.2 Overview

The fundamental goal of the Algorithms group is to develop phylogenetic reconstruction algorithms that will scale to the millions of taxa required for the Tree of Life. In addition, we have specific interest in developing methods that will be able to take advantage of a variety of data (mostly sequence data, but also whole-genome data, and non-molecular data), as well as to investigate issues in reticulate evolution (evolution caused by events such as hybridization or lateral gene transfer that does not fit with the linear descent model represented by trees). The research activity in the algorithms group has two complementary directions:

- the development of fundamental theory about phylogeny reconstruction methods, especially in terms of computational complexity, approximability, and theoretical performance guarantees under Markov models of evolution, and
- the development of novel reconstruction methods which have demonstrable advantages over existing phylogeny reconstruction methods, with respect to topological accuracy and/or computational complexity.

In general, the Berkeley group focuses on research of the first type, while the other researchers focus on research of the second type; however, most researchers in this group do both types of research, and the researchers interact with each other and with other members of CIPRES.

The majority of the funding for Algorithms research was used during the first three years of the grant (2003-2006), during which time UC Berkeley researchers were contributing to the CIPRES research very actively. Ongoing algorithms research during the last years of the grant has largely been done without CIPRES financial support, but has led to additional new advances in fundamental theory and in improved software, some of which has been added to the CIPRES software distribution, and made available to the public through the CIPRES portal.

Some of the highlights of the Algorithms research group activity are:

1. The development of new heuristics for maximum parsimony and maximum likelihood, that can analyze very large datasets much faster than existing methods. This component of the project includes new DCM-boosted versions of the PAUP* ratchet and of RAxML, and a new version of RAxML that includes bootstrapping. The initial development of DCM-boosters for maximum parsimony pre-dates the grant, but the specific application of this methodology for use with the CIPRES parsimony ratchet search took place during the first few years of the grant. The development of DCM-boosters for maximum likelihood (and in particular for use with RAxML), and of the new fast version of RAxML that includes bootstrapping, took place in the last two years of the grant. These heuristics are on the CIPRES portal, and available in the CIPRES software distribution.

2. Improved MCMC methods, and basic theory related to MCMC methods. Some of this research resulted in improvements to MrBayes, and the new version of MrBayes is part of the CIPRES portal and software distribution.

3. Fundamental theory about existing and novel phylogeny reconstruction methods under Markov models of evolution (developed by Mossel, Rao, Warnow, and students at Berkeley). Some of this research produced new methods with improved sequence length requirements. This work took place in the first three years of the grant.

4. Improved methods for detecting and reconstructing reticulate evolution (developed by Nakhleh, Moret, Warnow, Karp, and students). Most of this work took place in the first three years of the grant.

5. Improved methods for constructing phylogenetic trees from gene order and content data (developed by Moret, Warnow, Tang, Wang, and students). All of this work took place in the first three years of the grant.

6. New methods for multiple sequence alignment, for simultaneous estimation of alignments and trees, and for estimating the "indel history" of a set of sequences on a given phylogeny (work by Linder, Myers, Pachter, Roshan, Warnow, Wheeler, with collaborators and students) Included in this collection is a recently released new version of POY, which is both faster and more accurate than earlier versions, several new multiple sequence alignment methods, and several simulation studies. Most of this work was done in the last two years of the grant, although the work on POY has been ongoing throughout the grant. During the 2008-2009 year, the Warnow-Linder laboratory developed a new method, called SATé, for Simultaneous Estimation of Trees and Alignments. This method appeared in Science (101), and is able to construct trees and alignments from very large (up to 1000 sequences) DNA datasets in 24 hours, improving upon the best current two-phase methods. Subsequent research has identified the key design issues that yield improvements, and has produced a new variant of SATé with improved accuracy.

7. New supertree methods. Supertree methods estimate trees on large sets of taxa by combining trees on subsets of the taxa. While MRP (Matrix Representation with Parsimony) is the most popular supertree method, other methods have been proposed. The Warnow-Linder laboratory developed the SuperFine method, a novel supertree method that uses two steps to produce a highly accurate supertree. They showed (190) that SuperFine produces more accurate supertrees than MRP and other supertree methods, and completes on a fraction of the time used by MRP on large supertree datasets.

8. DACTAL. The Warnow-Linder laboratory also developed a new method for phylogeny estimation that can compute trees from molecular datasets without ever constructing a multiple sequence alignment on the entire dataset. This method, DACTAL (147), for "Divide-And-Conquer Trees without ALignments", runs quickly, and can compute trees with higher accuracy than even SATé.

5 Software Development and Central Resource

5.1 Personnel

Focus group leaders:

- Software development: Mark Holder, Mark Miller, Dave Swofford, and Wayne Maddison.
- Central Resource: Mark Miller.

Senior Personnel

- Francine Berman; University of California, San Diego, San Diego Supercomputer Center.
- Mark Holder; University of Kansas, Department of Ecology and Evolution.
- Mark Miller; University of California, San Diego, San Diego Supercomputer Center.
- Paul Lewis; University of Connecticut, Departments of Ecology and Evolutionary Biology.
- David Swofford; Duke University, National Evolutionary Synthesis Center (NESCent).
- Wayne Maddison; University of British Columbia, Departments of Zoology and Botany.

Other Personnel

- Jin Ruan; University of California, San Diego, San Diego Supercomputer Center.
- Madhusudan; University of California, San Diego, San Diego Supercomputer Center.
- Tracy Zhao; University of California, San Diego, San Diego Supercomputer Center.
- Can Van Tran; University of California, San Diego, San Diego Supercomputer Center.
- Adam Lathers; University of California, San Diego, San Diego Supercomputer Center.
- T. Phong Dinh; University of California, San Diego, San Diego Supercomputer Center.
- Dana Jermanis; University of California, San Diego, San Diego Supercomputer Center.
- Ashton Taylor; University of California, San Diego, San Diego Supercomputer Center.
- Brendan White; University of California, San Diego, San Diego Supercomputer Center.
- Terri Liebowitz; University of California, San Diego, San Diego Supercomputer Center.
- Lucie Chan; University of California, San Diego, San Diego Supercomputer Center.

- Paul Hoover; University of California, San Diego, San Diego Supercomputer Center.
- Alex Borchers; University of California, San Diego, San Diego Supercomputer Center.
- David Stockwell; University of California, San Diego, San Diego Supercomputer Center.
- Rahul Suri, Staff at UT-Austin.

Postdoctoral and Graduate Students:

- Peter Midford; (postdoctoral fellow, Mark Holder supervisor), University of Kansas, Department of Ecology and Evolution.
- Rutger Vos; (postdoctoral fellow, Wayne Maddison supervisor), University of British Columbia, Department of Zoology.
- Jeet Sukumaran; (graduate student, Mark Holder supervisor), University of Kansas, Department of Ecology and Evolution.

Intern Students:

- Matthew Ackerman, PhD student, The Google Summer of Code, NESCent. Google paid his stipend, but he was supervised by CIPRES project participants.
- Jorge Alva, high school student, not paid by CIPRES
- Maud Artaud; University of California, San Diego, San Diego Supercomputer Center (not paid by CIPRES).
- Jason Caravas (Google Summer of Code participant) was mentored by Rutger Vos in a 2007 project to develop nexml and Perl support for phyloinformatics (https://www.nescent.org/wg_phyloinformatics/PhyloSoC:Phylogenetic_XML). Not paid by CIPRES.
- Melanie Langlois; University of California, San Diego, San Diego Supercomputer Center. Not paid by CIPRES.
- David Suarez Pascal (Google Summer of Code participant) was mentored by Mark Holder in a project to support the NEXUS parsing library, NCL, in scripting languages (https://www.nescent.org/wg_phyloinformatics/PhyloSoC:Multi-language_bindings_to_the_NEXUS_Class_Library). Not paid by CIPRES.
- Kavya Rao, UCSD freshman, not paid by CIPRES
- Jennifer Vo, UCSD freshman, not paid by CIPRES

5.2 Overview

This portion of the CIPRES report covers two main activities: the Software Development effort, which is handled by a distributed group, and the Central Resource, which is the responsibility of the SDSC group led by Mark Miller. Because of the close connections between these two activities, we provide a merged report.

The group we refer to here as the "Software Group" includes all the CIPRES members at the San Diego Supercomputer Center (SDSC), as well as the members of the software development group who are located at different universities in the USA and Canada. This group began with three interdependent goals:

- Establishment of a computational platform to allow systematists to perform phylogenetic analyses of large datasets,
- Development of new open-source software to improve phylogenetic reconstruction and post-tree analyses, freely distributed to the scientific community, and
- Development of open-source freely distributed software libraries to provide a framework for programmers, to enable the creation and integration of community software into a central package that is deployable and scalable.

Initially the Central Resource group was led by Francine Berman, with Mark Miller as co-PI. In 2005 the PI role was assumed by Miller, as Berman moved on to lead other projects at SDSC. The SDSC Central Resource focused on several specific goals:

1. Providing professional software developers for the software focus group,
2. Providing professional software developers for the TreeBASE2 effort in the database focus group,
3. Creating and maintaining a public face for the CIPRES project, and
4. Installing and maintaining a computational resource for the CIPRES project and its constituency

Progress in achieving the first two goals is described under the sections describing activities of the software and database focus groups. The progress of the central resource in achieving the last two goals is presented below.

During the first two years, the two groups worked together on developing an appropriate architecture, purchasing and installing the compute cluster, initiating our library development, and using the libraries to create fast methods for large-scale phylogenetic analysis. However, community use of the libraries and the standalone software was not what we desired. Therefore, during the third year of the grant, we added a fourth goal:

- Development of the CIPRES Portal, to enable systematists to obtain highly accurate phylogenetic analyses of their datasets using CIPRES software on the CIPRES computer

Soon after its creation, the CIPRES Portal became a highly used resource, exceeding our expectations. As a result of the response, the focus of the software development group changed. We are actively engaged in extending the portal, adding functionality (in particular, enabling multiple sequence alignment, simulation tools, and algorithms for simultaneous estimation of alignments and trees). While we continued to create new software for phylogenetic construction and post-tree analyses, we scaled down the effort to create software libraries for programmers.

6 Modeling and Simulations

6.1 Personnel

Focus group leader: Junhyong Kim

Senior Personnel

- Junhyong Kim, University of Pennsylvania, Department of Biology.
- David Hillis, University of Texas, Section of Integrative Biology.
- Susan Davidson, University of Pennsylvania, Department of Computer and Information Science.
- Sampath Kannan, University of Pennsylvania, Department of Computer and Information Science.
- Spencer Muse, North Carolina State University, Department of Statistics.
- Lauren Ancel Meyers, University of Texas, Section of Integrative Biology.
- Paul Turner, Yale University, Department of Ecology and Evolutionary Biology.

Students and postdocs supported by CIPRES

- Sheng Guo (Graduate Student of J. Kim), University of Pennsylvania, Develop RNA simulator for macro-evolution.
- Stephen Fisher (Postdoctoral Fellow of J. Kim), University of Pennsylvania, Worked on CRIMSON system,
- Yifeng Zheng (Graduate Student of S. Davidson), University of Pennsylvania, Worked on CRIMSON system,
- Andrew McGregor (Graduate Student of S. Kannan), University of Pennsylvania, Ancestral state reconstruction.
- Derrick Zwickl (Graduate Student of D. Hillis), UT Austin, Key molecule simulation parameter estimation.
- Tracy Heath (Graduate Student of D. Hillis), UT Austin, Develop complex branching process simulators.
- Errol Strain (Graduate Student of S. Muse), North Carolina State, Key molecule simulation.
- Frank Mannino (Graduate Student of S. Muse), North Carolina State, Key molecule simulation and revision of HYPHY.

- Eric Miller (Graduate Student of Ancel-Meyers), UT Austin, RNA population level simulation.
- David Kysela (Graduate Student of P. Turner), Yale, RNA virus experimental evolution.

Students and postdocs not supported by CIPRES

- Sergei Kosakovsky Pond (Research Collaborator of Spencer Muse at NC State), Kosakovsky Pond scaled up the HhPHY simulator for very large scale trees and develop parallel implementation.
- Li-San Wang, (Postdoctoral fellow of J. Kim), University of Pennsylvania, Wang worked with Kim to develop a RNA simulator, and worked with Warnow on sequence alignment problems.
- Fan Ge (Graduate Student of J. Kim), University of Pennsylvania, Worked on collating empirical datasets.
- Kevin Bullaughey (Undergrad Student of J. Kim), University of Pennsylvania, Worked on statistics of tree comparison metrics

6.2 Overall goals

The Simulation and Modeling Team consisted of five groups led by Junhyong Kim (University of Pennsylvania), David Hillis (UT-Austin), Lauren Ancel Meyers (UT-Austin), Spencer Muse (NC State), and Paul Turner (Yale), with overall project directed by J. Kim. The stated goal at the beginning of the CIPRES project was to:

- Curate phylogenetically relevant key data from molecular databases (in collaboration with ATOL-Sanderson team)
- Statistically characterize key molecules (Muse, Hillis)
- Develop data management strategies for simulated datasets of several million branches (Kim, Davidson)
- Develop computational strategies for scalable simulation (Kim, Kannan, Moret, Warnow)
- Develop models of molecular evolution for key molecules (Muse, Hillis).
- Curate and analyze empirical datasets for benchmark purposes (Turner)

6.3 Accomplishments

The group activity was very successful, in large part achieving the stated goals and exceeding them in some ways. The major highlights of the group activity are:

- Development of novel tree branching simulation model that much more closely reproduces the statistical characteristics of empirical phylogenies than standard branching models,
- Development of an inhomogeneous fitness-dependent molecular evolution simulation that more closely reproduces the statistical characteristics of empirical phylogenies than standard stochastic models of sequence evolution,
- An (unprecedented) one-million taxon simulation dataset that can test full scalability of algorithms up to the size of Tree of Life,
- A new database and computational experiment system (which we call CRIMSON) that automates taxon-sampling and experimental design for algorithm studies, and
- Refinement of HyPHY for large-scale molecular evolution studies and use of experimental evolution to guide molecular simulations.

7 Outreach Activity

7.1 Personnel

Focus leader: Brent Mishler (UC Berkeley).

Senior Personnel:

- Michael Donoghue (PI at Yale subcontract, US citizen, white male)
- Brent Mishler (UC Berkeley, white Male, US citizen).
- Ward Wheeler (PI at AMNH subcontract, US citizen, white male).

Students and Staff paid by CIPRES

- Kirsten Fischer, graduate student at UC Berkeley, Female
- Denise Green, graduate student at UC Berkeley, Female
- Ruth Kirkpatrick, graduate student at UC Berkeley, Female
- Anna Larsen, Student, Female
- Cynthia Perrine, Staff at UC Berkeley, Female
- Rebecca Shapley, graduate student at UC Berkeley, Female

Description of student and staff activities Graduate student Kirsten Fisher worked in Mishler's lab (Spring Semester 2004 through Fall Semester 2004) on preparing the website materials and coordinating the workshops. Additionally, Kirsten prepared an educational public display for the Valley Life Sciences Building, and assisted Cynthia Perrine in public education through the Jepson Herbarium. She also completed her PhD dissertation, and is working to submit three manuscripts derived from dissertation. After graduation, she continued working with CIPRES in a temporary staff position, and then obtained a postdoctoral fellowships at NESCent. She has just been selected for a tenure-track faculty position at California State University, Los Angeles.

Graduate student Anna Larsen assisted Cynthia in a number of of field workshops, and after she completed her PhD in 2007, she succeeded Cynthia as Coordinator of Public Programs in the Jepson Herbarium, and has designed the last season of Tree of Life workshops.

Graduate students Rebecca Shapley and Denise Green worked together on methods for visualizing phylogenies and presenting them on the web to various audiences (ranging from the general public, though students and teachers, to professional researchers). Rebecca's online report on her work is at:

http://www.sims.berkeley.edu/~rebecca/cipres/index.htm. Rebecca and graduate student Denise Green interviewed representatives of these audiences in Spring 2005 to determine what features they most want from such websites, and what formats and displays they find most useful. They maintain an updated progress report at: http://groups.sims.berkeley.edu/TOL/index.htm, and completed their final written report for their Masters Degrees from the School for Information Management and Systems (SIMS), which they received in May 2005. We are proud that their project was awarded the 2005 "James R. Chen Award in Understanding People Using Technology" at the SIMS graduation ceremony. Both plan to go on to information technology careers in the private sector; Rebecca has obtained a very nice position at Google, where she is still involved in evaluating ways to share phylogenetic and biodiversity data on the web.

Graduate student Ruth Kirkpatrick worked in Mishler's lab, and led the design of the module on the early evolution of Cacti.

Cynthia Perrine is Coordinator of Public Programs in the Jepson Herbarium, and is supported partially through the CIPRES grant. She participated with Kirsten with planning the above activities, and also prepared the schedule for our Weekend Workshop series, which includes the public workshops on reconstructing the tree of life.

Students, postdocs, and staff, not paid by CIPRES All of these individuals worked on the outreach component at AMNH.

- Adriana Aquino, Staff
- Daniel Aviv, Staff
- Joel Cracraft, Staff
- Louise Crowley, Student
- Mohammad Faiz, Staff
- Megan Harrison, Postdoc
- Jay Holmes, Staff
- Daniel Janies, OSU-Staff
- Maritza Macdonald, Staff
- Mordecai MacLow, Staff
- Mark Norell, Staff
- Susan Perkins, Staff
- Paola Predraza, Postdoc
- Lorenzo Prendini, Staff
- David Randle, Staff

- Zobar Ris, Staff
- Monique Scott, Staff
- Cate Starr, Staff
- Ellen Trimarco, Staff

7.2 Overview of Activities

The outreach activity is primarily focused at three institutions UC Berkeley (and the Jepson Herbarium), Yale University (and the Peabody Museum), and the American Museum of Natural History. This activity includes museum exhibits (at both the Jepson Herbarium and the Peabody Museum), web-based activity, teacher training (at the AMNH), courses for adults in the general public (at the Jepson Herbarium), and K-12 activity (at the Peabody Museum and at the AMNH). The major activities included the following:

Jepson Herbarium
- Web site created by the Jepson Herbarium at UC Berkeley, and which is now accessed through the CIPRES homepage
- Public outreach activities organized at several locations by members of the Jepson Herbarium
- Workshops at the Jepson Herbarium for the educated public

Yale University
- The "Travels in the Great Tree of Life" museum exhibit, at the Peabody Museum of Natural History

American Museum of Natural History
- The "Tree of Life Institutes" to educate high school teachers and students in the New York metropolitan area

8 Databases

8.1 Personnel

Focus leader: Val Tannen

Senior Personnel

- Val Tannen, University of Pennsylvania.
- William H. Piel, Yale University.
- Susan Davidson, University of Pennsylvania.
- Mark Miller, San Diego Supercomputer Center.
- Michael Donoghue, Yale University.
- Brent Mishler, UC Berkeley.
- Dan Miranker, UT Austin.

Research Programmers, Students and Postdocs

- Jin Ruan, Senior Software Developer, Database Lead, San Diego Supercomputer Center.
- Mark J. Dominus, Senior Software Developer, University of Pennsylvania. Funded since 2006.
- Madhu Madhusudan, Senior Software Developer, San Diego Supercomputer Center. Funded since 2007.
- Lucie Chan, Senior Software Developer, San Diego Supercomputer Center. Funded until 2006.
- Shirley Cohen, UT Austin, then University of Pennsylvania. Database coordinator, then PhD student. Funded 2004-2006.
- Yifeng Zheng, PhD student, University of Pennsylvania.
  Funded 2004-2005.

8.2 Overall goals

The major objective of the Database Focus of the CIPRES project is the development of TreeBASE II. In addition, this focus has supported related research having to do with the storage and querying of the large phylogenetic trees constructed in the Simulation Focus of the project and with the provenance data needed used by workflow frameworks in phyloinformatics.

TreeBASE II is being developed as a robust, scalable, and versatile re-design and re-engineering of TreeBASE (which we shall call TreeBASE I in this report), a 10+ years old data resource whose capabilities are being overtaken by demands. As such, TreeBASE II will become a major resource for biological and biomedical research.

8.2.1 TreeBASE I

The advent of personal computers and PCR techniques have led to a proliferation of phylogenetic knowledge. By 1989, publications of new phylogenies were growing at a rate of 15 to 20% per year while showing no indication of slowing down. In addition, applications of phylogenies were extending beyond the normal confines of systematics into many diverse areas of science, including health and medical sciences, biotechnology, agriculture, fisheries, forestry, conservation, land and water management, ecotourism, and basic biological research. In response to this growth, TreeBASE, a curated database of phylogenetic data, was established in 1994. Designed to serve as a public repository of trees and character data, database submissions could include a broad range of phylogenetic studies – organismal, comparative, coevolution, and supertree analyses – which in turn could be based on a broad set of data types, including molecular, morphological, and paleontological. The only restriction is that submissions needed to be published in a peer-reviewed scientific journal before appearing on TreeBASE. As of 2004, TreeBASE has been actively searched by computers with over 60,000 unique IP numbers and has accepted over 1,300 submissions that map to over 3,700 trees and 60,000 distinct taxon strings.

The original impetus for the design of TreeBASE I was to meet the needs of researchers from both traditional systematics and molecular biology backgrounds who are concentrating on a series of focused experiments in the lab. Users in this category include those who periodically seek online representations of individual phylogenies for research and educational purposes.

A key component of creating a public archive of phylogenetic data is the efficient capture and curation of this data. Data processing consists of data deposition, annotation, and validation. The data collected from depositors consist of taxon labels, character and state labels, alignments, trees, tree inference methods, and citations. Submitters describe and submit all data in Nexus format with the exception of the citation, entered using free text fields. The data are decomposed into relations by the backend DBMS. Once processed, the data are searchable and accessible from the web site.

TreeBASE can be searched in seven ways: (1) by taxon: search is based on the taxonomic name; those names can be of taxa at the leaves of the phylogeny or of internal nodes, (2) by author: search is based on the last name of authors of phylogenetic studies, (3) by citation: search is based on

words that appear in the full reference, such as the title or journal name, (4) by study accession number: search is based on a unique code that is assigned to a phylogenetic study, (5) by matrix accession number: search is based on a unique code that is assigned to a particular matrix, (6) by structure: search is based on the topology and names of some taxa; the query retrieves all trees in which either all or parts of these trees match a *pattern* for both the taxa and the pattern of relationships among them (wildcards can be used here as part of the query tree: '*' denotes zero or more branches, and '?' denotes zero or one branches), or (7) once a tree is retrieved, "surf" the "neighboring trees," i.e., trees that differ by 1–3 "degrees of separation" from the initially retrieved tree.

8.2.2 New categories of users and new aims for TreeBASE II

We identify at least three new audiences in addition to the first one mentioned in 8.2.1.

In addition to the audience mentioned above, a second audience consists of researchers that want to run meta-analyses on large collections of trees. For example, they could be trying to identify patterns in trees that result from one type of analysis over another; or perhaps visualizing large collections of trees; or studying collaborative networks among phylogeneticists.

A third audience is phyloinformaticians who seek to make large-scale inference using synthetic methods applied to large collections of trees. For example, they may want to assemble a supertree for a large branch of the Tree of Life; or they might want to mine data in search of conflicting phylogenetic signal; they might want to examine the evolution of genes and genomes in a comparative context.

A fourth audience is bioinformaticians who conduct simulation studies. Frequently, simulation studies use simple models, such as the Kimura 2-Parameter and Jukes-Cantor that are not believed to be biologically realistic. Finding realistic evolutionary models, using real data, and carrying out simulation studies are some of the main goals of this group.

TreeBASE's aim is to deliver useful information to these four audiences, beyond simply publishing the submitted phylogenetic trees, characters, and matrices. TreeBASE II will add value to the submissions in the following ways:

- Intuitive access to data through a user-friendly web interface, including a new tree viewer (largely completed as of August 2008).
- A robust service layer and LSIDs to allow external tools and services to interface with the database (partially completed as of August 2008).
- Taxonomic intelligence for leaf and node labels (completed as of August 2008).
- Storage of LSIDs and foreign handles to better integrate with external data services, such as gene names, taxon names, and museum specimen IDs. (This has been completed for taxon names and partially completed for museum specimen IDs and genenames.)

- Ability to store geographic coordinates to support phylogeographic data visualization and analysis (partially completed as of August 2008).

8.2.3 TreeBASE II

As part of the CIPRES project, TreeBASE I was redesigned for the querying and distribution of structure data. The re-engineered system, TreeBASE II, expands the functionality of the existing system by providing improved submission, query, and curation tools. We describe briefly TreeBASE II capabilities:

Submission Process In the new system, depositions will be made easier, and annotations will be more automated. Users will be prompted for the same type of information as before, but they will be "helped" along the way. For example, the system will try to figure out the Genbank accession number for a given gene sequence by communicating with external data sources. It will also provide better error checking by, for example, matching taxon labels in trees with those in character matrices that the trees are paired with. All assistance features will be opt-in oriented and can be turned off by the user.

The steps involved in the submission process are as follows: After authenticating from the TreeBASE II web site, a preview page is presented to the user with an outline of the steps that follow. The information about citations is entered and loaded into the database in the next step. A citation entry includes the author(s), title, year, and page numbers of a publication. It also includes a publication type (book or article), an abstract if available, the ISBN/ISSN, DOI, and Pubmed id. The user can choose to proceed to the next step even if the citation information is incomplete.

The next step is to enter the character matrix information for the current study. Users can upload new character blocks. The character types can be morphological (discrete or continuous) molecular, or distance. The characters can be manually annotated with additional information such as gene regions, specimen ids, taxon labels, and images (although the system will do its best to detect the Genbank accession number based on the gene sequence). Once the user has reviewed the character information, the system checks it for consistency to catch any syntax and formating errors. The user can choose to proceed to the next step even if the character information is incomplete.

After entering the character information, the user proceeds to submit the trees for the current study. Trees can either be uploaded individually or in batches. Additionally, the user has the option to edit the internal structure of a tree through a friendly graphical interface, ATV (http://www.genetics.wustl.edu/eddy/atv). For example, he may choose to re-root the tree or relabel the leafs of the tree. It is important to note that every leaf of the tree needs to be accounted for by the character matrix from the current step. In practice, this means that the leaf and taxon labels need to match exactly. The system will detect and notify the user of any mismatches. The user can choose to proceed to the next step even if the tree information is incomplete.

The last step in the submission process is to enter the analysis information for the current study. This step is equivalent to the methods section of a scientific publication. In the TreeBASE II

model, a study is divided into one or more analysis steps where each analysis step constitutes a phylogenetic algorithm implemented by a piece of software such as PAUP*, and a set of inputs and outputs to the algorithm. Both the algorithm and the piece of software are entered by the user and the inputs and outputs to the algorithm are selected from a list of available inputs and outputs. These inputs and outputs were defined in previous steps assuming that the user has entered the complete character and tree information. Each step allows the storage of inference software command strings such as the contents of a PAUP block or MrBayes block.

Curation Study submission to TreeBASE will be normally done in conjunction with submitting a journal publication.

A reviewer or journal editor can view non-published study data in TreeBASE. After the submitted study is accepted/published in a journal, the submitter requests the study to be published to the TreeBASE. After a study has been flagged to request publication, the TreeBASE editor can publish the study thus making it visible to search/query users.

Since an article reviewer is anonymous, s/he cannot be identified or authenticated. TreeBASE II provides a process to authorize a reviewer to access a non-published study without the authentication.

In TreeBASE II, an editor can edit study related information, for example, to correct author, citation, or other data.

An editor can edit the study data. For example, s/he can correct the taxon names to match the tree node to the data matrix row, or can remove orphan data. We will provide an interface with access to uBio (`www.ubio.org`) for taxonomic services in order to facilitate this curation capability.

Search, Querying, and Browsing. TreeBASE II will offer both a search/query capability for the interactive browser-based user and a bulk query API for interoperation with various analysis software packages, as well as special CORBA-based interface for the integration tool CIPRES developed in this project.

Interactive User Search GUI. The GUI will be configurable to retrieve sets of studies, sets of matrices, or sets of trees. Different search criteria will be available in each of these subcases, and the GUI will allow further configuration to use a subset of these criteria for a search:

Study Search By:

- Author(s) last names (implemented as of August 2008)
- Citation title/abstract matches given keyword(s) (implemented as of August 2008)
- Contains analysis/analysis step such that:

  - Name matches given keyword(s)
  - Uses given algorithm (under implementation as of August 2008)
  - Uses given software package (under implementation as of August 2008)
  - Input and/or output data contains given set of taxa (implemented as of August 2008)
  - Input and/or output data contains tree that matches given tree pattern (implemented as of August 2008)
  - Input and/or output data contains matrices satisfying given search criteria (same as below)

Tree Search By:

- Tree id number
- Appears in a study satisfying given search criteria (same as above)
- Appears in an analysis/analysis step satisfying given search criteria (same as above)
- Contains given set of taxa (implemented as of August 2008)
- Matches given tree pattern (implemented as of August 2008)

Matrix Search By:

- Uses given set of taxa (implemented as of August 2008)
- Uses given set of character names (under implementation as of August 2008)
- Is a sequence matrix that uses a certain kind of biomolecular information
- Contains given specimen(s)

We will also add the ability for an interactive user to submit a bulk query and visualize/browse the result by feeding from the GUI into the API (see below). This will be used by sophisticated users who need to run complex queries (e.g., aggregate queries) whose answers are nonetheless relatively small.

Interactive User Browsing GUI (implemented as of August 2008). Here there are three subcases corresponding to the kind of information retrieved:

1. Visualize and browse the retrieved set of studies. This will be essentially a list of links, and the user can open each one to reveal substructure.
2. Visualize and browse retrieved set of matrices. This will be a list of links, and the user can open each and visualize matrix.

3. Visualize and browse the retrieved set of trees. This will be a list of links, and the user can open each with third-party tree visualization software. Here we will have connections to third-party software used for surfing of neighboring trees, and for constructing supertrees.

In each of these cases, the user can download search results on user's machine, through the browser, a file, or set of files containing the set of studies, set of matrices, or set of trees obtained from the search. The user can choose Nexus or nexml (see below) file formats.

Bulk Query API (under implementation as of August 2008). XML-based query interface used primarily by other tools that interoperate with TreeBASE II. The input to this API will use a query "language", closely based on the TreeBASE Domain Model and corresponding semantically to a simple subset of standard query languages such as SQL or ODMG/OQL (`www.odmg.org`). For the queries we will use XML syntax both for simplicity as well as to allow easy conversion into a Web service.

We will use the XML format nexml (`www.nexml.org`) for query output. Query results are returned in this XML format that covers Nexus data as well as the additional data to be found in TreeBASE II.

9 Contributions

CIPRES is an interdisciplinary research project with funding from CISE and AToL; hence, we describe our contributions to both of these scientific communities.

Major contributions:

- The CIPRES software distribution, which includes several novel algorithms for large-scale phylogenetic reconstruction (e.g., Rec-I-DCM3 versions of RAxML and PAUP*, GARLI, RAxML with bootstrapping, and Phycas) that enables phylogenies on very large datasets to be estimated with a higher level of accuracy than previously, and in reasonable time periods (see http://www.phylo.org/sub_sections/software.html/),
- The CIPRES portal, which allows systematists to obtain these analyses without having to download and install software (and which is available through the CIPRES Science Gateway of the Teragrid (see http://www.phylo.org/sub_sections/portal/),
- The open-source and freely available CIPRES libraries, which enables programmers to develop their own software (see http://www.phylo.org/sub_sections/software),
- The CIPRES compute cluster, which we made freely available,
- New software for large-scale phylogenetic estimation, including some methods that are available through the CIPRES software distribution, but also SATé (101), SuperFine (187), DACTAL (147), Mega-phylogeny (175), and other methods,
- TreeBASE-II (206), a greatly improved version of TreeBASE, enabling biologically important querying, and greater ease of use (see http://www.treebase.org/treebase-web/home.html),
- The million taxon, million site simulation of the RNASim (63) group, enabling the research community to do careful and extensive testing of phylogeny estimation methods (see http://kim.bio.upenn.edu/software/csd.shtml),
- New mathematical theory related to phylogenetic estimation, including results regarding sequence length requirements of different methods under Markov models of evolution, robustness to model violations, estimations of trees from gene order data, estimations of supertrees, estimations of reticulate evolution, and a host of other questions,
- Outreach to the lay public, through our museum partners (Yale Peabody, the Jepson Herbarium, and the American Museum of Natural History), and
- Human resource development, with 16 postdoctoral researchers and 73 graduate students participating in training activities. Most of the students were from the mathematics, computer science, or statistics disciplines, and it is clear that this project directly brought new research questions of interest to computer science, mathematics, and probability theory, greatly enriching the field of computational and mathematical phylogenetics.

Human Resource Development The CIPRES project involved a large number of students and postdocs, some with financial support, We had 16 postdoctoral fellows (6 funded by CIPRES, and 10 funded by other sources), 73 graduate students (41 funded by CIPRES and 31 by other sources), and several undergraduates (none of whom were funded by CIPRES). Our students and postdocs have done remarkably well. For example:

- Of the 39 PhD students funded by the project:
  - 32 finished their doctorates, 5 are making progress towards finishing their dissertations, and 2 left the program without finishing.
  - Of the 32 who finished their PhDs, 10 are in postdoctoral positions, 8 are tenure-track faculty (2 of these are already tenured), and 11 are in industry.
  - Five of our PhD graduates have won dissertation awards: Costis Daskalakis won the 2008 ACM Doctoral Dissertation Award, Radu Mihaescu won the Bernard Friedman Memorial Prize for outstanding thesis in applied mathematics at UC Berkeley, Luay Nakhleh won the Best Dissertation Award in Science and Engineering at the University of Texas, Sebastien Roch won the Erich Lehmann Award for Outstanding Dissertation in Theoretical Statistics at UC Berkeley, and Shel Swenson won a best dissertation award in the Mathematics Department of the University of Texas at Austin.
  - Four of our PhD graduates (Ganapathy, Zwickl, Fisher, and Smith) have had NESCENT (The National Evolutionary Synthesis Center) postdoctoral fellowships.
  - Two of our PhD graduates have received Sloan Foundation fellowships (Daskalakis and Nakhleh).
- Of our 6 postdoctoral fellows funded by the project:
  - Three (Alfaro, Holder, and Snir) are now tenure track faculty members
  - Two (Swenson and Vos) are in second postdoctoral positions
  - One (Midford) works for NESCENT

Other participants began as Assistant Professors and are now tenured (e.g. Linder, Meyers, Mossel, and Turner), or began as Associate Professors and are now Full Professors (Warnow). It is very clear that CIPRES has had a very good impact on the education and careers of our participants (including our senior faculty, for that matter).

Below, we provide information below about the current positions for the funded participants (6 postdocs and 41 graduate students), to give an indication of the impact of the project on their careers.

CIPRES-funded Masters students

1. Denise Green, UC Berkeley (worked with Brent Mishler). Denise finished her M.S. degree in 2005, in the School for Information Management and Systems (SIMS) at UC Berkeley (dissertation title: "Teaching with a Visual Tree of Life" (56).

2. Rebecca Shapley, UC Berkeley (worked with Brent Mishler). Rebecca completed her M.S. degree in the School for Information Management and Systems (SIMS) in 2005 at UC Berkeley ("Teaching with a Visual Tree of Life" (172)), and received the 2005 "James R. Chen Award in Understanding People Using Technology" at the SIMS graduation ceremony. Rebecca works at Google, where she is involved in evaluating ways to share phylogenetic and biodiversity data on the web.

CIPRES-funded PhD students

1. François Barbançon, UT Austin (student of Dan Miranker).
   PhD in Computer Science. Dissertation title: "Active learning and compilation of higher order schema integration queries" (13). François is now on the staff at Microsoft.

2. Nicholas Bray, UC Berkeley (student of Lior Pachter). Nick is still a Mathematics PhD student at Berkeley, working on population genetics.

3. Kevin Chen, UC Berkeley (student of Lior Pachter and Satish Rao). PhD Mathematics, January 2005. Dissertation title: "Three variations on the theme of comparative genomics: metagenomics, mitochondrial gene rearrangements and micrornas" (25). Chen had a post-doc funded by NIH under his own research grant, and began a tenure track appointment in Fall 2009 at Rutgers University in the Department of Genetics.

4. Shirley Cohen, UPenn (PhD student in Computer and Information Sciences of Val Tannen and Susan Davidson). Shirley left the program without finishing her degree.

5. Costis Daskalakis. PhD Computer Science, UC Berkeley (student of Satish Rao). Costis finished his dissertation in 2008, "The Complexity of Nash Equilibria" (31), and is now an Assistant Professor of Computer Science at MIT. He was awarded a 2010 Sloan Foundation fellowship, NSF Career Award, 2008 ACM Doctoral Dissertation Award, the 2008 Game Theory and Computer Science Prize, and a 2007 Microsoft Research Fellowship.

6. Nick Eriksson, PhD Mathematics, UC Berkeley (student of Bernd Sturmfels). PhD. May 2006. "Algebraic combinatorics for computational biology" (44). Nick took an NSF Postdoctoral Fellowship in the Statistics Department at the University of Chicago, and is now a statistical geneticist in the genetics biotech company 23andMe.

7. Yu Fan, PhD Ecology and Evolutionary Biology, U. Conn (student of Paul Lewis). Yu Fan is still a student, and should graduate in 2011.

8. Kirsten Fisher, PhD Integrative Biology, UC Berkeley (student of Brent Mishler). Kirsten finished her PhD in 2004, with the dissertation "Systematics and evolution of the moss family Calymperaceae" (48). Kirsten was a postdoctoral fellow at NESCENT, and is now an Assistant Professor at California State University, Los Angeles.

9. Ganesh Ganapathy, PhD Computer Sciences, University of Texas at Austin (student of Tandy Warnow and Vijaya Ramachandran). Dissertation title: "Algorithms and heuristics for combinatorial optimization in phylogeny" (50). Ganesh was a postdoctoral fellow at NESCent, and is now a postdoctoral fellow of Erich Jarvis at Duke University.

10. Sheng Guo, PhD, Department of Biology, Univ. Penn (student of Junhyong Kim). Sheng received his PhD in 2008 (dissertation title "Molecular evolution of Drosophila odorant receptors" (62)) and went onto Boehringer Ingelheim as a staff scientist.

11. Tracy Heath, PhD, Program in Ecology, Evolution, and Behavior, University of Texas at Austin (student of David Hillis). Tracy's work for the Simulations and Modelling component of the CIPRES project is the subject of her dissertation, "Understanding the Importance of Taxonomic Sampling for Large-scale Phylogenetic Analyses by Simulating Evolutionary Processes under Complex Models" (65), for which she received a PhD. Tracy is now a postdoctoral fellow at the University of Kansas with Mark Holder.

12. Cameron Hill, PhD Mathematics, UC Berkeley, dissertation title "Geometric model theory in efficient computability" (72) 2010. Cameron is now a postdoctoral fellow at Notre Dame University.

13. David Kysela, PhD, Ecology and Evolutionary Biology, Yale University (student of Paul Turner). Kysela completed his degree in 2008 (91), with dissertation titled "Bacteriophage response to the dynamic host environment: resistance, aging, and quorum sensing." He is now a postdoc at Indiana University.

14. Ruth Kirkpatrick, PhD, Integrative Biology, University of California at Berkeley (student of Brent Mishler). Ruth finished her PhD in 2007, title: "Systematics and evolution of the fern genus Pellaea" (87). Ruth is now an instructor at Santa Rosa Junior College.

15. Henry Lin, PhD, Computer Science, University of California at Berkeley, (student of Satish Rao and Christos Papadimitriou). Henry received his PhD in 2009, title "Internet Routing and Internet Service Provision" (96). Henry had a postdoctoral fellowship at the Institute for Theoretical Computer Science at Tsinghua University for 2009-2010, and is now a postdoctoral researcher in the Center for Bioinformatics and Computational Biology at the University of Maryland.

16. Kevin Liu, PhD, Computer Science, University of Texas at Austin (student of Tandy Warnow and Randy Linder). Kevin is still a student, working on simultaneous estimation of alignments and trees.

17. Wenguo Liu, Computer Science, UT Austin (student of Dan Miranker). Wenguo left the doctoral program without completing his degree.

18. Andrew McGregor, PhD, Computer and Information Sciences, Univ. of Pennsylvania (student of Sampath Kannan). Andrew finished his Ph.D. in 2007 (Dissertation Title: "Processing Data Streams" (111)). Andrew was a postdoc at the Information Theory Institute in San Diego, and then a postdoc at Microsoft Research, Silicon Valley. He is now a tenure-track Assistant Professor at the University of Massachusetts.

19. Frank Mannino, PhD, Statistics Department, NCSU (student of Spencer Muse). Frank Mannino completed his PhD in 2006 in the PhD program in Bioinformatics in the Department of Statistics (title: "Site-to-site variation in protein coding genes" (105)). He is now working as a bioinformatician for Glaxo Smith Kline in Philadelphia.

20. Rui Mao, PhD, Computer Sciences, University of Texas at Austin (student of Dan Miranker). Dissertation "Distance-Based Indexing and Its Applications in Bioinformatics" (107). He is now employed at Oracle.

21. Radu Mihaescu, PhD, Mathematics, UC Berkeley (student of Lior Pachter and Satish Rao). Dissertation awarded 2008, title: "Distance Methods for Phylogeny Reconstruction" (113), (winner of the Bernard Friedman Memorial Prize for an outstanding thesis in applied mathematics). Radu was a postdoc in the Department of Mathematics, University of California, Berkeley, and is now Assistant Vice President, Knight Capital Group.

22. Eric Miller, PhD, Ecology, Evolution, and Behavior program, University of Texas at Austin (student of Lauren Meyers). Eric is still a student.

23. Luay Nakhleh, PhD Computer Science, University of Texas at Austin (student of Tandy Warnow) Dissertation title: "Phylogenetic Networks"(135), awarded the Best Dissertation Award in Science and Engineering at the University of Texas. Luay is now an Associate Professor of Computer Science (with tenure) at Rice University. Luay received a Sloan Foundation fellowship, an NSF CAREER award, and a DOE CAREER award.

24. Manikandan Narayanan, PhD Computer Science, UC Berkeley (student of Dick Karp). PhD. Fall 2007. "Comparative and Evolutionary Analysis of Cellular Pathways" (145). Mani is now a Sr. Research Scientist at Merck Research Labs in Boston.

25. Serita Nelesen, PhD, Computer Sciences, University of Texas at Austin (student of Tandy Warnow and Warren Hunt). PhD Summer 2009. "Improved methods for phylogenetics" (147). Serita is now a tenure-track professor of Computer Science at Calvin College.

26. Smriti Ramakrishnan, PhD, Computer Science, University of Texas at Austin (student of Dan Miranker). Smriti finished her PhD in 2010; her dissertation title was "A Systems Approach to Computational Protein Identification" (156). She now has a position at Oracle.

27. Samantha Riesenfeld, PhD, Department of Computer Science, University of California at Berkeley (student of Dick Karp). PhD, Summer 2007. "Optimization and Reconstruction over Graphs" (159). Sam is a postdoctoral fellow working with Katie Pollard, of the UC Davis Genome Institute and Department of Statistics.

28. Sébastien Roch, Department of Statistics, UC Berkeley (student of Elchanan Mossel). PhD. Summer 2007. Dissertation title:"Markov Models on Trees: Reconstruction and Applications" (161). He received the Erich Lehmann Award for Outstanding Dissertation in Theoretical Statistics at UC Berkeley for his dissertation. Sébastien was a postdoctoral fellow at Microsoft Research in Cambridge, and is now an assistant professor of mathematics at UCLA.

29. Usman Roshan, Department of Computer Sciences, University of Texas at Austin (student of Tandy Warnow). Usman finished his dissertation, "Algorithmic techniques for improving the speed and accuracy of phylogenetic methods" (162), in 2004, and is now an Associate Professor (with tenure) at the New Jersey Institute of Technology.

30. Ariel Schwartz, PhD, Computer Science Department, UC Berkeley. Ariel finished his PhD in 2007. Dissertation title: "Posterior Decoding Methods for Optimization and Accuracy Control of Multiple Alignments" (168). Ariel took a postdoctoral position with Trey Idekker at UCSD, and is now Bioinformatics Scientist at Synthetic Genomics (since May 2008).

31. Stephen Smith, Department of Ecology and Evolutionary Biology, Yale University (student of Michael Donoghue). Stephen finished his PhD in 2008. Dissertation title: "Evolving biogeography: New methods and their application in the plant clade Lonicera" (174). Stephen Smith was a postdoc at NESCent, and is now a postdoctoral researcher for iPlant.

32. Errol Strain, Bioinformatics program, Department of Statistics, NCSU (student of Spencer Muse). PhD awarded 2006, title: "Plant Molecular Evolution" (183).

33. Jeet Sukumaran, Department of Ecology and Evolution, University of Kansas (student of Mark Holder). Jeet is still a student.

34. Shel Swenson, Department of Mathematics, University of Texas at Austin (student of Tandy Warnow and Randy Linder). PhD awarded 2009, title: "Supertree methods" (187). Shel received a best dissertation award from the Mathematics Department at UT-Austin. She was a postdoc for Warnow (partially funded by CIPRES) for 2009-2010 and is now leaving to take a postdoc at Georgia Tech, Mathematics.

35. Kunal Talwar, Department of Computer Science, UC Berkeley (student of Christos Papadimitriou and Satish Rao). PhD. 2005. Dissertation title: "Metric Methods in Approximation Algorithms" (194). Kunal is currently a Research Scientist at Microsoft Research.

36. Andres Varón, CUNY (student of Ward Wheeler). Dissertation title: "Algorithms and hypothesis selection in dynamic homology phylogenetic analysis" (202), awarded 2010.

37. Rutger Vos, Department of Biological Sciences, Simon Fraser University (student of Wayne Maddison). Rutger completed his PhD in 2006 at Simon Fraser University, for his dissertation "Inferring large phylogenies: the big tree problem" (205). Rutger was a postdoctoral fellow of Wayne Maddison, and currently has a postdoctoral position at the University of Reading with Mark Pagel.

38. Yifeng Zheng, Department of Computer and Information Sciences, Univ of Pennsylvania (student of Susan Davidson and Junhyong Kim). Yifeng completed his PhD in 2008 (Dissertation title: "Efficient Scientific Data Management Over Trees" (223)). Yifeng is now working for Google.

39. Derrick Zwickl, Program in Evolution, Ecology, and Behavior, University of Texas at Austin (student of David Hillis). Derrick received his PhD in 2006 for his dissertation "Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion" (225). Derrick had a postdoctoral position at NESCent, and is now a postdoctoral fellow of Mark Holder at the University of Kansas.

CIPRES funded postdoctoral fellows

1. Michael Alfaro, UCSD (postdoctoral fellow of John Huelsenbeck). Michael is a tenure track Assistant Professor in the Department of Ecology and Evolution at UCLA.

2. Mark Holder (postdoctoral fellow of Dave Swofford while at FSU). Mark was a postdoc at UConn with Paul Lewis, then at Florida State University with David Swofford, and is now a tenure-track Assistant Professor at the University of Kansas.

3. Peter Midford, UBC (postdoc of Wayne Maddison). Peter Midford currently works for NESCent.

4. Sagi Snir, UC Berkeley (postdoc of Lior Pachter). Sagi is currently an Assistant Professor in the school of Computer Science and Mathematics at Netanya Academic College, and a senior fellow at the Institute of Evolution at Haifa University

5. Shel Swenson (postdoc of Tandy Warnow). Shel finished her postdoctoral position with Tandy Warnow, working on supertree estimation methods and mentoring students from Huston-Tillotson (a historically black college in Austin). She will begin her second postdoctoral position at Georgia Tech, in the Mathematics Department, in January 2011.

6. Rutger Vos, Simon Fraser U. (research fellow with Wayne Maddison). Rutger was a postdoctoral fellow of Wayne Maddison, and currently has a postdoctoral position at the University of Reading with Mark Pagel.

10 CIPRES publications

References

[1] M. E. Alfaro and M.T. Holder. The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology and Systematics*, 37:19–42, 2006.

[2] M. E. Alfaro and J. P. Huelsenbeck. Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst. Biol.*, 55:89–96, 2006.

[3] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20:407–415, 2004.

[4] S. Aluru, N. Amato, D.A. Bader, S. Bhandarkar, L. Kale, and D. Marinescu. Parallel computational biology. In M.H. Heroux, P. Raghavan, and H.D. Simon, editors, *Frontiers of Scientific Computing*. SIAM Press, 2005.

[5] S. Angelov. *Pattern Discovery in Biological Datasets*. PhD dissertation, University of Pennsylvania, 2007.

[6] S. Angelov, B. Harb, S. Kannan, S. Khanna, and J. Kim. Efficient enumeration of phylogenetically informative substrings. In *10th Annual Intl. Conf. Research in Computational Molecular Biology (RECOMB)*, pages 248–264, 2006.

[7] S. Angelov, B. Harb, S. Kannan, S. Khanna, J. Kim, and L.-S. Wang. Genome identification and classification by short oligo arrays. In *Proc. 4th Workshop Algorithms in Bioinformatics (WABI'04)*, volume 3240 of *Lecture Notes in Computer Science*, pages 400–411. Springer-Verlag, 2004.

[8] A. Bachrach, K. Chen, C. Harrelson, R. Mihaescu, S. Rao, and A. Shah. Lower bounds for maximum parsimony with gene order data. In *Proc. 3rd RECOMB Workshop on Comparative Genomics*, Lecture Notes in Computer Science. Springer-Verlag, 2005.

[9] D.A. Bader. Computational biology and high-performance computing. *Communications of the ACM*, 47(11):35–40, 2004.

[10] D.A. Bader, V. Chandu, and M. Yan. ExactMP: An efficient parallel exact solver for phylogenetic tree reconstruction using maximum parsimony. In *35th International Conference on Parallel Processing (ICPP)*, pages 65–73. Columbus, OH, 2006.

[11] D.A. Bader, U. Roshan, and A. Stamatakis. Computational grand challenges in assembling the tree of life: Problems and solutions. In Chau-Wen Tzeng, editor, *Advances in Computers*, volume 68, pages 128–179, 2006.

[12] D.A. Bader and M. Yan. High performance algorithms for phylogeny reconstruction with maximum parsimony. In S. Aluru, editor, *Handbook of Computational Molecular Biology*, chapter 22, pages 1–19. Chapman & Hall / CRC Computer and Information Science Series, 2006.

[13] F. Barbançon. *Active Learning and Compilation of Higher Order Schema Integration Queries*. PhD dissertation, University of Texas, 2005.

[14] T. Batu, S. Kannan, S. Khanna, and A. McGregor. Reconstructing strings from random traces. In *Proc. Symp. Discrete Algorithms, SODA '04*, pages 910–918, 2004.

[15] N. Beerenwinkel, N. Eriksson, and B. Sturmfels. Evolution on distributive lattices. *Journal of Theoretical Biology*, 242(2):409–420, 2006.

[16] T.Y. Berger-Wolf. Online consensus of phylogenetic trees. In *Proc. 4th Workshop Algorithms in Bioinformatics (WABI'04)*, volume 3240 of *Lecture Notes in Computer Science*, pages 350–361. Springer-Verlag, 2004.

[17] S. Cohen Boulakia and V. Tannen, editors. *Data Integration in the Life Sciences*. Springer, 2007. LNCS 4544 (LNBI).

[18] S. Bowers, T. McPhillips, B. Ludascher, S. Cohen, and S. B. Davidson. A model for user-oriented data provenance in pipelined scientific workflows. In *Proceedings of IPAW'06 International Provenance and Annotation Workshop*, 2006.

[19] R.S. Boyer, W.A. Hunt Jr., and S.M. Nelesen. A compressed format for collections of phylogenetic trees and improved consensus performance. In *Proc. 5th Workshop Algorithms in Bioinformatics (WABI'05)*, Lecture Notes in Computer Science, pages 353–364. Springer-Verlag, 2005.

[20] J.E. Bradner, N. West, M.L. Grachan, E. Greenberg, S.J. Haggarty, T. Warnow, and R. Mazitschek. Chemical phylogenetics of histone deacetylases. *Nature Chemical Biology*, 6:238–243, 2010.

[21] N. Bray and L. Pachter. MAVID: constrained ancestral alignment of multiple sequences. *Genome Research*, 14:693–699, 2004.

[22] B.W.Chen, W. H. Piel, L. Gui, E. Bruford, and A. Monteiro. The hsp90 family of genes in the human genome: insights into their divergence and evolution. *Genomics*, 86:627–637, 2005.

[23] K. Chaudhuri, K. Chen, R. Mihaescu, and S. Rao. On the tandem duplication-random loss model of genome rearrangement. In *Proceedings 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2006.

[24] K. Chen and L. Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol.*, 1(2):e24, 2005.

[25] K. C. Chen. *Three variations on the theme of comparative genomics: Metagenomics, mitochondrial gene rearrangements and microRNAs.* PhD dissertation, EECS Department, University of California, Berkeley, 2005.

[26] B. Chor, A. Khetan, and S. Snir. Maximum likelihood on molecular clock comb: Analytic solutions. *Journal of Computational Biology (JCB)*, 13(3):819–837, 2006.

[27] B. Chor and S. Snir. Analytical solutions of maximum likelihood on forks of four taxa. *Mathematical Biosciences*, 208(2):347–358, 2007.

[28] C. Coarfa, Y. Dotsenko, J. Mellor-Crummey, L. Nakhleh, and U. Roshan. PRec-I-DCM3: A parallel framework for fast and accurate large scale phylogeny reconstruction. In *Proceedings of the First IEEE Workshop on High Performance Computing in Medicine and Biology (HiPCoMB2005)*, volume 2, pages 346–350, 2005. Best Paper award.

[29] S. Cohen, S. Cohen-Boulakia, and S. B. Davidson. Towards a model of provenance and user views in scientific workflows. In *Proceedings of DILS'06 Data Integration for the Life Sciences*, 2006.

[30] L. Cui, J.H. Leebens-Mack, L-S.Wang, J. Tang, L. Rymarquis, D.B. Stern, and C.W. dePamphilis. Adaptive evolution of chloroplast genome structure. *Biomedical Central Evolutionary Biology*, 6:13, 2006.

[31] C. Daskalakis. *The Complexity of Nash Equilibria.* PhD dissertation, Computer Science Division, UC-Berkeley, 2008.

[32] C. Daskalakis, C. Hill, A. Jaffe, R. Mihaescu, E. Mossel, and S. Rao. Maximal accurate forests from distance matrices. In Alberto Apostolico, Concettina Guerra, Sorin Istrail, Pavel A. Pevzner, and Michael S. Waterman, editors, *Research in Computational Molecular Biology, 10th Annual International Conference, RECOMB 2006, Venice, Italy, April 2-5, 2006, Proceedings (RECOMB 2006)*, volume 3909 of *Lecture Notes in Computer Science*, pages 281–295. Springer, 2006.

[33] C. Daskalakis, E. Mossel, and S. Roch. Optimal phylogenetic reconstruction. In *Proc. 38th Ann. ACM Symp. Theory of Comput. (STOC'06)*, pages 159–168, New York, 2006. ACM.

[34] C. Daskalakis, E. Mossel, and S. Roch. Phylogenies without branch bounds: Contracting the short, pruning the deep. *SIAM J. Discrete Mathematics*, 2010. To appear.

[35] S. Davidson, J. Kim, and Y. Zheng. Efficiently storing and extracting phylogenetic trees for simulation. In *Proceedings of Greater Philadelphia Bioinformatics Alliance Retreat*, 2004.

[36] S.B. Davidson, J. Kim, and Y. Zheng. Efficiently supporting structure queries on phylogenetic trees. In *Proc. 17th Scientific and Statistical Database Conf. SSDC'05*, 2005.

[37] A. Dornburg, F. Santini, and M.E. Alfaro. The influence of model averaging on clade posteriors: an example using the triggerfishes (family balistidae). *Syst. Biol.*, 57:905–919, 208.

[38] Y. Dotsenko, C. Coarfa, J. Mellor-Crummey, L. Nakhleh, and U. Roshan. PRec-I-DCM3: A parallel framework for fast and accurate large scale phylogeny reconstruction. *International Journal on Bioinformatics Research and Applications (IJBRA)*, 2(4):407–419, 2006.

[39] Z. Du, F. Lin, and U. Roshan. Reconstruction of large phylogenetic trees: a parallel approach. *Computational Biology and Chemistry*, 29(4):273–280, 2005.

[40] Z. Du, A. Stamatakis, F. Lin, U. Roshan, and L. Nakhleh. Parallel divide-and-conquer phylogeny reconstruction by maximum likelihood. In *Proceedings of the 2005 International Conference on High Performance Computing and Communications (HPCC 05)*, volume 2, pages 346–350, 2005.

[41] J.V. Earnest-DeYoung, E. Lerat, and B.M.E. Moret. Reversing gene erosion: reconstructing ancestral bacterial genomes from gene-content and gene-order data. In *Proc. 4th Workshop*

Algorithms in Bioinformatics (WABI'04), volume 3240 of *Lecture Notes in Computer Science*, pages 1–13. Springer-Verlag, 2004.

[42] R. Edgar and G. Myers. Identification and classification of repeated genomic elements. In *Proc. 13th Conf. on Intelligent Systems for Molecular Biology (ISMB'05)*, 2005.

[43] N. Eriksson. Tree construction using singular value decomposition. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 347–358. Cambridge University Press, 2005.

[44] N. Eriksson. *Algebraic combinatorics for comutational biology.* PhD dissertation, Mathematics Department, University of California at Berkeley, 2006.

[45] S.N. Evans and T. Warnow. Unidentifiable divergence times in rates-across-sites models. *IEEE Trans. Comput. Biol. and Bioinformatics*, 1:130–134, 2005.

[46] S.N. Evans, T. Warnow, and D. Ringe. Inference of divergence times as a statistical inverse problem. In P. Foster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 119–129. Cambridge University Press, 2004.

[47] Y. Fan, R. Wu, M.-H. Chen, L. Kuo, and P.O. Lewis. Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution*, page (advanced access), 2010.

[48] K. Fisher. *Systematics and evolution of the moss family Calymperaceae.* PhD dissertation, University of California at Berkeley, 2004.

[49] J.N. Foster, T.J. Green, and V. Tannen. Annotated xml: Queries and provenance. In *Proceedings PODS 2008*, 2008.

[50] G. Ganapathy. *Algorithms and heuristics for combinatorial optimization in phylogeny.* PhD dissertation, Computer Science Department, UT-Austin, 2006.

[51] G. Ganapathy, B. Goodson, R. Jansen, H. Le, V. Ramachandran, and T. Warnow. Pattern identification in biogeography. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4):334–346, 2006.

[52] G. Ganapathy, B. Goodson, R. Jansen, V. Ramachandran, and T. Warnow. Pattern identification in biogeography: metrics and algorithms for comparing area cladograms. In *Proc. 5th Workshop Algorithms in Bioinformatics (WABI'05)*, Lecture Notes in Computer Science, pages 116–127. Springer-Verlag, 2005.

[53] G. Ganapathy, V. Ramachandran, and T. Warnow. On contract-and-refine-transformations between phylogenetic trees. In *Proc. 15th ACM/SIAM Symp. Discrete Algs. (SODA'04)*, pages 893–902. SIAM Press, 2004.

[54] F. Ge, L.-S. Wang, and J. Kim. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *Public Library of Science Biology*, 3(10):e316, 2005.

[55] H. Glenner, A.J. Hansen, M.V. Sorensen, F. Ronquist, J.P. Huelsenbeck, and E. Willerslev. Bayesian inference of the metazoan phylogeny: A combined analysis. *Curr. Biol.*, 14:1644–1649, 2004.

[56] D. Green. *Teaching with a Visual Tree of Life.* Masters thesis, University of California at Berkeley, 2005. See also http://groups.sims.berkeley.edu/TOL/.

[57] T. J. Green, G. Karvounarakis, N. Taylor, O. Biton, Z. Ives, and Val Tannen. Orchestra: Facilitating collaborative data sharing. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2007.

[58] T.J. Green, G. Karvounarakis, Z. Ives, and V. Tannen. Update exchange with mappings and provenance. In *Proceedings VLDB 2007*, pages 675–686, 2007.

[59] T.J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *Proceedings PODS*, 2007.

[60] T.J. Green and V. Tannen. Models for incomplete and probabilistic information. *IEEE Data Engineering Bull*, 29:17–24, 2006.

[61] S. Guindon, A.G. Rodrigo, K.A. Dyer, and J.P. Huelsenbeck. Modeling the site-specific variation of selection patterns along lineages. *Proc. Nat'l Acad. Sci., USA*, 101(35):12957–12962, 2004.

[62] S. Guo. *Molecular evolution of Drosophila odorant receptors*. PhD dissertation, The University of Pennsylvania, 2008.

[63] S. Guo and J. Kim. Macroevolution simulation using a sequence-structure fitness model reveals statistical complexity of empirical data, 2008. Arxiv: http://arxiv.org/abs/0912.2326.

[64] B. Harb, S. Kannan, and A. McGregor. Approximating the best-fit tree under L_p norms. In *Proc. 8th Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'05)*, pages 123–133, 2005.

[65] T. Heath. *Understanding the Importance of Taxonomic Sampling for Large-scale Phylogenetic Analyses by Simulating Evolutionary Processes under Complex Models*. PhD dissertation, University of Texas at Austin, 2008.

[66] T.A. Heath, S. M. Hedtke, and D. M. Hillis. Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*, 46(3):239–257, 2008.

[67] T.A. Heath, D. J. Zwickl, J. Kim, and D. M. Hillis. Taxon sampling affects inferences of macro-evolutionary processes from phylogenetic trees. *Systematic Biology*, 57(1):160–166, 2008.

[68] M.D. Hendy and S. Snir. Hadamard conjugation for the Kimura-3ST model: Combinatorial proof using pathsets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, pages 461–471, July 2008.

[69] K.G. Herbert, N. H. Gehani, W. H. Piel, J. T. L. Wang, and C. H. Wu. BIO-AJAX: An extensible framework for biological data cleaning. *ACM SIGMOD Record, Special Section on Data Engineering for Life Sciences*, 33(2):51–57, 2004.

[70] K.G. Herbert, J. Spirollari, J. T. L. Wang, W. H. Piel, J. Westbrook, W. C. Barker, Z. Z. Hu, and C. H. Wu. Bioinformatic databases. In C. Craig, editor, *Encyclopedia of Computer Science and Engineering*. John Wiley and Sons, Ltd., 2007.

[71] K.G. Herbert, J. T. L. Wang, S. Pusapati, and W. H. Piel. Lineage path integration for phylogenetic resources. In *Proceedings of the 17th International Conference on Scientific and Statistical Database Management*, pages 117–120, 2005.

[72] C. Hill. *Geometric model theory in efficient computability.* PhD dissertation, The University of California at Berkeley, 2010.

[73] M.T. Holder, P.O. Lewis, and D.L. Swofford. The Akaike Information Criterion will not choose the no common mechanism model. *Systematic Biology*, 59(4):477–485, Jul 2010.

[74] M.T. Holder, P.O. Lewis, D.L. Swofford, and B. Larget. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. *Syst. Biol.*, 54(6):961–965, 2005.

[75] J.P. Huelsenbeck, B. Larget, and M.E. Alfaro. Bayesian phylogenetic model selection using reversible jump Markov Chain Monte Carlo. *Mol. Biol. Evol.*, 21(6):1123–1133, 2004.

[76] J.P. Huelsenbeck and B. Rannala. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.*, 53(6):904–913, 2004.

[77] J.P. Huelsenbeck and R. Ronquist. Bayesian analysis of molecular evolution using MrBayes. In R. Nielsen, editor, *Statistical Methods in Molecular Evolution*. 2005.

[78] S. Janson and E. Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.*, 32:2630–2649, 2004.

[79] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, 23(2):e123–e128, 2007.

[80] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Inferring phylogenetic networks by the maximum parsimony criterion: A case study. *MBE*, 24(1):324–337, 2007. http://www.cs.rice.edu/~nakhleh/Papers/MBE06.pdf.

[81] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 23(8):1046–1047, 2007.

[82] 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. In *ISBRA07*, volume 4463 of *LNCS*, pages 61–72. springer, 2007. http://www.cs.rice.edu/~nakhleh/Papers/isbra07.pdf.

[83] G. E. Jordan and W. H. Piel. Phylowidget: web-based visualizations for the tree of life. *Bioinformatics*, 24(14):1641–1642, 2008.

[84] D. Shasha J.T. L. Wang, H. Shan and W. H. Piel. Treerank: A similarity measure for nearest neighbor searching in phylogenetic databases. In *Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM 2003)*, pages 171–180, 2003.

[85] I. Kanj, L. Nakhleh, and G. Xia. Reconstructing evolution of natural languages: Complexity and parameterized algorithms. In *Proceedings of the 12th Annual International Computing and Combinatorics Conference (COCOON 06)*, 2006.

[86] S. Kannan and A. McGregor. More on reconstructing strings from random traces: Insertions and deletions. In *Proc. Intl. Symp. Information Theory, ISIT'05*, pages 297–301, 2005.

[87] Ruth Kirkpatrick. *Systematics and evolution of the fern genus Pellaea.* PhD dissertation, The University of California at Berkeley, 2007.

[88] I. Koffina, G. Serfiotis, V. Christophides, and V. Tannen. Mediating rdf/s queries to relational and XML sources. *Int'l Journal on Semantic Web and Information Systems*, 2:68–91, 2006.

[89] S.L. Kosakovsky Pond and S.V. Muse. Modeling heterogeneity of synonymous and nonsynonymous substitution rates across sites. *Mol. Biol. Evol.*, 22(12):2375–2385, 2005.

[90] M. Kulkarni and B. Moret. Consensus methods using phylogenetic databases. In *Proceedings of the Computational Systems Bioinformatics (CSB) Conference*, 2005.

[91] D.T. Kysela. *Bacteriophage response to the dynamic host environment: resistence, aging, and quorum sensing*. PhD dissertation, Yale University, 2008.

[92] D.T. Kysela and P.E. Turner. Optimal bacteriophage mutation rates for phage therapy. *Journal of Theoretical Biology*, 249:411–421, 2007.

[93] D.T. Kysela and P.E. Turner. Host aging promotes virulent parasite transmission, 2008. In preparation.

[94] H. Lapp, S. Bala, J.P. Balhoff, A. Bouck, N. Goto, M.T. Holder, R. Holland, A. Holloway, T. Katayama, P.O. Lewis, A. Mackey, B.I. Osborne, W.H. Piel, S. L. Kosakovsky Pond, A. Poon, W-G Qiu, J.E. Stajich, A. Stoltzfus, T. Thierer, A.J. Vilella, R.A. Vos, C.M. Zmasek, D. Zwickl, and T.J. Vision. The 2006 NESCent phyloinformatics hackathon: A field report. *Evolutionary Bioinformatics*, 3:357–366, 2007.

[95] P.O. Lewis, M.T. Holder, and K.E. Holsinger. Polytomies and Bayesian phylogenetic inference. *Syst. Biol.*, 54:241–253, 2005.

[96] H. Lin. *Internet Routing and Internet Service Provision*. PhD dissertation, The University of California at Berkeley, 2009.

[97] C.R. Linder, B.M.E. Moret, L. Nakhleh, and T. Warnow. Network (reticulated) evolution: Biology, models, and algorithms, 2004. Tutorial available at `compbio.unm.edu/papers.html`.

[98] C.R. Linder and L. Rieseberg. Reconstructing patterns of reticulate evolution in plants. *American J. Botany*, 91(10):1700–1708, 2004.

[99] C.R. Linder and T. Warnow. An overview of phylogeny reconstruction. In S. Aluru, editor, *Handbook of Computational Molecular Biology*. Chapman & Hall, 2005.

[100] K. Liu, S. Nelesen, S. Raghavan, C. R. Linder, and T. Warnow. Barking up the wrong treelength: the impact of gap penalty on alignment and tree accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:7–21, 2008.

[101] K. Liu, S. Nelesen, S. Raghavan, C. R. Linder, and T. Warnow. Rapid and accurate largescale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):561–1564, 2009.

[102] K. Liu, T.J. Warnow, M.T. Holder, S. Nelesen, J. Yu, A. Stamatakis, and C.R. Linder. SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, 2010. In review.

[103] T. Liu, J. Tang, and B.M.E. Moret. Quartet methods for phylogeny reconstruction from gene orders. In *Proc. 11th Conf. Computing and Combinatorics (COCOON'05)*, volume 3595 of *Lecture Notes in Computer Science*, pages 63–73. Springer-Verlag, 2005.

[104] W. Maddison, P. E. Midford, and S. E Otto. Estimating a binary character's effect on speciation and extinction. systematic biology. *Systematic Biology*, 56(5):701–710, 2007.

[105] F.V. Mannino. *Site-to-Site Rate Variation in Protein Coding Genes*. PhD dissertation, North Carolina State University, 2006.

[106] F.V. Mannino and S.V. Muse. Extensive site-to-site variability of synonymous substitution rates in mitochondrial genomes. *Genetics (in press)*, 2006.

[107] R. Mao. *Distance-Based Indexing and Its Applications in Bioinformatics*. PhD dissertation, University of Texas at Austin, 2007.

[108] R. Mao, W. Xu, N. Singh, and D.P. Miranker. An assessment of a metric space database index to support sequence homology. In *Proc. 3rd IEEE Symp. on Bioinformatics and Bioengineering BIBE'03*, pages 375–382. IEEE Press, 2003.

[109] M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and insertions. *Theor. Comput. Sci.*, 325(3):347–360, 2004.

[110] R. McBride, S. Duffy, R. Montville, Y. Yang, S.-J. Lee, J. Kim, and P.E. Turner. Experimental phylogenetics of RNA phage phi-6, 2008. In preparation.

[111] A. McGregor. *Processing Data Streams*. PhD dissertation, University of Pennsylvania, 2007.

[112] R. Mihaescu, D. Levy, and Lior Pachter. Why neighbor-joining works. http://lanl.arxiv.org/abs/cs.DS/0602041, 2006. Paper presented at the 3rd International Conf. in Phylogenomics, in Sainte Adele, Canda.

[113] R. H. Mihaescu. *Distance Methods in Phylogeny*. PhD dissertation, Mathematics Department, University of California, Berkeley, 2008.

[114] B. Milch, B. Marthi, D. Sontag, S. Russell, D.L. Ong, and A. Kolobov. Approximate inference for infinite contingent Bayesian networks. In *Proc. 10th Workshop on Artificial Intelligence and Statistics*, 2005.

[115] B. Milch, B. Marthi, D. Sontag, S. Russell, D.L. Ong, and A. Kolobov. BLOG: Probabilistic models with unknown objects. In *Proc. Int'l Joint Conf. on Artificial Intelligence IJCAI'05*, 2005.

[116] D. Miranker, W. Xu, and R. Mao. A metric-space DBMS to support biological discovery. In *Proceedings of the 15th International Conference on Statistical and Scientific Database Management*, pages 241–244, 2003.

[117] B.A. Moore and M.J. Donoghue. A Bayesian approach for evaluating the impact of historical events on rates of diversification. *Proc. Nat. Acad. Sci. USA*, 106:4307–4312, 2009.

[118] B.R. Moore, S. A. Smith, and M. J. Donoghue. Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Syst. Biol.*, 55:662–676, 2006.

[119] B.R. Moore, S. A. Smith, R. H. Ree, and M. J. Donoghue. Incorporating fossil data in biogeographic inference: a likelihood approach. *Evolution*, 2008. (in press).

[120] S. Moran, S. Rao, and S. Snir. Using semi-definite programming to enhance supertree resolvability. In *Proc. 5th Workshop Algorithms in Bioinformatics (WABI'05)*, pages 89–103, 2005.

[121] S. Moran and S. Snir. Efficient approximation of convex recolorings. *Journal of Computer and System Sciences*, 73:1078–1089, 2007. An earlier version appeared in APPROX/RANDOM 2005.

[122] S. Moran and S. Snir. Convex recolorings of strings and trees: Definitions, hardness results and algorithms. *J. Comput. Syst. Sci.*, 74(5):850–869, 2008.

[123] S. Moran, S. Snir, and W.-K. Sung. Partial convex recolorings of trees and galled networks: Tight upper and lower bounds, 2007. http://www.cs.technion.ac.il/~moran/r/PS/gnets-TOA-7Feb2007.pdf.

[124] B.M.E. Moret, J. Tang, and T. Warnow. Reconstructing phylogenies from gene-content and gene-order data. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 321–352. Oxford University Press, 2005.

[125] B.M.E. Moret and T. Warnow. Advances in phylogeny reconstruction from gene order and content data. In E.A. Zimmer and E.H. Roalson, editors, *Molecular Evolution: Producing the Biochemical Data, Part B*, volume 395 of *Methods in Enzymology*, pages 673–700. Elsevier North Holland, 2005.

[126] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

[127] E. Mossel. Survey: Information flow on trees. In J. Nestril and P. Winkler, editors, *Graphs, Morphisms and Statistical Physics. DIMACS series in discrete mathematics and theoretical computer science*. AMS Press, 2004.

[128] E. Mossel. Distorted metrics on trees and phylogenetic forests. *IEEE Trans. Comput. Biol. and Bioinformatics*, 4:106–116, 2006.

[129] E. Mossel and E.Vigoda. Limitations of Markov Chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Annals of Applied Probability*, 16(4):2215–2234, 2006.

[130] E. Mossel and S. Roch. Learning Nonsingular Phylogenies and Hidden Markov Models. In *Proc. 37th Symp. on the Theory of Computing (STOC'05)*, pages 366–376. 2005. To appear in the Annals of Probability.

[131] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.*, 16(2):583–614, 2006.

[132] E. Mossel and M. Steel. A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.*, 187(2):189–203, 2004.

[133] E. Mossel and M. Steel. How much can evolved characters tell us about the tree that generated them? In Olivier Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 384–412. Oxford University Press, 2005.

[134] E. Mossel and E. Vigoda. Limitations of Markov Chain Monte Carlo algorithms for Bayesian inference of phylogeny (short report). *Science*, 309:2207–2209, 2005.

[135] L. Nakhleh. *Phylogenetic Networks*. PhD dissertation, University of Texas, 2005.

[136] L. Nakhleh, G. Jin, F. Zhao, and J. Mellor-Crummey. Reconstructing phylogenetic networks using maximum parsimony. In *Proc. 4th Computational Systems Biology Conf. (CSB'05)*. IEEE Press, 2005.

[137] L. Nakhleh, D. Miranker, F. Barbancon, W. Piel, and M. Donoghue. Requirements of phylogenetic databases. In *Proc. 3rd IEEE Symp. on Bioinformatics and Bioengineering BIBE'03*, pages 141–148. IEEE Press, 2003.

[138] L. Nakhleh, D. Ruths, and H. Innan. Gene trees, species trees, and species networks. In R. Guerra and D. Allison, editors, *Meta-analysis and Combining Information in Genetics*. Chapman & Hall, CRC Press, 2005.

[139] L. Nakhleh, D. Ruths, and L.-S. Wang. RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. In *Proc. 11th Conf. Computing and Combinatorics (COCOON'05)*, volume 3595 of *Lecture Notes in Computer Science*, pages 84–93. Springer-Verlag, 2005.

[140] L. Nakhleh, J. Sun, T. Warnow, R. Linder, B.M.E. Moret, and A. Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proc. 8th Pacific Symp. on Biocomputing (PSB'03)*, pages 315–326. World Scientific Pub., 2003.

[141] L. Nakhleh and L.-S. Wang. Phylogenetic networks: properties and relationship to trees and clusters. *Transactions on Computational Systems Biology II*, pages 82–99, 2005. LNCS#3680.

[142] L. Nakhleh and L.-S. Wang. Phylogenetic networks, trees, and clusters. In *Proceedings of the 2005 International Workshop on Bioinformatics Research and Applications (IWBRA 05)*, pages 919–926, 2005. LNCS #3515.

[143] L. Nakhleh, T. Warnow, and C.R. Linder. Reconstructing reticulate evolution in species—theory and practice. In *Proc. 8th Conf. Comput. Mol. Biol. (RECOMB'04)*, pages 337–346. ACM Press, 2004.

[144] L. Nakhleh, T. Warnow, C.R. Linder, and K. St. John. Reconstructing reticulate evolution in species – theory and practice. *Journal of Computational Biology*, 12(6–7):796–811, 2005. Special issue for selected papers from RECOMB 2004.

[145] M. Narayanan. *Comparative and Evolutionary Anaylsis of Cellular Pathways*. PhD dissertation, EECS Department,University of California, Berkeley, 2007.

[146] M. Narayanan and R. M. Karp. Comparing protein interaction networks via a graph match-and-split algorithm. *Journal of Computational Biology*, 14(7):892–907, 2007.

[147] S. Nelesen. *Improved methods for phylogenetics*. PhD dissertation, The University of Texas at Austin, 2009.

[148] S. Nelesen, K. Liu, D. Zhao, C. R. Linder, and T. Warnow. The effect of the guide tree on muultiple sequence alignment and subsequent phylogenetic analyses. In *Proceedings of the 2008 Pacific Symposium on Biocomputing*, 2008.

[149] N.D. Pattengale, E.J. Gottlieb, and B.M.E. Moret. Efficiently computing the Robinson-Foulds metric. *Journal of Computational Biology*, 14:724–735, 2007.

[150] N.D. Pattengale and B.M.E. Moret. A sublinear-time randomized approximation scheme for the Robinson-Foulds metric. In *Proc. 10th Int'l Conf. on Research in Comput. Molecular Biol. RECOMB'06*, volume 3909 of *Lecture Notes in Computer Science*, pages 221–230, 2006.

[151] W.H. Piel. Phyloinformatics and tree networks. In C.H. Wu, P. Wang, and J. T. L. Wang, editors, *Computational Biology and Genome Informatics*. World Scientific Press, 2003.

[152] W.H. Piel, L. Chan, M.J. Dominus, J. Ruan, R.A. Vos, and V. Tannen. TreeBASE Version 2: A database of phylogenetic knowledge. In *e-Biosphere 09 International Conference on Biodiversity Informatics*, 2009.

[153] S. Kosakovsky Pond and S.V. Muse. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*, 22(12):2375–2385, 2005.

[154] S.L. Kosakovsky Pond, S.D. Frost, and S.V. Muse. HyPhy: A platform for molecular evolutionary analysis. *Bioinformatics*, 21:676–679, 2005.

[155] S.L. Kosakovsky Pond, F.V. Mannino, M.B. Gravenor, S.V. Muse, and S.D.W. Frost. Evolutionary model selection with a genetic algorithm: a case study using stem RNA. *Mol. Biol. Evol.*, 24:159–170, 2006.

[156] S. Ramakrishnan. *A Systems Approach to Computational Protein Identification*. PhD dissertation, The University of Texas at Austin, 2010.

[157] H.R. Ree, B. R. Moore, C. Webb, and M. J. Donoghue. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*, 59:2299–2311, 2005.

[158] R.H. Ree and S. A. Smith. Maximum-likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology*, 57:4–14, 2008.

[159] S. Riesenfeld. *Optimization and Reconstruction over Graphs*. PhD dissertation, EECS Department,University of California, Berkeley, 2007.

[160] S. Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 3(1):92–94, 2006.

[161] S. Roch. *Markov Models on Trees: Reconstruction and Applications*. PhD thesis, University of California, Berkeley, 2007.

[162] U. Roshan. *Algorithmic techniques for improving the speed and accuracy of phylogenetic methods*. PhD thesis, The University of Texas at Austin, 2004.

[163] U. Roshan, B. M. E. Moret, T. L. Williams, and T. Warnow. Performance of supertree methods on various dataset decompositions. In O. R. P. Bininda-Emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, volume 3 of *Computational Biology*, pages 301–328. Kluwer Academics, 2004. (Dress, A. series ed.).

[164] U. Roshan, B.M.E. Moret, T.L. Williams, and T. Warnow. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proc. 3rd Computational Systems Biology Conf. (CSB'04)*, pages 98–109. IEEE Press, 2004.

[165] D. Ruths and L. Nakhleh. Recombination and phylogeny: effects and detection. *International Journal on Bioinformatics Research and Applications*, 1(2):202–212, 2005.

[166] D. Ruths and L. Nakhleh. RECOMP: A parsimony-based method for detecting recombination. In *Proceedings of the 4th Asia Pacific Bioinformatics Conference*, pages 59–68, 2006.

[167] D. Ruths and L. Nakhleh. Techniques for assessing phylogenetic branch support: A performance study. In *Proceedings of the 4th Asia Pacific Bioinformatics Conference*, pages 187–196, 2006.

[168] A.S. Schwartz. *Posterior Decoding Methods for Optimization and Accuracy Control of Multiple Alignments*. PhD thesis, EECS Department, University of California, Berkeley, Mar 2007.

[169] A.S. Schwartz, E.W. Myers, and L. Pachter. Alignment Metric Accuracy. Electronic paper, available at www.arXiv:q-bio.QM/0510052.

[170] A.S. Schwartz and L. Pachter. Multiple alignment by sequence annealing. In *Proceedings of the European Conference on Computational Biology (ECCB 2006)*, 2006. Received best paper award.

[171] A.S. Schwartz and L. Pachter. Multiple alignment by sequence annealing. *Bioinformatics*, 23(2):e24–29, 2007.

[172] R. Shapley. *Teaching with a Visual Tree of Life*. Masters thesis, University of California at Berkeley, 2005. See also http://groups.sims.berkeley.edu/TOL/.

[173] R. Sharan and G. Myers. A motif-base framework for recognizing sequence families. In *Proc. 13th Conf. on Intelligent Systems for Molecular Biology (ISMB'05)*, 2005.

[174] S.A. Smith. *Evolving biogeography: New methods and their application in the plant clade Lonicera*. PhD dissertation, Yale University, 2008.

[175] S.A. Smith, J.M. Beaulieu, and M.J. Donoghue. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol Bio*, 9(37), 2009.

[176] S.A. Smith and M. J. Donoghue. Rates of molecular evolution are linked to life history in flowering plants. *Science*, 322(5898):88–89, 2008.

[177] S.A. Smith, R. H. Ree, and M. J. Donoghue. Accuracy of maximum-likelihood inferences of geographic range reconstructions and parameter estimates: a simulation study, 2008. In preparation.

[178] S. Snir and L. Pachter. Phylogenetic profiling of insertions and deletions in vertebrate genomes. In *Proc. 10th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB'06)*, pages 265–280, 2006.

[179] S. Snir and S. Rao. Using max cut to enhance rooted trees consistency. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 3(4):323–333, 2006.

[180] S. Snir and T. Tuller. The Net-HMM: a HMM-based likelihood model for evolutionary networks. In *Proc. 8th Int'l Workshop Algs. in Bioinformatics (WABI'08)*, 2008.

[181] S. Snir, T. Warnow, and S. Rao. Short quartet puzzling: A new quartet-based phylogeny reconstruction algorithm. *Journal of Computational Biology*, 15(1):91–103, 2008.

[182] A. Stamatakis, P. Hoover, and J. Rougemont. A rapid bootstrap algorithm for the RAxML Web-Servers. *Systematic Biology*, 2008. in press.

[183] E. Strain. *Plant molecular evolution.* PhD dissertation, North Carolina State University, 2006.

[184] J. Sukumaran and M.T. Holder. DendroPy: A python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

[185] K.M. Swenson, M. Marron, J.V. Earnest-DeYoung, and B.M.E. Moret. Approximating the true evolutionary distance between two genomes. In *Proc. 6th Workshop on Algorithm Engineering and Experiments (ALENEX'05)*, pages 121–129. SIAM Press, 2005.

[186] K.M. Swenson, N.D. Pattengale, and B.M.E. Moret. A framework for orthology assignment from gene rearrangement data. In *Proc. 3rd RECOMB Workshop on Comparative Genomics*, Lecture Notes in Computer Science, pages 153–166. Springer-Verlag, 2005.

[187] M.S. Swenson. *Supertree methods.* PhD dissertation, University of Texas at Austin, 2009.

[188] M.S. Swenson, F. Barbançon, C.R. Linder, and T. Warnow. A simulation study comparing supertree and combined analysis methods using smidgen. In *Proceedings of WABI (Workshop on Algorithms for Bioinformatics)*, 2009.

[189] M.S. Swenson, F. Barbançon, C.R. Linder, and T. Warnow. A simulation study comparing supertree and combined analysis methods using smidgen. *Algorithms for Molecular Biology*, 2009. Special issue of selected papers from WABI 2009.

[190] M.S. Swenson, R. Suri, C.R. Linder, and T. Warnow. Superfine: fast and accurate supertree estimation. *Systematic Biology*, 2010. in review.

[191] M.S. Swenson, R. Suri, C.R. Linder, and T. Warnow. Using Quartets MaxCut to improve supertree estimation. In *Proceedings of WABI (Workshop on Algorithms for Bioinformatics) 2010*, 2010.

[192] M.S. Swenson, R. Suri, C.R. Linder, and T. Warnow. Using Quartets MaxCut to improve supertree estimation. *Algorithms for Molecular Biology*, 2010. Special issue of selected papers from WABI 2010.

[193] V. Le Sy, A. Varon, and W. C. Wheeler. Pairwise alignment with rearrangement. *Genome Informatics*, 17(2):141–151, 2006.

[194] A. Talwar. *Metric Methods in Approximation Algorithms.* PhD dissertation, EECS Department,University of California, Berkeley, 2005.

[195] J. Tang. *Large Scale Phylogenetic Reconstruction from Arbitrary Gene-order Data.* PhD dissertation, Computer Science Department, University of New Mexico, 2004.

[196] J. Tang and B.M.E. Moret. Linear programming for phylogenetic reconstruction based on gene rearrangements. In *Proc. 16th Symp. on Combinatorial Pattern Matching (CPM'05)*, volume 3537 of *Lecture Notes in Computer Science*, pages 406–416. Springer-Verlag, 2005.

[197] J. Tang, B.M.E. Moret, L. Cui, and C.W. dePamphilis. Phylogenetic reconstruction from arbitrary gene-order data. In *Proc. 4th IEEE Symp. on Bioinformatics and Bioengineering BIBE'04*, pages 592–599. IEEE Press, 2004.

[198] J. Tang and L. Wang. Improving genome rearrangement phylogeny using sequence-style parsimony. In *Proc. 5th IEEE Conf. on Bioinformatics and Bioengineering*, pages 137–144, 2005.

[199] D. J. Taylor and W. H. Piel. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Molecular Biology and Evolution*, 21(8):1534–1537., 2004.

[200] R. Timme. *Reticulate Evolution in Helianthus (Asteraceae); Comparative Chloroplast Genomics of Helianthus and Lactuca*. PhD dissertation, University of Texas at Austin, 2006.

[201] S. Tringe, C. von Mering, A. Kobayashi, A. Salamov, K. Chen, and et al. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005.

[202] A. Varón. *Algorithms and hypothesis selection in dynamic homology phylogenetic analysis*. PhD dissertation, City University of New York, 2010.

[203] A. Varón, L.S. Vinh, I. Bomash, and W.C. Wheeler. Poy 4.0 beta 2398. http://research.amnh.org/scicomp/projects/poy.php, 2007.

[204] L.S. Vinh, A. Varón, D. Janies, and W.C. Wheeler. Towards phylogenomic reconstructions. *BIOCOMP 2007, the 2007 International Conference on Bioinformatics & Computational Biology*, pages 98–104, 2007.

[205] R. Vos. *Inferring large phylogenies: the big tree problem*. PhD dissertation, Simon Fraser University, 2006. http://ir.lib.sfu.ca/handle/1892/3503.

[206] R.A. Vos, H. Lapp, W. Piel, and V. Tannen. TreeBASE2: Rise of the machines. *Nature Precedings*, 2010. Peer-reviewed for iEvoBio [ievobio.org].

[207] L.-S. Wang, J. Leebens-Mack, P.K. Wall, K. Beckmann, C.W. dePamphilis, and T. Warnow. The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE Transactions on Computational Biology and Bioinformatics*, 2009. in press.

[208] L.-S. Wang and T. Warnow. Distance-based genome rearrangement phylogeny. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 353–383. Oxford University Press, 2005.

[209] L.-S. Wang and T. Warnow. Reconstructing chromosomal evolution. *SIAM Journal on Computing*, 36:99–131, 2006.

[210] L.-S. Wang, T. Warnow, B.M.E. Moret, R.K. Jansen, and L.A. Raubeson. Distance-based genome rearrangement phylogeny. *Journal of Molecular Evolution*, 63:473–83, 2006.

[211] T. Warnow. Large scale phylogenetic analysis. In S. Aluru, editor, *Handbook of Computational Molecular Biology*. Chapman & Hall, 2005.

[212] T. Warnow, S.N. Evans, D. Ringe, and L. Nakhleh. A stochastic model of language evolution that incorporates homoplasy and borrowing. In P. Foster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*. Cambridge University Press, 2004.

[213] M. J. Sanderson W.H. Piel and M. J. Donoghue. The small-world dynamics of tree networks and data mining in phyloinformatics. *Bioinformatics*, 19(9):1162–1168, 2003.

[214] W.C. Wheeler. Chromosomal character optimization. *Molecular Phylogenetics and Evolution*, 2007.

[215] W.C. Wheeler, L. Aagesen, C.P. Arango, J. Faivovich, T. Grant, C.A. D'Haese, D. Janies, W.L. Smith, A. Varon, and G. Giribet. *Dynamic Homology and Phylogenetic Systematics: A unified approach using POY*. American Museum of Natural History, 2006.

[216] T.L. Williams, D.A. Bader, M. Yan, and B.M.E. Moret. High-performance phylogeny reconstruction under maximum parsimony. In A. Zomaya, editor, *Parallel Computing for Bioinformatics and Computational Biology*, chapter 16. John Wiley & Sons, 2006.

[217] T.L. Williams and M. L. Smith. Phylospaces: Evolutionary trees and tuple space. In *Proc. Fifth IEEE International Workshop on High Performance Computational Biology (HiCOMB 2006)*, 2006.

[218] T.L. Williams and M.L. Smith. Cooperative-Rec-I-DCM3: A population-based approach for reconstructing phylogenies. In *Proc. Third IEEE Symp. on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'05)*, pages 127–134, 2005.

[219] T.L. Williams and M.L. Smith. The role of diverse populations in phylogenetic analysis. In *Genetic and Evolutionary Computation Conference (GECCO-2006)*, 2006.

[220] W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M-H Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology*, 2010. In press.

[221] W. Xu, W. Briggs, J. Padolina, W. Liu, C.R. Linder, and D.P. Miranker. Using MoBIos scalable genome join to find conserved primer pair candidates between two genomes. In *Proc. 12th Conf. on Intelligent Systems for Molecular Biology (ISMB'04)*, volume 20, pages i355–i362, 2004.

[222] S. Zhang, K. G. Herbert, J. T. L. Wang, W. H. Piel, and D. R. B. Stockwell. PhyloMiner: A tool for evolutionary data analysis. In *18th International Conference on Scientific and Statistical Database Management*, pages 129–132, July 2006.

[223] Y. Zheng. *Efficient Scientific Data Management Over Trees*. PhD dissertation, University of Pennsylvania, 2006.

[224] R.A. Zufall, C.l. McGrath, S.V. Muse, and L.A. Katz. Genome architecture drives protein evolution in ciliates. *Mol. Biol. Evol.*, 23:1681–1687, 2006.

[225] D. J. Zwickl. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, The University of Texas at Austin., 2006.