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

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 highlights of the final year's activities in Section **??**; detailed information about these activities appears in later relevant sections of this report. We continue in Section 2 with the list of participants, including faculty, students, postdocs, professional staff, and organizational partners. In Section **??** 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 3), software and the central resource (Section 4), simulation and modelling (Section 5), outreach activity (Section 6), and databases (Section 7). We summarize our contributions to science, human resources, and broader impact, in Section 8; details of the research contributions are provided in the relevant sections. Section 9 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. We indicate ethnicity, citizenship and permanent residency, gender, and disability status, where known. We also indicate whether each worked for 160 hours or more in any grant year.

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, white, male, no disabilities. Home Page: http://www.cc.gatech.edu/~bader/.
- Francine Berman, Prof. of Computer Science and Director of the San Diego Supercomputing Center, UCSD. US Citizen, white, female, no disabilities. Home Page: http://www.sdsc.edu/about/Director.html
- Susan Davidson, Prof. of Computer and Information Sciences, U. Penn. US citizen, white, female, no disabilities. Home Page: http://www.cis.upenn.edu/~susan/.
- Michael Donoghue, Prof. of Ecology and Evolutionary Biology, Yale. US citizen, white, male, no disabilities. Home Page: http://www.yale.edu/eeb/donoghue.
- Steven Evans, Prof. of Statistics (joint appointment in Mathematics), UC Berkeley. US citizen, white, male, no disabilities. Home Page: http://www.stat.berkeley.edu/~evans/
- David Hillis, Prof. of Integrative Biology, UT Austin. US citizen, white, male, no disabilities. 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. US citizen, white, male, no disabilities. Home Page: http://people.ku.edu/~mtholder/
- John Huelsenbeck, Prof. of Integrative Biology, UC Berkeley. US citizen, white, male, no disabilities. Home Page: http://ib.berkeley.edu/people/faculty/person_detail.php?person=319
- Warren Hunt, Prof. of Computer Sciences, UT Austin. US citizen, white, male, no disabilities. Home Page: http://www.cs.utexas.edu/~hunt/
- Robert Jansen, Prof. of Integrative Biology, UT Austin. US citizen, white, male, no disabilities. Home Page: http://www.biosci.utexas.edu/ib/faculty/jansen.htm
- Sampath Kannan, Prof. of Computer and Information Sciences, U. Penn. US citizen, Asian, male, no disabilities. Home Page: http://www.cis.upenn.edu/~kannan/
- Richard Karp, Prof. of Computer Science, UC Berkeley. US citizen, white, male, no disabilities. Home Page: http://www.cs.berkeley.edu/~karp/
- Junhyong Kim, Prof. of Biology, U. Penn. US permanent resident, Asian, male, no disabilities. Home Page: http://www.bio.upenn.edu/faculty/kim/
- Paul Lewis, Prof. of Ecology and Evolutionary Biology, U. Conn. US citizen, white, male, no disabilities. Home Page: http://www.eeb.uconn.edu/people/plewis/
- C. Randal Linder, Associate Prof. of Integrative Biology, UT Austin. US citizen, white, male, no disabilities. Home Page: http://www.biosci.utexas.edu/IB/faculty/linder.htm.

- David Maddison, Prof. of Entomology, U. Arizona. US citizen, white, male, no disabilities. Home Page: http://david.bembidion.org/
- Wayne Maddison, Prof. of Zoology, U. British Columbia. US citizen, white, male, no disabilities. Home Page: http://salticidae.org/wpm/home.html
- Lauren Ancel Meyers, Assoc. Prof. of Integrative Biology, UT Austin. US citizen, white, female, no disabilities. Home Page: http://cluster3.biosci.utexas.edu/research/meyers/LaurenM/Lauren_M.html
- Daniel Miranker, Prof. of Computer Sciences, UT Austin. US citizen, white, male, no disabilities. Home Page: http://www.cs.utexas.edu/users/miranker.
- Brent Mishler, Prof. of Integrative Biology, UC Berkeley. US citizen, white, male, no disabilities. 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). US citizen, white, male, no disabilities. Home Page: http://people.epfl.ch/bernard.moret
- Elchanan Mossel, Assoc. Prof. of Statistics and Computer Science, UC Berkeley. US permanent resident, white, male, no disabilities. Home Page: http://www.stat.berkeley.edu/~mossel/
- Spencer Muse, Associate Prof. of Statistics, NCSU. US citizen, white, male, no disabilities. 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). US citizen, white, male, no disabilities. 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. White male, US citizen, no disabilities. Home Page: http://www.cs.berkeley.edu/~christos/
- William Piel, Assoc. Director of Evolutionary Bioinformatics, Peabody Museum of Natural History, Yale University. US citizen, white, male, no disabilities. Home Page: http:www.treebase.org/~piel.
- Satish Rao, Prof. of Computer Science, UC Berkeley. US citizen, Asian, male, no disabilities. Home Page:http://www.cs.berkeley.edu/~satishr
- Usman Roshan, Associate Prof. of Computer Science, NJIT. US Permanent Resident, white, male, no disabilities. Home Page: http://cs.njit.edu/usman/
- Stuart Russell, Prof. of Computer Science, UC Berkeley. US citizen, white, male, no disabilities. Home Page: http://www.cs.berkeley.edu/~russell.
- David L. Swofford, Senior Research Scientist, Duke Institute for Genome Sciences and Policy. US citizen, white, male, no disabilities. Home Page: http://www.genome.duke.edu/centers/ceg/swofford/.

- Jijun Tang, Associate Prof. of Computer Science and Engineering, U. South Carolina. U.S. Permanent Resident, Asian, male, no disabilities. Home Page: http://www.cse.sc.edu/~jtang/
- Val Tannen, Prof. of Computer and Information Sciences, UPenn. US citizen, white, male, no disabilities. Home Page: http:www.cis.upenn.edu/~val.
- Paul Turner, Assoc. Prof. of Ecology and Evolution, Yale. US citizen, African-Americn, male, no disabilities. Home Page: http://www.yale.edu/turner/home/index.htm
- Tandy Warnow, Prof. of Computer Sciences, UT Austin. US citizen, white, female, no disabilities. Home Page: http://www.cs.utexas.edu/users/tandy
- Ward C. Wheeler, Curator of Invertebrates, AMNH. White male, US citizen, no disabilities. Home Page: http://www.amnh.org/science/divisions/invertzoo/bio.php?scientist=wheeler
- Tiffani Williams, Assistant Prof. of Computer Science, Texas A&M. US Citizen, African American, female, no disabilities. 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. US citizen, white, male, no disabilities.
- Adam Cathers, San Diego Supercomputing Center, UCSD. US citizen, white, male, no disabilities.
- Lucie Chan, Staff Member, San Diego Supercomputing Center, UCSD. US citizen, Asian, female, no disabilities.
- April Davidson, Project Coordinator, UNM. US Citizen, white, female, no disabilities.
- T. Phong Dinh, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, Asian, male, no disabilities.
- Mark Dominus, Staff Member, U. Penn. US Citizen, white, male, no disabilities.
- Kevin Fowler, San Diego Supercomputing Center, UCSD.
- Paul Hoover, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, white, male, no disabilities.

- Dana Jermanis, Senior Software Developer, San Diego Supercomputing Center, UCSD. US Citizen, White, female, no disabilities.
- Terri Liebowitz, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, white, female, no disabilities.
- Brian Lucena, Visiting Faculty, UC Berkeley. US citizen, white, male, no disabilities.
- Madhu Madhusudan, Staff Member, San Diego Supercomputing Center, UCSD. US permanent resident, Asian, male, no disabilities.
- Tim McPhillips, San Diego Supercomputing Center, UCSD. US citizen, white, male, no disabilities.
- Mark Miller, Team Leader, San Diego Supercomputing Center, UCSD. US citizen, white, male, no disabilities. Home Page: http://www.sdsc.edu/~mmiller.
- Erica Ocegueda, Staff member, UNM. US citizen, white, female, no disabilities.
- Cynthia Perrine, Education Coordinator, UC Berkeley. US citizen, white, female, no disabilities.
- Jin Ruan, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, white, male, no disabilities.
- David Stockwell, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, white, male, no disabilities.
- Rahul Suri, Staff Member, UT-Austin. US citizen, Asian, male, no disabilities.
- Ashton Taylor, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, white, male, no disabilities. Home page: http://www.digitalmudstudios.com.
- Can Tran, Staff Member, San Diego Supercomputing Center, UCSD. US permanent resident, Asian, male, no disabilities.
- Brandan White, San Diego Supercomputing Center, UCSD. US Citizen, white, male, no disabilities.
- Tracy Zhao, Staff Member, San Diego Supercomputing Center, UCSD. US Citizen, female, no disabilities.

Staff that are not paid by CIPRES

- Laurie Alvarez, Staff member, UT-Austin. Part Native American, part white, US citizen, female, no disabilities.
- Adriana Aquino, Staff. US citizen, female, Hispanic, no disabilities. Worked less than 160 hours in all grant years.
- Daniel Aviv, Staff. Male, US citizen, no disabilities. worked less than 160 hours in all grant years.

- Joel Cracraft, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Louise Crowley, Student, Female, US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Mohammad Faiz, Staff. Male, US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Megan Harrison, Postdoc, Female US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Jay Holmes, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Daniel Janies, OSU-Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Maritza Macdonald, Staff. Female, Hispanic US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Mordecai MacLow, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Mark Norell, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Susan Perkins, Staff. Female US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Paola Predraza, Postdoc, Female, Hispanic US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Lorenzo Prendini, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- David Randle, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Zobar Ris, Staff. Male US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Monique Scott, Staff. Female, African American, US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Cate Starr, Staff. Female, US citizen, no disabilities. Worked less than 160 hours in all grant years.
- Ellen Trimarco, Staff. Female, US citizen, no disabilities. 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). US permanent resident, Hispanic/Latino, Male, no disabilities. Home Page: http://www.eeb.ucla.edu/indivfaculty.php?FacultyKey=10361
- Mark Holder (postdoctoral fellow of Dave Swofford while at FSU). US citizen, White, Male, no disabilities.
- Peter Midford, UBC (postdoc of Wayne Maddison). US citizen, White Male, no disabilities. Home Page: http://mesquiteproject.org/midford/.
- Sagi Snir, UC Berkeley (postdoc of Lior Pachter). Not US citizen/permanent resident, white, male, no disabilities. Home Page: http://math.berkeley.edu/~ssagi/
- Shel Swenson, UT Austin (postdoc of Tandy Warnow). US citizen, white, female, no disabilities.
- Rutger Vos, Simon Fraser U. (research fellow with Wayne Maddison). Not US citizen/permanent resident, white, male, no disabilities. Home Page: http://rutgervos.blogspot.com/.

Postdoctoral fellows not funded by CIPRES

- François Barbançon, UT Austin (postdoctoral fellow of Tandy Warnow). US citizen, Male, White/Asian, no disability.
- Tanya Berger-Wolf, UNM (postdoctoral fellow of Bernard Moret). US citizen, Female, White, no disability.
- Sarah Cohen-Boulakia, U. Penn (postdoctoral fellow of Val Tannen)
- Steve Fisher, U. Penn (postdoctoral fellow of Junhyong Kim). US citizen/permanent resident, white, male, no disabilities.
- Fan Ge, University of Pennsylvania. Not US citizen/permanent resident, Asian, male, no disabilities.
- Sergei Kosakovsky Pond, UCSD (research fellow working with Spencer Muse), US citizen/permanent resident, white, male, no disabilities.
- Yelena Shvets, UC Berkeley (postdoc of Steve Evans and Monty Slatkin). White female, no disabilities.
- Li-San Wang, U. Penn (postdoctoral fellow of Junhyong Kim), US citizen/permanent resident, Asian, male, no disabilities.
- Cam Webb, Yale U. (postdoctoral fellow of Michael Donoghue). US citizen, White, Male, no disabilities.
- Tiffani Williams, UNM (postdoctoral fellow of Bernard Moret). US Citizen, African American, female, no disabilities.

2.5 Graduate Students.

Students paid (mostly partially) through CIPRES:

- François Barbançon, UT Austin (student of Dan Miranker). Not US Citizen/permanent resident, white, male, no disabilities.
- Nicholas Bray, UC Berkeley (student of Lior Pachter). US citizen, white, male, no disabilities.
- Kevin Chen, UC Berkeley (student of Lior Pachter and Satish Rao). US citizen, Asian, male, no disabilities.
- Shirley Cohen, UPenn (student of Val Tannen and Susan Davidson). White female, US citizen, no disabilities. Home Page: http://www.seas.upenn.edu/~shirleyc.
- Costis Daskalakis, UC Berkeley (student of Satish Rao). Not US citizen/permanent resident, white, male, no disabilities.
- Nick Eriksson, UC Berkeley (student of Bernd Sturmfels). US citizen, white, male, no disabilities.
- Yu Fan, U. Conn (student of Paul Lewis). Chinese citizen (not permanent resident of the US), Asian male, no disabilities.
- Kirsten Fisher, UC Berkeley (student of Brent Mishler). US Citizen, white, female, no disabilities.
- Ganesh Ganapathy, UT Austin (student of Tandy Warnow). Not US Citizen/permanent resident, Asian, male, no disabilities.
- Denise Green, UC Berkeley (worked with Brent Mishler). US citizen, white, female, no disabilities.
- Sheng Guo, U. Penn (student of Junyong Kim). Not US citizen or permanent resident, Asian Male, no disabilities.
- Tracy Heath, UT Austin (student of David Hillis). US Citizen, white, female, no disabilities.
- Cameron Hill, UC Berkeley (student in the Mathematics Department). US citizen, African American, male, no disabilities.
- David Kysela, Yale University (student of Paul Turner). US Citizen, white, male, no disabilities.
- Ruth Kirkpatrick, UC Berkeley (worked with Brent Mishler). US citizen, white, female, no disabilities.
- Henry Lin, UC Berkeley (student of Satish Rao). US Citizen, Asian, male, no disabilities.
- Kevin Liu, UT Austin (student of Tandy Warnow). US Citizen, Asian, male, no disabilities.
- Wenguo Liu, UT Austin (student of Dan Miranker). Not US Citizen/Permanent Resident, Asian, male, no disabilities.

- Andrew McGregor, U. Penn (student of Sampath Kannan). White male, not US citizen or permanent resident, no disabilities.
- Frank Mannino, NCSU (student of Spencer Muse). US Citizen, white, male, no disabilities.
- Rui Mao, UT Austin (student of Dan Miranker). Not US Citizen/Permanent Resident, Asian, female, no disabilities.
- Radu Mihaescu, UC Berkeley (student of Lior Pachter and Satish Rao). US citizen, white, male, no disabilities.
- Eric Miller, UT Austin (student of Lauren Meyers). US citizen, white, male, no disabilities.
- Luay Nakhleh, UT Austin (student of Tandy Warnow). Not US Citizen/permanent resident, white, male, no disabilities.
- Manikandan Narayanan, UC Berkeley (student of Dick Karp). Not US citizen/permanent resident, Asian, male, no disabilities.
- Serita Nelesen, UT Austin (student of Warren Hunt). US Citizen, white, female, no disabilities.
- Smriti Ramakrishnan, UT Austin (student of Dan Miranker). Not US Citizen, Asian, female, no disabilities.
- Samantha Riesenfeld, UC Berkeley (student of Dick Karp). US Citizen, white, female, no disabilities.
- Sébastien Roch, UC Berkeley (student of Elchanan Mossel). Not US citizen/permanent resident, white, male, no disabilities.
- Usman Roshan, UT Austin (student of Tandy Warnow). Not US citizen/permanent resident, white, male, no disabilities.
- Ariel Schwartz, UC Berkeley (student of Gene Myers and Lior Pachter). US citizen, white, male, no disabilities.
- Rebecca Shapley, UC Berkeley (worked with Brent Mishler). US citizen, white, female, no disabilities.
- Stephen Smith, Yale (student of Michael Donoghue). US citizen, white, male, no disabilities.
- Errol Strain, NCSU (student of Spencer Muse). US citizen, white, male, no disabilities.
- Jeet Sukumaran, Kansas (student of Mark Holder). US citizen, white, male, no disabilities.
- Shel Swenson, UT Austin (student of Tandy Warnow). US citizen, white, female, no disabilities.
- Kunal Talwar, UC Berkeley (student of Christos Papadimitriou and Satish Rao). Not US Citizen/permanent resident, Asian, male, no disabilities.
- Andres Varón, CUNY (student of Ward Wheeler). Male Hispanic/Latino, Not US citizen or permanent resident, no disabilties.

- Rutger Vos, U. British Columbia (student of Wayne Maddison). Not US citizen/permanent resident, white, male, no disabilities.
- Yifeng Zheng, U. Penn (student of Susan Davidson and Junhyong Kim). US citizen, Asian, male, no disabilities. Home Page: http://www.cis.upenn.edu/~yifeng.
- Derrick Zwickl, UT Austin (student of David Hillis). US Citizen, white, male, no disabilities.

Graduate students paid through other grants:

- Matt Ackerman; PhD student at Missouri State University, US Citizen, white, male, no disabilities. Funded by Google (Summer of Code participant), supervised by Mark Holder.
- Dan Adkins, UC Berkeley (student of Satish Rao). US Citizen, white, male, no disabilities.
- Stanislav Angelov, U. Penn (student of Sanjeev Khanna and Sampath Kannan). Not US citizen or permanent resident, White Male, no disabilities.
- Maud Artaud, UCSD (visiting student from France). Not US citizen/permanent resident, white, female, no disabilities.
- Jason Caravas; PhD student at Wayne State University, US Citizen, white, male, no disabilities. Funded by Google (Summer of Code participant), supervised by Rutger Vos.
- Guojing Cong, UNM (student of David Bader). Not US Citizen/permanent resident, Asian, male, no disabilities.
- Siobain Duffy, Yale U. (student of Paul Turner). US Citizen, white, male, no disabilities.
- Fan Ge, U. Penn (student of Junhyong Kim). Not US Citizen, Asian, male, no disabilities.
- Eric Gottlieb, M.S., UNM CS, of Bernard Moret. US Citizen, white, male, no disabilities.
- Boulos Harb, U. Penn (student of Sampath Kannan). Not US Citizen, white, male, no disabilities.
- Chris Harrelson, UC Berkeley (student of Satish Rao). US Citizen, white, male, no disabilities.
- Kris Hildrum, UC Berkeley (student of Satish Rao). US Citizen, white, female, no disabilities.
- Bonnie Kirkpatrick, UC Berkeley (student of Steve Evans). US Citizen, white, female, no disabilities.
- Mahesh Kulkarni, UNM (student of Bernard Moret). Not US citizen/permanent resident, Asian, male, no disabilities.
- Melanie Langlois, UCSD (visiting student from France). Not US citizen, white, female, no disabilities.
- Richard Liang, UC Berkeley (student of Steve Evans). US Citizen, Asian, male, no disabilities.
- Brian Moore, Yale U. (student of Michael Donoghue). US citizen, white, male, no disabilities.

- Monique Morin, UNM (student of Bernard Moret). US citizen, white, female, no disabilities.
- Anneke Padolina, UT-Austin (student of Randy Linder). US citizen, white, female, no disabilities.
- David Suarez Pascal; PhD student at UNAM, Not US Citizen, Latino Male (citizen and resident of Mexico), no disabilities. Funded by Google (Summer of Code participant), supervised by Mark Holder.
- Nicholas Pattengale, UNM (student of Bernard Moret). US Citizen, white, male, no disabilities.
- Sindhu Raghavan, UT-Austin (student of Tandy Warnow). Not US citizen, Asian, male, no disabilities.
- Peter Ralph, UC Berkeley (student of Steve Evans). US Citizen, white, male, no disabilities.
- Derek Ruths, Rice (student of Luay Nakhleh). US Citizen, white, male, no disabilities.
- Allan Sly, UC Berkeley (student of Steve Evans). US Citizen, white, male, no disabilities.
- Krister Swenson, UNM (student of Bernard Moret). US Citizen, white, male, no disabilities.
- Jijun Tang, UNM (student of Bernard Moret). Not US Citizen/permanent resident, Asian, male, no disabilities.
- Ruth Timme, UT Austin (student of Randy Linder). US Citizen, white, female, no disabilities.
- Fang Yue, U. South Carolina (student of Jijun Tang). Not US Citizen, Asian, male, no disabilities.
- David Zhao, UT Austin (student of Tandy Warnow). US Citizen, Asian, male, no disabilities.
- Lijuan Zhao, UNM (student of Bernard Moret). Not US Citizen/permanent resident, Asian, male, no disabilities.

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, US Citizen, white, male, no disabilities.
- Kevin Bullaughey, U Penn, Not US Citizen, white, male, no disabilities.
- Chris Crutchfield, UC Berkeley, US Citizen, white, male, no disabilities.
- Alex Jaffe, UC Berkeley, US Citizen, white, male, no disabilities.
- Sun Jin Lee, Yale U., Not US Citizen, Asian, male, no disabilities.

- Ving Ian Lei, UT Austin (student of Dan Miranker). Not US Citizen, Asian, female, no disabiites.
- Jenny Liu, UC Berkeley, US citizen, Asian, female, no disabilities.
- Erik Lewis, UC Berkeley, US citizen, white, male, no disabilities.
- Zack Mahdavi, UT Austin, US citizen, white, male, no disabilities.
- Diana Miachalek, UC Berkeley, US citizen, white, female, no disabilities.
- Kavya Rao, UCSD (intern at SDSC), US citizen, Asian, female, no disabilities.
- Apurva Shah, UC Berkeley, US citizen, Asian, male, no disabilities.
- Jennifer Vo, UCSD (intern at SDSC). US citizen, white, female, no disabilities.
- Yul Yang, Yale U. US citizen, Asian, male, no disabilities.

2.7 Other students

- Jorge Alva, high school student, intern at SDSC. US Citizen, Hispanic, male, no disabilities.

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 4.

3 Algorithms

3.1 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.

3.2 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 **?**, 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 **?** 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 **?**, for "Divide-And-Conquer Trees without ALignments", runs quickly, and can compute trees with higher accuracy than even SATé.

Below, we discuss each of the research contributions in turn, beginning with the highlights given above, and then continuing with the other contributions.

4 Software Development and Central Resource

4.1 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

4.2 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

5 Modeling and Simulations

5.1 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

5.2 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)

5.3 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.

6 Outreach Activity

6.1 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

6.2 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

7 Databases

7.1 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.

7.2 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.

7.2.1 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 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 **?**, SuperFine **?**, DACTAL **?**, Mega-phylogeny **?**, and other methods,
- TreeBASE-II **?**, 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 **?** 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" **?**.

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" **?**), 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" **?**. 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" **?**. 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" **?**, 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" **?**. 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" **?**. 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" **?**. 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" **?**) 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" **?**, 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" **?** 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 **?**, 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" **?**. 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" **?**. 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" **?**). 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" **?**). 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" **?**. 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" **?**, (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"**?**, 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" **?**. 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" **?**. 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" **?**. 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" **?**. 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" **?**. 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" **?**, 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" **?**. 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" **?**. 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" **?**.

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" **?**. 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" **?**. 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" **?**, 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" **?**. 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" **?**). 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" **?**. 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.

9 CIPRES publications

References

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

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.

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

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

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

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

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/.

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

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.

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.

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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/.

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

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.

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

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

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

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

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

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

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

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

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.