Report from the 2007 NESCent Hackathon: Comparative Phylogenetic Methods in R

Event dates: December 10-14, 2007

Participants (* event organizer):
On-site
Michael Alfaro, Washington State University
Charles Bell, University of New Orleans
Ben Bolker, University of Florida
Peter Cowan, University of California - Berkeley
Damien de Vienne, University of Paris
Richard Desper, National Center for Biotechnology Information
Luke Harmon, University of Idaho
Christoph Heibl, Ludwig-Maximilians Universitat, Munchen
Andrew Hipp, Morton Arboretum
Gene Hunt, Smithsonian Institution
Thibaut Jombart, University of Lyon
Steve Kembel*, University of California - Berkeley
Hilmar Lapp*, NESCent
Wayne Maddison, University of British Coiumbia
Peter Midford, University of Kansas
Brian O'Meara*, NESCent
David Orme, Imperial College
Emmanuel Paradis, Research Institute for Development
Sam Price*, NESCent
Dan Rabosky, Cornell University
Ryan Scherle, NESCent
Brian Sidlauskas, NESCent
Stacey Smith, Duke University
Dave Swofford, Duke University
Todd Vision*, NESCent
Peter Waddell, University of South Carolina
Amy Zanne*, NESCent

By teleconference
Marguerite Butler, University of Hawaii
Joe Felsenstein, University of Washington

Project wiki (public): http://hackathon.nescent.org/R_Hackathon_1.

Goals

The R statistical analysis package has emerged as a popular platform for implementation of powerful comparative phylogenetic methods to understand the evolution of organismal traits. This event was designed to bring together active R developers as well as end-users working on the integration of comparative phylogenetic methods within R to actively address issues of data exchange standards, code interoperability, usability, documentation quality, and the breadth of functionality for comparative methods available within R. The idea originated from a whitepaper submitted by NESCent postdocs Amy Zanne and Sam Price.

Summary of activities

Prior to the hackathon, calls for input and applications to participate in the meeting were solicited from the end-user community. A variety of information was obtained from developers and end-user prior to the event and organized on the wiki, including: What new software projects are underway in R or in the area of comparative phylogenetic methodology? What are the inputs and outputs to each of the packages that participants are involved in? What formats do they support or require? What are the internal data representation models? What technical training do developers need that can be provided by on-site tutorials? What are the top priority goals on which to focus at the hackathon? Through the wiki, a mailing list, and conference calls, participants had extensive communications prior to the event, and a considerable amount of preparation was done to ensure efficient software development activities at the event.

The meeting opened with introductions, presentations on needs as perceived by end-user perspective, the overlap and missing functionality among the available packages, and current and possible future data-representation standards. The participants then decided upon six subgroups with specific charges. Three were devoted to implementing missing functionality for estimation of (1) divergence times, or (2) rates of clade diversification, and (3) patterns of trait evolution. Two were focused on improving data representation and exchange through (4) design of a new internal data representation standard and (5) work on input and output data formats. One group focused on (6) interoperability with the Mesquite software package. The final group was tasked with (7) end-user documentation, including a help wiki for newbies, an R taskvew specifically for comparative phylogenetic analysis, and various specialized howto documents. A number of "bootcamps", short training sessions, were held on technologies that were relevant to the hackathon: (a) the use S4 classes in R, (b) the use of the R package Sweave for writing "vignettes" (R tutorials containing working code), (c) version control, and (d) numerical optimization techniques, and (e) communication from and to Mesquite using R. Most of the remainder of the meeting was devoted to work on the specific goals of each subgroup. Short standups were held each morning in which each subgroup reported on progress and roadblocks to the other groups. A discussion section on the penultimate day focused on future plans.

Outcomes

Diversification subgroup

The diversification subgroup implemented the binary-state speciation / extinction (BiSSE) model and key innovation test (Ree 2005), which were previously unavailable in R. Work on these methods is ongoing but will be incorporated into an existing R package. Additionally, major updates were made to the laser package, including the addition of a set of functions to test for diversification rate shifts using combined phylogenetic and taxonomic data. Several vignettes were written to illustrate diversification rate analyses.

Divergence time estimation subgroup

This subgroup compared existing methods for divergence time estimation in R, as well as implementing new methods. The penalized likelihood method in ape was recoded, and new models of rate change other than the Poisson approximation were implemented. Existing R code was implemented in C to allow cross-checking. Other outcomes include new tree parameterization methods, and a howto documentation file for estimating divergence times in R.

Trait evolution subgroup

This subgroup compared existing methods for continuous character analysis in R, finding that most methods produced similar results, with some caveats that are documented on the wiki. Numerous improvements to functionality were implemented and have been released in the geiger package, including new models of character evolution and improved reliability of results.

Internal data representation standards and I/O subgroup

A prerelease version of the new R phylobase package was released in January. This takes advantage of the work done at the hackathon on the S4 class, as well as basic functionality for multiple trees, and is intended to eventually include more comprehensive and robust I/O functionality than current packages. It obviates the need for ad hoc tracking of leaf labels in analysis methods. One can identify nodes graphically and query trees independent of the underlying representation. Most of the current functionality consists of wrappers around ade4 and ape functions, but in time this will be moved to native phylobase functions. The package is hosted at Rforge (https://r-forge.r-project.org/projects/phylobase/) and the developers have a dedicated mailing list (phylobase-devel@nescent.org). This will be announced on evoldir and R mailing lists once it has undergone initial testing.

Interoperability between R and Mesquite subgroup

Substantial progress was made in enabling users to call Mesquite through R via a new Mesquite wrapper library (rmlink), which is installed by a new RMesquite package (this has dependencies on rJava and ape). Newick strings and phylo objects can be be converted to Mesquite Tree objects, R arrays and matrices to Mesquite character matrices, and Mesquite Tree and Character data objects to R phylo and matrix objects. NEXUS files can be read into a Mesquite "Project" from within R. And R users now have access to Mesquite's BiSSE likelihood calculation and ancestral state reconstruction methods, with standard R objects as arguments.

Documentation subgroup subgroup

Initial content was developed for the help wiki (which will eventually be hosted at http://r-phylo.org), vignettes, and a task view. During the course of the hackathon the documentation subgroup worked with various other subgroups to improve documentation and develop tutorials for existing code.

Cross-cutting accomplishments subgroup

A poster describing the hackathon and promoting the use of R for comparative phylogenetic analysis is being submitted by Brian O'Meara for the 2008 Evolution meetings.

Future plans

Hackathon participants continue to communicate regularly using the NESCent mailing list. Additionally, a special interest group mailing list (R-sig-phylo) has been established for developers and end-users to field and answer questions on all the relevant packages and methods. Several of the packages are continuing to use the NESCent source code repository for version control. Participants from the first three subgroups plan to use this repository for sharing benchmark data files to allow cross-package testing. Users group meetings are planned for the 2009 SICB meetings (to be organized by Alfaro) and the 2009 Evolution meetings (to be organized by Harmon).

A proposal was submitted to NSF to establish a summer course in R for phyloinformatics in Idaho (instructors would include Alfaro, Butler, Harmon and Rabosky), and NESCent provided a letter offering financial support should the proposal be funded. Discussions are underway to include an R module in the NESCent phyloinformatics summer course. NESCent anticipates offering one or more student summer projects related to the hackathon goals should NESCent be a mentoring organization in the 2009 Google Summer of Code.

Plans for publications are still in flux, but participants discussed several possibilities: a PLoS computational biology primer on the comparative phylogenetic methods, a more technical paper about the phylobase class for Evolutionary Bioinformatics, and a Bioinformatics application note about the R/Mesquite interaction when that is more mature, and a review of the wide variety of (poorly characterized) methods for trait evolution that includes some benchmark comparisons.