rexpokit and cladoRcpp: R packages integrating FORTRAN and C++ for faster matrix exponentiation and likelihood calculations in historical biogeography

Nicholas J. Matzke, Postdoctoral Fellow, NIMBioS (National Institute of Mathematical, www.nimbios.org)
iEvoBio 2014, Raleigh, NC, Tuesday, June 24, 2014
Software Bazaar, Room 402, 3:15-5:00 pm
Coauthor: Drew Schmidt, University of Tennessee

DEC (LnL=-34.5)

DEC+J (LnL=-20.9)
BioGeoBEARS: testing biogeographical models

LAGRANGE or DIVA

Founder-event

Vicariance

Range expands
Test biogeographic processes in BioGeoBEARS

I implemented this method in *three* R packages:

- **Package #1**, `rexpokit`, handles anagenic events requiring matrix exponentiation (FORTRAN EXPOKIT library, C++, Rcpp)
Test biogeographic processes in BioGeoBEARS

I implemented this method in three R packages:

Package #1, \texttt{rexpokit}, handles anagenic events requiring matrix exponentiation (FORTRAN \texttt{EXPOKIT} library, C++, Rcpp)

Package #2, \texttt{cladoRcpp}, handles cladogenesis events (C++, Rcpp)

\texttt{Test biogeographic processes in BioGeoBEARS}
Test biogeographic processes in BioGeoBEARS

I implemented this method in *three* R packages:

- **Package #3, BioGeoBEARS**, does the ML/Bayesian searches, model testing, etc.

**BioGeographic Bayesian Evolutionary Analysis with R Scripts**

TRY IT YOURSELF!

http://phylo.wikidot.com/biogeobears
Dealing with slowness #1: Matrix exponentiation

\[ e^{Q\nu} = P(\nu) \]

\( \nu \) = branch length
\( Q = Q \) matrix
\( P = \text{Probability transition matrix} \)
Dealing with slowness #1: Matrix exponentiation

\[ e^{Q\nu} = P(\nu) \]

DNA: 4x4 Q matrix
Amino acids: 20x20 matrix
Biogeography: # ranges x # ranges matrix
Number of states in biogeography

\# states = 2^{\text{number of areas}}

2 areas = 2^2 = 4 states
4 areas = 2^4 = 16 states
8 areas = 2^8 = 256 states
10 areas = 2^{10} = 1024 states

Exponentiating a 1024x1024 matrix is slow
Dealing with slowness #1: Matrix exponentiation

Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*

Cleve Moler†
Charles Van Loan‡

Abstract. In principle, the exponential of a matrix could be computed in many ways. Methods involving approximation theory, differential equations, the matrix eigenvalues, and the matrix characteristic polynomial have been proposed. In practice, consideration of computational stability and efficiency indicates that some of the methods are preferable to others, but that none are completely satisfactory.
Most of this paper was originally published in 1978. An update, with a separate bibliography, describes a few recent developments.

Key words. matrix, exponential, roundoff error, truncation error, condition

AMS subject classifications. 15A15, 65F15, 65F30, 65L99

PII. S0036144502418010
Package #1, `rexpokit`, handles anagenic events requiring matrix exponentiation (FORTRAN EXPOKIT library, C++, Rcpp)
R Package: rexpokit

rexpokit: R wrappers for EXPOKIT; other matrix functions

This package wraps some of the matrix exponentiation utilities from EXPOKIT (http://www.maths.uq.edu.au/expokit/), a FORTRAN library that is widely recommended for matrix exponentiation (Sidje RB, 1998. "Expokit: A Software Package for Computing Matrix Exponentials." ACM Trans. Math. Softw. 24(1): 130-156). EXPOKIT includes functions for exponentiating both small, dense matrices, and large, sparse matrices (in sparse matrices, most of the cells have value 0). Rapid matrix exponentiation is useful in phylogenetics when we have a large number of states (as we do when we are inferring the history of transitions between the possible geographic ranges of a species), but is probably useful in other ways as well.

Version: 0.24.1
Depends: methods, SparseM, Rcpp (≥ 0.9.10)
LinkingTo: Rcpp
Published: 2013-07-15
Author: Nicholas J. Matzke [aut, cre, cph], Roger B. Sidje [aut, cph]
Maintainer: Nicholas J. Matzke <matzke at berkeley.edu>
License: GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL: http://phylo.wikidot.com/rexpokit
NeedsCompilation: yes
Citation: rexpokit citation info
Materials: NEWS
CRAN checks: rexpokit results

Downloads:

Reference manual: rexpokit.pdf
Package source: rexpokit_0.24.1.tar.gz
Windows binaries: r-devel: rexpokit_0.24.1.zip, r-release: rexpokit_0.24.1.zip, r-oldrel: rexpokit_0.24.1.zip
OS X Snow Leopard binaries: r-release: rexpokit_0.24.1.tgz, r-oldrel: rexpokit_0.24.1.tgz
Expokit is a software package that provides matrix exponential routines for small dense or very large sparse matrices, real or complex. Here you will find the source code in Fortran and Matlab. The native Fortran version is embeddable in C/C++.

**Understanding Expokit**

1. \(w(t) = \exp(tA)v\) is the analytic solution of the homogeneous ODE problem: \(w'(t) = Aw(t), w(0) = v\).

2. \(w(t) = \exp(tA)v + t\phi(tA)u\), where \(\phi(x) = (\exp(x)-1)/x\), is the analytic solution of the nonhomogeneous ODE problem: \(w'(t) = Aw(t) + u, w(0) = v\).

Expokit handles both cases.

**From Computing Reviews:** "This amazingly complete paper gives the theoretical background of the methods, presents applications, provides computational results for a standard set of pathological matrices, and closes with a mini-user's manual (...) It is entirely self-contained (...) Here we are given several sure ways to compute (...) exponentials, and these with rather sharp error bounds."

**Moler and Van Loan:** "The most extensive software for computing the matrix exponential that we are aware of is Expokit." [Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review*, Vol 45, No 1, March 2003.]"
R Package: rexpokit

Dense matrix exponentiation:
EXPOKIT's DGPADM —
(faster than the other R options, esp. for large matrices)

Sparse matrix exponentiation:
EXPOKIT’s DMEXPV

(DMEXPV’s imprecision affected ML analysis, needs more exploration)
R Package: rexpokit

Use of multicore mapply() to farm out the matrix exponentiations on the branches
Dealing with slowness #2: conditional probabilities at cladogenesis
Number of possible cladogenetic range-inheritance events

\[ \# \text{ states} = 2^{\text{number of areas}} \]

2 areas = 4 states = 4-1 = 3 non-null states

At every node,
3 possible ancestor states,
3 possible left descendants,
3 possible right descendants

\[ = 3 \times 3 \times 3 = 27 \text{ combinations to assign probabilities to} \]
Number of possible cladogenetic range-inheritance events

\[ \# \text{states} = 2^{\text{number of areas}} \]

9 areas = 512 states = 512 - 1 = 511 non-null states

At every node,

511 possible ancestor states,
511 possible left descendants,
511 possible right descendants

\[ = 511 \times 511 \times 511 = 133,432,831 \text{ combinations to assign probabilities to} \]
Problem can be simplified, according to your cladogenesis model

E.g., range AB giving rise to:

left = CDEFG, right = DEFGH

…should get probability 0.

But, searching intelligently involves lots of nested for-loops
R Package: cladoRcpp

cladoRcpp: C++ implementations of phylogenetic cladogenesis calculations

This package implements in C++/Rcpp various cladogenesis-related calculations that are slow in pure R. These include the calculation of the probability of various scenarios for the inheritance of geographic range at the divergence events on a phylogenetic tree, and other calculations necessary for models which are not continuous-time markov chains (CTMC), but where change instead occurs instantaneously at speciation events. Typically these models must assess the probability of every possible combination of (ancestor state, left descendent state, right descendent state). This means that there are up to (# of states)^3 combinations to investigate, and in biogeographical models, there can easily be hundreds of states, so calculation time becomes an issue. C++ implementation plus clever tricks (many combinations can be eliminated a priori) can greatly speed the computation time over naive R implementations. CITATION INFO: This package is the result of my Ph.D. research, please cite the package if you use it! Type: citation(package="cladoRcpp") to get the citation information.

Version: 0.14.4
Imports: Rcpp
LinkingTo: Rcpp, RcppArmadillo
Published: 2014-05-18
Author: Nicholas J. Matzke [aut, cre, cph]
Maintainer: Nicholas J. Matzke <matzke at berkeley.edu>
License: GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL: http://phylo.wikidot.com/biogeobears
NeedsCompilation: yes
Materials: NEWS
CRAN checks: cladoRcpp results

Downloads:
Reference manual: cladoRcpp.pdf
Package source: cladoRcpp_0.14.4.tar.gz
Windows binaries: r-devel: cladoRcpp_0.14.4.zip, r-release: cladoRcpp_0.14.4.zip, r-oldrel:
Package #2, `cladoRcpp`, handles cladogenesis events (C++, Rcpp)
BioGeoBEARS speed

Small analysis (4 areas, 16 states)

C++ Lagrange: < 1 second
Python Lagrange: a few sec.
BioGeoBEARS: ~10 seconds
BioGeoBEARS speed

Large analysis (10 areas, 6 max, 848 states)

C++ Lagrange: 208 min.
Python Lagrange: won’t run

BioGeoBEARS single-core: 187 minutes

BioGeoBEARS twelve cores: 22 minutes
r expokit improvements

Coauthor Drew Schmidt, University of Tennessee

Gets a 40% speed improvement via various fixes in the R/FORTRAN interaction

New version of r expokit going on CRAN this week

See his talk at UseR conference, June 30-July 3, UCLA
BioGeoBears

UPDATE 6/23/2013: I have just uploaded my Evolution 2013 talk, which was Sunday, 11:45 am at the Ernst Mayr symposium. The PDF can be found at the bottom of the file. I have also uploaded a zipped source package for BioGeoBEARS. This is more up-to-date than github. It is the same as what I just submitted to CRAN as 0.1.1.

See also by iEvoBio Lightning Talk, 1:30 pm Tuesday I think.

The goal of my R package "BioGeoBEARS" — short for "BioGeography with Bayesian Evolutionary Analysis in R Scripts", if you were wondering — is to infer the history of species' geographic ranges on phylogenetic trees. It implements the LAGRANGE model (2 free parameters) as well as models with fewer or more free parameters. Standard model-testing procedures may then be applied.

The default implementation is actually ML (maximum likelihood), for comparison with LAGRANGE, but there is a Bayesian option making use of LaplacesDemon — contact me at matzke@berkeley.edu, if you can't figure it out from the functions.

See also my IBS poster.

Author: Nicholas J. Matzke

Download BioGeoBEARS

A draft version of BioGeoBEARS is available for download on github at: https://github.com/nmatzke/BioGeoBEARS. It is easier to download it from within R (see the install_github command, below).
Accurate when simulated under Yule process

Yule simulations ($\lambda=0.3, \mu=0, \alpha=0, \omega=0$)

d

true parameter value

DEC inference

DEC+J inference

j

Proportion correct

$\Delta$AICc

Accurate when simulated under Yule process
Accurate even with simulation under SSE

SSE simulations ($\lambda=0.3$, $\mu=0.3$, $\alpha=1$, $\omega=-1$)

- True parameter value
- DEC inference
- DEC+J inference

Proportion correct

$\Delta$AICc

True model: +J +J +J DEC +J +J +J DEC +J +J +J
Diversification of an Afro-Asian songbird clade (Erythropygia-Copyscyphus) reveals founder-event speciation via trans-oceanic dispersals and a southern to northern colonization pattern in Africa

Gary Voolker A1, Joshua V. Pralhala1, Jerry W. Huntley2 and Rauni C.K. Bowie1*

1Department of Wildlife and Human Sciences and Redwood Research and Teaching Collections, Texas A&M University, College Station, TX 77843, USA
2Perry R. Hapeman Institute, UCT-CNSE Centre of Excellence, University of Cape Town, Rondebosch 7701, South Africa

ABSTRACT

Erythropygia scrubels and their allies are distributed throughout Africa, from Southern Africa, Asia, Madagascar and the Seychelles. This broad distribution, as well as the distribution of Erythropygia taxa across Asia, presents an interesting opportunity to explore the mechanisms by which this biogeographic distribution was achieved. Multilocus sequence data (3150 base pairs from two mitochondrial and two nuclear genes) were generated for all species of Erythropygia and Copyscyphus scrubels, as well as four genera previously known to include Erythropygia parabulchera. Using model-based phylogenetic approaches implemented in BEAST and BEGMA, we constructed a time-calibrated molecular phylogenetic hypothesis for the lineage. Molecular area reconstructions were performed on the phylogeny using probabilistic approaches implemented in Landmark and BEGMA. Our results confirm that Erythropygia is not monophyletic, with southern African and Indian Ocean Islands distributed species. Overall, the Erythropygia and allies clad emerged in Africa in the last 7.0 Ma. Subsequent waves of trans-oceanic dispersals occurred in an initial colonization of Southeast Asia, and an eastward migration of colonizations from Southeast Asia to the Seychelles, from there to Madagascar, and from Indian Ocean Islands back to Southeast Asia. Within the two clades of Erythropygia, ancestral area reconstructions within Africa indicate a Southern African origin, with subsequent lineage divergence in each clad radiating northwest colonization. Overall, this clad of two-branch origins shows a remarkable number of trans-oceanic colonization events, that were possibly facilitated by wind-driven dispersal, repeated Africa to Asia colonizations, and the ability for birds to disperse to new areas, are exceptional rare in birds. Also now is our finding that colonization patterns in Africa indicate a southern to northern progression.© 2014 Elsevier Inc. All rights reserved.
Model Selection in Historical Biogeography Reveals that Founder-event Speciation is a Crucial Process in Island Clades

See also:

In review: email me at matzke@nimbios.org for this

Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing

Nicholas Joseph Matzke

PhD Thesis, Department of Integrative Biology, University of California, Berkeley. 3060 Valley Life Sciences Building, Berkeley, CA, 94720; matzke@berkeley.edu.


Abstract. Historical biogeography has been characterized by a large diversity of methods and unresolved debates about which processes, such as dispersal or vicariance, are most important for explaining distributions. A new R package, BioGeoBEARS, implements many models in a common likelihood framework, so that standard statistical model selection procedures can be applied to let the data choose the best model. Available models include a likelihood version of DIVA ("DIVAlike"), LAGRANGE’s DEC model, and BAYAREA, as well as "+I" versions of these models which include founder-event speciation, an important process left out of most inference methods. I use BioGeoBEARS on a large sample of island and non-island clades (including two fossil clades) to show that founder-event speciation is a crucial process in almost every clade, and that most published datasets reject the non-I models currently in widespread use. BioGeoBEARS is open-source and freely available for installation at the Comprehensive R Archive Network at http://CRAN.R-project.org/package=BioGeoBEARS. A step-by-step tutorial is available at http://phylo.wikidot.com/biogeobears.

Keywords. cladogenesis, DIVA, founder-event speciation, historical biogeography, jump dispersal, LAGRANGE, phylogenetics
Thanks! I welcome questions/comments/collaborations at: matzke@nimbios.org (also: seeking a job!)

Thanks especially to:

NIMBioS
Brian O’Meara
Jeremy Beaulieu
Katie Massana
Michael Landis

Ph.D. committee
John Huelsenbeck
Tony Barnosky
David Jablonski
Roger Byrne

Systematic Biology
editors & reviewers

Funding: NIMBioS
NSF “Bivalves in Time and Space”
UC Berkeley Wang Fellowship
UC Berkeley Tien Fellowship
Google Summer of Code

NIMBioS

TRY IT YOURSELF AT:
http://phylo.wikidot.com/biogeobears
DEC: Dispersal-extinction cladogenesis

Cladogenesis -- range evolution at speciation events (LAGRANGE model)

Probabilities per speciation event:

| Left:  | A | A | A | A | A | B | B | B | B | B | C | C | C | C | C | C | C | C | B | B | B | B | B |
| Right: A | B | C | AB | BC | AC | ABC | A | B | C | AB | BC | AC | ABC | A | B | C | AB | BC | AC | ABC | A | B | C | AB | BC | AC | ABC |

<table>
<thead>
<tr>
<th>Starting C</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>BC</th>
<th>AC</th>
<th>ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ending range</td>
<td>sym</td>
<td>sym</td>
<td>sym</td>
<td>sym</td>
<td>sym</td>
<td>sym</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>AC</th>
<th>ABC</th>
<th>AB</th>
<th>BC</th>
<th>AC</th>
<th>ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
</tr>
<tr>
<td>B</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
</tr>
<tr>
<td>AB</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
</tr>
<tr>
<td>BC</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
</tr>
<tr>
<td>AC</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
</tr>
<tr>
<td>ABC</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
<td>sub</td>
<td>vic</td>
</tr>
</tbody>
</table>

Count: 1 1 6 6 6 12
DEC+J: Dispersal-extinction cladogenesis PLUS founder-event speciation

Probabilities per speciation event:

- Sympatric
- Subset
- Vicariance
- Founder Event

Probabilities per speciation event:

<table>
<thead>
<tr>
<th>starting range</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>AC</th>
<th>ABC</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>end range</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>ABC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
</tbody>
</table>

DEC+J: Dispersal-extinction cladogenesis PLUS founder-event speciation
BayArea cladogenesis model

Cladogenesis -- “no event” during speciation (i.e., pure continuous time)

Probabilites per speciation event:

Sympatric speciation (range duplication)