Probabilistic Programming for Statistical Phylogenetics

Fredrik Ronquist
Swedish Museum of Natural History, Stockholm, Sweden
Phylogenetics

- Modeling and inference involving evolutionary trees
- Widely used across the life sciences
- Example applications:
  - Virus transmission pathways
  - Epidemiological models including genetic information
  - Identification of pathogens
  - Relationships among organisms
  - Divergence time estimation
  - Dynamics of molecular evolution
  - Positive selection analysis
  - Center of origin analysis
  - Patterns of diversification and extinction
  - Biogeography
Million years ago
Table 1. Most-cited articles over the period 2001-2013.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Article cited</th>
<th>Times cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sheldrick, G.M. (2008). A short history of SHELX. <em>Acta Crystallographica Section A</em>, 64, 112-122.</td>
<td>34,533</td>
</tr>
</tbody>
</table>
Statistical Phylogenetics

- Statistical approaches increasingly important:
  - Difficult problems requiring accurate and unbiased inference (e.g., structure of rapid radiations)
  - More aspects of the evolutionary process being examined (structural dependencies, biogeography etc)
  - Combination of background knowledge and sequence information (e.g., divergence time estimation)
- Modeling explosion, especially in the Bayesian context
- Challenging for empiricists to communicate and correctly understand models
- Challenging for developers of inference software
MrBayes: Bayesian Inference of Phylogeny

MrBayes is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters.

Program features include:

- A common command-line interface across Macintosh, Windows, and UNIX operating systems;
- Extensive help available from the command line;
- Analysis of nucleotide, amino acid, restriction site, and morphological data;
- Mixing of data types, such as molecular and morphological characters, in a single analysis;
- Easy linking and unlinking of parameters across data partitions;
- An abundance of evolutionary models, including 4x4, doublet, and codon models for nucleotide data and many of the standard rate matrices for amino acid data;
- Estimation of positively selected sites in a fully hierarchical Bayesian framework;
- Full integration of the BEST algorithms for the multi-species coalescent.
- Support for complex combinations of positive, negative, and backbone constraints on topologies;
MrBayes script

#NEXUS

begin mrbayes;

execute data.nex;

outgroup Ibalia;
charset morphology = 1-166;
charset molecules = 167-3246;
charset COI = 167-1244;
charset EF1a = 1245-1611;
charset LWRh = 1612-2092;
charset 28S = 2093-3246;
partition favored= 5: morphology, COI, EF1a, LWRh, 28S;

set partition=favored;
lset applyto=(1) rates=gamma;
lset applyto=(2,3,4,5) rates=invgamma nst=mixed;
unlink revmat=(all) pinvar=(all) shape=(all) statefreq=(all);
prset ratepr=variable;

end;
### Models supported by MrBayes 3 (simplified)

#### Data Type: Restriction 0 - 1
- **State Frequencies (Substitution Rates):** fixed/estimated (Dirichlet) prset statefreqpr
- **Across-Site Rate Variation:** equal/gamma lset rates
- **Coding Bias:** all/variable/nopresentsites/noabsencesites lset coding

#### Data Type: Standard 0 - 9
- **State Frequencies (Substitution Rates):** equal/estimated (SymmDir) prset symdirihyperpr
- **Across-Site Rate Variation:** equal/gamma lset rates
- **Coding Bias:** all/variable/informative lset coding
- **Misc.:** unordered/ordered ctype

#### Data Type: DNA
- **Model Type:** 4by4 lset nucmodel
  - **State Frequencies:** fixed/est.(Dirichlet) prset statefreqpr
  - **Substitution Rates:** F81/HKY/GTR lset nst=1/2/6
  - **Across-Site Rate Variation:** equal/gamma/propinv/invgamma/adggamma lset rates
  - **Across-Tree Rate Variation:** yes/no lset covarion

- **Model Type:** doublet lset nucmodel
  - **State Frequencies:** fixed/est. (Dirichlet) (over 16 states) prset statefreqpr
  - **Substitution Rates:** F81/HKY/GTR lset nst=1/2/6
  - **Across-Site Rate Variation:** equal/gamma/propinv/invgamma lset rates

- **Model Type:** codon lset nucmodel
  - **State Frequencies:** fixed/est. (Dirichlet) (over 61 states) prset statefreqpr
  - **Substitution Rates:** F81/HKY/GTR lset nst=1/2/6
  - **Across-Site Rate Variation:** equal/Ny98/M3 lset omegavar
- **Across-Site Omega Variation**
## Models supported by MrBayes 3 (simplified)

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Model Type</th>
<th>State Frequencies</th>
<th>Substitution Rates</th>
<th>Across-Site Rate Variation</th>
<th>Across-Tree Rate Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein A - Y</td>
<td>Equalin/GTR prset aamodelpr</td>
<td>fixed/est. (Dirichlet) prset statefreqpr</td>
<td>fixed/est. (Dirichlet) prset aarevmatpr</td>
<td>equal/gamma/propinv/invgamma/adgamma</td>
<td>yes/no Iset covarion</td>
</tr>
<tr>
<td></td>
<td>Poisson/Jones/Dayhoff/Mtrev/Mtmam/Wag/Rtrev/Cprev/Vt/Blossum/mixed prset aamodelpr</td>
<td>fixed/mixed</td>
<td>fixed/mixed</td>
<td>equal/gamma/propinv/invgamma/adgamma</td>
<td>yes/no Iset covarion</td>
</tr>
</tbody>
</table>

### Parameter Variation Across Partitions

- **Topology Models**
  - unconstrained/constraints/fixed constraint prset topologypr
  - Fixed prset brlenspr
  - Unconstrained prset brlenspr
  - Clock prset brlenspr

- **Brlens Type**
  - Fixed prset brlenspr
  - Unconstrained prset brlenspr

- **Brlens Prior**
  - Exponential/Uniform prset brlenspr
  - Uniform prset brlenspr
  - Coalescence prset brlenspr
  - Birth-Death prset brlenspr

- **Inferring Site Parameters**
  - ancstates/possel/siteomega/siterate report

- **Additional Parameters**
  - prset (Iset for diploidy)
  - Treeheight
  - Coalescence
  - Speciation Extinction
  - Treeheight
  - Sampleprob

- **Clockrate Variation**
  - strict/cpp/ibr/bm prset clockratepr

- **Dating Constraints**
  - unconstrained/calibrated prset nodeagepr
  - prset treeagepr
CAN'T YOU DO ANYTHING RIGHT?!?
Hierarchical Normal Model

\[ \mu \sim \text{Unif}(a, b) \]
\[ \sigma \sim \text{Exp}(c) \]
\[ x \sim \text{iid Norm}(\mu, \sigma) \]
Tree model

Birth-death model:

- Birth rate $\lambda$
- Death rate $\mu$
- Root time $t_{\text{Mrca}} = t_{11}$

The birth-death model induces a probability distribution on

- Topology $\mathcal{T}$
- Speciation times $t$

Given a substitution rate $r$, branch lengths $b$ are given by

$$b_i = r(t_{a_i} - t_i)$$
Substitution model

Observation error usually ignored, that is, it is assumed that $y_i = s_i$ for all leaves $i$. 

Ancestral states

Tip states

Substitution model
DNA sequences are drawn iid from a discrete-state continuous-time Markov chain over four nucleotides: A, C, G, T

\[ Q = \begin{pmatrix}
    - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\
    \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\
    \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\
    \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & -
\end{pmatrix} \]

Instantaneous rate matrix for the General Time Reversible (GTR) substitution model

\[ \pi \quad \text{Stationary state frequencies} \]

\[ r \quad \text{Exchangeability rates} \]
Deterministic node
\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]

\[ v \]

\[ v \]

\[ \alpha \]

\[ \pi \]

\[ r \]

\[ Q \]

\[ S_{i,j} \]
GTR Phylogeny Model
Tree Plate Representation
Modular Representation

Pivot variable
RevBayes Project

- Interactive computing environment intended primarily for Bayesian phylogenetic inference
- Uses a probabilistic programming language (PPL), Rev, for constructing probabilistic phylogenetic and evolutionary graphical models interactively, step by step
- Rev is similar to the modeling languages of BUGS, JAGS and STAN
- RevBayes provides generic computing machinery for simulation, MCMC inference and Bayesian model testing
RevBayes

Bayesian graphical models

About
Software
Introduction
Docs
Tutorials
Validation
Forum
Citation

http://www.revbayes.com

https://github.com/revbayes
Basic properties of the Rev language

# There are three kinds of statements in the language creating model variables

# 1. Arrow assignment (value assignment, create constant nodes)

> a <- 4          # Give a the value 4
> b <- sqrt(a)    # Give b the value of sqrt(a), that is, 2
> b              # Print the value of b
2

# 2. Equation assignment (create deterministic nodes)

> c := sqrt(a)    # Make c a dynamic function node evaluating sqrt(a)
> c              # Print the value of c
2
> a <- 9          # Give a the value 9
> b              # Print the value of b
2
> c              # Print the value of c
3
Basic properties of the Rev language

# 3. Tilde assignment (create stochastic variables (nodes))

> a ~ dnExp( rate = x )                # a is drawn from exp dist with rate = x
Basic properties of the Rev language

# -------------------------------
# Declaring and defining functions
# -------------------------------

> function foo ( x ) { x * x }
> foo( 2 )
4

# If you wish, you can specify types as well

> function PosReal foo ( Real x ) { x * x }

# Functions can be used to define deterministic nodes

> a := foo( b )

# -------------------------------
# Declaring and defining new types
# -------------------------------

> class myclass : Move {
+     Real myTuningParam;
+     procedure Real move( Real x ) { myTuningParam * x }
+ }

# Inheritance, function overriding and overloading
A complete MCMC analysis in Rev

\[
a \leftarrow -1.0 \\
b \leftarrow 1.0
\]

\[
\mu \sim \text{dnUnif}(a, b) \\
\sigma \sim \text{dnExp}(1.0)
\]

\[
\text{for } (i \text{ in } 1:10) \{ \\
    x[i] \sim \text{dnNorm}(\mu, \sigma) \\
    x[i].\text{clamp}(0.5)
\}
\]

\[
\text{mymodel} = \text{model}(\mu) \quad \# \text{Any stochastic node in the model works}
\]

\[
\text{mymcmc} = \text{mcmc(mymodel)}
\]

\[
\text{mymcmc.run(1000)}
\]
Interactive environment

> a
-1

> mu
-0.003889918

> str(a)

   _variable   = a
   _dagType    = Constant DAG node
   _children   = [ mu ]

> str(mu)

   _variable   = mu
   _dagType    = Stochastic DAG node
   _clamped    = FALSE
   _lnProb     = -0.693147
   _parents    = [ a, b ]
   _children   = [ x[1], x[2], x[3], x[4], x[5], x[6], x[7], x[8],
                  x[9], x[10] ]
Definition of a new phylogenetic model

Appr. 20 lines
# Read in data
myData <- read("data.nex")

# Apply model
myModel = zihengGTR(myData)

# Construct mcmc
myMCMC = mcmc(myModel)

# Run mcmc
myMCMC.run(10000)
Canarian endemic radiations

Millipedes of the genus *Dolichoiulus* (Diplopoda, Julida, Julidae, Pachyulinae)

46 endemic species in the Canary Islands
Model

Inference

Bayesian inference using MCMC sampling, accommodating uncertainty in all model parameters
Challenges

- No automated mechanism for generating message-passing for variable elimination of ancestral states (essential for MCMC performance) -> Delayed sampling
- No automated mechanisms for “looking into” and generating MCMC inference machinery for more generic constructs, like stochastic branching generating a phylogenetic tree
“Black-box” creation of tree model

#################################################
# PhyloCTMC Model #
#################################################

# The sequence evolution model
seq ~ dnPhyloCTMC(tree=psi, Q=Q, siteRates=sr, pInv=p_inv, type="DNA")

# Attach the data
seq.clamp(data)
Higher-order PPL for phylogenetics?

- Solves representation problem for tree topology as a random variable, and for models integrating tree-generating process and substitution process

- Challenges:
  - Static, compile-time variable elimination
  - Generating computationally efficient inference machinery

- Opportunities:
  - Generic simulation and inference machinery for a much wider class of models
  - New algorithms, like SMC, PMCMC etc

- Test case: BAMM model
BAMM Model (Bayesian Analysis of Macroevolutionary Mixtures): birth and death rates vary across the tree under a compound Poisson process.

No software for asymptotically exact sampling presented to date.

We are implementing the model using three higher-order probabilistic programming languages: WebPPL, Anglican and Birch.