National Institute of Allergy and Infectious Diseases

Phylogenetics and Sequence Analysis

Lecture 7: Molecular Evolutionary Analysis of Pathogens Using BEAST

Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biology Branch (BCBB)
Fall 2015

NIAID

Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biology Branch (BCBB)

Course Organization

- Building a clean sequence
- Collecting homologs
- Aligning your sequences
- Building trees
- **Further Analysis**

NIAID

Lecture Organization

- What's so special about pathogens?
- What is BEAST?
- A short tour of Bayesian MCMC analysis
- An overview of the BEAST package
- **BEAST Analysis Demo**
- Odds and ends

NIAID

What's so special about pathogens?

- Short generation time
- Rapid evolution
- Genotypes - easy, phenotypes - hard
- Large populations
- Structured populations
- Rigorous temporal sampling of genotypes

NIAID

What is BEAST?

- **B**ayesian **E**volutionary **A**nalysis **S**ampling **T**rees
- A collection of programs for performing Bayesian MCMC analysis of molecular sequences
- Can incorporate sample time information
- Can perform a broad range of other evolutionary analyses using sequence data.

NIAID

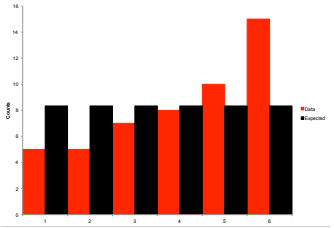
What is Bayesian analysis?

- Calculation of the probability of parameters (tree, substitution model) given the data (sequence alignment)
- $p(\theta|D)$ = (Likelihood x prior)/probability of the data
- $p(\theta|D) = p(D|\theta)p(\theta)/p(D)$

NIAID

What is Bayesian analysis?

Likelihood that this die is unbiased?



NIAID

Bayesian Analysis

Exploring the posterior probability distribution

Posterior probabilities of trees and parameters are approximated using Markov Chain Monte Carlo (MCMC) sampling

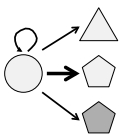
Markov Chain: A statement of the probability of moving from one state to another

NIAID

What is MCMC?

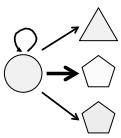
Markov Chain Monte Carlo

Markov chain



One link in the chain

Monte Carlo

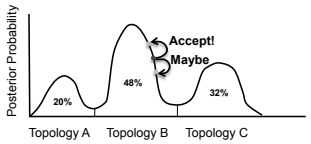


Choosing a link

NIAID

What is MCMC?

Markov Chain Monte Carlo: accept or reject?



$$\Pr(\text{accept}) = \min\left(1, \frac{\Pr(a)}{\Pr(A)} \times \frac{\Pr(X|a)}{\Pr(X|A)} \times \frac{\Pr(A|a)}{\Pr(a|A)}\right)$$

NIH National Institute of Allergy and Infectious Diseases

NIAID

What is BEAST?

The Programs:

- BEAUti - Creating XML input files
- BEAST - MCMC analysis of molecular sequences
- Tracer - Viewing MCMC output
- LogCombiner - Combining output files
- TreeAnnotator - Generate the consensus tree
- FigTree - Drawing a tree

NIH National Institute of Allergy and Infectious Diseases

NIAID

Different types of BEAST analyses

- Calculating a Bayesian coalescent phylogeny
- **Calculating a Time-Stamped Bayesian coalescent**
- Phylogeographic analysis (time and location data)
- Estimated population dynamics (Bayesian skyride and skygrid)
- Combined gene and species phylogeny estimate (*BEAST)

NIH National Institute of Allergy and Infectious Diseases

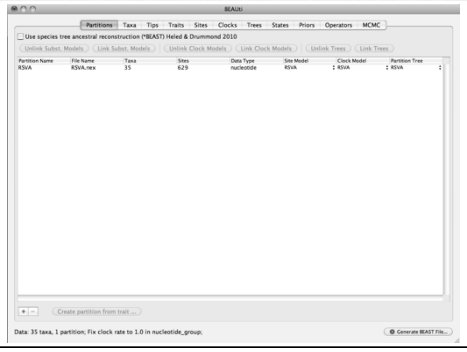
NIAID

Defining your analysis

- Prior knowledge of tree?
- Calibrating nodes?
- Substitution model?
- Effective population sizes?
- What priors to use?

NIAID

Setting up the analysis: BEAUTi



NIAID

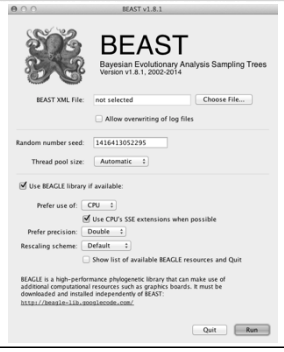
Setting up the analysis: BEAUTi

- Incorporate sample times
- Substitution model parameters
- Strict or relaxed clock?
- Tree prior
- Substitution model priors
- Adjustments from previous runs (operators)
- Setting the chain

NIAID

Running the analysis: BEAST

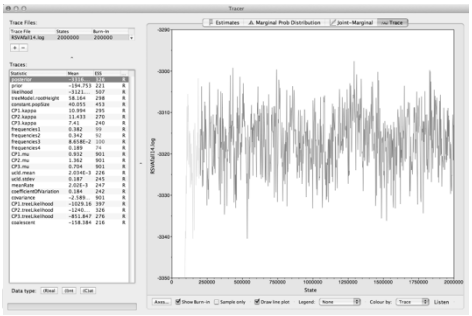
- Load your input file
- Use BEAGLE?
- That's it



National Institute of Allergy and Infectious Diseases

NIAID

Evaluating the analysis: Tracer



NIAID

Evaluating the analysis: Tracer

- Check for convergence
- Viewing parameter estimates from multiple runs
- Extract parameter estimates and statistics

National Institute of Allergy and Infectious Diseases

NIAID

Evaluating the analysis: Tracer

- What if my analysis didn't converge?
- Can I make multiple simultaneous runs?
  - Swarm on Biowulf

NIAID

Merging output files: LogCombiner

NIAID

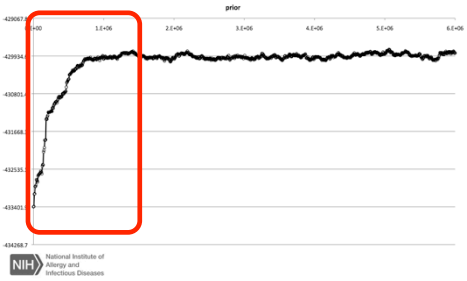
Merging output files: LogCombiner

- Log files vs Tree files
- Selecting files
- Specifying burn-in (number of states)
- Specifying subsampling
- Specifying output file

NIAID

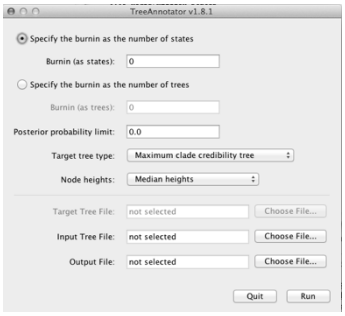
Merging output files: LogCombiner

- Burn in?



NIAID

Calculating the tree: TreeAnnotator



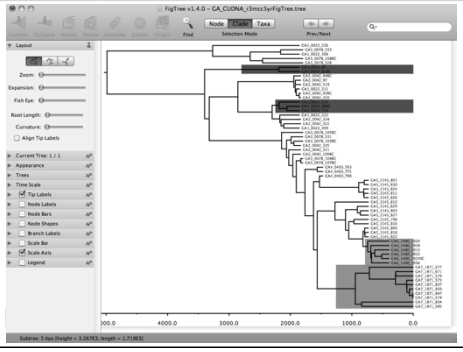
NIAID

Calculating the tree: TreeAnnotator

- Burn in? Number of trees or the number of steps.
- Tree Type: MCC, Max sum of CC, or target
- Node heights: target, mean, or median
- Specify input and output files

NIAID

Drawing trees: FigTree



NIAID

Drawing trees: FigTree

- Specifying additional values (esp. posterior probabilities)
- Tree appearance
- Ordering branches
- Re-rooting
- Exporting graphics

NIH National Institute of Allergy and Infectious Diseases

NIAID

Running BEAST

DEMO

NIH National Institute of Allergy and Infectious Diseases

NIAID

Running BEAST: Demo

- Site Models
- Substitution Models
  - HKY - Unequal base frequencies and transition/transversion rate ratio
    - Must specify prior and initial estimates for transition/transversion rate ratio
  - GTR - Unequal base frequencies and each substitution has its own rate parameter
    - Must specify prior and initial estimates for each substitution rate (relative to C-T rate)

NIH National Institute of Allergy and Infectious Diseases

NIAID

Running BEAST: Demo

- Site Models
  - Substitution models
  - Site heterogeneity models
- Get estimates from the program jModelTest

```
Model selected: TVM+I+G
 -lnL   =   1676.8109
  K     =   9
  AIC   =   3371.6218

  Base frequencies:
    freqA =              0.2259
    freqC =              0.3199
    freqG =              0.2405
    freqT =              0.2137
  Substitution model:
    Rate matrix
    R(a) [A-C] =         0.2494
    R(b) [A-G] =         4.8655
    R(c) [A-T] =         0.7435
    R(d) [C-G] =         0.3907
    R(e) [C-T] =         4.8655
    R(f) [G-T] =         1.0000
  Among-site rate variation
    Proportion of invariable sites (I) =       0.6508
    Variable sites (G)
     Gamma distribution shape parameter =      0.5913
```

NIH National Institute of Allergy and Infectious Diseases

NIAID

Running BEAST: Demo

Site Models

- Site heterogeneity models
  - Gamma
    - Modeling rate of change using a discrete gamma distribution
  - Invariant
    - Percent of non-variable sites in the data

NIH National Institute of Allergy and Infectious Diseases

NIAID

Running BEAST: Demo

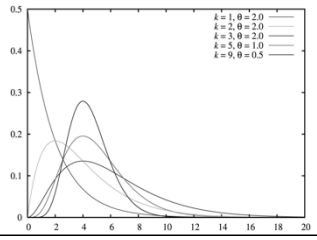
Site Models

- Site heterogeneity models
  - Gamma

Mean = kθ

Shape parameter = θ

Coefficient of Variation = 1/√θ

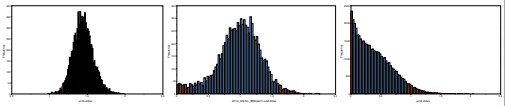


NIAID

Setting up the analysis: Models

Testing Models and Priors

Does the relaxed clock fit the data?



Examining the ucld.stdev distribution

NIAID

Running BEAST

Testing Models and Priors

Path Sampling/Stepping Stone analysis

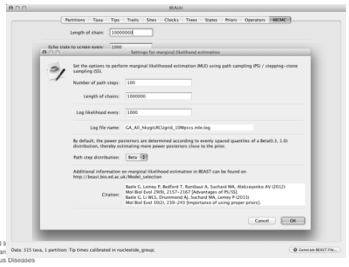
- Estimation of marginal likelihoods under different analysis parameters.
- Invoke on MCMC tab in BEAUti.
- Separate runs necessary for each changed parameter.
- Runs a complete MCMC analysis, then the *X* PS/SS iterations.

NIAID

Running BEAST

Testing Models and Priors

Path Sampling/Stepping Stone analysis



NIAD

Running BEAST

Testing Models and Priors

Path Sampling/Stepping Stone analysis
log marginal likelihoods

	Path Sampling	Stepping Stone
HKY/strict clock	-4725.85	-4728.68
HKY+gi/strict	-4515.99	-4518.05
HKY+gi/LN relaxed	-4436.10	-4438.75
GTR/strict clock	-4746.62	-4749.14
GTR+gi/strict	-4526.87	-4529.05
GTR+gi/LN relaxed	-4548.39	-4551.22

NIAD

Running BEAST: swarm on Biowulf

- Requires a .swarm file
  - A text file containing

```
beast beastJob_1.xml > beastJob_1out.txt
sleep 2; beast beastJob_2.xml > beastJob_2out.txt
sleep 4; beast beastJob_3.xml > beastJob_3out.txt
sleep 6; beast beastJob_4.xml > beastJob_4out.txt
sleep 8; beast beastJob_5.xml > beastJob_5out.txt
```

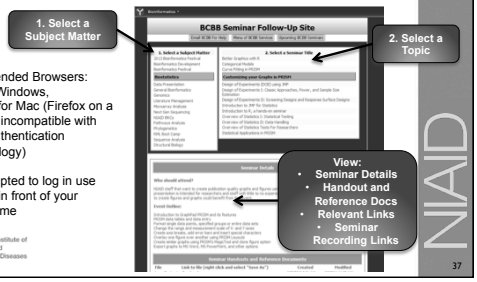
- Run in command line

```
[username]$ swarm -f beastInput.swarm –module BEAST
```

NIAD

Seminar Follow-Up Site

- For access to past recordings, handouts, slides visit this site from the NIH network: http://collab.niaid.nih.gov/sites/research/SIG/Bioinformatics/



Recommended Browsers:
- IE for Windows,
- Safari for Mac (Firefox on a Mac is incompatible with NIH Authentication technology)

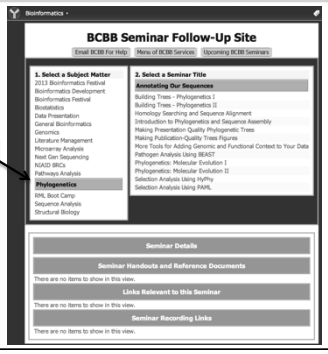
Login
- If prompted to log in use "NIH\" in front of your username

NIH National Institute of Allergy and Infectious Diseases

NIAID

37

Retrieving Slides/Handouts

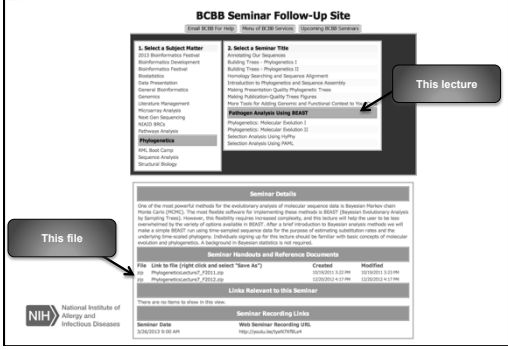


NIH National Institute of Allergy and Infectious Diseases

NIAID

38

Retrieving Slides/Handouts



NIH National Institute of Allergy and Infectious Diseases

NIAID

39

Questions?

Consultation & Advice | Software Development | Biocomputing Resources

ScienceApps@niaid.nih.gov

BCBB Bioinformatics and Computational Biosciences Branch
NIAID Office of Cyber Infrastructure and Computational Biology

NIH National Institute of Allergy and Infectious Diseases

NIAID

40

NIH National Institute of Allergy and Infectious Diseases

NIAID

41