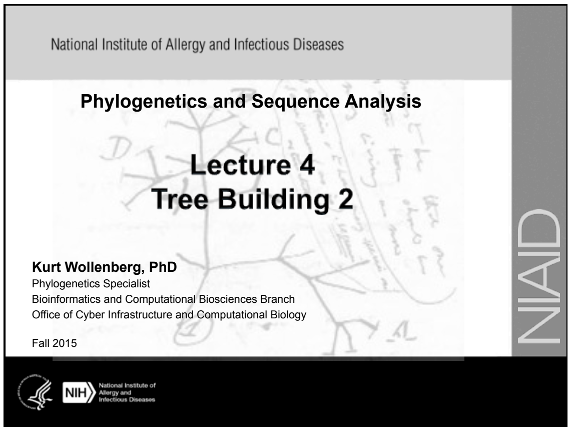


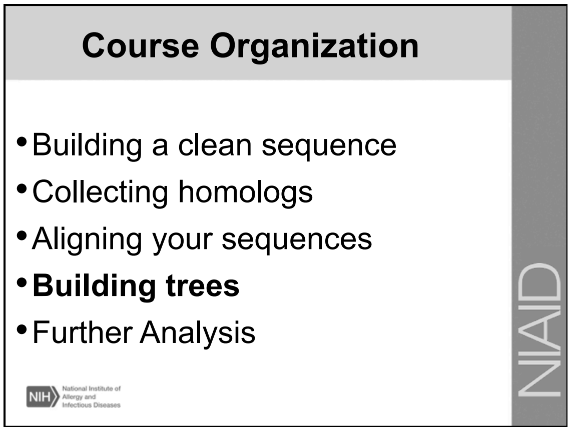
National Institute of Allergy and Infectious Diseases

Phylogenetics and Sequence Analysis

Lecture 4
Tree Building 2

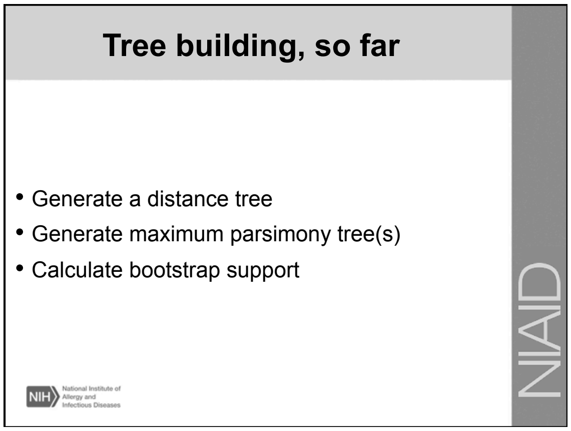
Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biosciences Branch
Office of Cyber Infrastructure and Computational Biology

Fall 2015



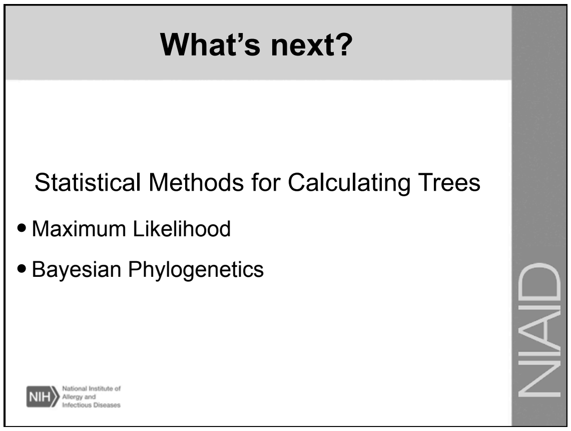
Course Organization

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



Tree building, so far

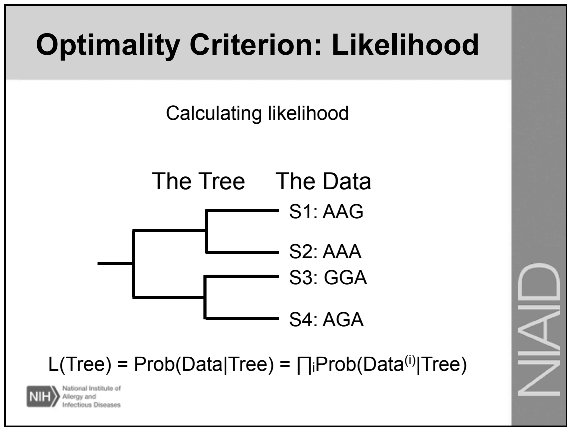
- Generate a distance tree
- Generate maximum parsimony tree(s)
- Calculate bootstrap support



What's next?

Statistical Methods for Calculating Trees

- Maximum Likelihood
- Bayesian Phylogenetics



Optimality Criterion: Likelihood

Calculating likelihood

The Tree The Data

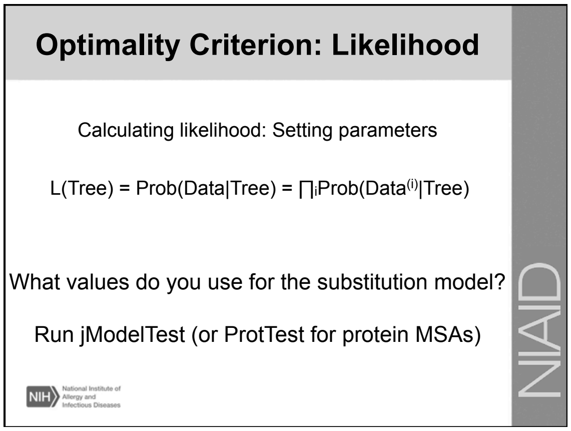
S1: AAG

S2: AAA

S3: GGA

S4: AGA

$L(\text{Tree}) = \text{Prob}(\text{Data}|\text{Tree}) = \prod_i \text{Prob}(\text{Data}^{(i)}|\text{Tree})$



Optimality Criterion: Likelihood

Calculating likelihood: Setting parameters

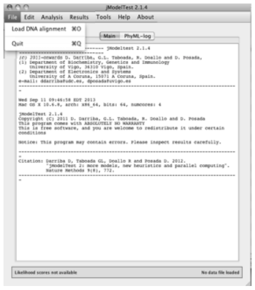
$L(\text{Tree}) = \text{Prob}(\text{Data}|\text{Tree}) = \prod_i \text{Prob}(\text{Data}^{(i)}|\text{Tree})$

What values do you use for the substitution model?

Run jModelTest (or ProtTest for protein MSAs)

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest



```
---------------------------- jModeltest 2.1.4
----------------------------
(c) 2011-onwards D. Darriba, G.L. Taboada, R. Doallo and D. Posada,
(1) Department of Biochemistry, Genetics and Immunology
    University of Vigo, 36310 Vigo, Spain.
(2) Department of Electronics and Systems
    University of A Coruna, 15071 A Coruna, Spain.
e-mail: ddarriba@udc.es, dposada@uvigo.es
--------------------------------------------------------------------------------
-

Wed Sep 11 09:46:58 EDT 2013
Mac OS X 10.6.8, arch: x86_64, bits: 64, numcores: 4

jModelTest 2.1.4
Copyright (C) 2011 D. Darriba, G.L. Taboada, R. Doallo and D. Posada
This program comes with ABSOLUTELY NO WARRANTY
This is free software, and you are welcome to redistribute it under certain
conditions

Notice: This program may contain errors. Please inspect results carefully.

--------------------------------------------------------------------------------
-
Citation: Darriba D, Taboada GL, Doallo R and Posada D. 2012.
          "jModelTest 2: more models, new heuristics and parallel computing".
          Nature Methods 9(8), 772.
--------------------------------------------------------------------------------
-
```

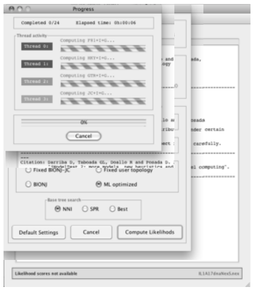
Likelihood scores not available — No data file loaded

NIH National Institute of Allergy and Infectious Diseases

NIAID

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest



```
Completed 0/24      Elapsed time: 0h:00:06
Computing F81+I+G...
Computing HKY+I+G...
Computing GTR+I+G...
Computing JC+I+G...
```

○ Fixed BIONJ-JC ○ Fixed user topology

○ BIONJ ● ML optimized

● NNI ○ SPR ○ Best

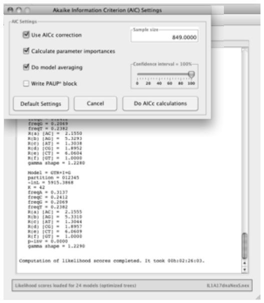
Likelihood scores not available — IL1A17dnaNexS.nex

NIH National Institute of Allergy and Infectious Diseases

NIAID

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest Results



☑ Use AICc correction

☑ Calculate parameter importances

☑ Do model averaging

☐ Write PAUP* block

Sample size: 849.0000

Confidence interval = 100%

```
   freqG = 0.2069
   freqT = 0.2382
   R(a) [AC] =  2.1550
   R(b) [AG] =  5.3293
   R(c) [AT] =  1.3038
   R(d) [CG] =  1.8952
   R(e) [CT] =  6.0604
   R(f) [GT] =  1.0000
   gamma shape = 1.2280

   Model = GTR+I+G
   partition = 012345
   -lnL = 5915.3868
   K = 42
   freqA = 0.3137
   freqC = 0.2412
   freqG = 0.2069
   freqT = 0.2382
   R(a) [AC] =  2.1555
   R(b) [AG] =  5.3310
   R(c) [AT] =  1.3064
   R(d) [CG] =  1.8957
   R(e) [CT] =  6.0609
   R(f) [GT] =  1.0000
   p-inv = 0.0000
   gamma shape = 1.2290

Computation of likelihood scores completed. It took 00h:02:26:03.
```

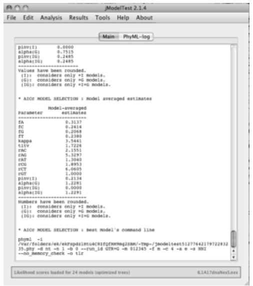
Likelihood scores loaded for 24 models (optimized trees) — IL1A17dnaNexS.nex

NIH National Institute of Allergy and Infectious Diseases

NIAID

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest Results



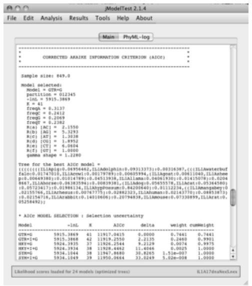
```
pinv(I)       0.0000
alpha(G)      0.7515
pinv(IG)      0.2485
alpha(IG)     0.2485
----------------------
Values have been rounded.
 (I):  considers only +I models.
 (G):  considers only +G models.
 (IG): considers only +I+G models.


* AICc MODEL SELECTION : Model averaged estimates

           Model-averaged
Parameter       estimates
--------------------------
fA                0.3137
fC                0.2414
fG                0.2068
fT                0.2380
kappa             3.5441
titv              1.7226
rAC               2.1551
rAG               5.3297
rAT               1.3040
rCG               1.8953
rCT               6.0605
rGT               1.0000
pinv(I)           0.2134
alpha(G)          1.2281
pinv(IG)          0.0000
alpha(IG)         1.2291
--------------------------
Numbers have been rounded.
 (I):  considers only +I models.
 (G):  considers only +G models.
 (IG): considers only +I+G models.


* AICc MODEL SELECTION : Best Model's command line

phyml  -i
/var/folders/ek/ekFspdzlEtu4C8IfQfRMMmg2SBM/-Tmp-/jmodeltest51277642179722832
35.phy -d nt -n 1 -b 0 --run_id GTR+G -m 012345 -f m -c 4 -a e -s NNI
--no_memory_check -o tlr
```

Likelihood scores loaded for 24 models (optimized trees)   IL1A17dnaNex5.nex

Substitution models
http://www.molecularevolution.org/resources/models/nucleotide

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest Results

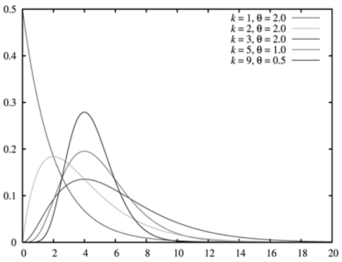


```
---------------------------------------------------------------
*                                                             *
*        CORRECTED AKAIKE INFORMATION CRITERION (AICc)        *
*                                                             *
---------------------------------------------------------------

 Sample size: 849.0

 Model selected:
   Model = GTR+G
   partition = 012345
   -lnL = 5915.3869
   K = 41
   freqA = 0.3137
   freqC = 0.2412
   freqG = 0.2069
   freqT = 0.2382
   R(a) [AC] =  2.1550
   R(b) [AG] =  5.3293
   R(c) [AT] =  1.3038
   R(d) [CG] =  1.8952
   R(e) [CT] =  6.0604
   R(f) [GT] =  1.0000
   gamma shape = 1.2280

Tree for the best AICc model =
((((((((IL1Apig:0.06956462,IL1Adolphin:0.09313373):0.00316387,(((IL1Awaterbuf
falo:0.01747010,IL1Acow1:0.00179789):0.00605994,(IL1Agoat:0.00611040,IL1Ashee
p:0.00649380):0.01014789):0.04513938,IL1Allama:0.04061930):0.01415078):0.0204
8667,IL1Ahorse:0.06383594):0.00839381,(IL1Adog:0.05655578,IL1Acat:0.05364580)
:0.05723417):0.01986134,IL1Ahypossum:0.84200640):0.01112234,((IL1Amangabey:0
.02155766,IL1Arhesus:0.00767775):0.02882323,IL1Ahuman:0.02143770):0.08853873)
:0.02154716,IL1Arabbit:0.14010606):0.20794838,IL1Amouse:0.07330899,IL1Arat:0.
05258492);


* AICc MODEL SELECTION : Selection uncertainty

Model              -lnL    K        AICc       delta      weight cumWeight
----------------------------------------------------------------------------
GTR+G         5915.3869   41  11917.0415      0.0000      0.7441    0.7441
GTR+I+G       5915.3868   42  11919.2550      2.2135      0.2460    0.9901
HKY+G         5924.3935   37  11926.2544      9.2129      0.0074    0.9975
HKY+I+G       5924.3934   38  11928.4462     11.4046      0.0025    1.0000
SYM+G         5934.1044   38  11947.8680     30.8265   1.51e-007    1.0000
SYM+I+G       5934.1049   39  11950.0664     33.0249   5.02e-008    1.0000
```

Likelihood scores loaded for 24 models (optimized trees)   IL1A17dnaNex5.nex

Substitution models
http://www.molecularevolution.org/resources/models/nucleotide

The Gamma Distribution



Mean = kθ   Shape parameter = θ
Coefficient of Variation = 1/√θ

Calculating likelihood: Programs

PAUP* – Commercial, NIH Biowulf, or NIAID HPC
DNA only

PHYLIP – Download, NIH Biowulf, or NIAID HPC
dnaml and proml programs

PAML – Download, NIH Biowulf, or NIAID HPC

RaxML – Download or NIH BioWulf or webserver

PhyML – Download or NIAID HPC or webserver

GARLi – Download or NIAID HPC or webserver

MEGA6 - Download

Generally the user has more flexibility with a local program.
But local programs can hog your computer.

National Institute of Allergy and Infectious Diseases

NIAID

Calculating likelihood: Programs

GARLi – Genetic Algorithm for Rapid Likelihood Inference

Input data format:

PHYLIP or nexus **<u>SEQUENTIAL</u>**

Output files:

filename.best.tree

filename.best.all.tree

filename.*.log – intermediate run output

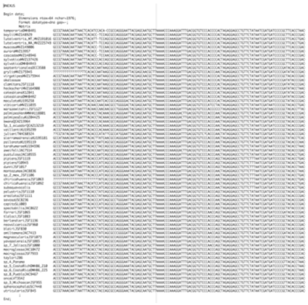
National Institute of Allergy and Infectious Diseases

NIAID

Calculating likelihood: GARLi

Input data format:
PHYLIP or Nexus
<u>SEQUENTIAL</u>
<u>not</u>
<u>INTERLEAVED</u>

and now fasta



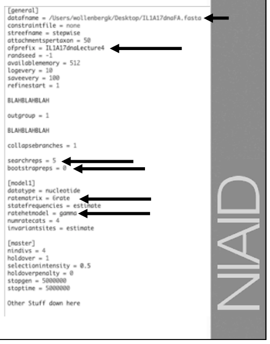
```
#NEXUS

Begin data;
        Dimensions ntax=64 nchar=1976;
        Format datatype=dna gap=-;
        Matrix
[illegible]
        ;
End;
```

National Institute of Allergy and Infectious Diseases

NIAID

Calculating likelihood: GARLi

Specifying input parameters:
garli.conf

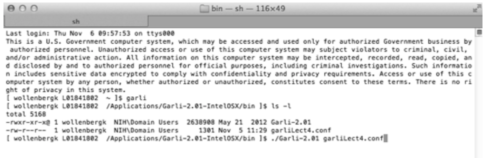


```
[general]
datafname = /Users/wollenbergk/Desktop/IL1A17dnaFA.fasta
constraintfile = none
streefname = stepwise
attachmentspertaxon = 50
ofprefix = IL1A17dnaLecture4
randseed = -1
availablememory = 512
logevery = 10
saveevery = 100
refinestart = 1

BLAHBLAHBLAH

outgroup = 1

BLAHBLAHBLAH

collapsebranches = 1

searchreps = 5
bootstrapreps = 0

[model1]
datatype = nucleotide
ratematrix = 6rate
statefrequencies = estimate
ratehetmodel = gamma
numratecats = 4
invariantsites = estimate

[master]
nindivs = 4
holdover = 1
selectionintensity = 0.5
holdoverpenalty = 0
stopgen = 5000000
stoptime = 5000000

Other Stuff down here
```

www.nescent.org/wg_garli/GARLI_Configuration_Settings

Calculating likelihood: GARLi

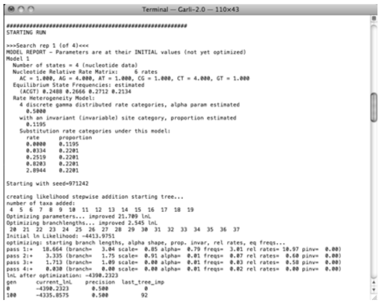
Running the program



```
Last login: Thu Nov  6 09:57:53 on ttys000
This is a U.S. Government computer system, which may be accessed and used only for authorized Government business by
 authorized personnel. Unauthorized access or use of this computer system may subject violators to criminal, civil,
and/or administrative action. All information on this computer system may be intercepted, recorded, read, copied, an
d disclosed by and to authorized personnel for official purposes, including criminal investigations. Such informatio
n includes sensitive data encrypted to comply with confidentiality and privacy requirements. Access or use of this c
omputer system by any person, whether authorized or unauthorized, constitutes consent to these terms. There is no ri
ght of privacy in this system.
[ wollenbergk L01841802  ~ ]$ garli
[ wollenbergk L01841802  /Applications/Garli-2.01-IntelOSX/bin ]$ ls -l
total 5168
-rwxr-xr-x@ 1 wollenbergk  NIH\Domain Users  2638908 May 21  2012 Garli-2.01
-rw-r--r--  1 wollenbergk  NIH\Domain Users     1301 Nov  5 11:29 garliLect4.conf
[ wollenbergk L01841802  /Applications/Garli-2.01-IntelOSX/bin ]$ ./Garli-2.01 garliLect4.conf
```

Calculating likelihood: GARLi

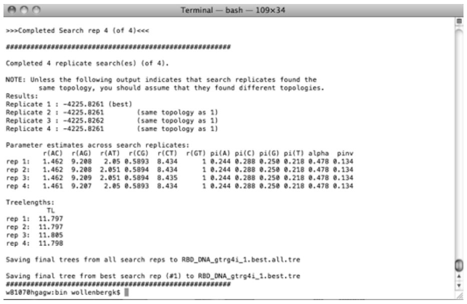
Running the program



```
########################################################
STARTING RUN

>>>Search rep 1 (of 4)<<<
MODEL REPORT - Parameters are at their INITIAL values (not yet optimized)
Model 1
  Number of states = 4 (nucleotide data)
  Nucleotide Relative Rate Matrix:     6 rates
    AC = 1.000, AG = 4.000, AT = 1.000, CG = 1.000, CT = 4.000, GT = 1.000
  Equilibrium State Frequencies: estimated
    (ACGT) 0.2488 0.2666 0.2712 0.2134
  Rate Heterogeneity Model:
    4 discrete gamma distributed rate categories, alpha param estimated
      0.5000
    with an invariant (invariable) site category, proportion estimated
      0.1195
    Substitution rate categories under this model:
      rate      proportion
      0.0000    0.1195
      0.0334    0.2201
      0.2519    0.2201
      0.8203    0.2201
      2.8944    0.2201

Starting with seed=971242

creating likelihood stepwise addition starting tree...
number of taxa added:
 4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19
Optimizing parameters... improved 21.709 lnL
Optimizing branchlengths... improved 2.545 lnL
 20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37
Initial ln Likelihood: -4413.9751
optimizing: starting branch lengths, alpha shape, prop. invar, rel rates, eq freqs...
pass 1:+   18.664 (branch=   3.04 scale=  0.85 alpha=  0.79 freqs=  3.01 rel rates= 10.97 pinv=  0.00)
pass 2:+    3.335 (branch=   1.75 scale=  0.91 alpha=  0.01 freqs=  0.07 rel rates=  0.60 pinv=  0.00)
pass 3:+    1.713 (branch=   1.09 scale=  0.00 alpha=  0.01 freqs=  0.03 rel rates=  0.58 pinv=  0.00)
pass 4:+    0.030 (branch=   0.00 scale=  0.00 alpha=  0.01 freqs=  0.02 rel rates=  0.00 pinv=  0.00)
lnL after optimization: -4390.2323
gen      current_lnL    precision  last_tree_imp
0        -4390.2323       0.500          0
100      -4335.8575       0.500          92
```

Calculating likelihood: GARLi

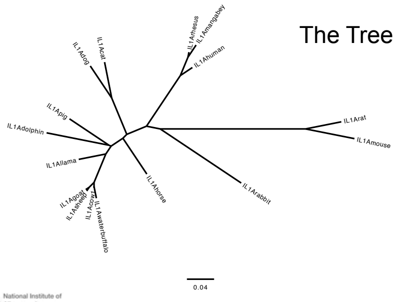
Running the program



```
>>>Completed Search rep 4 (of 4)<<<

#######################################################

Completed 4 replicate search(es) (of 4).

NOTE: Unless the following output indicates that search replicates found the
        same topology, you should assume that they found different topologies.
Results:
Replicate 1 : -4225.8261 (best)
Replicate 2 : -4225.8261        (same topology as 1)
Replicate 3 : -4225.8262        (same topology as 1)
Replicate 4 : -4225.8261        (same topology as 1)

Parameter estimates across search replicates:
         r(AC)  r(AG)  r(AT)  r(CG)  r(CT)  r(GT) pi(A) pi(C) pi(G) pi(T) alpha  pinv
rep 1:   1.462  9.208   2.05 0.5893  8.434      1 0.244 0.288 0.250 0.218 0.478 0.134
rep 2:   1.462  9.208  2.051 0.5894  8.434      1 0.244 0.288 0.250 0.218 0.478 0.134
rep 3:   1.462  9.209  2.051 0.5894  8.435      1 0.244 0.288 0.250 0.218 0.478 0.134
rep 4:   1.461  9.207   2.05 0.5893  8.434      1 0.244 0.288 0.250 0.218 0.478 0.134

Treelengths:
          TL
rep 1:  11.797
rep 2:  11.797
rep 3:  11.805
rep 4:  11.798

Saving final trees from all search reps to RBD_DNA_gtrg4i_1.best.all.tre

Saving final tree from best search rep (#1) to RBD_DNA_gtrg4i_1.best.tre
#######################################################
w81070hgagw:bin wollenbergk$
```

Calculating likelihood: GARLi

The Tree



Calculating likelihood: GARLi

Bootstrap results?

GARLi **does not** calculate a bootstrap consensus!

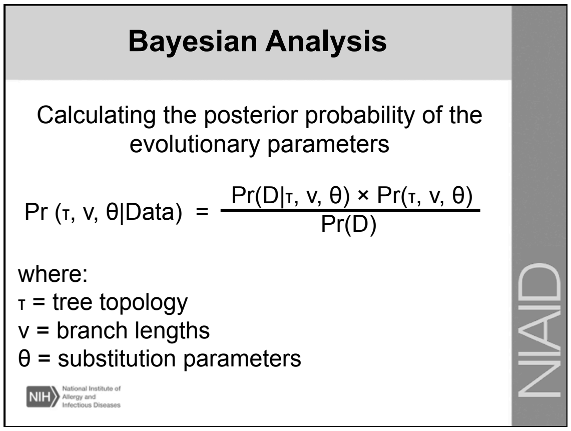
You must use another piece of software for this.

Dendropy:sumtrees – a very efficient python package for summarizing collections of phylogenetic trees

Dendropy – Freely distributed from:

https://pythonhosted.org/DendroPy

sumtrees has also been implemented on the NIAID HPC

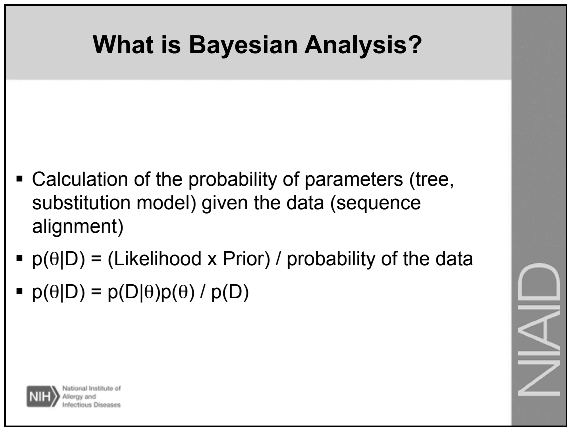


Bayesian Analysis

Calculating the posterior probability of the evolutionary parameters

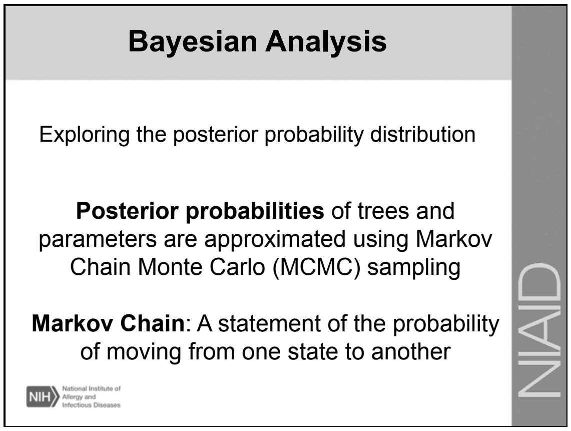
$$\Pr(\tau, v, \theta|\text{Data}) = \frac{\Pr(D|\tau, v, \theta) \times \Pr(\tau, v, \theta)}{\Pr(D)}$$

where:
τ = tree topology
v = branch lengths
θ = substitution parameters



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
- $p(\theta|D) = p(D|\theta)p(\theta) / p(D)$

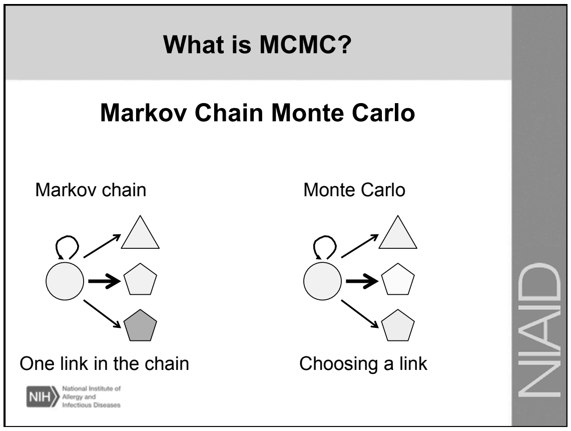


Bayesian Analysis

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



What is MCMC?

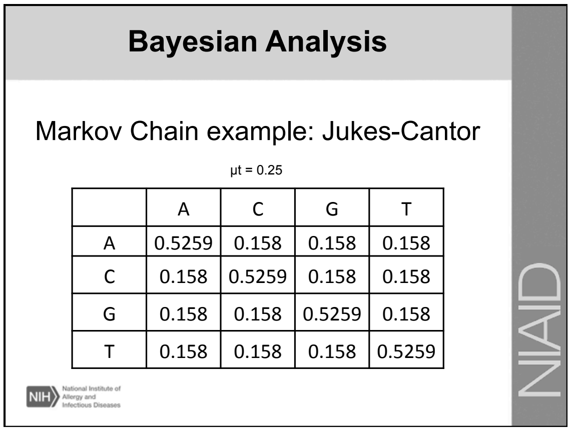
Markov Chain Monte Carlo

Markov chain

Monte Carlo

One link in the chain

Choosing a link

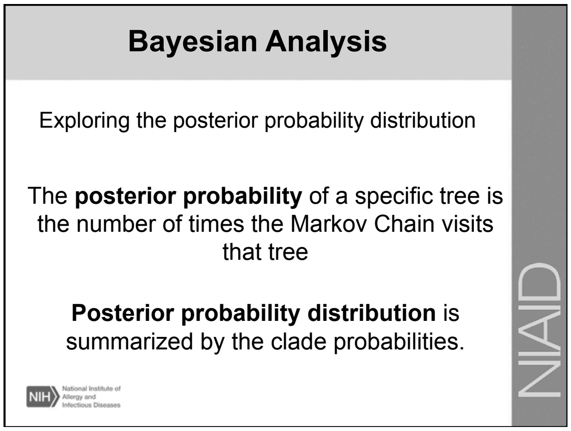


Bayesian Analysis

Markov Chain example: Jukes-Cantor

μt = 0.25

	A	C	G	T
A	0.5259	0.158	0.158	0.158
C	0.158	0.5259	0.158	0.158
G	0.158	0.158	0.5259	0.158
T	0.158	0.158	0.158	0.5259

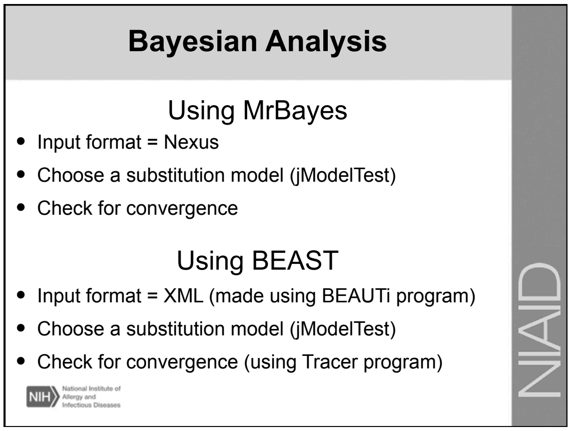


Bayesian Analysis

Exploring the posterior probability distribution

The **posterior probability** of a specific tree is the number of times the Markov Chain visits that tree

Posterior probability distribution is summarized by the clade probabilities.



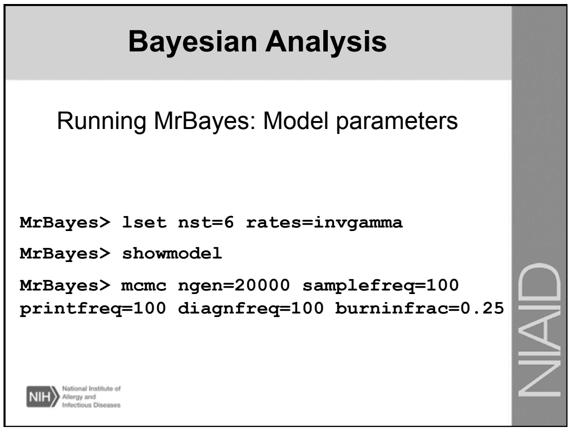
Bayesian Analysis

Using MrBayes

- Input format = Nexus
- Choose a substitution model (jModelTest)
- Check for convergence

Using BEAST

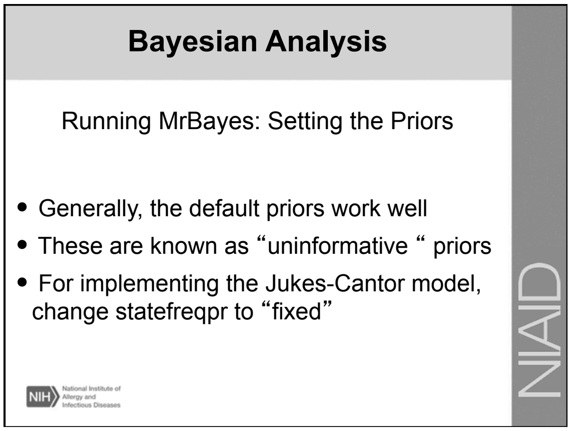
- Input format = XML (made using BEAUTi program)
- Choose a substitution model (jModelTest)
- Check for convergence (using Tracer program)



Bayesian Analysis

Running MrBayes: Model parameters

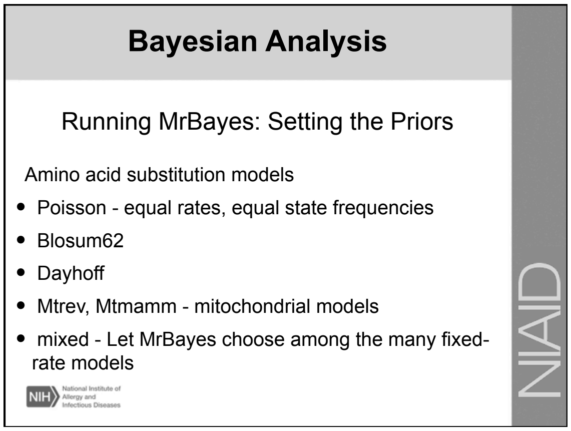
```
MrBayes> lset nst=6 rates=invgamma
MrBayes> showmodel
MrBayes> mcmc ngen=20000 samplefreq=100
printfreq=100 diagnfreq=100 burninfrac=0.25
```



Bayesian Analysis

Running MrBayes: Setting the Priors

- Generally, the default priors work well
- These are known as “uninformative “ priors
- For implementing the Jukes-Cantor model, change statefreqpr to “fixed”

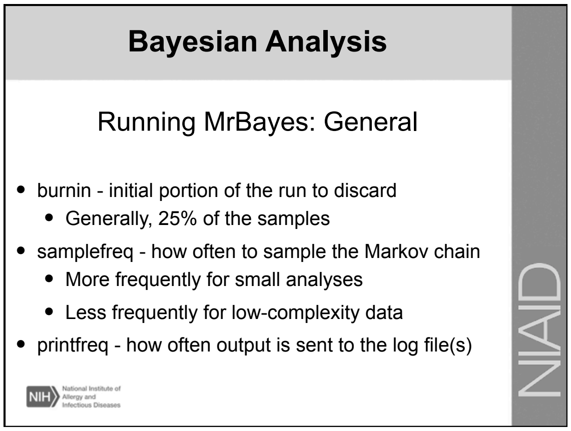


Bayesian Analysis

Running MrBayes: Setting the Priors

Amino acid substitution models

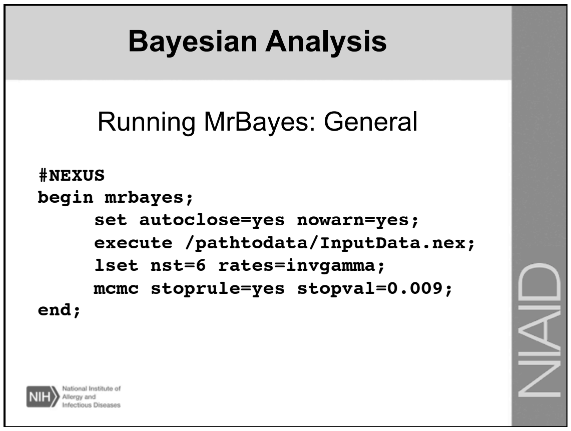
- Poisson - equal rates, equal state frequencies
- Blosum62
- Dayhoff
- Mtrev, Mtmamm - mitochondrial models
- mixed - Let MrBayes choose among the many fixed-rate models



Bayesian Analysis

Running MrBayes: General

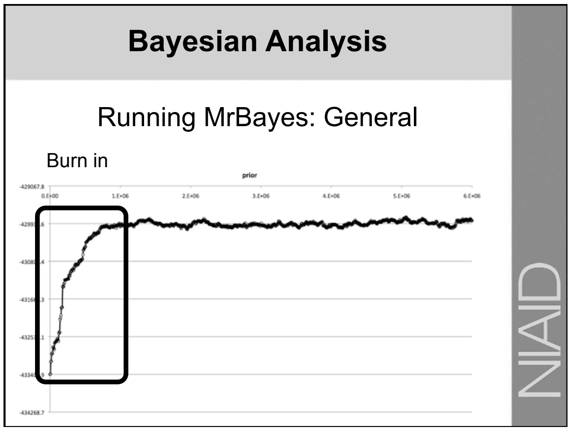
- burnin - initial portion of the run to discard
  - Generally, 25% of the samples
- samplefreq - how often to sample the Markov chain
  - More frequently for small analyses
  - Less frequently for low-complexity data
- printfreq - how often output is sent to the log file(s)



Bayesian Analysis

Running MrBayes: General

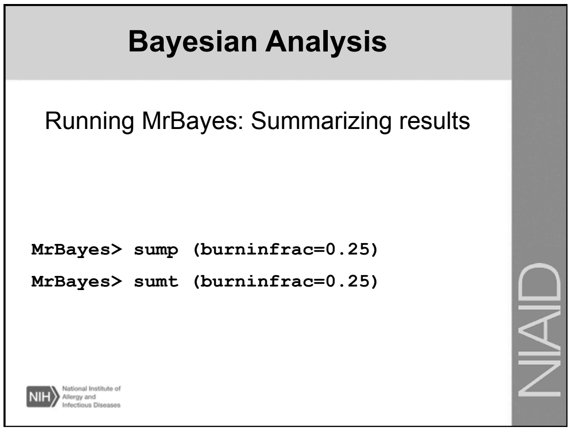
```
#NEXUS
begin mrbayes;
     set autoclose=yes nowarn=yes;
     execute /pathtodata/InputData.nex;
     lset nst=6 rates=invgamma;
     mcmc stoprule=yes stopval=0.009;
end;
```



Bayesian Analysis

Running MrBayes: General

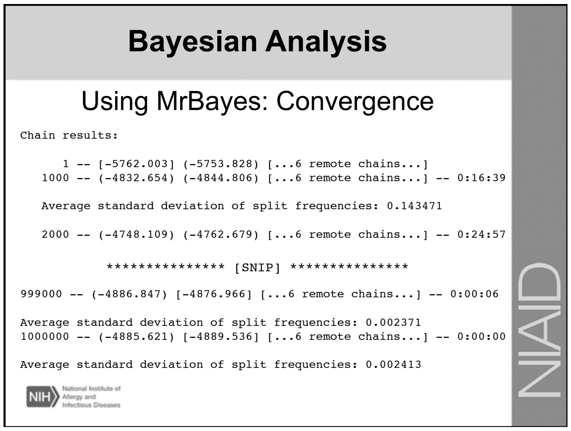
Burn in



Bayesian Analysis

Running MrBayes: Summarizing results

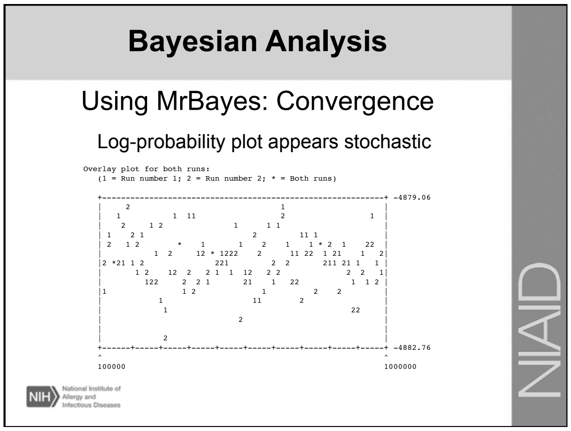
```
MrBayes> sump (burninfrac=0.25)
MrBayes> sumt (burninfrac=0.25)
```



Bayesian Analysis

Using MrBayes: Convergence

```
Chain results:

      1 -- [-5762.003] (-5753.828) [...6 remote chains...]
   1000 -- (-4832.654) (-4844.806) [...6 remote chains...] -- 0:16:39

   Average standard deviation of split frequencies: 0.143471

   2000 -- (-4748.109) (-4762.679) [...6 remote chains...] -- 0:24:57

            *************** [SNIP] ***************

999000 -- (-4886.847) [-4876.966] [...6 remote chains...] -- 0:00:06

Average standard deviation of split frequencies: 0.002371
1000000 -- (-4885.621) [-4889.536] [...6 remote chains...] -- 0:00:00

Average standard deviation of split frequencies: 0.002413
```

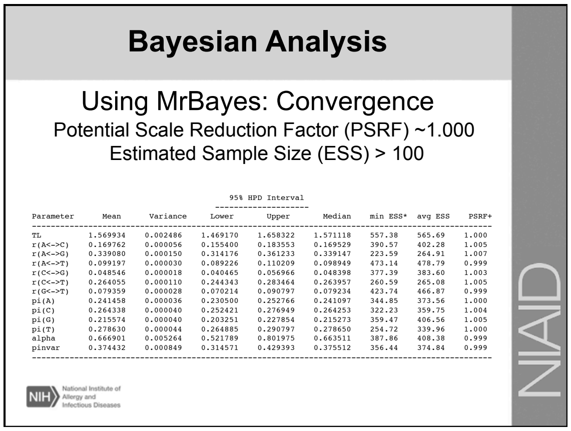


Bayesian Analysis

Using MrBayes: Convergence

Log-probability plot appears stochastic

```
Overlay plot for both runs:
   (1 = Run number 1; 2 = Run number 2; * = Both runs)

   +------------------------------------------------------------+ -4879.06
   |     2                                1                     |
   |   1           1  11                  2                  1  |
   |    2     1 2               1      1 1                      |
   | 1    2 1                       2         11 1              |
   | 2   1 2        *    1       1    2    1    1 * 2  1    22  |
   |           1  2     12 * 1222    2      11 22  1 21    1   2|
   |2 *21 1 2               221         2  2       211 21 1   1 |
   |       1 2    12  2   2 1  1  12   2 2              2  2   1|
   |         122     2  2 1       21    1   22           1  1 2 |
   |1                1 2              1          2    2         |
   |            1                   11        2                 |
   |             1                                       22     |
   |                             2                              |
   |                                                            |
   |             2                                              |
   +------+-----+-----+-----+-----+-----+-----+-----+-----+-----+ -4882.76
   ^                                                            ^
   100000                                                       1000000
```

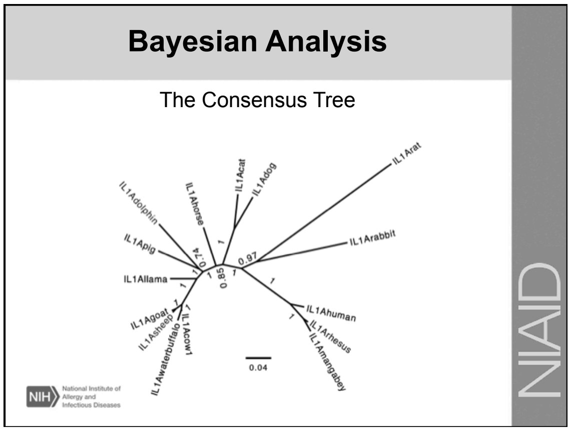


Bayesian Analysis

Using MrBayes: Convergence

Potential Scale Reduction Factor (PSRF) ~1.000
Estimated Sample Size (ESS) > 100

```
                                   95% HPD Interval
                                 --------------------
Parameter      Mean      Variance     Lower       Upper       Median    min ESS*  avg ESS    PSRF+
--------------------------------------------------------------------------------------------------
TL           1.569934    0.002486    1.469170    1.658322    1.571118    557.38    565.69    1.000
r(A<->C)     0.169762    0.000056    0.155400    0.183553    0.169529    390.57    402.28    1.005
r(A<->G)     0.339080    0.000150    0.314176    0.361233    0.339147    223.59    264.91    1.007
r(A<->T)     0.099197    0.000030    0.089226    0.110209    0.098949    473.14    478.79    0.999
r(C<->G)     0.048546    0.000018    0.040465    0.056966    0.048398    377.39    383.60    1.003
r(C<->T)     0.264055    0.000110    0.244343    0.283464    0.263957    260.59    265.08    1.005
r(G<->T)     0.079359    0.000028    0.070214    0.090797    0.079234    423.74    466.87    0.999
pi(A)        0.241458    0.000036    0.230500    0.252766    0.241097    344.85    373.56    1.000
pi(C)        0.264338    0.000040    0.252421    0.276949    0.264253    322.23    359.75    1.004
pi(G)        0.215574    0.000040    0.203251    0.227854    0.215273    359.47    406.56    1.005
pi(T)        0.278630    0.000044    0.264885    0.290797    0.278650    254.72    339.96    1.000
alpha        0.666901    0.005264    0.521789    0.801975    0.663511    387.86    408.38    0.999
pinvar       0.374432    0.000849    0.314571    0.429393    0.375512    356.44    374.84    0.999
--------------------------------------------------------------------------------------------------
```

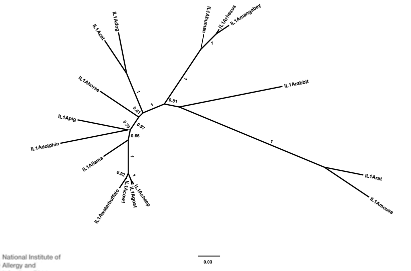


Bayesian Analysis

The Consensus Tree

Likelihood Analysis

The Consensus Tree

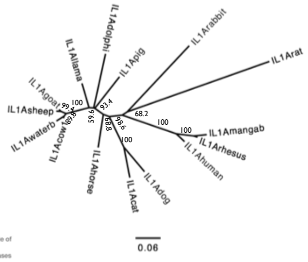


NIH National Institute of Allergy and Infectious Diseases

NIAID

Distance Analysis

The Neighbor-Joining Tree

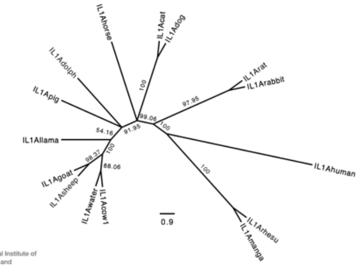


NIH National Institute of Allergy and Infectious Diseases

NIAID

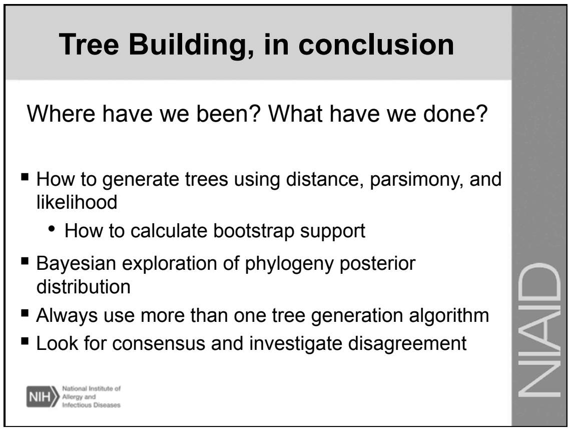
Parsimony Analysis

The Bootstrap Consensus Tree



NIH National Institute of Allergy and Infectious Diseases

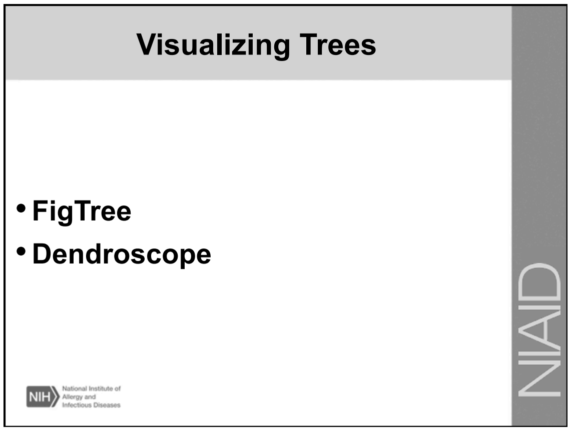
NIAID



Tree Building, in conclusion

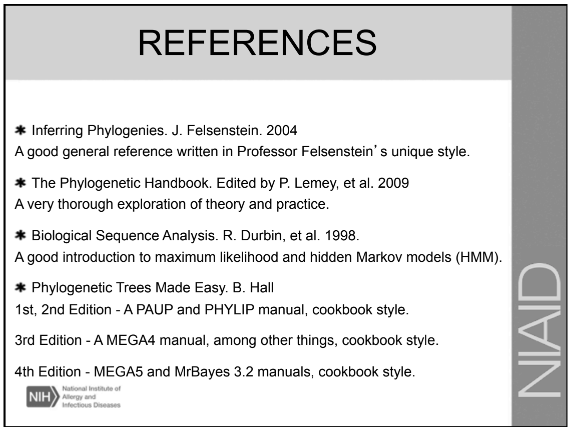
Where have we been? What have we done?

- How to generate trees using distance, parsimony, and likelihood
  - How to calculate bootstrap support
- Bayesian exploration of phylogeny posterior distribution
- Always use more than one tree generation algorithm
- Look for consensus and investigate disagreement



Visualizing Trees

- **FigTree**
- **Dendroscope**



REFERENCES

✱ Inferring Phylogenies. J. Felsenstein. 2004
A good general reference written in Professor Felsenstein's unique style.

✱ The Phylogenetic Handbook. Edited by P. Lemey, et al. 2009
A very thorough exploration of theory and practice.

✱ Biological Sequence Analysis. R. Durbin, et al. 1998.
A good introduction to maximum likelihood and hidden Markov models (HMM).

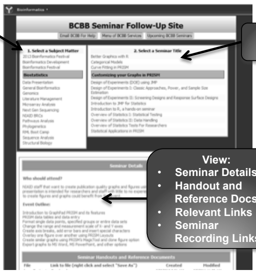
✱ Phylogenetic Trees Made Easy. B. Hall
1st, 2nd Edition - A PAUP and PHYLIP manual, cookbook style.

3rd Edition - A MEGA4 manual, among other things, cookbook style.

4th Edition - MEGA5 and MrBayes 3.2 manuals, cookbook style.

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/



1. Select a Subject Matter

2. Select a Topic

BCBB Seminar Follow-Up Site

Email BCBB For Help | Menu of BCBB Services | Upcoming BCBB Seminars

1. Select a Subject Matter
2013 Bioinformatics Festival
Bioinformatics Development
Bioinformatics Festival
Biostatistics
Data Presentation
General Bioinformatics
Genomics
Literature Management
Microarray Analysis
Next Gen Sequencing
NIAID BRCs
Pathways Analysis
Phylogenetics
RML Boot Camp
Sequence Analysis
Structural Biology

2. Select a Seminar Title
Better Graphics with R
Categorical Models
Curve Fitting in PRISM
Customizing your Graphs in PRISM
Design of Experiments (DOE) using JMP
Design of Experiments I: Classic Approaches, Power, and Sample Size Estimation
Design of Experiments II: Screening Designs and Response Surface Designs
Introduction to JMP for Statistics
Introduction to R, a hands-on seminar
Overview of Statistics I: Statistical Testing
Overview of Statistics II: Data Handling
Overview of Statistics Tests For Researchers
Statistical Applications in PRISM

Seminar Details

Who should attend?

NIAID staff that want to create publication quality graphs and figures using [illegible] presentation is intended for researchers and staff with little to no experience [illegible] to create figures and graphs could benefit from [illegible]

Event Outline:

Introduction to GraphPad PRISM and its features
PRISM data tables and data entry
Format single data points, specified groups or entire data sets
Change the range and measurement scale of X- and Y-axes
Create axis breaks, add error bars and insert special characters
Overlay one figure over another using PRISM Layouts
Create similar graphs using PRISM's MagicTool and clone figure option
Export graphs to MS Word, MS PowerPoint, and other options

Seminar Handouts and Reference Documents

File | Link to file (right click and select "Save As") | Created | Modified

View:
- Seminar Details
- Handout and Reference Docs
- Relevant Links
- Seminar Recording Links

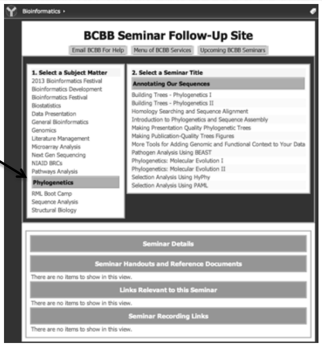
Recommended Browsers:
- IE for Windows,
- Safari for Mac (Firefox on a Mac is incompatible with NIH Authentication technology)

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

National Institute of Allergy and Infectious Diseases

46

Retrieving Slides/Handouts



This lecture series

Bioinformatics ›

BCBB Seminar Follow-Up Site

Email BCBB For Help | Menu of BCBB Services | Upcoming BCBB Seminars

1. Select a Subject Matter
2013 Bioinformatics Festival
Bioinformatics Development
Bioinformatics Festival
Biostatistics
Data Presentation
General Bioinformatics
Genomics
Literature Management
Microarray Analysis
Next Gen Sequencing
NIAID BRCs
Pathways Analysis
Phylogenetics
RML Boot Camp
Sequence Analysis
Structural Biology

2. Select a Seminar Title
Annotating Our Sequences
Building Trees - Phylogenetics I
Building Trees - Phylogenetics II
Homology Searching and Sequence Alignment
Introduction to Phylogenetics and Sequence Assembly
Making Presentation Quality Phylogenetic Trees
Making Publication-Quality Trees Figures
More Tools for Adding Genomic and Functional Context to Your Data
Pathogen Analysis Using BEAST
Phylogenetics: Molecular Evolution I
Phylogenetics: Molecular Evolution II
Selection Analysis Using HyPhy
Selection Analysis Using PAML

Seminar Details

Seminar Handouts and Reference Documents

There are no items to show in this view.

Links Relevant to this Seminar

There are no items to show in this view.

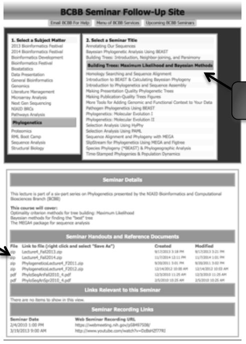
Seminar Recording Links

There are no items to show in this view.

National Institute of Allergy and Infectious Diseases

47

Retrieving Slides/Handouts



This lecture

This file

BCBB Seminar Follow-Up Site

Email BCBB For Help | Menu of BCBB Services | Upcoming BCBB Seminars

1. Select a Subject Matter
2013 Bioinformatics Festival
2014 Bioinformatics Festival
Bioinformatics Development
Bioinformatics Festival
Biostatistics
Data Presentation
General Bioinformatics
Genomics
Literature Management
Microarray Analysis
Next Gen Sequencing
NIAID BRCs
Pathways Analysis
Phylogenetics
Proteomics
RML Boot Camp
Sequence Analysis
Structural Biology

2. Select a Seminar Title
Annotating Our Sequences
Bayesian Phylogenetic Analysis Using BEAST
Building Trees: Introduction, Neighbor-joining, and Parsimony
Building Trees: Maximum Likelihood and Bayesian Methods
Homology Searching and Sequence Alignment
Introduction to BEAST & Calculating Bayesian Phylogeny
Introduction to Phylogenetics and Sequence Assembly
Making Presentation Quality Phylogenetic Trees
Making Publication-Quality Trees Figures
More Tools for Adding Genomic and Functional Context to Your Data
Pathogen Phylogenetics Using BEAST
Phylogenetics: Molecular Evolution I
Phylogenetics: Molecular Evolution II
Selection Analysis Using HyPhy
Selection Analysis Using PAML
Sequence Alignment and Phylogeny with MEGA
SlipStream for Phylogenetics Using MEGA and Figtree
Species Phylogeny (*BEAST) & Phylogeographic Analysis
Time-Stamped Phylogenies & Population Dynamics

Seminar Details

This lecture is part of a six-part series on Phylogenetics presented by the NIAID Bioinformatics and Computational Biosciences Branch (BCBB)

This course will cover:
Optimality criterion methods for tree building: Maximum Likelihood
Bayesian methods for finding the "best" tree
The MEGA4 package for sequence analysis

Seminar Handouts and Reference Documents

File	Link to file (right click and select "Save As")	Created	Modified
zip	Lecture4_Fall2013.zip	9/17/2013 3:18 PM	9/17/2013 3:21 PM
zip	Lecture4_Fall2014.zip	11/7/2014 12:11 PM	11/7/2014 1:01 PM
zip	PhylogeneticsLecture4_F2011.zip	9/20/2011 3:01 PM	9/20/2011 3:02 PM
zip	PhylogeneticsLecture4_F2012.zip	12/14/2012 10:00 AM	12/14/2012 10:03 AM
pdf	PhyloSeqAnFall2010_4.pdf	12/3/2010 11:25 AM	12/3/2010 11:25 AM
pdf	PhyloSeqAnSpr2010_4.pdf	2/5/2010 10:25 AM	2/5/2010 10:25 AM

Links Relevant to this Seminar

There are no items to show in this view.

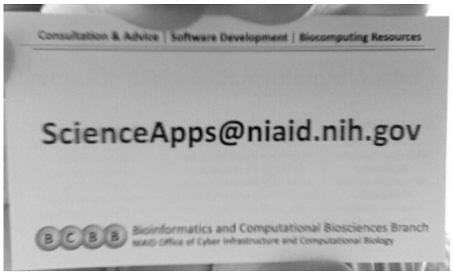
Seminar Recording Links

Seminar Date	Web Seminar Recording URL
2/4/2010 1:00 PM	https://webmeeting.nih.gov/p58497508/
3/19/2013 9:00 AM	http://www.youtube.com/watch?v=DcBeH2f77RI

National Institute of Allergy and Infectious Diseases

48

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

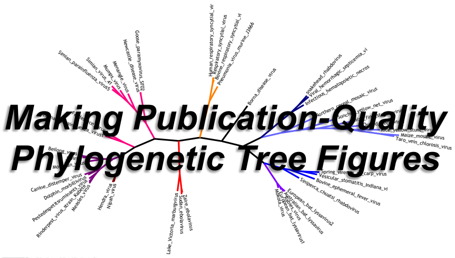
NIH National Institute of Allergy and Infectious Diseases

NIAID

49

Next

Tuesday, 10 November at 1300


Simian_parainfluenza_virus5
Simian_virus_41
Mumps_virus
Menangle_virus
Newcastle_disease_virus
Goose_paramyxovirus_SF02
Human_respiratory_syncytial_vir
Respiratory_syncytial_virus
Bovine_respiratory_syncytial_vi
Pneumonia_virus_murine_J3666
Borna_disease_virus
snakehead_rhabdovirus
Viral_hemorrhagic_septicemia_vi
Infectious_hematopoietic_necros
Northern_cereal_mosaic_virus
Sonchus_yellow_net_virus
Rice_yellow_stunt_virus
Maize_mosaic_virus
Taro_vein_chlorosis_virus
Spring_viremia_carp_virus
Vesicular_stomatitis_Indiana_vi
Bovine_ephemeral_fever_virus
Siniperca_chuatsi_rhabdovirus
European_bat_lyssavirus2
Australian_bat_lyssavirus
Rabies_virus
European_bat_lyssavirus1
Mokola_virus
Lake_Victoria_marburgvirus
Sudan_ebolavirus
Zaire_ebolavirus
Hendra_virus
Nipah_virus
Beilong_virus
Canine_distemper_virus
Dolphin_morbillivirus
Pestedespetitsruminants_virus
Rinderpest_virus_strain_Kabete
Measles_virus

Making Publication-Quality Phylogenetic Tree Figures

NIH National Institute of Allergy and Infectious Diseases

NIAID

50