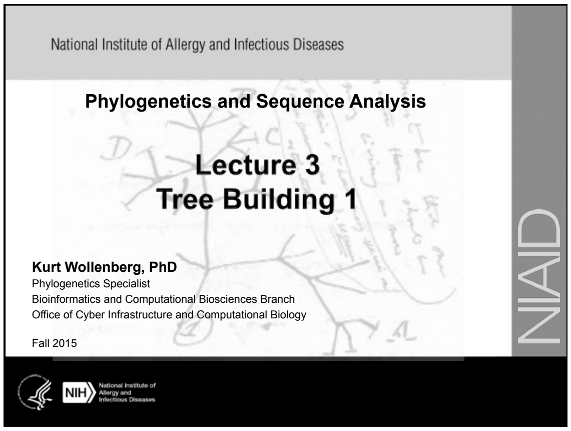


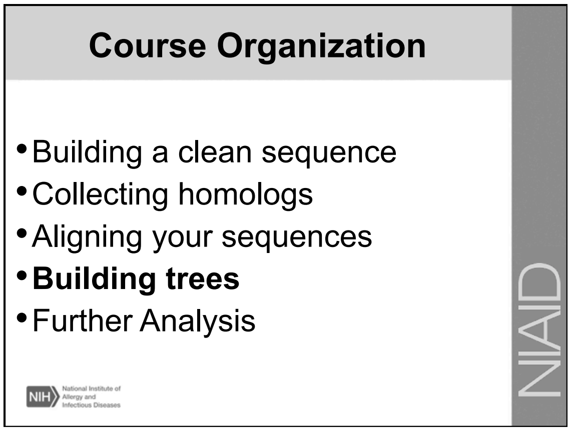
National Institute of Allergy and Infectious Diseases

Phylogenetics and Sequence Analysis

Lecture 3
Tree Building 1

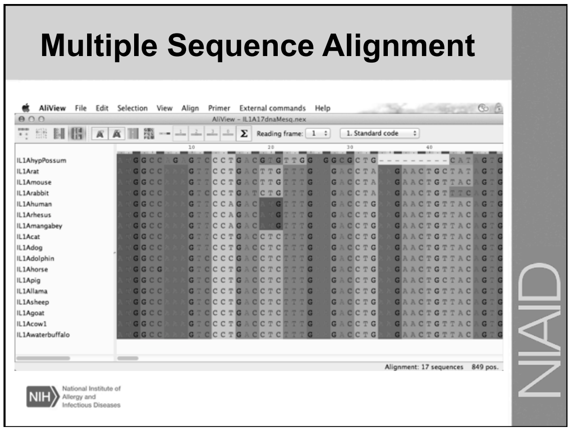
Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biosciences Branch
Office of Cyber Infrastructure and Computational Biology

Fall 2015



Course Organization

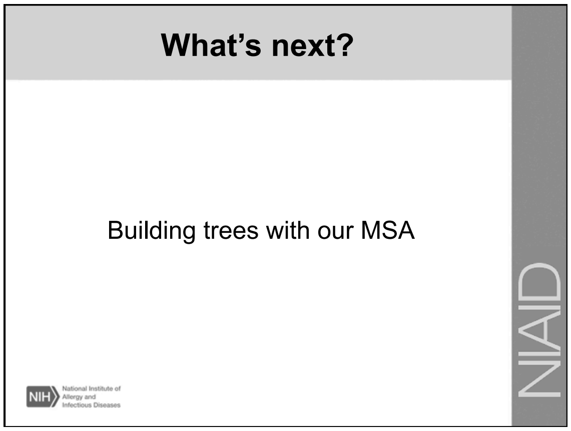
- Building a clean sequence
- Collecting homologs
- Aligning your sequences
- **Building trees**
- Further Analysis



Multiple Sequence Alignment

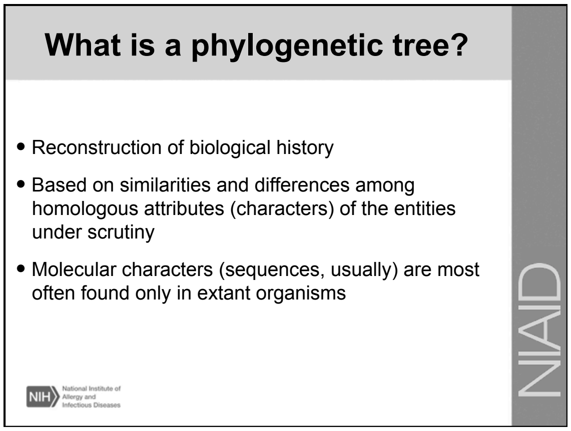
IL1AhypPossum
IL1Arat
IL1Amouse
IL1Arabbit
IL1Ahuman
IL1Arhesus
IL1Amangabey
IL1Acat
IL1Adog
IL1Adolphin
IL1Ahorse
IL1Apig
IL1Allama
IL1Asheep
IL1Agoat
IL1Acow1
IL1Awaterbuffalo

Alignment: 17 sequences 849 pos.



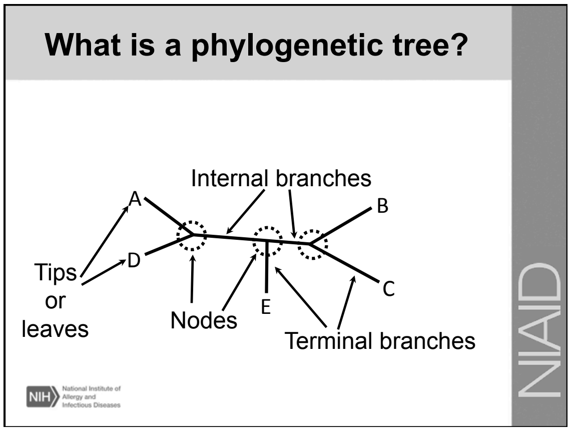
What's next?

Building trees with our MSA



What is a phylogenetic tree?

- Reconstruction of biological history
- Based on similarities and differences among homologous attributes (characters) of the entities under scrutiny
- Molecular characters (sequences, usually) are most often found only in extant organisms



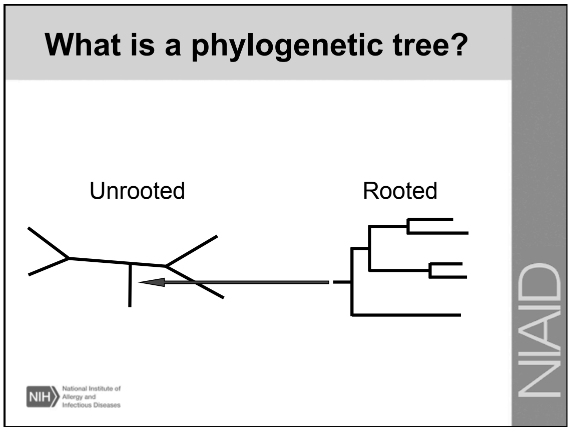
What is a phylogenetic tree?

Internal branches

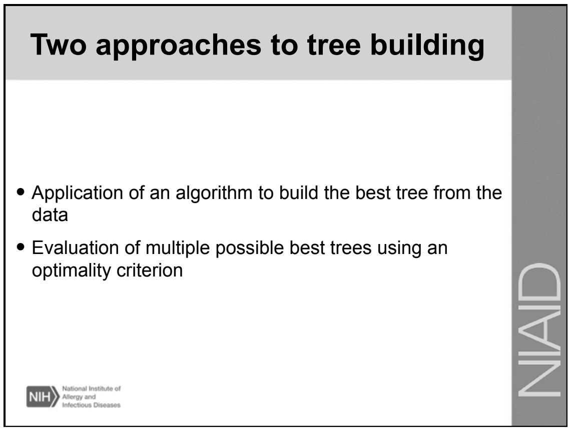
Tips or leaves

Nodes

Terminal branches

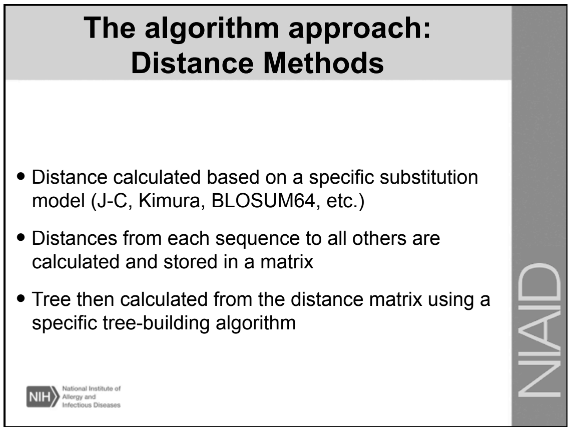


What is a phylogenetic tree?



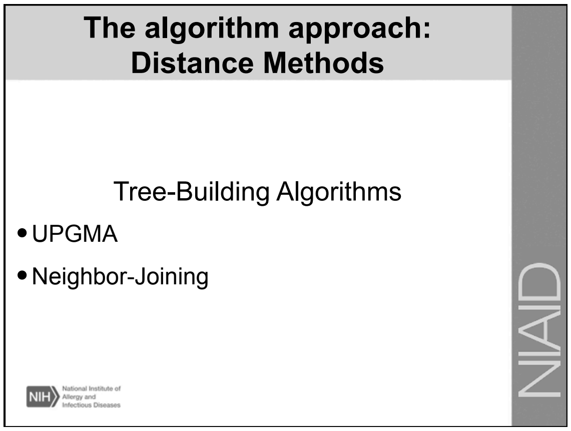
Two approaches to tree building

- Application of an algorithm to build the best tree from the data
- Evaluation of multiple possible best trees using an optimality criterion



The algorithm approach: Distance Methods

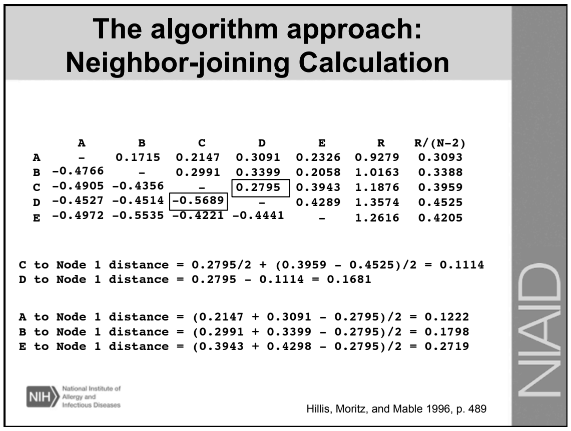
- Distance calculated based on a specific substitution model (J-C, Kimura, BLOSUM64, etc.)
- Distances from each sequence to all others are calculated and stored in a matrix
- Tree then calculated from the distance matrix using a specific tree-building algorithm



The algorithm approach: Distance Methods

Tree-Building Algorithms

- UPGMA
- Neighbor-Joining



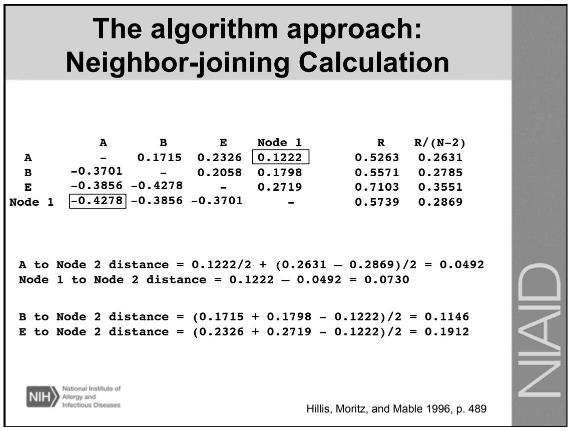
The algorithm approach: Neighbor-joining Calculation

	A	B	C	D	E	R	R/(N-2)
A	-	0.1715	0.2147	0.3091	0.2326	0.9279	0.3093
B	-0.4766	-	0.2991	0.3399	0.2058	1.0163	0.3388
C	-0.4905	-0.4356	-	$\boxed{0.2795}$	0.3943	1.1876	0.3959
D	-0.4527	-0.4514	$\boxed{-0.5689}$	-	0.4289	1.3574	0.4525
E	-0.4972	-0.5535	-0.4221	-0.4441	-	1.2616	0.4205

C to Node 1 distance = 0.2795/2 + (0.3959 - 0.4525)/2 = 0.1114
D to Node 1 distance = 0.2795 - 0.1114 = 0.1681

A to Node 1 distance = (0.2147 + 0.3091 - 0.2795)/2 = 0.1222
B to Node 1 distance = (0.2991 + 0.3399 - 0.2795)/2 = 0.1798
E to Node 1 distance = (0.3943 + 0.4298 - 0.2795)/2 = 0.2719

Hillis, Moritz, and Mable 1996, p. 489



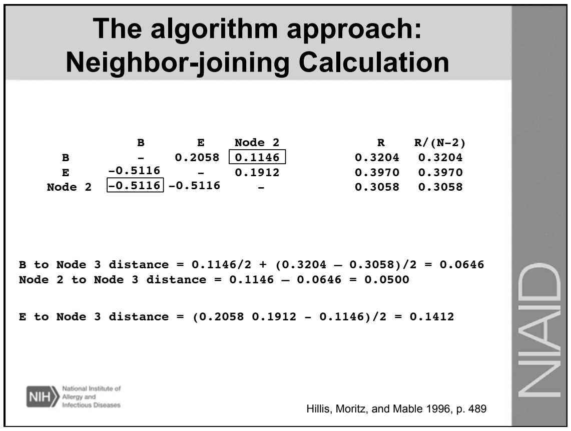
The algorithm approach: Neighbor-joining Calculation

	A	B	E	Node 1	R	R/(N-2)
A	-	0.1715	0.2326	$\boxed{0.1222}$	0.5263	0.2631
B	-0.3701	-	0.2058	0.1798	0.5571	0.2785
E	-0.3856	-0.4278	-	0.2719	0.7103	0.3551
Node 1	$\boxed{-0.4278}$	-0.3856	-0.3701	-	0.5739	0.2869

A to Node 2 distance = 0.1222/2 + (0.2631 – 0.2869)/2 = 0.0492
Node 1 to Node 2 distance = 0.1222 – 0.0492 = 0.0730

B to Node 2 distance = (0.1715 + 0.1798 - 0.1222)/2 = 0.1146
E to Node 2 distance = (0.2326 + 0.2719 - 0.1222)/2 = 0.1912

Hillis, Moritz, and Mable 1996, p. 489

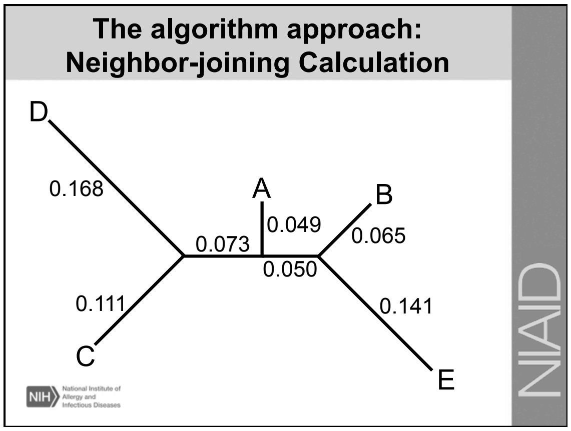


The algorithm approach: Neighbor-joining Calculation

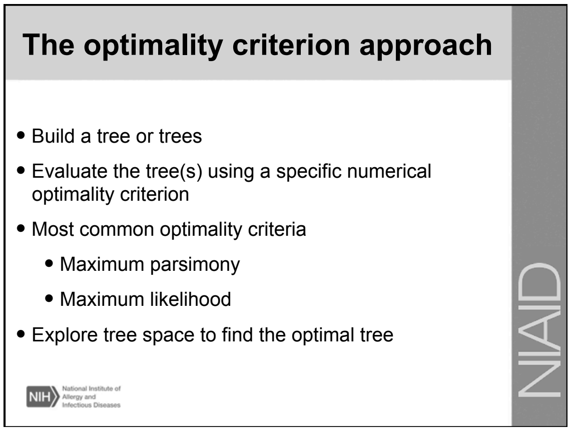
	B	E	Node 2	R	R/(N-2)
B	-	0.2058	$\boxed{0.1146}$	0.3204	0.3204
E	-0.5116	-	0.1912	0.3970	0.3970
Node 2	$\boxed{-0.5116}$	-0.5116	-	0.3058	0.3058

```
B to Node 3 distance = 0.1146/2 + (0.3204 – 0.3058)/2 = 0.0646
Node 2 to Node 3 distance = 0.1146 – 0.0646 = 0.0500

E to Node 3 distance = (0.2058 0.1912 - 0.1146)/2 = 0.1412
```

Hillis, Moritz, and Mable 1996, p. 489

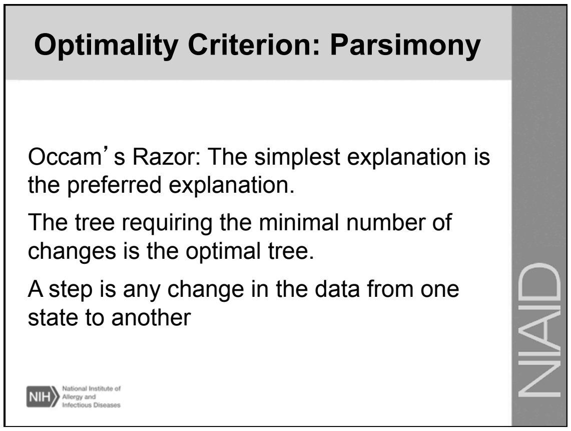


The algorithm approach: Neighbor-joining Calculation



The optimality criterion approach

- Build a tree or trees
- Evaluate the tree(s) using a specific numerical optimality criterion
- Most common optimality criteria
  - Maximum parsimony
  - Maximum likelihood
- Explore tree space to find the optimal tree

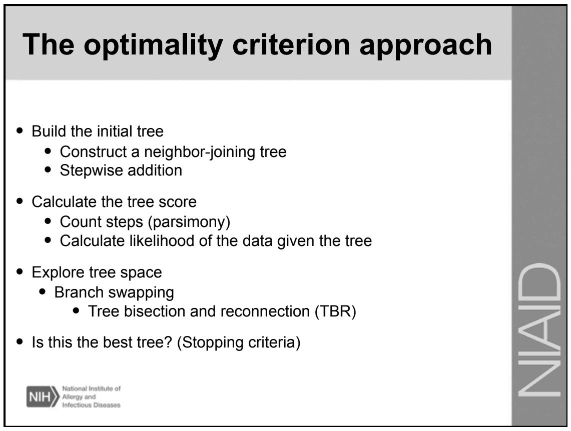


Optimality Criterion: Parsimony

Occam’s Razor: The simplest explanation is the preferred explanation.

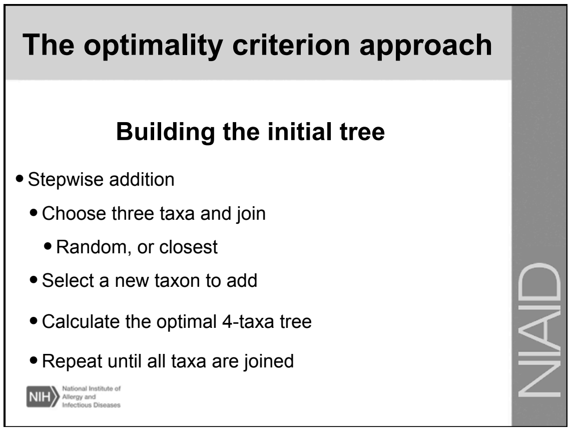
The tree requiring the minimal number of changes is the optimal tree.

A step is any change in the data from one state to another



The optimality criterion approach

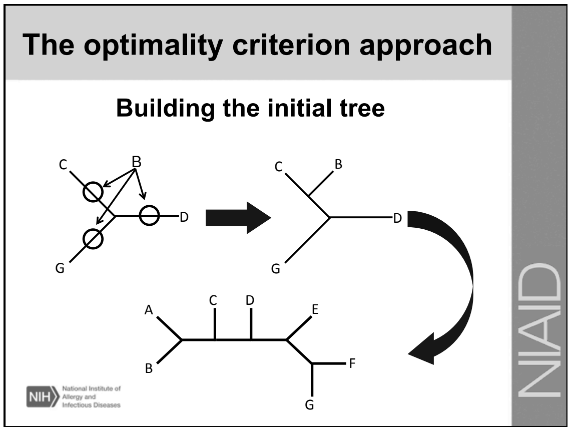
- Build the initial tree
  - Construct a neighbor-joining tree
  - Stepwise addition
- Calculate the tree score
  - Count steps (parsimony)
  - Calculate likelihood of the data given the tree
- Explore tree space
  - Branch swapping
    - Tree bisection and reconnection (TBR)
- Is this the best tree? (Stopping criteria)



The optimality criterion approach

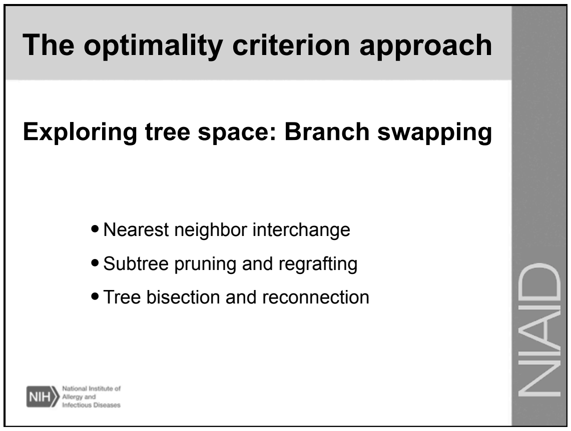
Building the initial tree

- Stepwise addition
  - Choose three taxa and join
    - Random, or closest
  - Select a new taxon to add
  - Calculate the optimal 4-taxa tree
  - Repeat until all taxa are joined



The optimality criterion approach

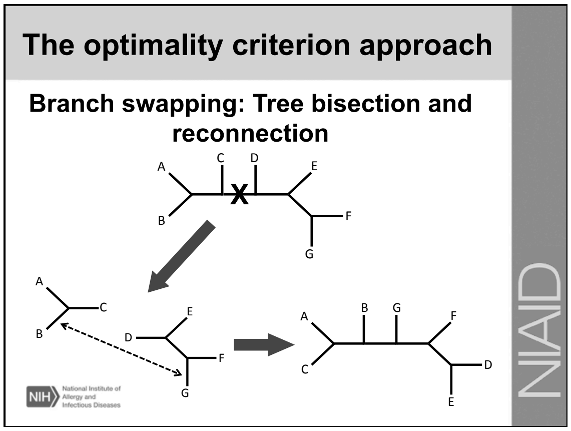
Building the initial tree



The optimality criterion approach

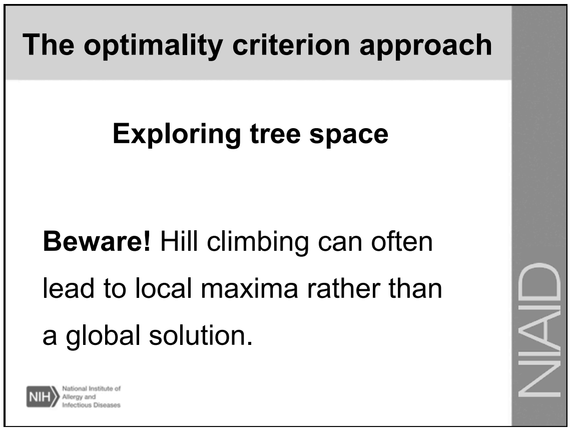
Exploring tree space: Branch swapping

- Nearest neighbor interchange
- Subtree pruning and regrafting
- Tree bisection and reconnection



The optimality criterion approach

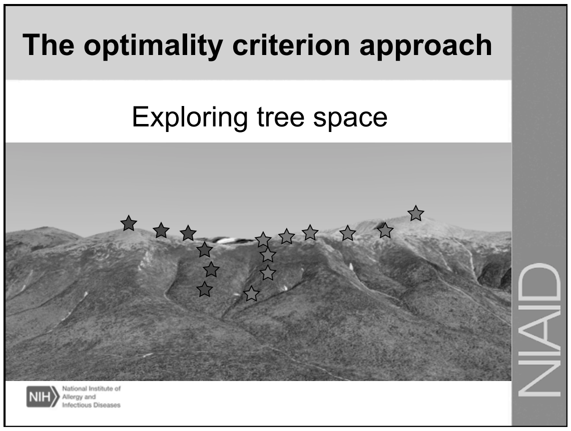
Branch swapping: Tree bisection and reconnection



The optimality criterion approach

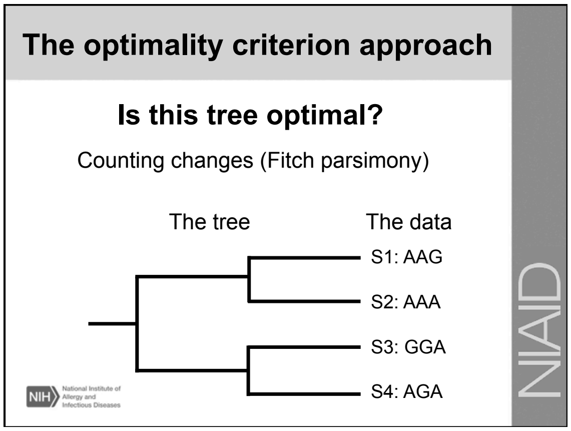
Exploring tree space

Beware! Hill climbing can often lead to local maxima rather than a global solution.



The optimality criterion approach

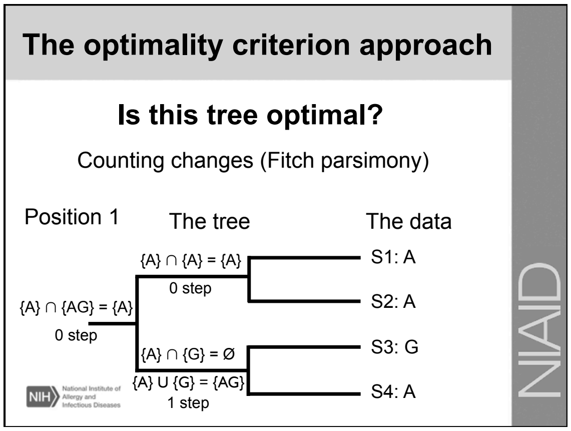
Exploring tree space



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

Position 1

The tree

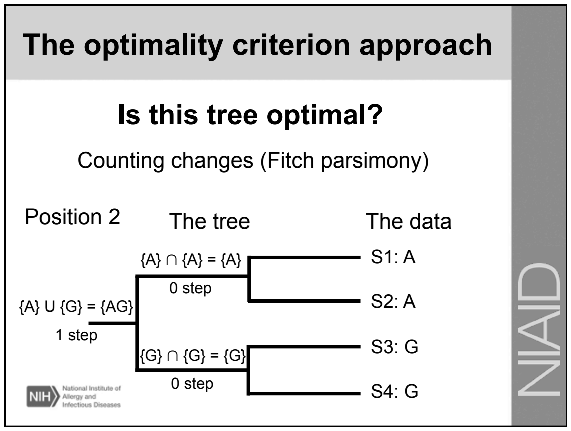
The data

{A} ∩ {A} = {A}
0 step

{A} ∩ {AG} = {A}
0 step

{A} ∩ {G} = Ø
{A} U {G} = {AG}
1 step

S1: A
S2: A
S3: G
S4: A



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

Position 2

The tree

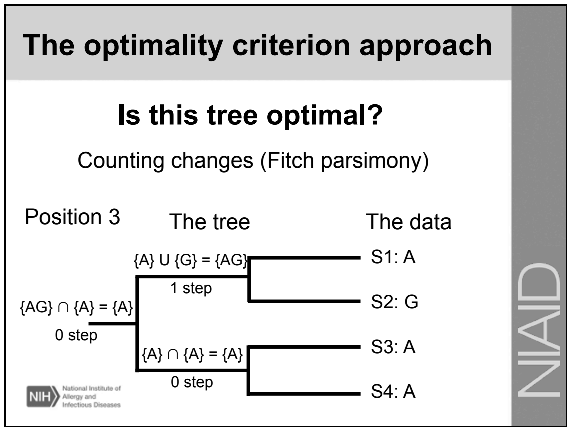
The data

{A} ∩ {A} = {A}
0 step

{A} U {G} = {AG}
1 step

{G} ∩ {G} = {G}
0 step

S1: A
S2: A
S3: G
S4: G



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

Position 3

The tree

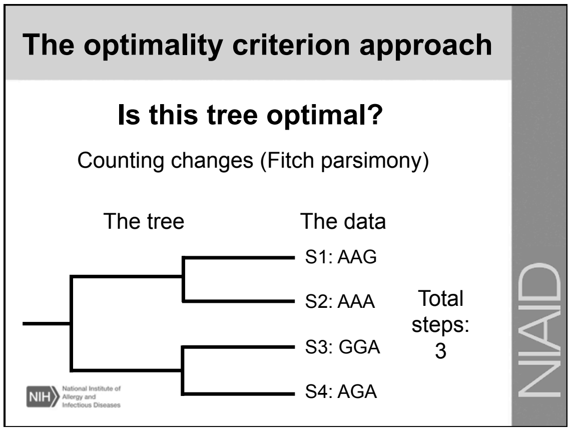
The data

{A} U {G} = {AG}
1 step

{AG} ∩ {A} = {A}
0 step

{A} ∩ {A} = {A}
0 step

S1: A
S2: G
S3: A
S4: A



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data

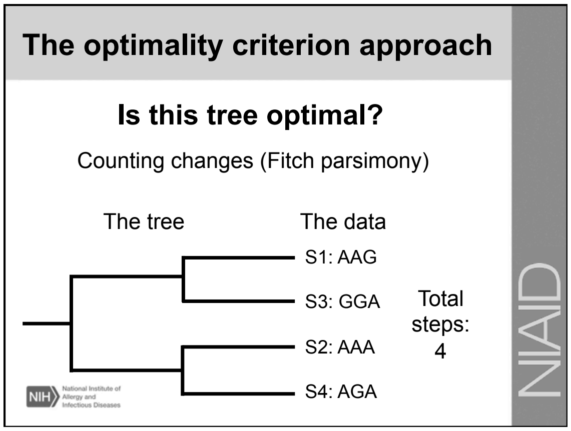
S1: AAG

S2: AAA

S3: GGA

S4: AGA

Total steps: 3



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data

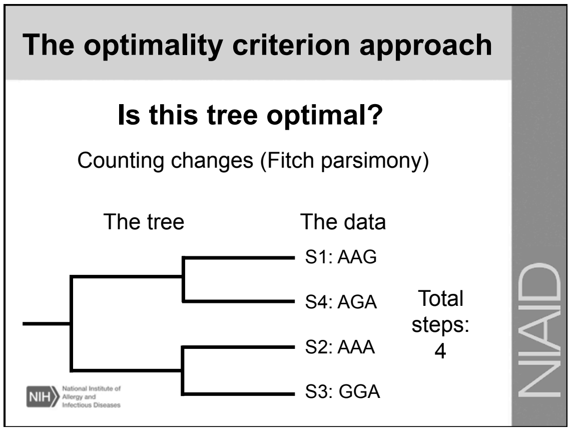
S1: AAG

S3: GGA

S2: AAA

S4: AGA

Total steps: 4



The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data

S1: AAG

S4: AGA

S2: AAA

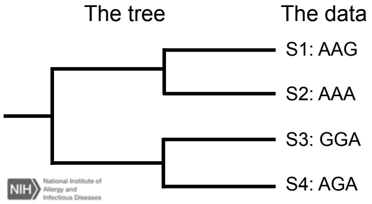
S3: GGA

Total steps: 4

The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)



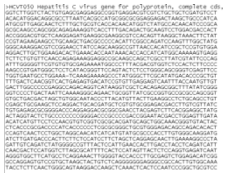
Optimal Tree!

Total steps: 3

Multiple Sequence Alignment

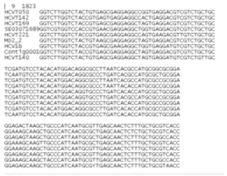
Export MSA as FASTA and PHYLIP formats

FASTA



```
>HCVT050 Hepatitis C virus gene for polyprotein, complete cds,
GGTCTTGGTCTACTGTGAGCGAGGAGGCCGGTGAGGACGTCGTCTGCTGCTCGATGTCCT
ACACATGGACAGGCGCCTTAATCACGCCATGCGGCGCGGAGGAGACTAAGCTGCCCATCA
ATGCGTTGAGCAACTCTTTGCTGCGTCACCACAACATGGTCTATGCCACAACATCCCGCA
GCGCAAGCCAGCGGCAGAGAAAGGTCACCTTTGACAGACTGCAAGTCCTGGACGACCACT
ACCGGGACGTGCTCAAGGAGATGAAGGCGAAGGCGTCCACAGTTAAGGCTAAACTTCTAT
CCGTAGAAGAAGCCTGCAAGCTGACGCCCCCACATTCGGCCAGATCCAAGTTTGGCTATG
GGGCAAAGGACGTCCGGAACCTATCCAGCAAGGCCGTTAACCACATCCGCTCCGTGTGGA
AGGACTTGCTGGAAGACACTGAAACACCAATAAACACCACCATCATGGCAAAAAGTGAGG
TCTTCTGTGTTCAACCAGAGAAAGGAGGCCGCAAGCCAGCTCGCCTTATCGTATTCCCAG
ATTTGGGGGTTCGTGTGTGCGAGAAAATGGCCCTTTACGACGTGGTCTCCACTCTTCCCC
AAGCCGTGATGGGCTCCTCATACGGGTTCCAGTACTCTCCTGGGCAGCGGGTCGAGTTCC
TGGTGAATGCCTGGAAA-TCAAAGAAAAGCCCTATGGGCTTCGCATATGACACCCGCTGT
TTTGACTCAACGGTCACTGAGAGTGACATCCGTGTTGAGGAGTCAATTTACCAATGTTGT
GACTTGGCCCCCGAGGCCAGACAGGTCATAAGGTCGCTCACAGAGCGGCTTTATATCGGG
GGTCCCCTGACTAATTCAAAAGGGCAGAACTGCGGTTATCGCCGGTGCCGCGCCAGCGGT
GTGCTGACGACTAGCTGTGGCAATACCCTTACATGTTACTTGAAGGCCTCTGCAGCCTGT
CGAGCTGCGAAGCTCCAGGACTGCACGATGCTCGTGTGCGGAGACGACCTTGTCGTTATC
TGTGAGAGCGCGGGGACCCAGGAGGACGCGGCGAACCTACGAGTCTTCACGGAGGCTATG
ACTAGGTACTCTGCCCCCCCCGGGGACCCGCCCCGACCGGAATACGACCTGGAGTTGATA
ACATCATGTTCCTCCAACGTGTCGGTCGCGCACGATGCAGCTGGCAAACGGGTGTACTAC
CTCACCCGCGACCCCATCACCCCCCTCGCGCGGGCTGCGTGGGAGACAGCCAGACACACT
CCAGTCAACTCCTGGCTAGGCAACATCATCATGTATGCGCCCACCTTGTGGGCAAGGATG
ATCTTGATGACCCACTTCTTCTCCATCCTTCTAGCTCAGGAGCAACTTGAAAAGGCCCTA
GATTGTCAGATCTATGGGGCCGTTTACTCCATTGAACCACTTGACCTACCTCAGATCATT
CAACGACTCCATGGTCTTAGCGCATTTTCACTCCATAGTTACTCTCCAGGTGAGATCAAT
AGGGTGGCTTCATGCCTCAGGAAACTTGGGGTACCACCCTTGCGAGTCTGGAGACATCGG
GCCAGGAGTGTCCGTGCTAAGCTACTGTCTCAGGGGGGGAGGGCCGCCACTTGTGGCAAA
TACCTCTTCAACTGGGCAGTAAGGACCAAGCTCAAACTCACTCCAATCCCGGCTGCGTCC
```

PHYLIP



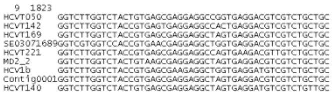
```
 9  1823
HCVT050   GGTCTTGGTCTACTGTGAGCGAGGAGGCCGGTGAGGACGTCGTCTGCTGC
HCVT142   GGTCTTGGTCTACCGTGAGTGAGGAGGCCACTGAGGACGTCGTCTGCTGC
HCVT169   GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
SE03071689GGTCGTGGTCCACCGTGAACGAGGAGGCTGGTGAGGACGTCGTCTGCTGC
HCVT221   GGTCTTGGTCTACCGTGAGCGAGGAGGCCAGTGAAGACGTTGTCTGCTGC
MD2_2     GGTCTTGGTCTACTGTAAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
HCV1b     GGTCTTGGTCTACCGTGAGCGAAGAGGCTGGTGAGGATGTCGTCTGCTGC
Contig0001GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
HCVT140   GGTCTTGGTCTACTGTGAGCGAGGAGGCTAGTGAGGATGTCGTCTGTTGC

TCGATGTCCTACACATGGACAGGCGCCTTAATCACGCCATGCGGCGCGGA
TCAATGTCCTACACATGGACAGGCGCCCTGATCACGCCATGCGCTGCGGA
TCGATGTCCTACACATGGACAGGCGCCTTAATCACGCCATGCGCCGCGGA
TCGATGTCCTACACATGGACAGGCGCCCTAATTACACCATGCGCCGCGGA
TCAATGTCCTACACATGGACAGGTGCCCTGATCACACCATGCGCTGCGGA
TCGATGTCCTACACATGGACAGGTGCCTTGATCACGCCATGCGCCGCGGA
TCGATGTCCTACACATGGACAGGCGCCCTGATCACACCATGCGCCGCGGA
TCGATGTCCTACACATGGACAGGCGCCTTGATCACACCATGCGCCGCGGA
TCAATGTCCTACACATGGACGGGCGCCCTGATCACACCATGCGCTGCGGA

GGAGACTAAGCTGCCCATCAATGCGTTGAGCAACTCTTTGCTGCGTCACC
GGAAAGCAAGCTGCCCATCAACGCGCTGAGCAACTCTCTGCTGCGTCACC
GGAGAGTAAGCTGCCCATTAATGCGCTGAGCAACTCTTTGCTGCGCCACC
GGAAAGCAAGTTGCCCATTAATGCGCTGAGCAACTCTTTGCTGCGTCACC
GGAAAGCAAGCTGCCCATCAATGCGTTGAGCAACTCTTTGCTGCGTCACC
GGAGAGCAAGCTGCCCATCAACGCGTTGAGCAACTCTTTGCTGCGTCACC
GGAAAGCAAGCTACCCATCAATGCGTTGAGCAACTCTTTGCTGCGCCACC
GGAGAGCAAGCTGCCCATCAATGCGYTGAGCAACTCTTTGCTGCGTAACC
```

Multiple Sequence Alignment

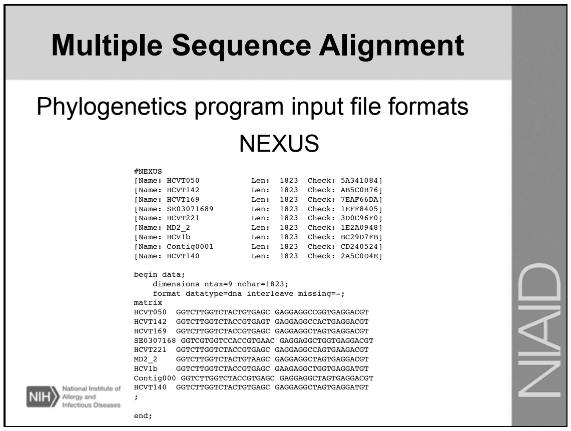
Phylogenetics program input file formats

PHYLIP

1st line: Number of sequences(space)Number of sites
2nd line: Sequence ID (10 characters max) Sequence



```
  9  1823
HCVT050   GGTCTTGGTCTACTGTGAGCGAGGAGGCCGGTGAGGACGTCGTCTGCTGC
HCVT142   GGTCTTGGTCTACCGTGAGTGAGGAGGCCACTGAGGACGTCGTCTGCTGC
HCVT169   GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
SE03071689GGTCGTGGTCCACCGTGAACGAGGAGGCTGGTGAGGACGTCGTCTGCTGC
HCVT221   GGTCTTGGTCTACCGTGAGCGAGGAGGCCAGTGAAGACGTTGTCTGCTGC
MD2_2     GGTCTTGGTCTACTGTAAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
HCV1b     GGTCTTGGTCTACCGTGAGCGAAGAGGCTGGTGAGGATGTCGTCTGCTGC
Contig0001GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
HCVT140   GGTCTTGGTCTACTGTGAGCGAGGAGGCTAGTGAGGATGTCGTCTGTTGC
```

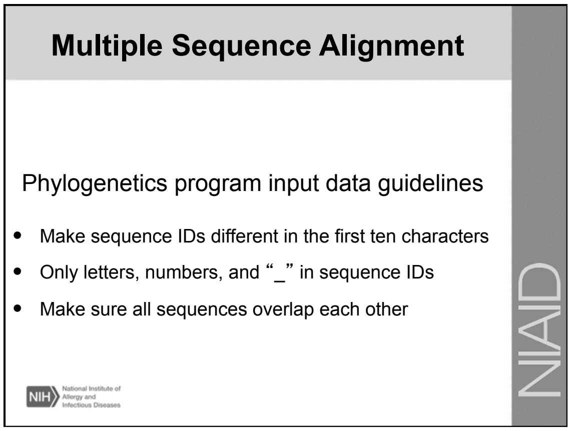

NIAID
NIH National Institute of Allergy and Infectious Diseases

Multiple Sequence Alignment

Phylogenetics program input file formats

NEXUS

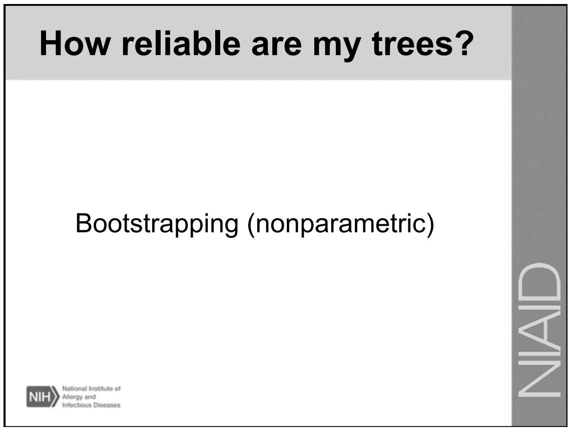
```
#NEXUS
[Name: HCVT050          Len:  1823  Check: 5A341084]
[Name: HCVT142          Len:  1823  Check: AB5C0B76]
[Name: HCVT169          Len:  1823  Check: 7EAF66DA]
[Name: SE03071689       Len:  1823  Check: 1EFF8405]
[Name: HCVT221          Len:  1823  Check: 3D0C96F0]
[Name: MD2_2            Len:  1823  Check: 1E2A0948]
[Name: HCV1b            Len:  1823  Check: BC29D7FB]
[Name: Contig0001       Len:  1823  Check: CD240524]
[Name: HCVT140          Len:  1823  Check: 2A5C0D4E]

begin data;
    dimensions ntax=9 nchar=1823;
    format datatype=dna interleave missing=-;
matrix
HCVT050  GGTCTTGGTCTACTGTGAGC GAGGAGGCCGGTGAGGACGT
HCVT142  GGTCTTGGTCTACCGTGAGT GAGGAGGCCACTGAGGACGT
HCVT169  GGTCTTGGTCTACCGTGAGC GAGGAGGCTAGTGAGGACGT
SE0307168 GGTCGTGGTCCACCGTGAAC GAGGAGGCTGGTGAGGACGT
HCVT221  GGTCTTGGTCTACCGTGAGC GAGGAGGCCAGTGAAGACGT
MD2_2    GGTCTTGGTCTACTGTAAGC GAGGAGGCTAGTGAGGACGT
HCV1b    GGTCTTGGTCTACCGTGAGC GAAGAGGCTGGTGAGGATGT
Contig000 GGTCTTGGTCTACCGTGAGC GAGGAGGCTAGTGAGGACGT
HCVT140  GGTCTTGGTCTACTGTGAGC GAGGAGGCTAGTGAGGATGT
;

end;
```


NIAID
NIH National Institute of Allergy and Infectious Diseases

Multiple Sequence Alignment

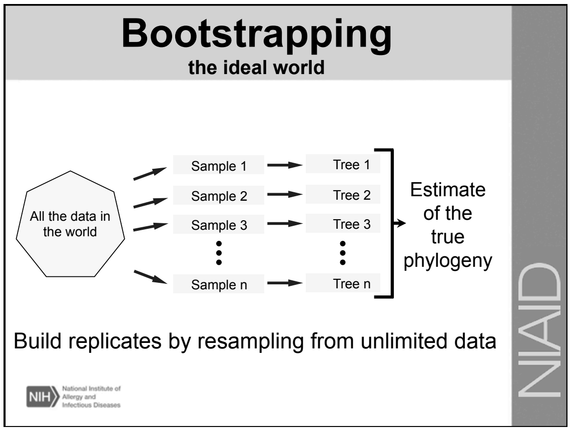
Phylogenetics program input data guidelines

- Make sequence IDs different in the first ten characters
- Only letters, numbers, and "_" in sequence IDs
- Make sure all sequences overlap each other


NIAID
NIH National Institute of Allergy and Infectious Diseases

How reliable are my trees?

Bootstrapping (nonparametric)



Bootstrapping

the ideal world

All the data in the world

Sample 1 → Tree 1

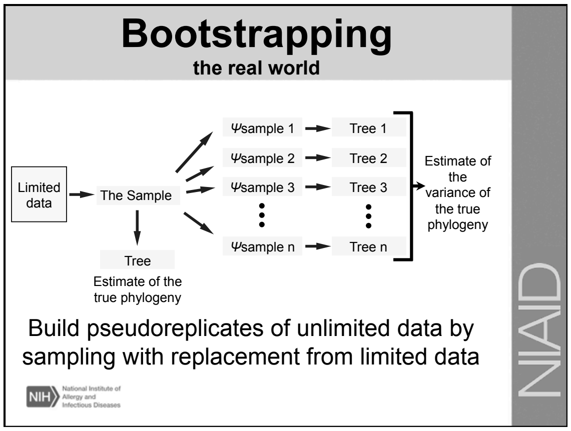
Sample 2 → Tree 2

Sample 3 → Tree 3

Sample n → Tree n

Estimate of the true phylogeny

Build replicates by resampling from unlimited data



Bootstrapping

the real world

Limited data → The Sample

Ψsample 1 → Tree 1

Ψsample 2 → Tree 2

Ψsample 3 → Tree 3

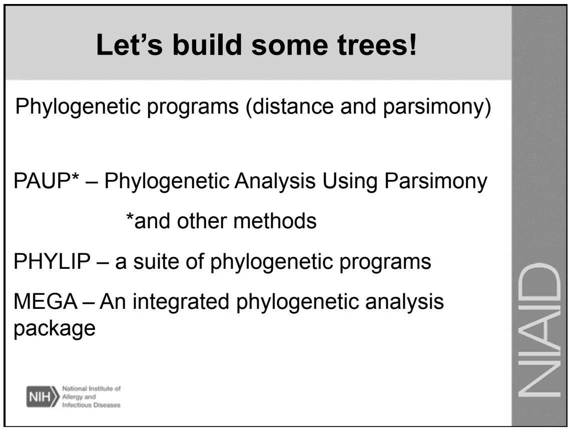
Ψsample n → Tree n

Estimate of the variance of the true phylogeny

Tree

Estimate of the true phylogeny

Build pseudoreplicates of unlimited data by sampling with replacement from limited data



Let's build some trees!

Phylogenetic programs (distance and parsimony)

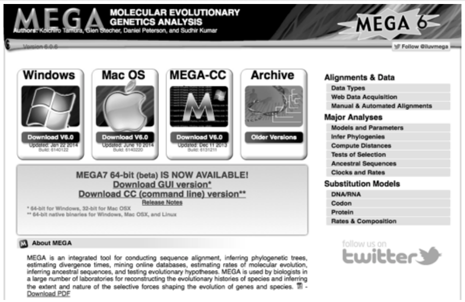
PAUP* – Phylogenetic Analysis Using Parsimony

*and other methods

PHYLIP – a suite of phylogenetic programs

MEGA – An integrated phylogenetic analysis package

Let’s build some trees!



MEGA MOLECULAR EVOLUTIONARY GENETICS ANALYSIS

Authors: Koichiro Tamura, Glen Stecher, Daniel Peterson, and Sudhir Kumar

Version 6.0.6

Windows	Mac OS	MEGA-CC	Archive
Download V6.0 Updated: Jan 22 2014 Build: 6140122	Download V6.0 Updated: June 10 2014 Build: 6140220	Download V6.0 Updated: Dec 11 2013 Build: 6131211	Older Versions

MEGA7 64-bit (beta) IS NOW AVAILABLE!
Download GUI version*
Download CC (command line) version**
Release Notes

* 64-bit for Windows, 32-bit for Mac OSX
** 64-bit native binaries for Windows, Mac OSX, and Linux

About MEGA

MEGA is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, estimating divergence times, mining online databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses. MEGA is used by biologists in a large number of laboratories for reconstructing the evolutionary histories of species and inferring the extent and nature of the selective forces shaping the evolution of genes and species. - Download PDF

Alignments & Data
- Data Types
- Web Data Acquisition
- Manual & Automated Alignments

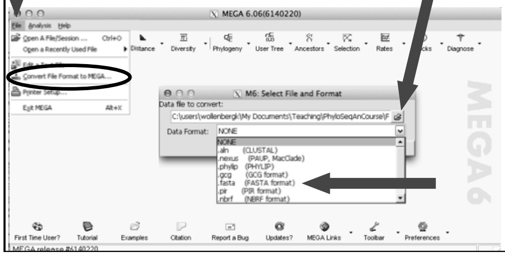
Major Analyses
- Models and Parameters
- Infer Phylogenies
- Compute Distances
- Tests of Selection
- Ancestral Sequences
- Clocks and Rates

Substitution Models
- DNA/RNA
- Codon
- Protein
- Rates & Composition

MEGA6

Importing a FASTA-formatted alignment

File > Convert to MEGA format
Navigate to FASTA file
Data Format > .fasta (FASTA format)



MEGA 6.06(6140220)

M6: Select File and Format

Data file to convert: C:\users\wollenbergk\My Documents\Teaching\PhyloSeqAnCourse\F

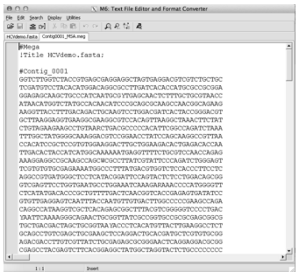
Data Format: NONE

```
NONE
.aln     (CLUSTAL)
.nexus   (PAUP, MacClade)
.phylip  (PHYLIP)
.gcg     (GCG format)
.fasta   (FASTA format)
.pir     (PIR format)
.nbrf    (NBRF format)
```

MEGA release #6140220

MEGA6

MEGA data file format

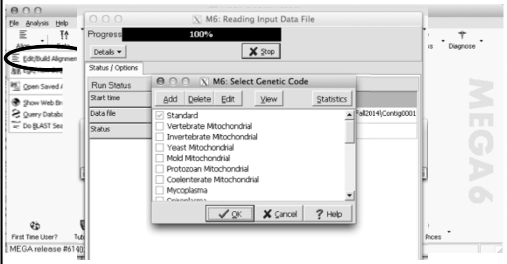


M6: Text File Editor and Format Converter

```
#Mega
!Title HCVdemo.fasta;

#Contig_0001
GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGC
TCGATGTCCTACACATGGACAGGCGCCTTGATCACACCATGCGCCGCGGA
GGAGAGCAAGCTGCCCATCAATGCGYTGAGCAACTCTTTGCTGCGTAACC
ATAACATGGTCTATGCCACAACATCCCGCAGCGCAAGCCAACGGCAGAAG
AAGGTTACCTTTGACAGACTGCAAGTCCTGGACGATCACTACCGGGACGT
GCTTAAGGAGGTGAAGGCGAAGGCGTCCACAGTTAAGGCTAAACTTCTAT
CTGTAGAAGAAGCCTGTAARCTGACGCCCCCACATTCGGCCAGATCTAAA
TTTGGCTATGGGGCAAAGGACGTCCGGAACCTATCCAGCAAGGCCGTTAA
CCACATCCGCTCCGTGTGGAAGGACTTGCTGGAAGACACTGAGACACCAA
TTGACACTACCATCATGGCAAAAAATGAGGTTTTCTGCGTCCAACCAGAG
AAAGGAGGCCGCAAGCCAGCWCGCCTTATCGTATTCCCAGATCTGGGAGT
TCGTGTGTGCGAGAAAATGGCCCTTTATGACGTGGTCTCCACCCTTCCTC
AGGCCGTGATGGGCTCCTCATACGGATTCCAGTACTCTCCTGGACAGCGG
GTCGAGTTCCTGGTGAATGCCTGGAAATCAAAGARAAACCCCATGGGGTT
CTCATATGACACCCGCTGTTTTGACTCAACGGTCACCGAGAGTGATATCC
GTGTTGAGGAGTCAATTTACCAATGTTGTGACTTGGCCCCCGAAGCCAGA
CAGGCCATAAGGTCGCTCACAGAGCGGCTTTACGTCGGGGGTCCCCTGAC
YAATTCAAAAGGGCAGAACTGCGGTTATCGCCGGTGCCGCGCGAGCGGCG
TGCTGACGACTAGCTGCGGTAAYACCCTCACATGTTACTTGAAGGCCTCT
GCAGCCTGTCGAGCTGCGAAGCTCCAGGACTGCACGATGCTCGTGTGCGG
AGACGACCTTGTCGTTATCTGCGAGAGCGCGGGAACTCAGGAGGACGCGG
CGAGCCTACGAGTCTTCACGGAGGCTATGGCTAGGTACTCTGCCCCCCCC
```

MEGA6

Imported sequence alignment: Alignment Explorer



M6: Reading Input Data File

Progress 100%

Run Status	
Start time	
Data file	Fall2014\Contig0001
Status	

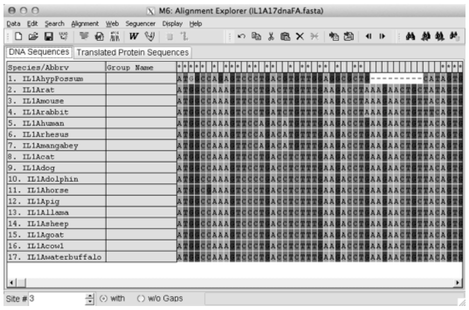
M6: Select Genetic Code

- ☑ Standard
- □ Vertebrate Mitochondrial
- □ Invertebrate Mitochondrial
- □ Yeast Mitochondrial
- □ Mold Mitochondrial
- □ Protozoan Mitochondrial
- □ Coelenterate Mitochondrial
- □ Mycoplasma

MEGA release #6140

MEGA6

Imported sequence alignment: Alignment Explorer



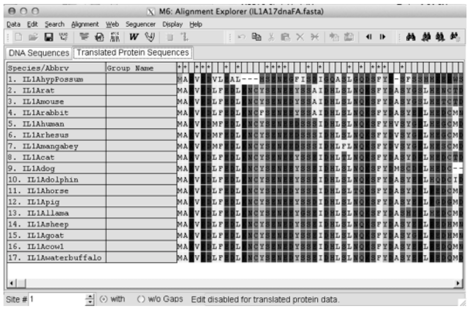
M6: Alignment Explorer (IL1A17dnaFA.fasta)

```
Species/Abbrv          Group Name
1. IL1AhypPossum                  ATGGCCAGAGTCCCTGACGTGTTGGAGGCGCTG---------CATAGTG
2. IL1Arat                        ATGGCCAAAGTTCCTGACTTGTTTGAAGACCTAAAGAACTGCTATAGTG
3. IL1Amouse                      ATGGCCAAAGTTCCTGACTTGTTTGAAGACCTAAAGAACTGTTACAGTG
4. IL1Arabbit                     ATGGCCAAAGTCCCTGATCTGTTTGAAGACCTAAAGAACTGTTTCAGTG
5. IL1Ahuman                      ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTG
6. IL1Arhesus                     ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTG
7. IL1Amangabey                   ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTG
8. IL1Acat                        ATGGCCAAAGTTCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
9. IL1Adog                        ATGGCCAAAGTTCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
10. IL1Adolphin                   ATGGCCAAAGTCCCCGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
11. IL1Ahorse                     ATGGCGAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
12. IL1Apig                       ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGCTACAGTG
13. IL1Allama                     ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
14. IL1Asheep                     ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
15. IL1Agoat                      ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
16. IL1Acow1                      ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
17. IL1Awaterbuffalo              ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
```

Site # 3 ● with ○ w/o Gaps

MEGA6

Imported sequence alignment: Alignment Explorer



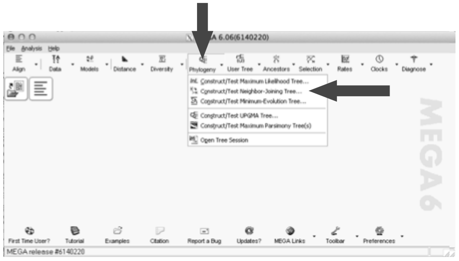
M6: Alignment Explorer (IL1A17dnaFA.fasta)

```
Species/Abbrv          Group Name
1. IL1AhypPossum
2. IL1Arat
3. IL1Amouse
4. IL1Arabbit
5. IL1Ahuman
6. IL1Arhesus
7. IL1Amangabey
8. IL1Acat
9. IL1Adog
10. IL1Adolphin
11. IL1Ahorse
12. IL1Apig
13. IL1Allama
14. IL1Asheep
15. IL1Agoat
16. IL1Acow1
17. IL1Awaterbuffalo
```

Site # 1 ● with ○ w/o Gaps Edit disabled for translated protein data.

MEGA6

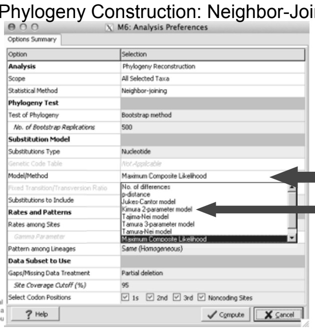
Phylogeny Construction: Neighbor-Joining



MEGA release #6140220

MEGA6

Phylogeny Construction: Neighbor-Joining



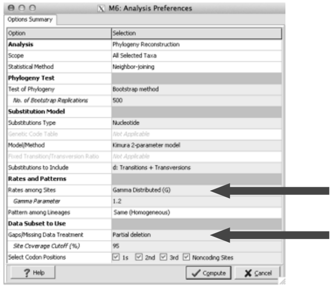
M6: Analysis Preferences

Options Summary

Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Maximum Composite Likelihood
Fixed Transition/Transversion Ratio	No. of differences p-distance Jukes-Cantor model Kimura 2-parameter model Tajima-Nei model Tamura 3-parameter model Tamura-Nei model Maximum Composite Likelihood
Substitutions to Include	
Rates and Patterns	
Rates among Sites	
Gamma Parameter	
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Select Codon Positions	☑ 1s ☑ 2nd ☑ 3rd ☑ Noncoding Sites

MEGA6

Phylogeny Construction: Neighbor-Joining



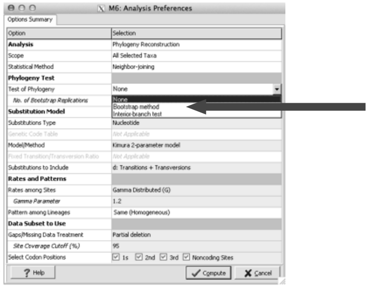
M6: Analysis Preferences

Options Summary

Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Kimura 2-parameter model
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	d: Transitions + Transversions
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
Gamma Parameter	1.2
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Select Codon Positions	☑ 1s ☑ 2nd ☑ 3rd ☑ Noncoding Sites

MEGA6

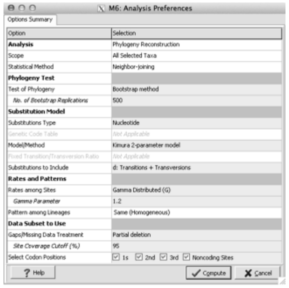
Phylogeny Construction: Neighbor-Joining



Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	None Bootstrap method Interior-branch test
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Kimura 2-parameter model
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	d: Transitions + Transversions
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
Gamma Parameter	1.2
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Select Codon Positions	☑ 1s ☑ 2nd ☑ 3rd ☑ Noncoding Sites

MEGA6

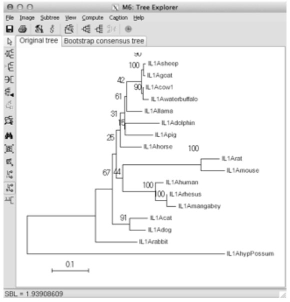
Phylogeny Construction: Neighbor-Joining



Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Kimura 2-parameter model
Fixed Transition/Transversion Ratio	*Not Applicable*
Substitutions to Include	d: Transitions + Transversions
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
Gamma Parameter	1.2
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Select Codon Positions	☑ 1s ☑ 2nd ☑ 3rd ☑ Noncoding Sites

MEGA6

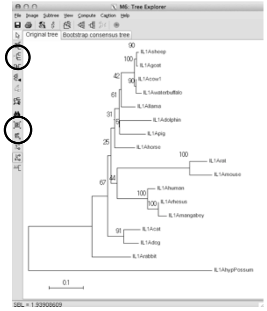
Phylogeny Construction: Neighbor-Joining



SBL = 1.93908609

MEGA6

Phylogeny Construction: Neighbor-Joining

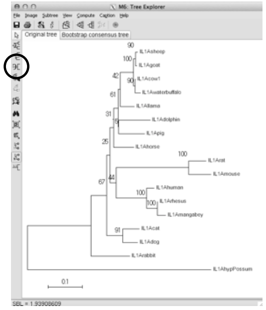


NIH National Institute of Allergy and Infectious Diseases

NIAID

MEGA6

Phylogeny Construction: Neighbor-Joining



NIH National Institute of Allergy and Infectious Diseases

NIAID

MEGA6

Phylogeny Construction: The Caption Option

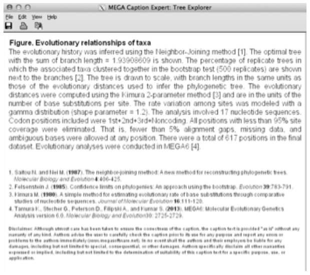


Figure. Evolutionary relationships of taxa

The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 1.93908609 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Kimura 2-parameter method [3] and are in the units of the number of base substitutions per site. The rate variation among sites was modeled with a gamma distribution (shape parameter = 1.2). The analysis involved 17 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 617 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 [4].

1. Saitou N. and Nei M. (**1987**). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.
2. Felsenstein J. (**1985**). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.
3. Kimura M. (**1980**). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111-120.
4. Tamura K., Stecher G., Peterson D., Filipski A., and Kumar S. (**2013**). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*30: 2725-2729.

Disclaimer: Although utmost care has been taken to ensure the correctness of the caption, the caption text is provided "as is" without any warranty of any kind. Authors advise the user to carefully check the caption prior to its use for any purpose and report any errors or problems to the authors immediately (www.megasoftware.net). In no event shall the authors and their employers be liable for any damages, including but not limited to special, consequential, or other damages. Authors specifically disclaim all other warranties expressed or implied, including but not limited to the determination of suitability of this caption text for a specific purpose, use, or application.

NIH National Institute of Allergy and Infectious Diseases

NIAID

Recapitulation

Where have we been? What have we done?

- What is a phylogenetic tree?
- How to build trees
  - Distance
  - Parsimony
- Calculated bootstrap support

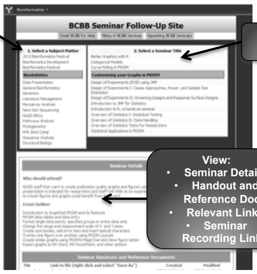
NIH National Institute of Allergy and Infectious Diseases

NIAID

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



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
- 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 usi
presentation is intended for researchers and staff with little to no experie
to create figures and graphs could benefit from this event.

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

2. Select a Topic

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)

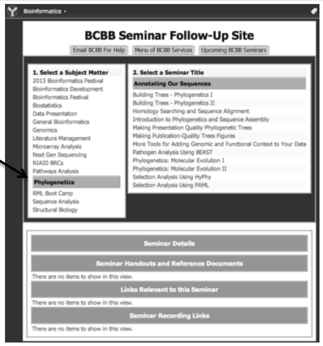
Login
- If prompted to log in use "NIH\" in front of your username

NIH National Institute of Allergy and Infectious Diseases

NIAID

56

Retrieving Slides/Handouts



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.

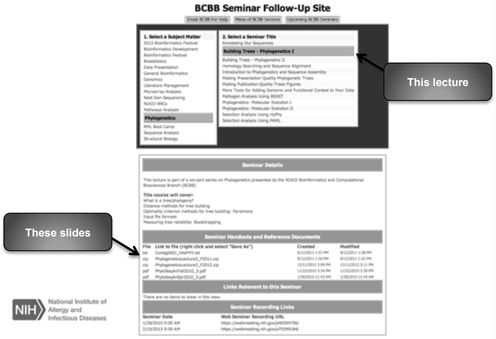
This lecture series

NIH National Institute of Allergy and Infectious Diseases

NIAID

57

Retrieving Slides/Handouts



BCBB Seminar Follow-Up Site

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

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:
What is a tree/phylogeny?
Distance methods for tree building
Optimality criterion methods for tree building: Parsimony
Input file formats
Measuring tree reliability: Bootstrapping

Seminar Handouts and Reference Documents

File	Link to file (right click and select "Save As")	Created	Modified
txt	Contig0001_hitsPHY.txt	9/12/2011 1:57 PM	9/12/2011 1:58 PM
zip	PhylogeneticsLecture3_F2011.zip	9/12/2011 1:50 PM	9/12/2011 1:52 PM
zip	PhylogeneticsLecture3_F2012.zip	12/11/2012 3:09 PM	12/11/2012 3:11 PM
pdf	PhyloSeqAnFall2010_3.pdf	11/22/2010 3:34 PM	11/22/2010 3:39 PM
pdf	PhyloSeqAnSpr2010_3.pdf	1/28/2010 11:43 AM	1/28/2010 11:43 AM

Links Relevant to this Seminar

There are no items to show in this view.

Seminar Recording Links

Seminar Date	Web Seminar Recording URL
1/28/2010 9:00 AM	https://webmeeting.nih.gov/p40294756/
2/19/2013 9:00 AM	https://webmeeting.nih.gov/p75599184/

This lecture

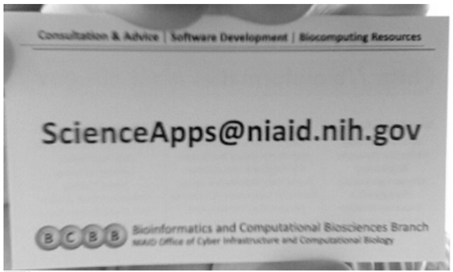
These slides

NIH National Institute of Allergy and Infectious Diseases

NIAID

58

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

59

Next

Tuesday, 17 November at 0930

jModeltest: Fitting analysis parameters to your data

GARLi: Genetic Algorithm for Rapid Likelihood inference

MrBayes: Robust statistical phylogeny inference

NIH National Institute of Allergy and Infectious Diseases

NIAID

60