National Institute of Allergy and Infectious Diseases

Virus Sequence Alignment and Phylogenetics

Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biology Branch (BCBB)
Office of Cyber Infrastructure and Computational Biology

NIAID

National Institute of
Allergy and
Infectious Diseases

21 June 2019

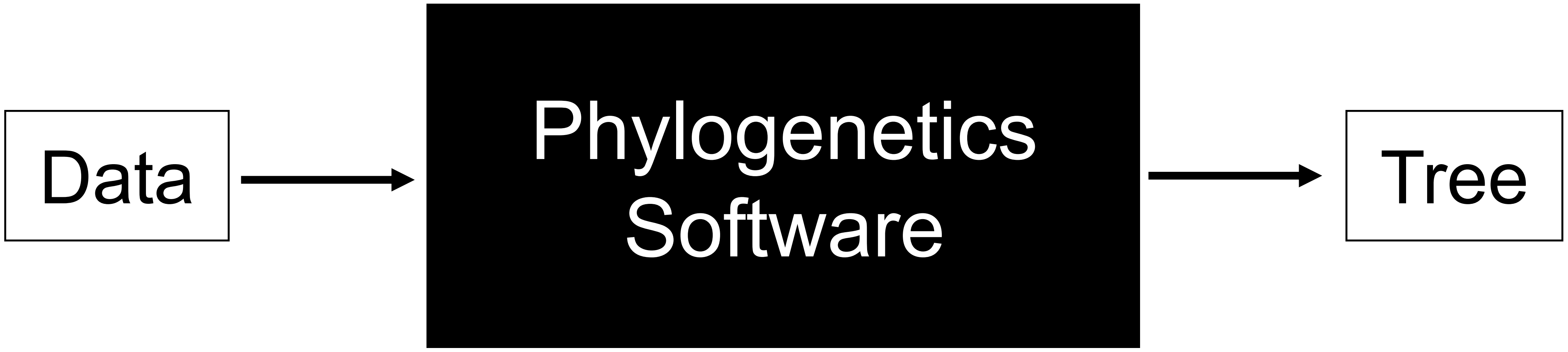
We Are BCBB!

- Group of (so far) 52
  - Bioinformatics Software Developers
  - Computational Biologists
  - Project Management & Analysis Professionals

NIH National Institute of Allergy and Infectious Diseases

NIAID

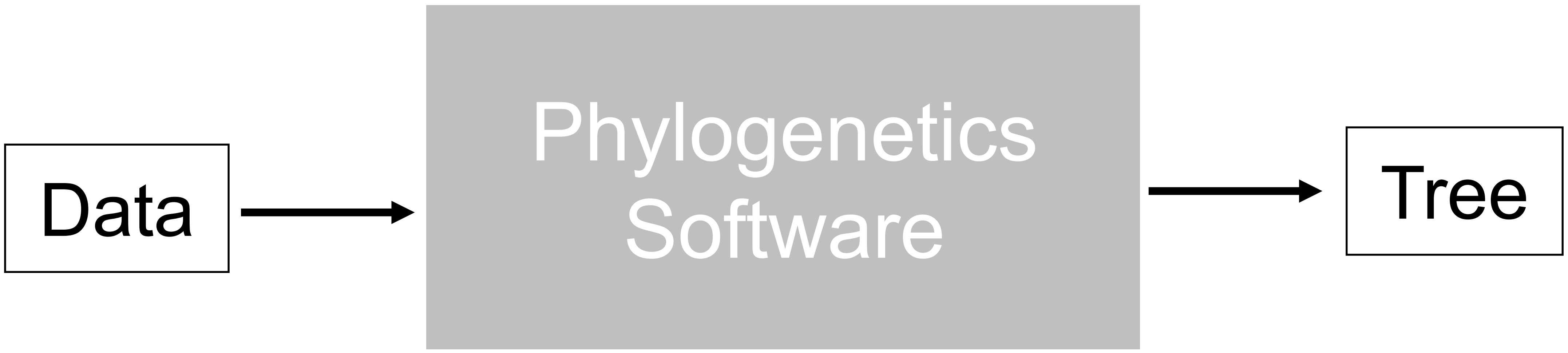
The Goal



Data

Phylogenetics Software

Tree



Data

Phylogenetics Software

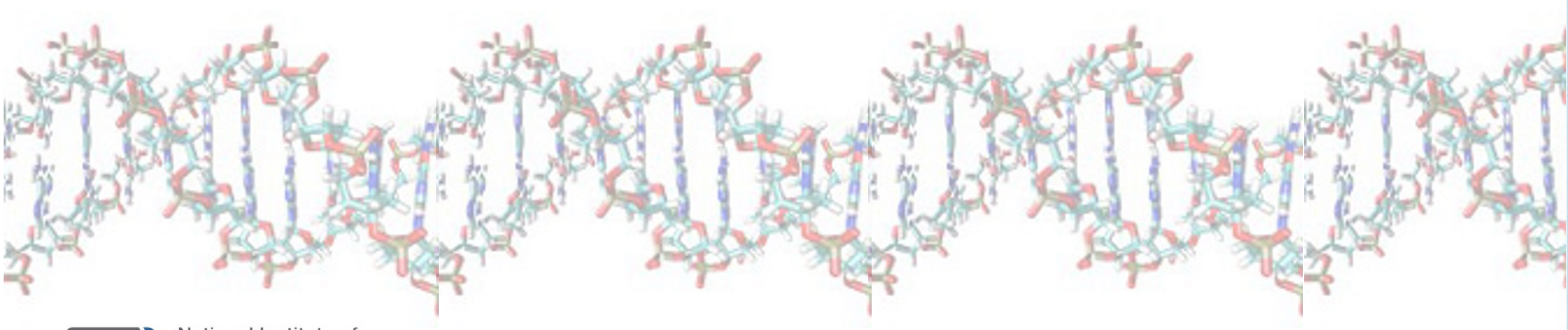
Tree

Biological sequences

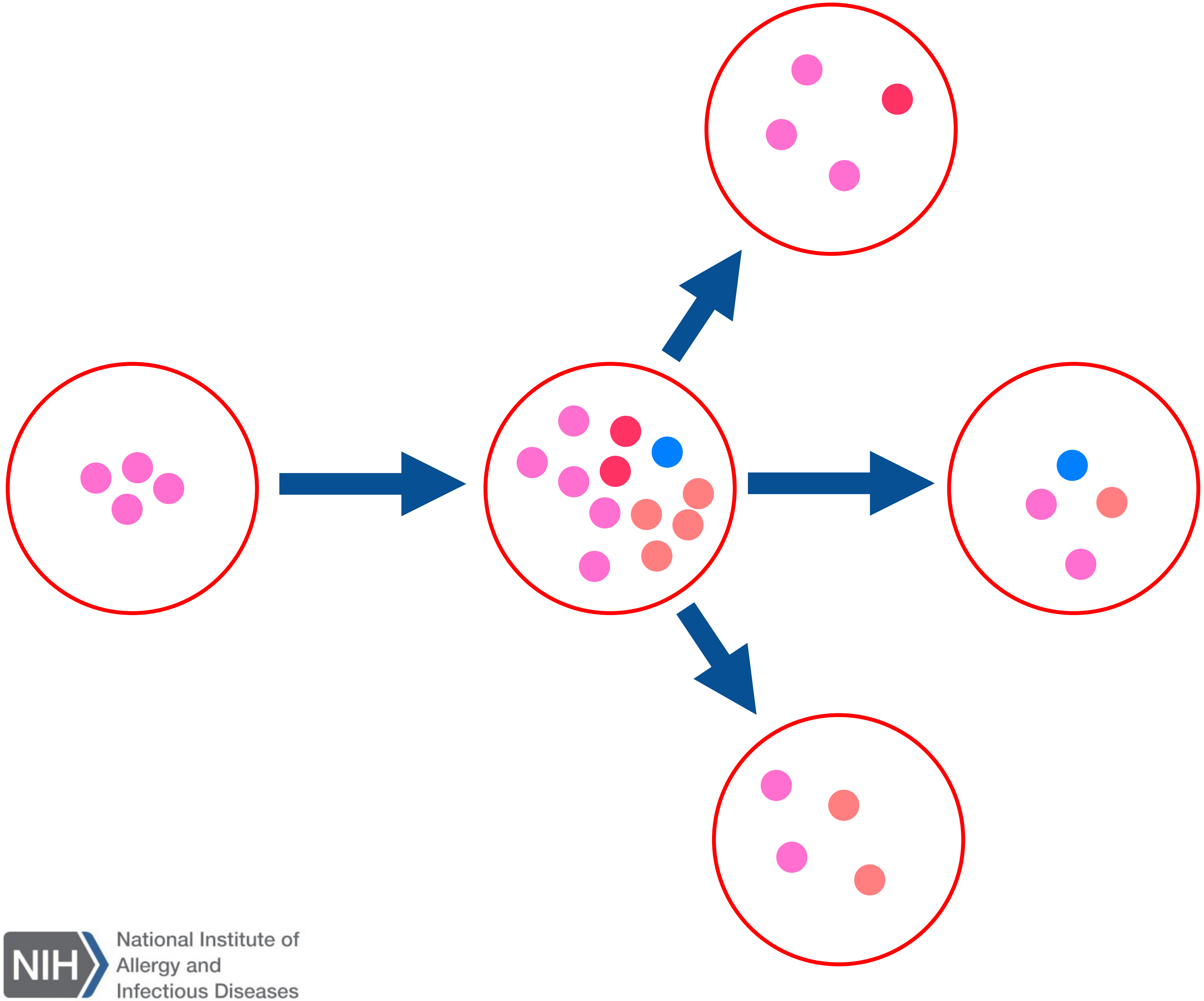
Why analyze biological sequences?

Biological sequences

- DNA contains the information basic to every process in a cell
- Proteins (and RNA) are the machines performing cellular processes
- Passed from one generation to the next



Sequence data are genealogical

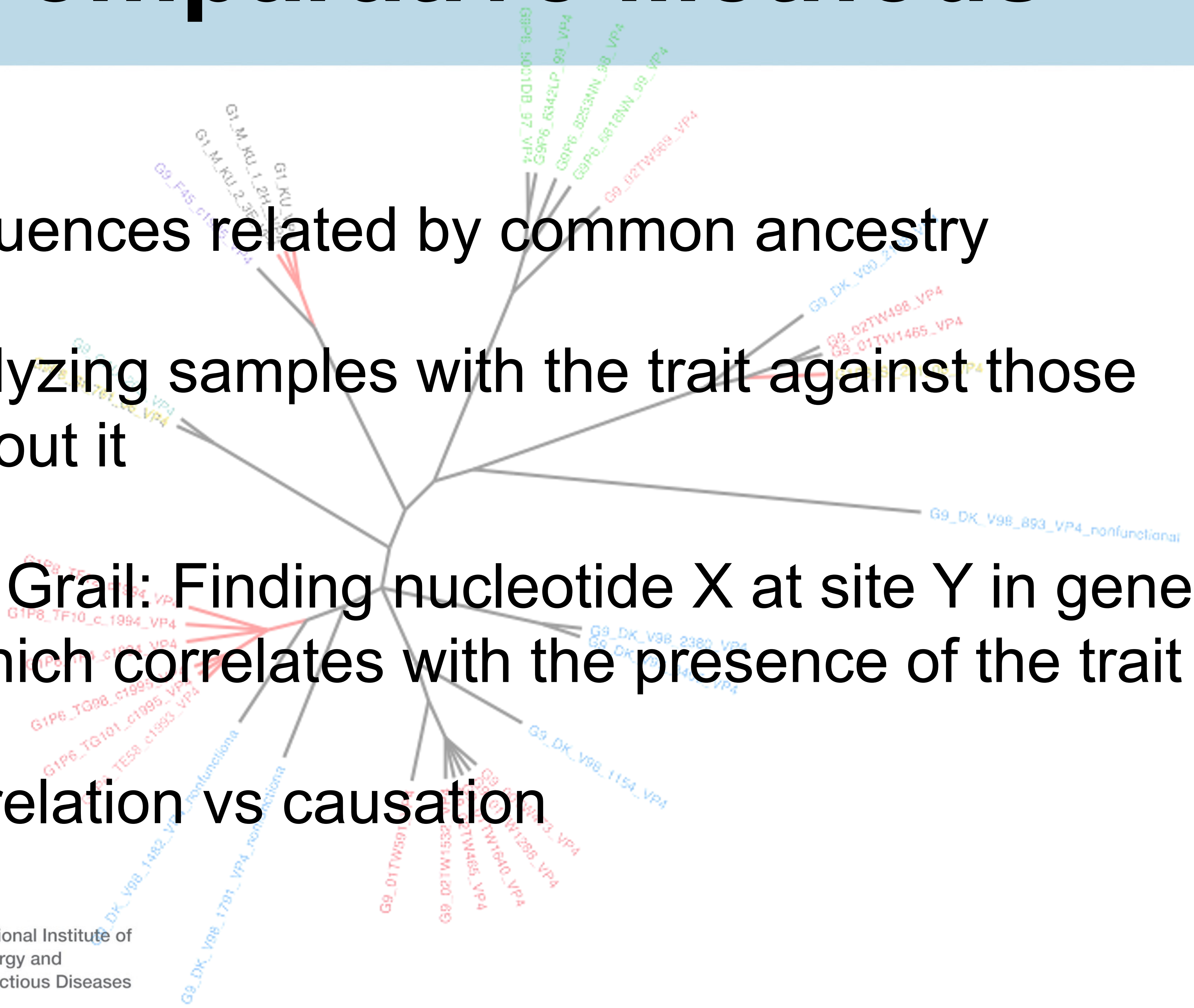


National Institute of Allergy and Infectious Diseases

Comparative Methods

Why analyze sequences using comparative methods?

Comparative Methods



- Sequences related by common ancestry
- Analyzing samples with the trait against those without it
- The Grail: Finding nucleotide X at site Y in gene Z which correlates with the presence of the trait
- Correlation vs causation

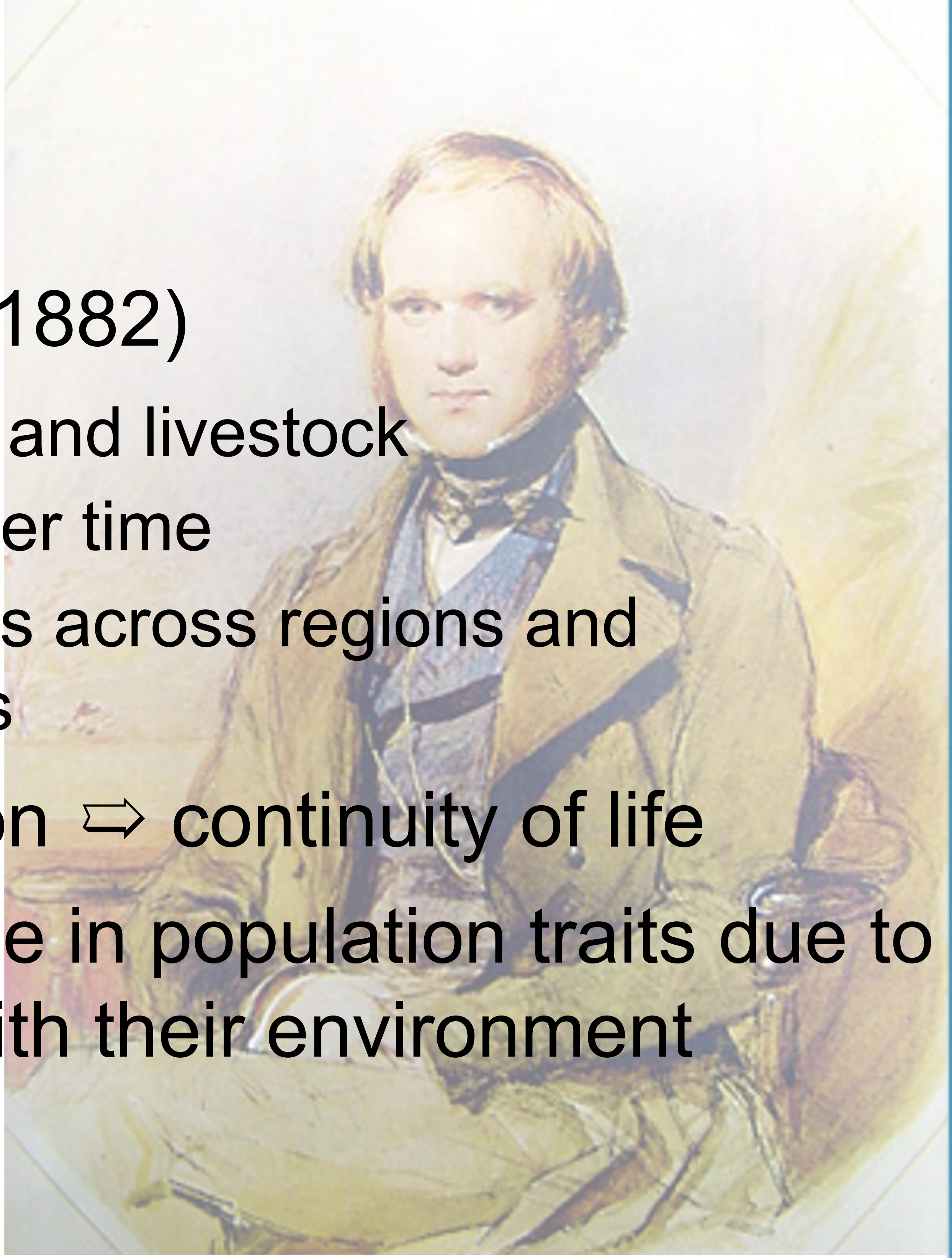
Hierarchy of Life

- Carl Linnaeus (1707 - 1778)
  - Swedish physician/naturalist
  - Hierarchical organization of life
  - Binomial system of scientific names

NIH National Institute of Allergy and Infectious Diseases

NIAID

Common Ancestry

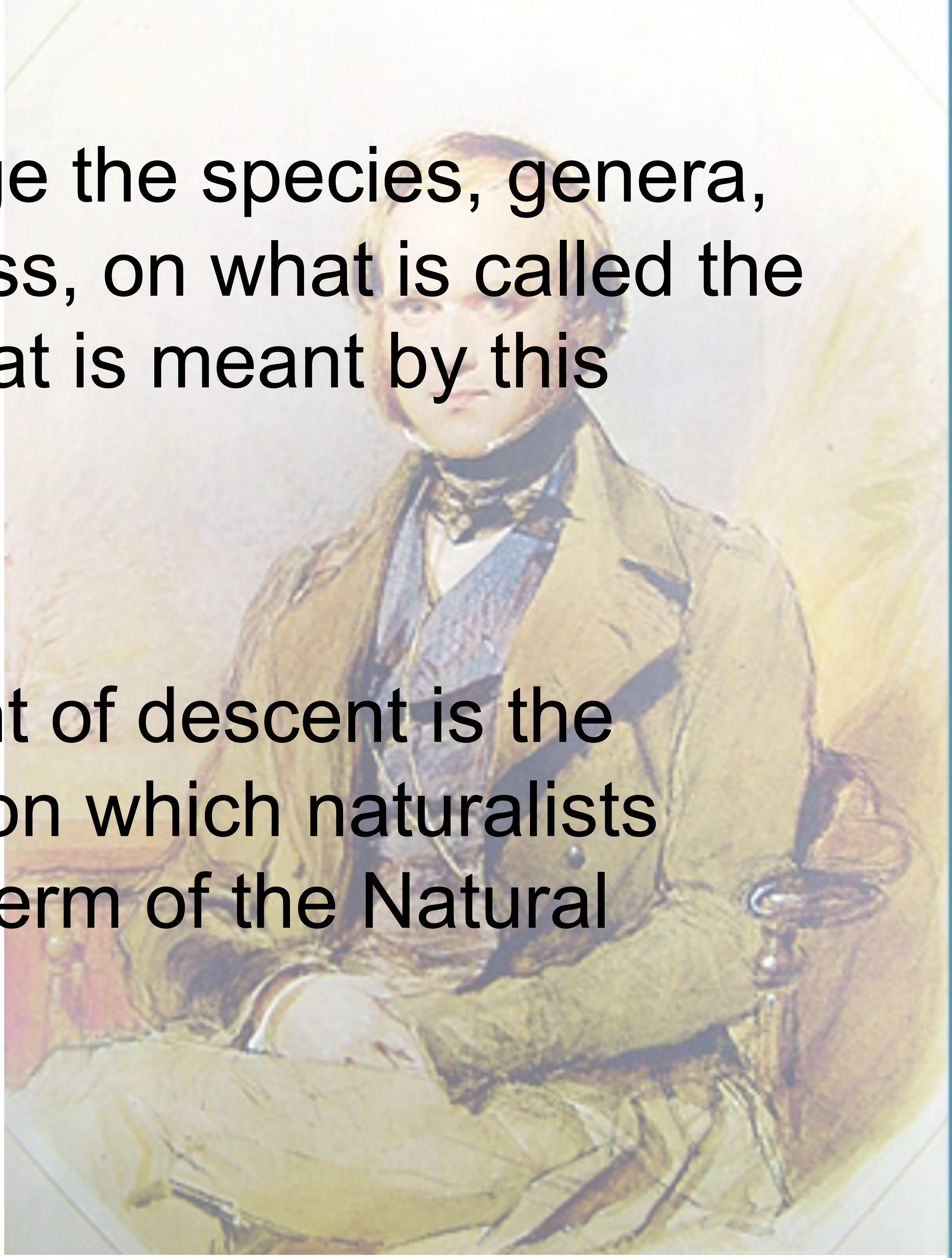


- Charles Darwin (1809 - 1882)
  - Artificial selection: crops and livestock
  - Fossil record: change over time
  - Biogeography: similarities across regions and differences within locales
- Descent with modification ⇨ continuity of life
- Natural selection: change in population traits due to individual interactions with their environment

NIH National Institute of Allergy and Infectious Diseases

NIAID

Common Ancestry



"Naturalists try to arrange the species, genera, and families in each class, on what is called the Natural System. But what is meant by this system?" p.413

"… I believe this element of descent is the hidden bond of connexion which naturalists have sought under the term of the Natural System" p. 433

NIH National Institute of Allergy and Infectious Diseases

NIAID

What’s so special about viruses?

- Short generation time
- Rapid evolution
- Genotypes - easy, phenotypes - hard
- Large populations
- Structured populations
- Rigorous temporal sampling of genotypes
- Shorter genomes → more WGS data

PAIRWISE ALIGNMENT

- Sequence Alignment: Assigning homology to sites among a group of known sequences
- BLAST: Alignment of one sequence with many unknown

HOMOLOGY vs. ANALOGY

common ancestry

convergence

University of Nebraska
Department of Entomology

NIH National Institute of Allergy and Infectious Diseases

NIAID

PAIRWISE ALIGNMENT

- Sequence Alignment: Assigning homology to sites among a group of known sequences
  - Alignment of single loci
    - Clustal(W,X,Omega), MUSCLE, TCoffee, MAFFT
  - Alignment of overlapping contigs
    - Sequencher, Lasergene
  - Alignment of short reads
    - BWA, Bowtie, SOAP, MAQ

PAIRWISE ALIGNMENT

- Single locus

```
>GeneA_Human
ATGGGCCTTATATGCGTGATGCTGAAAG
>GeneA_Gorilla
ATGGGACTTATCTGCGTGATGCTGACAG
>GeneA_Macaque
ATGGGTCTCATATGTGTGATGCTTACAG
>GeneA_Mouse
ATGGCCCTGATATGCGTGATGCTGAACG
>GeneA_Sheep
ATGGCCCTAATATGC---AGGCTGAACG
```

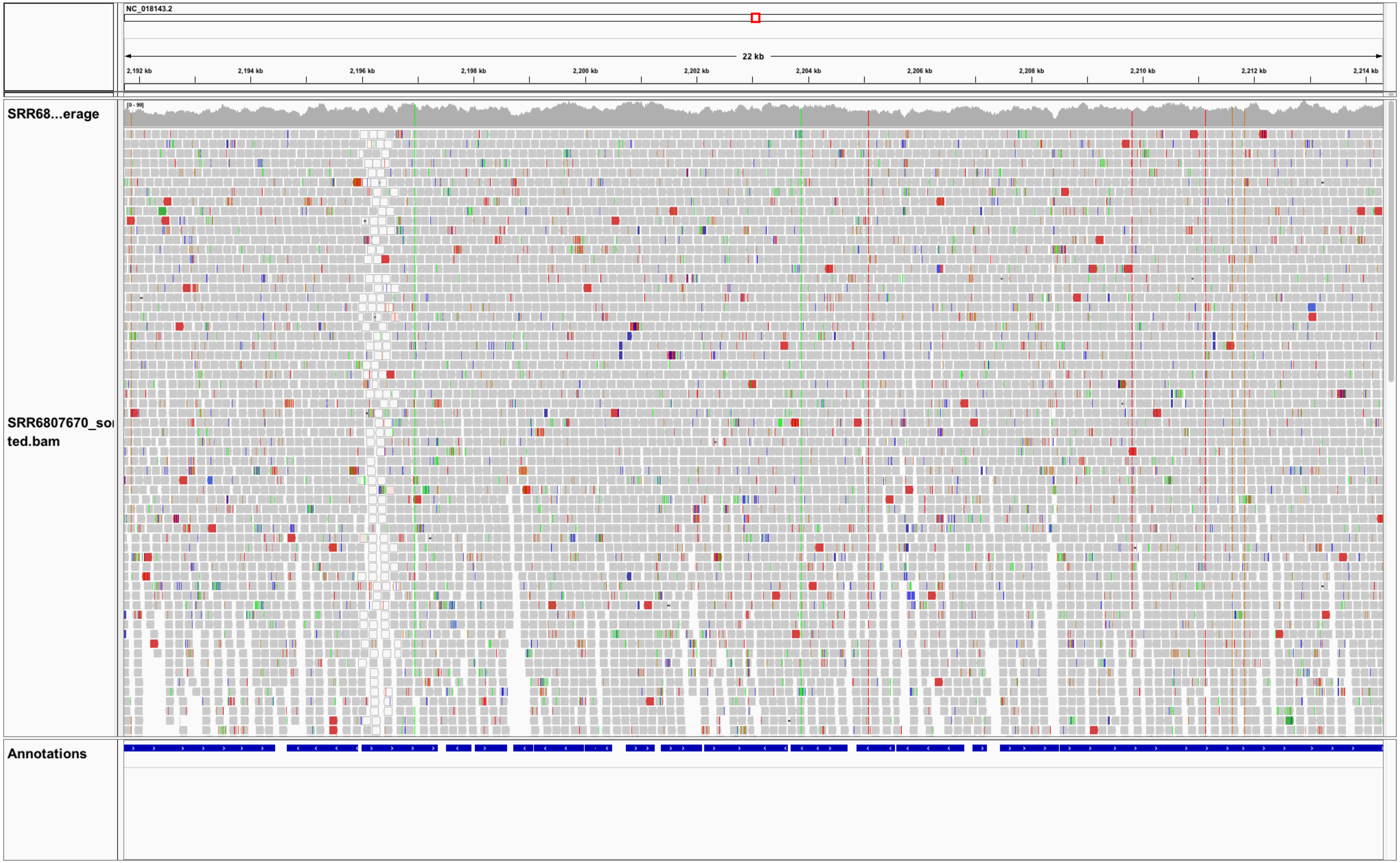
PAIRWISE ALIGNMENT

- Overlapping contigs

```
ATGGGCCTTATATGCGTGATGCTGAAAG
       TTATATGCGTGATGCTGAAAGGGCTTAG
         ATATGCGTGATGCTGAAAGGGCTTAGAAAT
            TGCGTGATGCTGAAAGGGCTTAGAAATT
                  ATGCTGAAAGGGCTTAGAAATTCGG
                       AAAGGGCTTAGAAATTGCGGCTAGGCCTCC
                                        CGGCTAGGCCTCCGAACGC
TACCCGGAATATACGCACTA
               CACTACGACTTTCCCGAATCTTTAAGCC
                       CTTTCCCGAATCTTTAAGCCGATCCGGA
```

PAIRWISE ALIGNMENT

- Short reads



PAIRWISE ALIGNMENT

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
           G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN  GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
           ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
           GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2   GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

National Institute of Allergy and Infectious Diseases

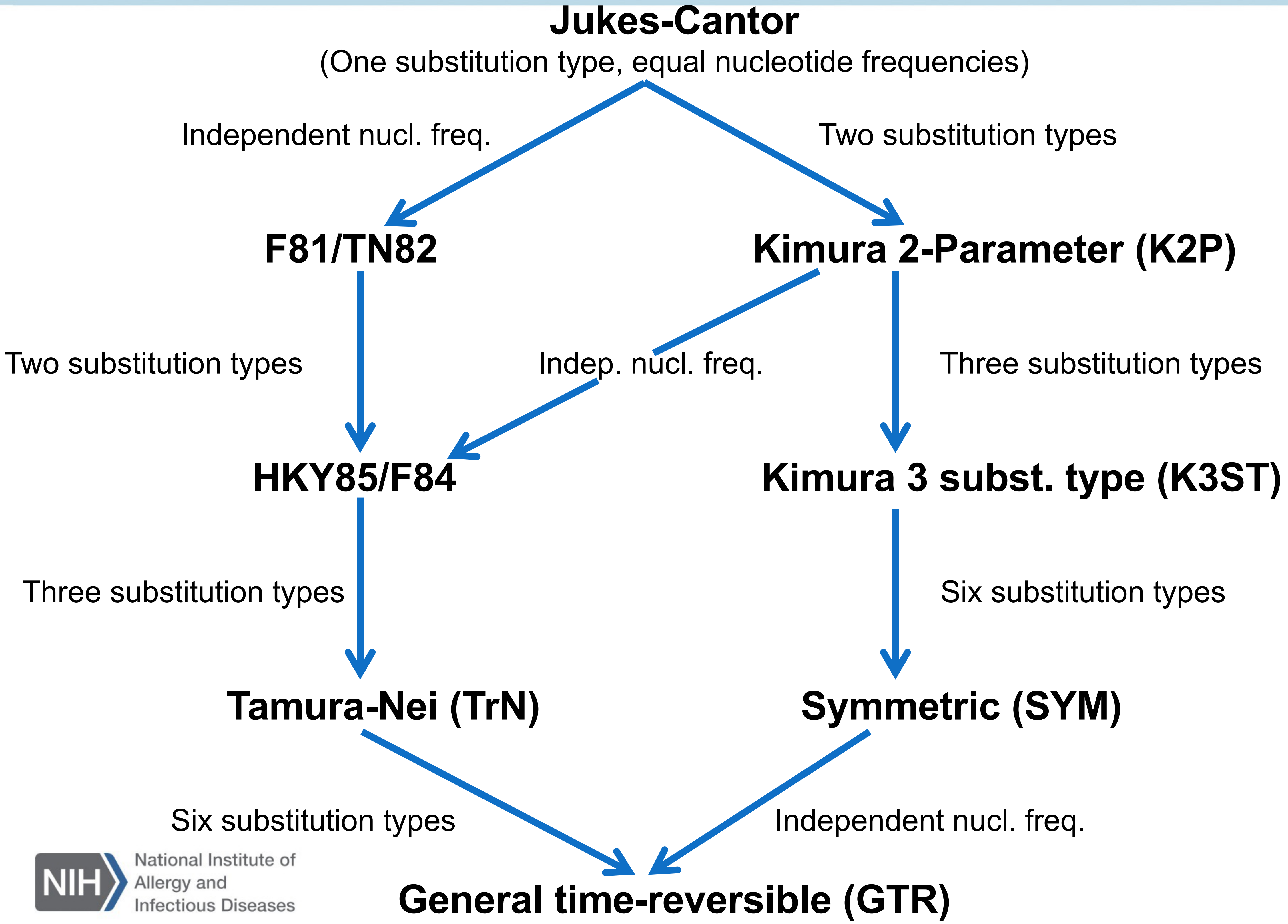
PAIRWISE ALIGNMENT

Jukes-Cantor Substitution Probabilities

μt = 0.25

	A	C	G	T
A	0.5259	0.1580	0.1580	0.1580
C	0.1580	0.5259	0.1580	0.1580
G	0.1580	0.1580	0.5259	0.1580
T	0.1580	0.1580	0.1580	0.5259

Substitution Models



Jukes-Cantor
(One substitution type, equal nucleotide frequencies)

Independent nucl. freq.

Two substitution types

F81/TN82

Kimura 2-Parameter (K2P)

Two substitution types

Indep. nucl. freq.

Three substitution types

HKY85/F84

Kimura 3 subst. type (K3ST)

Three substitution types

Six substitution types

Tamura-Nei (TrN)

Symmetric (SYM)

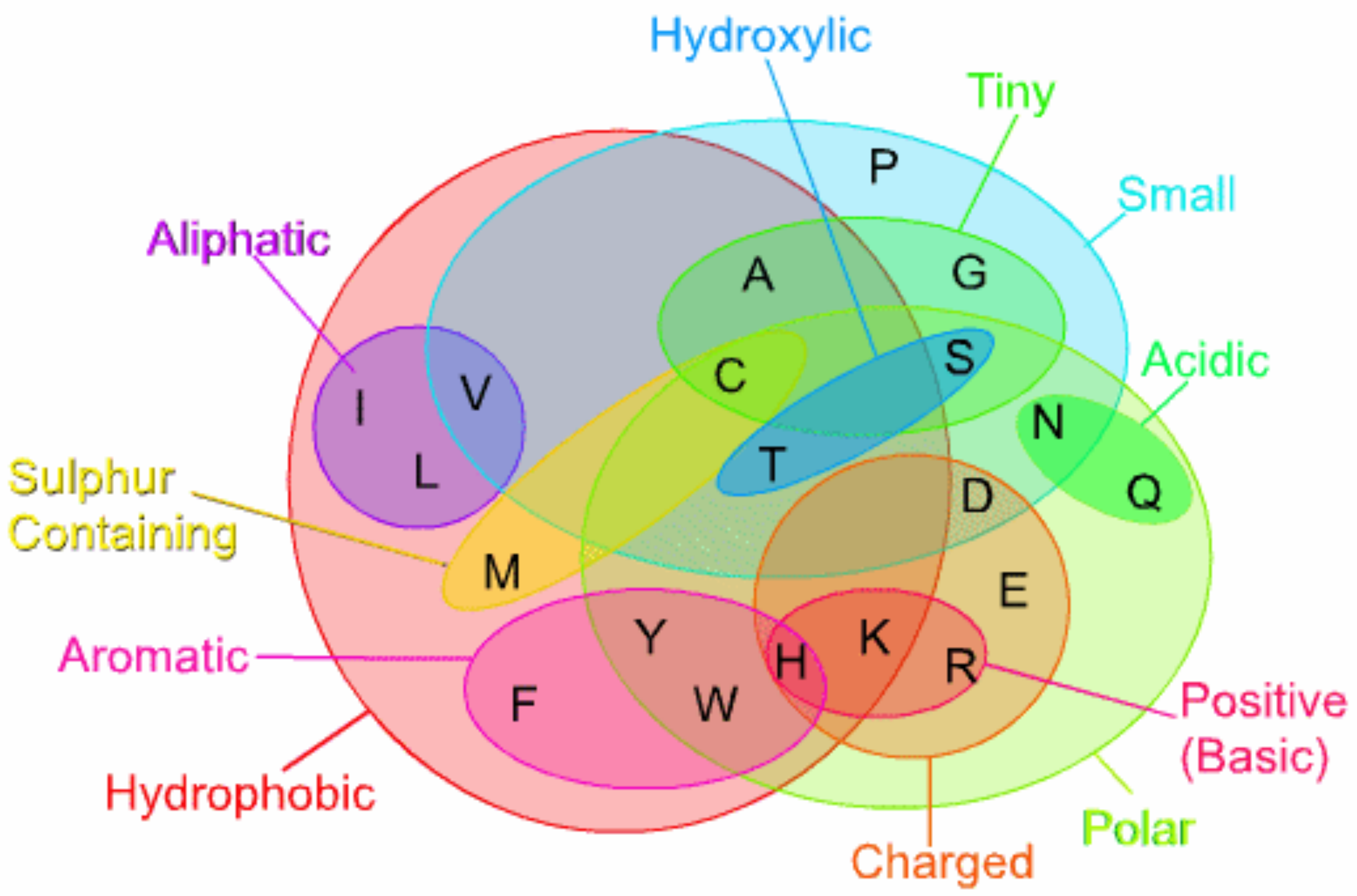
Six substitution types

Independent nucl. freq.

General time-reversible (GTR)

PAIRWISE ALIGNMENT

Protein Score Matrices
Similarity of Amino Acids



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

NIH National Institute of Allergy and Infectious Diseases

From Esquivel RO, et al.. 2013. Advances in Quantum Mechanics, Chapter 27 InTech.

PAIRWISE ALIGNMENT

Protein Score Matrices

- Derived from empirical data
- Account for depth of relationship among the data
- Expressed as log-odds ratio:
  - Logarithm of the ratio of the probabilities of two residues being aligned due to homology versus random chance

PAIRWISE ALIGNMENT

Protein Substitution Matrices

- PAM250: Based on phylogenies where all sequences differ by no more than 15%.
- BLOSUM62: Based on clusters of sequences with greater than 62% identical residues.

Protein Substitution Matrices

BLOSUM62

C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−2	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11
	C	*S*	*T*	*P*	*A*	*G*	*N*	*D*	*E*	*Q*	*H*	*R*	*K*	*M*	*I*	*L*	*V*	*F*	*Y*	*W*

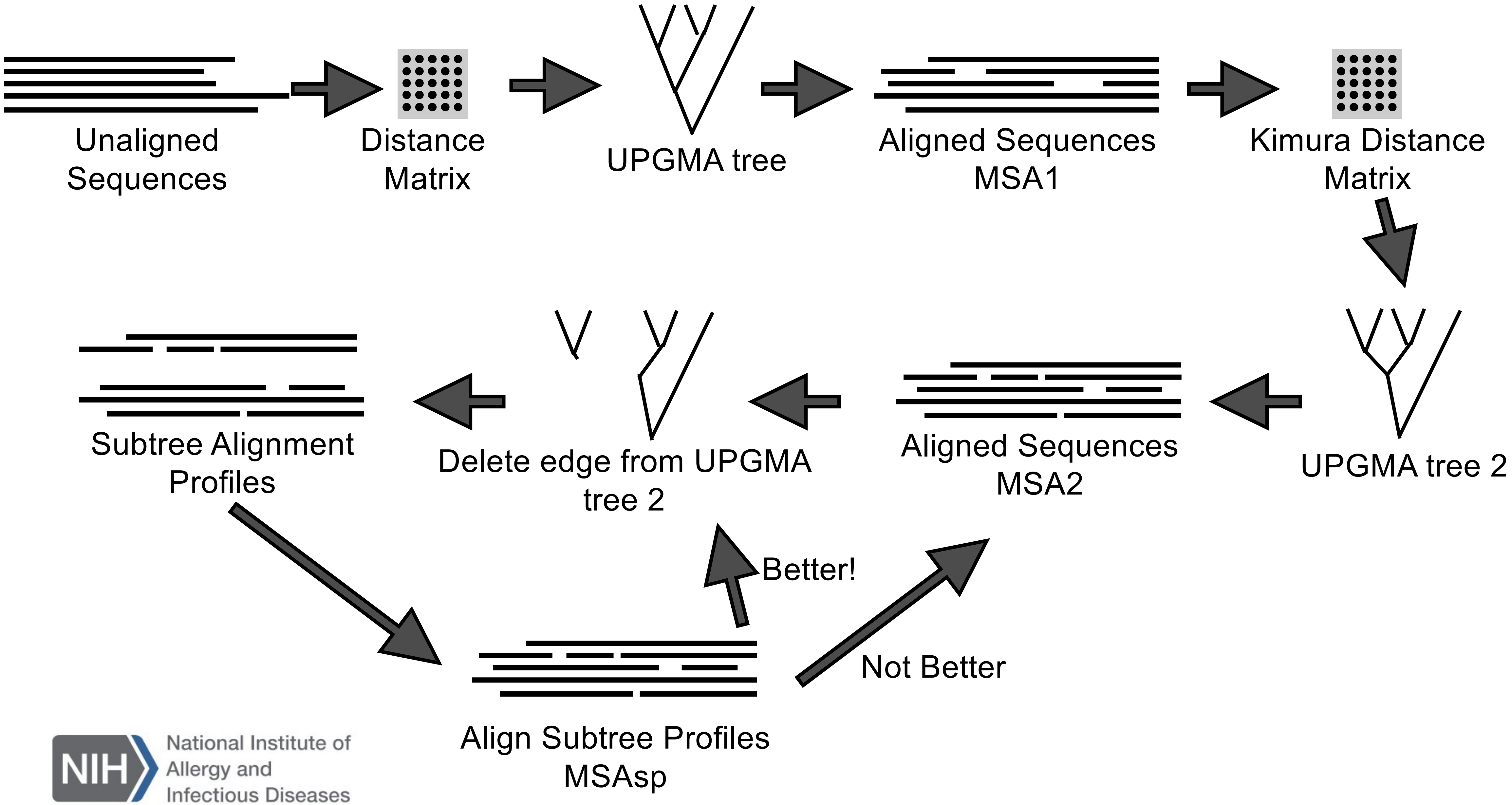
National Institute of
Allergy and
Infectious Diseases

Multiple Sequence Alignment

- Global alignment (Needleman-Wunsch)
  - Assign homology across the entire sequence
  - Clustal
- Local alignment (Smith-Waterman)
  - Assign homology for subsequences
  - MUSCLE and BLAST
  - Good for aligning very divergent sequences
- **Inspect and edit your alignment!**

Multiple Sequence Alignment

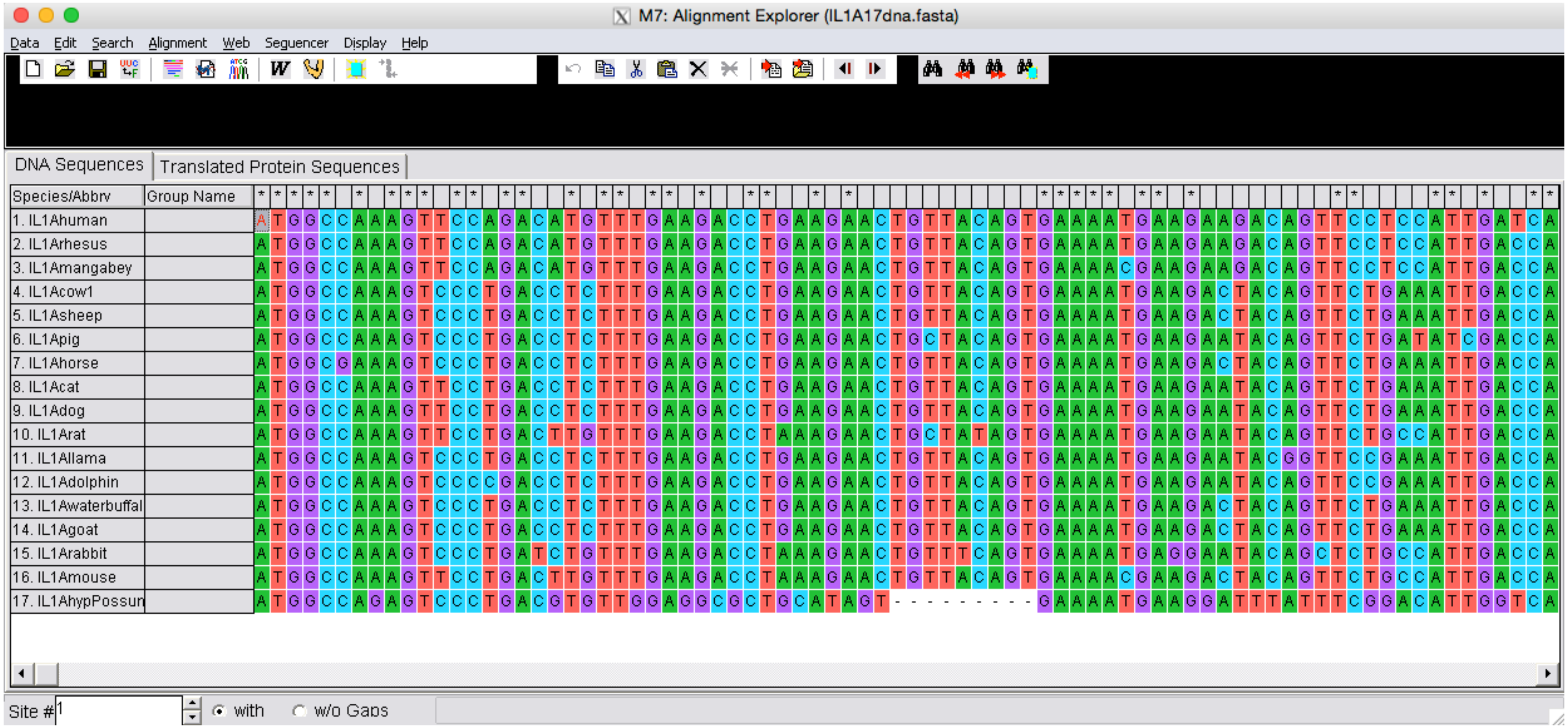
The Progressive Alignment Algorithm



From RC Edgar 2004, NAR 32: 1792-1797

Multiple Sequence Alignment

MSA with MEGA7

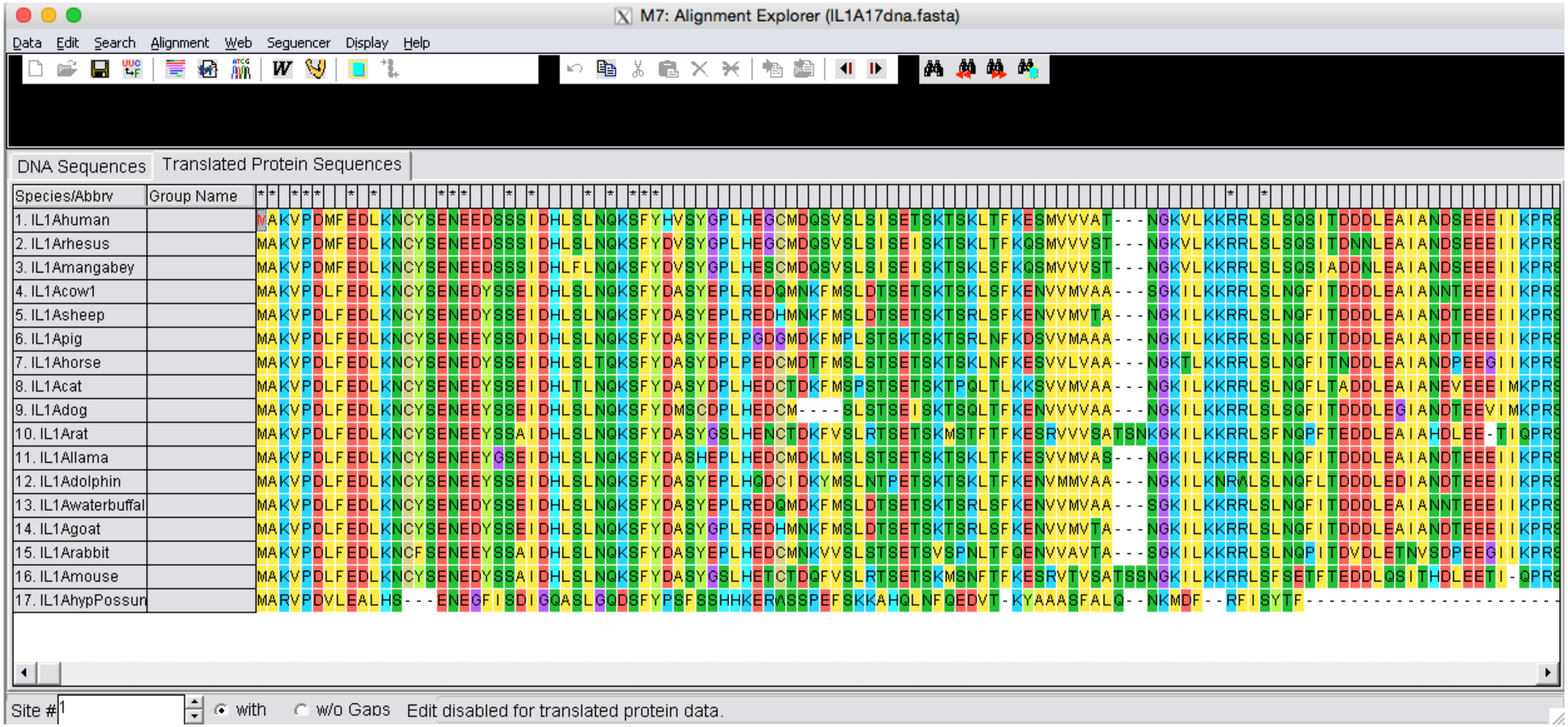

M7: Alignment Explorer (IL1A17dna.fasta)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
W
DNA Sequences
Translated Protein Sequences

```
Species/Abbrv         Group Name  ***** * *** ** **  * ** ** *  **  * *           ***** ** *        **    ** *  **
1. IL1Ahuman                      ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGAAGACAGTTCCTCCATTGATCA
2. IL1Arhesus                     ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGAAGACAGTTCCTCCATTGACCA
3. IL1Amangabey                   ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTGAAAACGAAGAAGACAGTTCCTCCATTGACCA
4. IL1Acow1                       ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGACTACAGTTCTGAAATTGACCA
5. IL1Asheep                      ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGACTACAGTTCTGAAATTGACCA
6. IL1Apig                        ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGCTACAGTGAAAATGAAGAATACAGTTCTGATATCGACCA
7. IL1Ahorse                      ATGGCGAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGACTACAGTTCTGAAATTGACCA
8. IL1Acat                        ATGGCCAAAGTTCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGAATACAGTTCTGAAATTGACCA
9. IL1Adog                        ATGGCCAAAGTTCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGAATACAGTTCTGAAATTGACCA
10. IL1Arat                       ATGGCCAAAGTTCCTGACTTGTTTGAAGACCTAAAGAACTGCTATAGTGAAAATGAAGAATACAGTTCTGCCATTGACCA
11. IL1Allama                     ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGAATACGGTTCCGAAATTGACCA
12. IL1Adolphin                   ATGGCCAAAGTCCCCGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGAATACAGTTCCGAAATTGACCA
13. IL1Awaterbuffal               ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGACTACAGTTCTGAAATTGACCA
14. IL1Agoat                      ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTGAAAATGAAGACTACAGTTCTGAAATTGACCA
15. IL1Arabbit                    ATGGCCAAAGTCCCTGATCTGTTTGAAGACCTAAAGAACTGTTTCAGTGAAAATGAGGAATACAGCTCTGCCATTGACCA
16. IL1Amouse                     ATGGCCAAAGTTCCTGACTTGTTTGAAGACCTAAAGAACTGTTACAGTGAAAACGAAGACTACAGTTCTGCCATTGACCA
17. IL1AhypPossun                 ATGGCCAGAGTCCCTGACGTGTTGGAGGCGCTGCATAGT---------GAAAATGAAGGATTTATTTCGGACATTGGTCA
```

Site # 1 ● with ○ w/o Gaps

Multiple Sequence Alignment

MSA with MEGA7


M7: Alignment Explorer (IL1A17dna.fasta)
Data
Edit
Search
Alignment
Web
Sequencer
Display
Help
DNA Sequences
Translated Protein Sequences

```
Species/Abbrv        Group Name  ** ***  * *     ***   * *    * * ***                                                  *  *
1. IL1Ahuman                     MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETSKTSKLTFKESMVVVAT---NGKVLKKRRLSLSQSITDDDLEAIANDSEEEIIKPRS
2. IL1Arhesus                    MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYDVSYGPLHEGCMDQSVSLSISEISKTSKLTFKQSMVVVST---NGKVLKKRRLSLSQSITDNNLEAIANDSEEEIIKPRS
3. IL1Amangabey                  MAKVPDMFEDLKNCYSENEEDSSSIDHLFLNQKSFYDVSYGPLHESCMDQSVSLSISEISKTSKLSFKQSMVVVST---NGKVLKKRRLSLSQSIADDNLEAIANDSEEEIIKPRS
4. IL1Acow1                      MAKVPDLFEDLKNCYSENEDYSSEIDHLSLNQKSFYDASYEPLREDQMNKFMSLDTSETSKTSKLSFKENVVMVAA---SGKILKKRRLSLNQFITDDDLEAIANNTEEEIIKPRS
5. IL1Asheep                     MAKVPDLFEDLKNCYSENEDYSSEIDHLSLNQKSFYDASYEPLREDHMNKFMSLDTSETSKTSRLSFKENVVMVTA---NGKILKKRRLSLNQFITDDDLEAIANDTEEEIIKPRS
6. IL1Apig                       MAKVPDLFEDLKNCYSENEEYSSDIDHLSLNQKSFYDASYEPLPGDGMDKFMPLSTSKTSKTSRLNFKDSVVMAAA---NGKILKKRRLSLNQFITDDDLEAIANDTEEEIIKPRS
7. IL1Ahorse                     MAKVPDLFEDLKNCYSENEDYSSEIDHLSLTQKSFYDASYDPLPEDCMDTFMSLSTSETSKTSKLNFKESVVLVAA---NGKTLKKRRLSLNQFITNDDLEAIANDPEEGIIKPRS
8. IL1Acat                       MAKVPDLFEDLKNCYSENEEYSSEIDHLTLNQKSFYDASYDPLHEDCTDKFMSPSTSETSKTPQLTLKKSVVMVAA---NGKILKKRRLSLNQFLTADDLEAIANEVEEEIMKPRS
9. IL1Adog                       MAKVPDLFEDLKNCYSENEEYSSEIDHLSLNQKSFYDMSCDPLHEDCM----SLSTSEISKTSQLTFKENVVVVAA---NGKILKKRRLSLSQFITDDDLEGIANDTEEVIMKPRS
10. IL1Arat                      MAKVPDLFEDLKNCYSENEEYSSAIDHLSLNQKSFYDASYGSLHENCTDKFVSLRTSETSKMSTFTFKESRVVVSATSNKGKILKKRRLSFNQPFTEDDLEAIAHDLEE-TIQPRS
11. IL1Allama                    MAKVPDLFEDLKNCYSENEEYGSEIDHLSLNQKSFYDASHEPLHEDCMDKLMSLSTSETSKTSKLTFKESVVMVAS---NGKILKKRRLSLNQFITDDDLEAIANDTEEEIIKPRS
12. IL1Adolphin                  MAKVPDLFEDLKNCYSENEEYSSEIDHLSLNQKSFYDASYEPLHQDCIDKYMSLNTPETSKTSKLTFKENVMMVAA---NGKILKNRWLSLNQFLTDDDLEDIANDTEEEIIKPRS
13. IL1Awaterbuffal              MAKVPDLFEDLKNCYSENEDYSSEIDHLSLNQKSFYDASYEPLREDQMDKFMSLDTSETSKTSRLSFKENVVMVAA---SGKILKKRRLSLNQFITDDDLEAIANNTEEEIIKPRS
14. IL1Agoat                     MAKVPDLFEDLKNCYSENEDYSSEIDHLSLNQKSFYDASYGPLREDHMNKFMSLDTSETSKTSRLSFKENVVMVTA---NGKILKKRRLSLNQFITDDDLEAIANDTEEEIIKPRS
15. IL1Arabbit                   MAKVPDLFEDLKNCFSENEEYSSAIDHLSLNQKSFYDASYEPLHEDCMNKVVSLSTSETSVSPNLTFQENVVAVTA---SGKILKKRRLSLNQPITDVDLETNVSDPEEGIIKPRS
16. IL1Amouse                    MAKVPDLFEDLKNCYSENEDYSSAIDHLSLNQKSFYDASYGSLHETCTDQFVSLRTSETSKMSNFTFKESRVTVSATSSNGKILKKRRLSFSETFTEDDLQSITHDLEETI-QPRS
17. IL1AhypPossun                MARVPDVLEALHS---ENEGFISDIGQASLGQDSFYPSFSSHHKERWSSPEFSKKAHQLNFQEDVT-KYAAASFALQ--NKMDF--RFISYTF-----------------------
```

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

NIH National Institute of Allergy and Infectious Diseases

Multiple Sequence Alignment

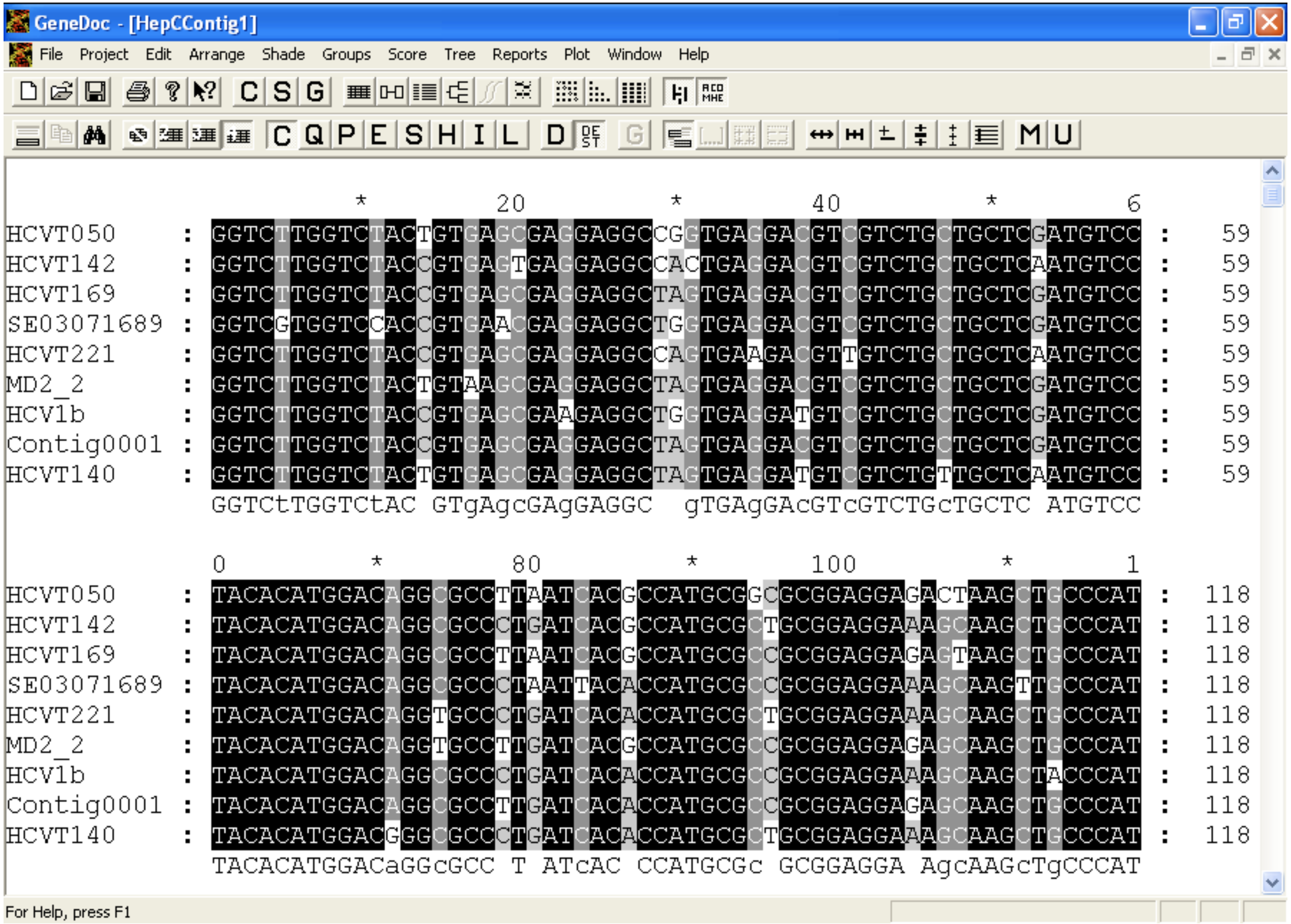
NEVER

directly input the output of a MSA program into an analysis program!

ALWAYS

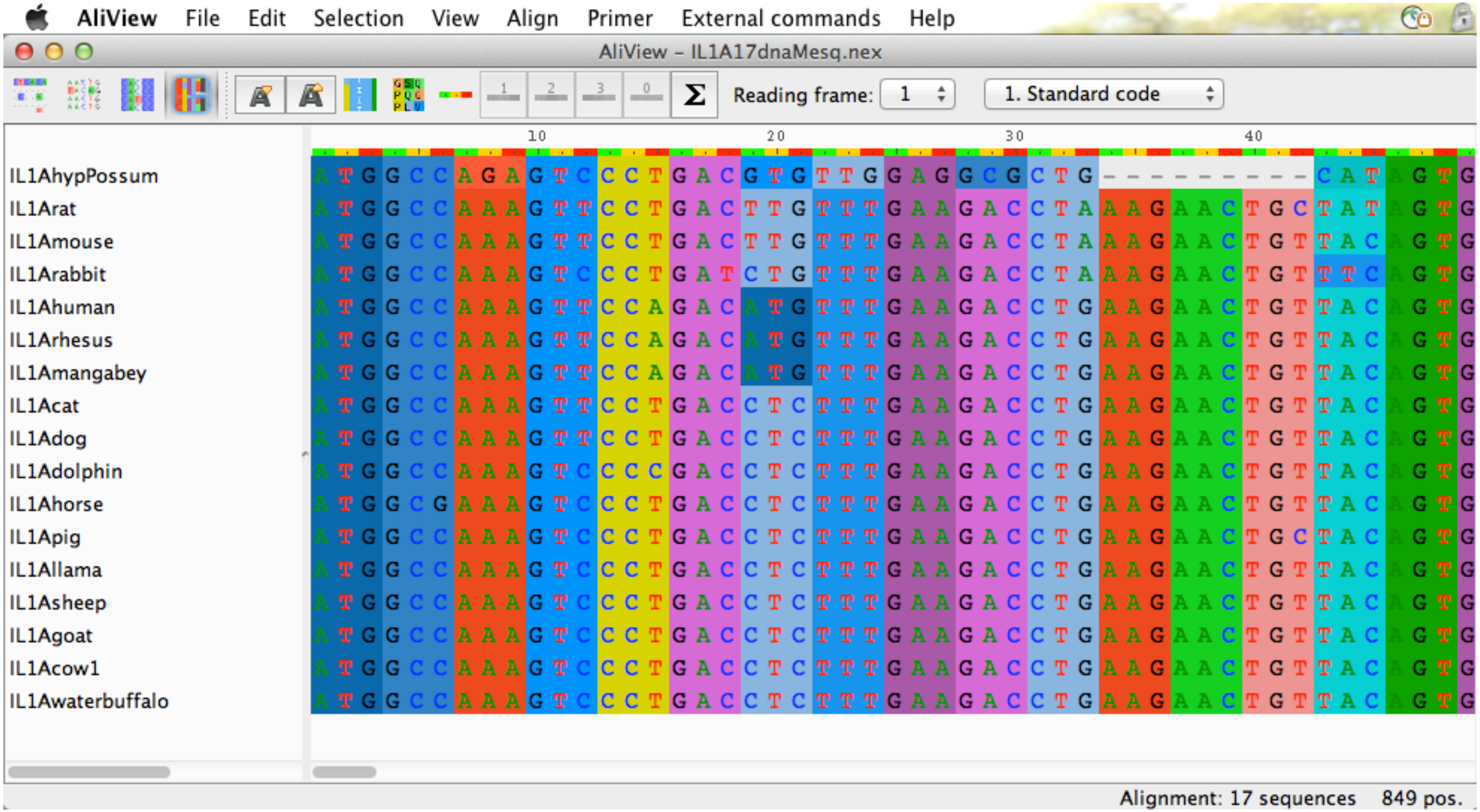
inspect the alignment to improve it.

Multiple Sequence Alignment



```
                      *        20         *        40         *        6
HCVT050    : GGTCTTGGTCTACTGTGAGCGAGGAGGCCGGTGAGGACGTCGTCTGCTGCTCGATGTCC :    59
HCVT142    : GGTCTTGGTCTACCGTGAGTGAGGAGGCCACTGAGGACGTCGTCTGCTGCTCAATGTCC :    59
HCVT169    : GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGCTCGATGTCC :    59
SE03071689 : GGTCGTGGTCCACCGTGAACGAGGAGGCTGGTGAGGACGTCGTCTGCTGCTCGATGTCC :    59
HCVT221    : GGTCTTGGTCTACCGTGAGCGAGGAGGCCAGTGAAGACGTTGTCTGCTGCTCAATGTCC :    59
MD2_2      : GGTCTTGGTCTACTGTAAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGCTCGATGTCC :    59
HCV1b      : GGTCTTGGTCTACCGTGAGCGAAGAGGCTGGTGAGGATGTCGTCTGCTGCTCGATGTCC :    59
Contig0001 : GGTCTTGGTCTACCGTGAGCGAGGAGGCTAGTGAGGACGTCGTCTGCTGCTCGATGTCC :    59
HCVT140    : GGTCTTGGTCTACTGTGAGCGAGGAGGCTAGTGAGGATGTCGTCTGTTGCTCAATGTCC :    59
             GGTCtTGGTCtAC GTgAgcGAgGAGGC  gTGAgGAcGTcGTCTGcTGCTC ATGTCC

             0         *        80         *       100         *       1
HCVT050    : TACACATGGACAGGCGCCTTAATCACGCCATGCGGCGCGGAGGAGACTAAGCTGCCCAT :   118
HCVT142    : TACACATGGACAGGCGCCCTGATCACGCCATGCGCTGCGGAGGAAAGCAAGCTGCCCAT :   118
HCVT169    : TACACATGGACAGGCGCCTTAATCACGCCATGCGCCGCGGAGGAGAGTAAGCTGCCCAT :   118
SE03071689 : TACACATGGACAGGCGCCCTAATTACACCATGCGCCGCGGAGGAAAGCAAGTTGCCCAT :   118
HCVT221    : TACACATGGACAGGTGCCCTGATCACACCATGCGCTGCGGAGGAAAGCAAGCTGCCCAT :   118
MD2_2      : TACACATGGACAGGTGCCTTGATCACGCCATGCGCCGCGGAGGAGAGCAAGCTGCCCAT :   118
HCV1b      : TACACATGGACAGGCGCCCTGATCACACCATGCGCCGCGGAGGAAAGCAAGCTACCCAT :   118
Contig0001 : TACACATGGACAGGCGCCTTGATCACACCATGCGCCGCGGAGGAGAGCAAGCTGCCCAT :   118
HCVT140    : TACACATGGACGGGCGCCCTGATCACACCATGCGCTGCGGAGGAAAGCAAGCTGCCCAT :   118
             TACACATGGACaGGcGCC T ATcAC CCATGCGc GCGGAGGA AgcAAGcTgCCCAT
```

Multiple Sequence Alignment



Reading frame: 1

1. Standard code

```
IL1AhypPossum      ATGGCCAGAGTCCCTGACGTGTTGGAGGCGCTG---------CATAGTG
IL1Arat            ATGGCCAAAGTTCCTGACTTGTTTGAAGACCTAAAGAACTGCTATAGTG
IL1Amouse          ATGGCCAAAGTTCCTGACTTGTTTGAAGACCTAAAGAACTGTTACAGTG
IL1Arabbit         ATGGCCAAAGTCCCTGATCTGTTTGAAGACCTAAAGAACTGTTTCAGTG
IL1Ahuman          ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Arhesus         ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Amangabey       ATGGCCAAAGTTCCAGACATGTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Acat            ATGGCCAAAGTTCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Adog            ATGGCCAAAGTTCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Adolphin        ATGGCCAAAGTCCCCGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Ahorse          ATGGCGAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Apig            ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGCTACAGTG
IL1Allama          ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Asheep          ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Agoat           ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Acow1           ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
IL1Awaterbuffalo   ATGGCCAAAGTCCCTGACCTCTTTGAAGACCTGAAGAACTGTTACAGTG
```

Alignment: 17 sequences   849 pos.

Multiple Sequence Alignment

Programs

- Clustal
  - Your own computer
  - Web Server
  - NIAID HPC cluster
- MUSCLE
  - Your own computer
  - Web Server
  - NIAID HPC cluster
- MAFFT
  - Web Server

Multiple Sequence Alignment

Multiple Sequence Alignment Editors

- Geneious
- MacVector
- MegAlign (Lasergene)
- AliView
- GeneDoc
- BioEdit

Web Resources

ClustalW

http://www.clustal.org/

Muscle

http://www.drive5.com/muscle/download3.6.html

MAFFT

http://mafft.cbrc.jp/alignment/server/

MEGA7

https://www.megasoftware.net/

AliView

http://www.ormbunkar.se/aliview/

GeneDoc

http://genedoc.software.informer.com/

BioEdit

http://www.mbio.ncsu.edu/BioEdit/bioedit.html

What's next?

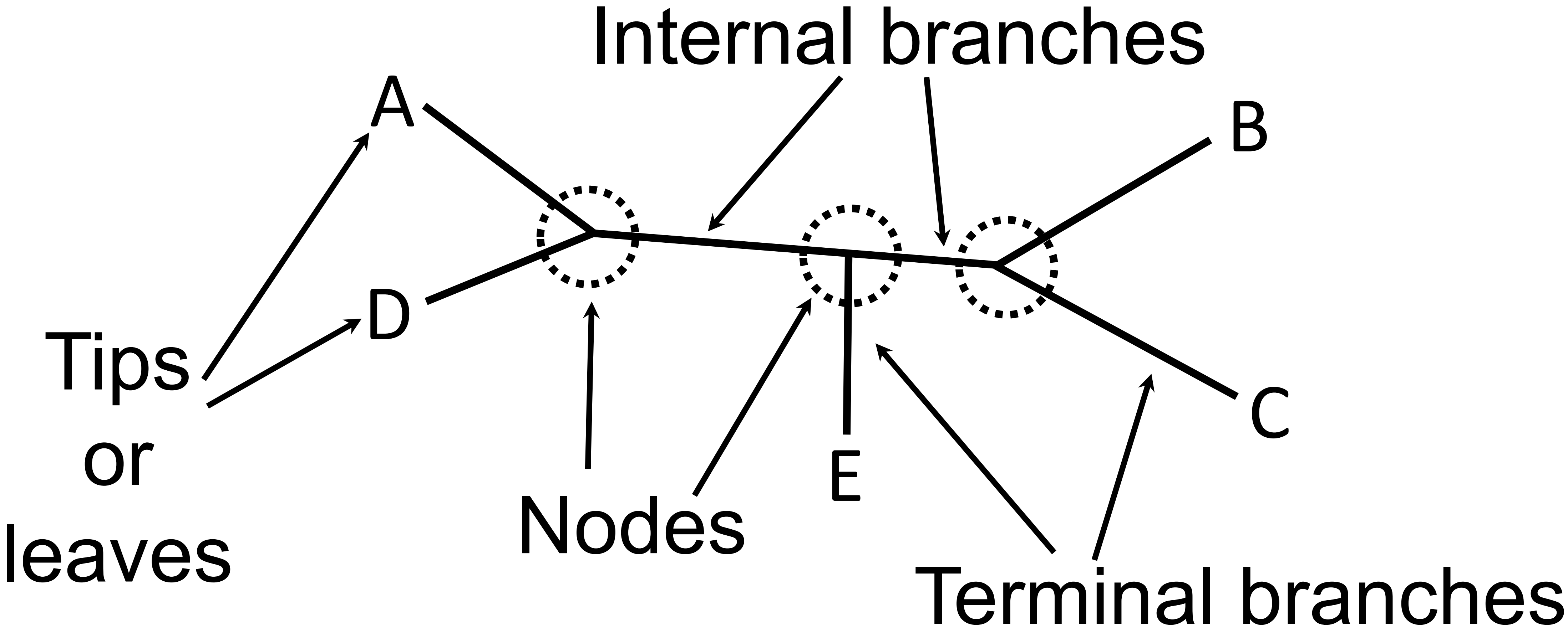
After the break

Building trees with our MSA

What is a phylogenetic tree?

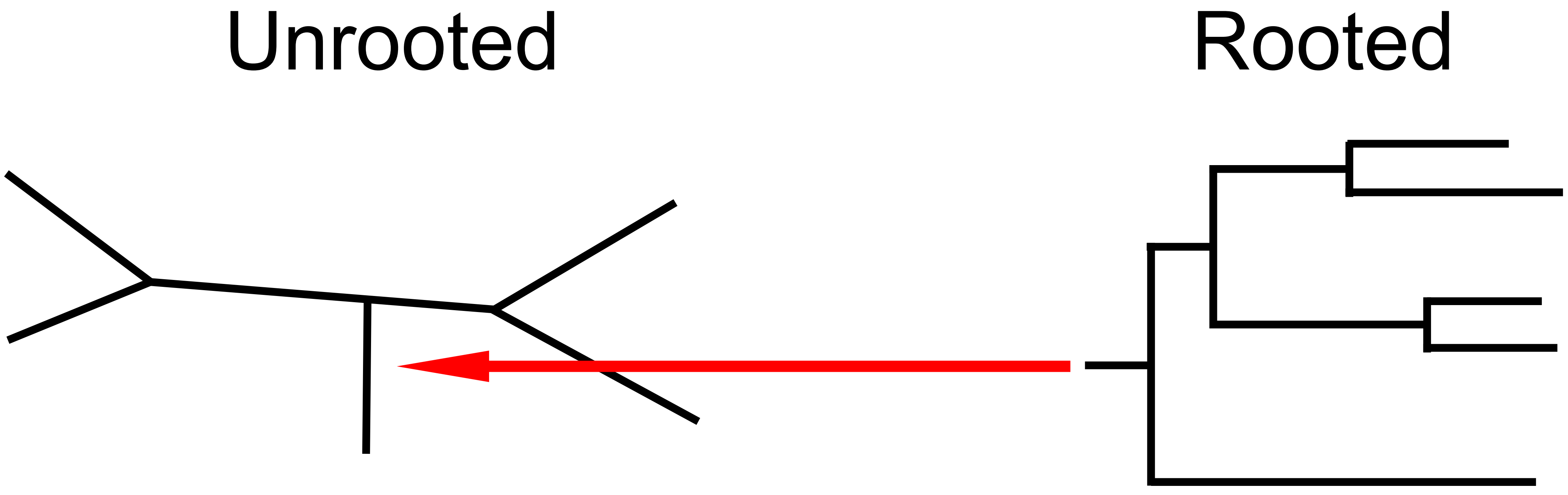
- Reconstruction of biological history
- Based on similarities and differences among homologous attributes (characters) of the entities under scrutiny
- Molecular characters (sequences, usually) are most often found only in extant organisms

What is a phylogenetic tree?



National Institute of
Allergy and
Infectious Diseases

What is a phylogenetic tree?



Unrooted

Rooted

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

SUBSTITUTION MODEL

Jukes-Cantor Substitution Probabilities

$\mu t = 0.25$

	A	C	G	T
A	0.5259	0.1580	0.1580	0.1580
C	0.1580	0.5259	0.1580	0.1580
G	0.1580	0.1580	0.5259	0.1580
T	0.1580	0.1580	0.1580	0.5259

SUBSTITUTION MODEL

BLOSUM62

C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−2	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11
	C	*S*	*T*	*P*	*A*	*G*	*N*	*D*	*E*	*Q*	*H*	*R*	*K*	*M*	*I*	*L*	*V*	*F*	*Y*	*W*

The algorithm approach: Distance Methods

Tree-Building Algorithms

- UPGMA
- Neighbor-Joining

The algorithm approach: Neighbor-joining Calculation

	A	B	C	D	E	R
A	–	0.1715	0.2147	0.3091	0.2326	0.9279
B	-0.4766	–	0.2991	0.3399	0.2058	1.0163
C	-0.4905	-0.4356	–	$\boxed{0.2795}$	0.3943	1.1876
D	-0.4527	-0.4514	$\boxed{-0.5689}$	–	0.4289	1.3574
E	-0.4972	-0.5535	-0.4221	-0.4441	–	1.2616

C to Node 1 distance = 0.2795/2 + (1.1876 – 1.3574)/6 = 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
A	-	0.1715	0.2326	$\boxed{0.1222}$	0.5263
B	-0.3701	-	0.2058	0.1798	0.5571
E	-0.3856	-0.4278	-	0.2719	0.7103
Node 1	$\boxed{-0.4278}$	-0.3856	-0.3701	-	0.5739

A to Node 2 distance = 0.1222/2 + (0.5263 – 0.5739)/4 = 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

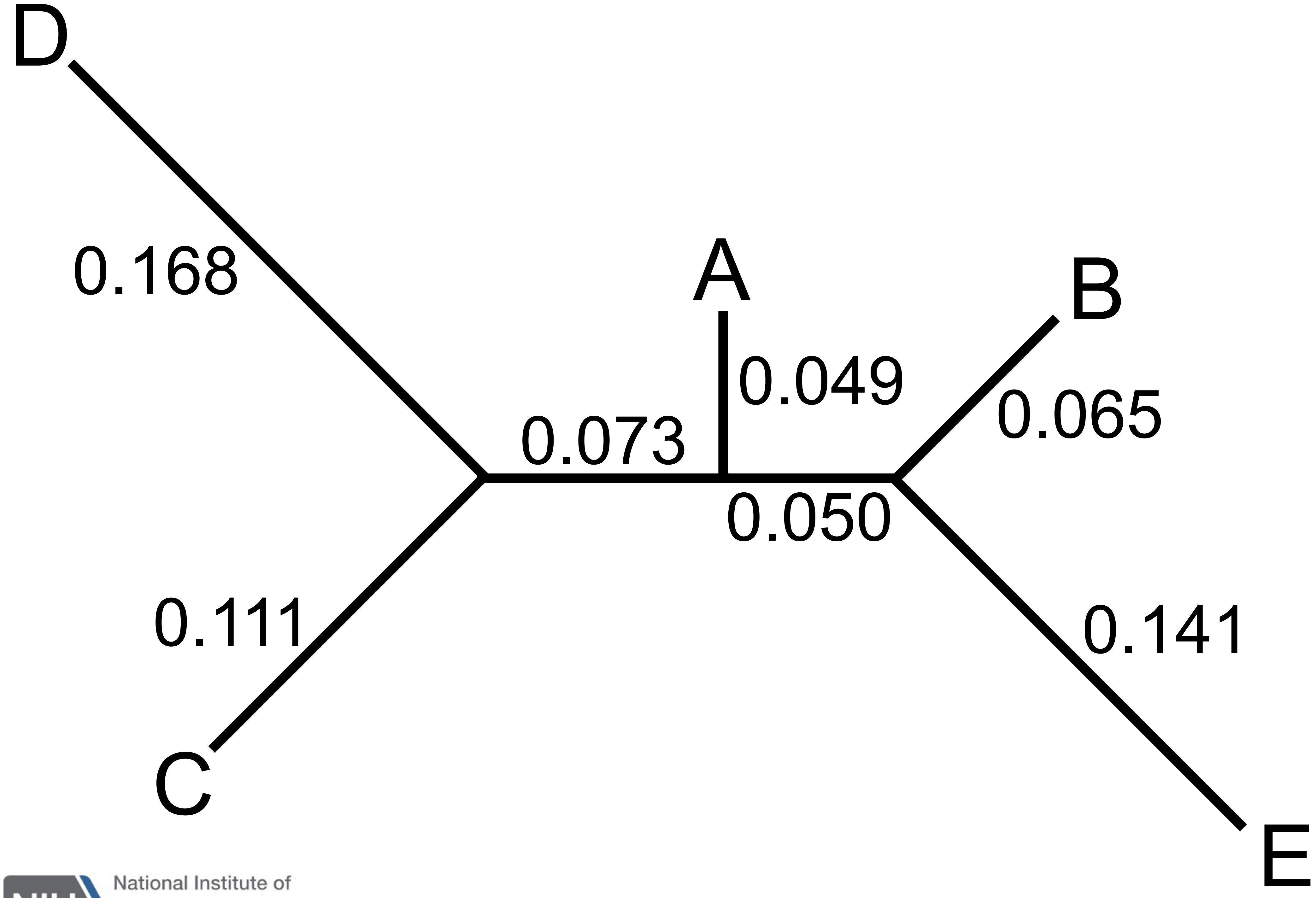
	B	E	Node 2	R
B	-	0.2058	$\boxed{0.1146}$	0.3204
E	-0.5116	-	0.1912	0.3970
Node 2	$\boxed{-0.5116}$	-0.5116	-	0.3058

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

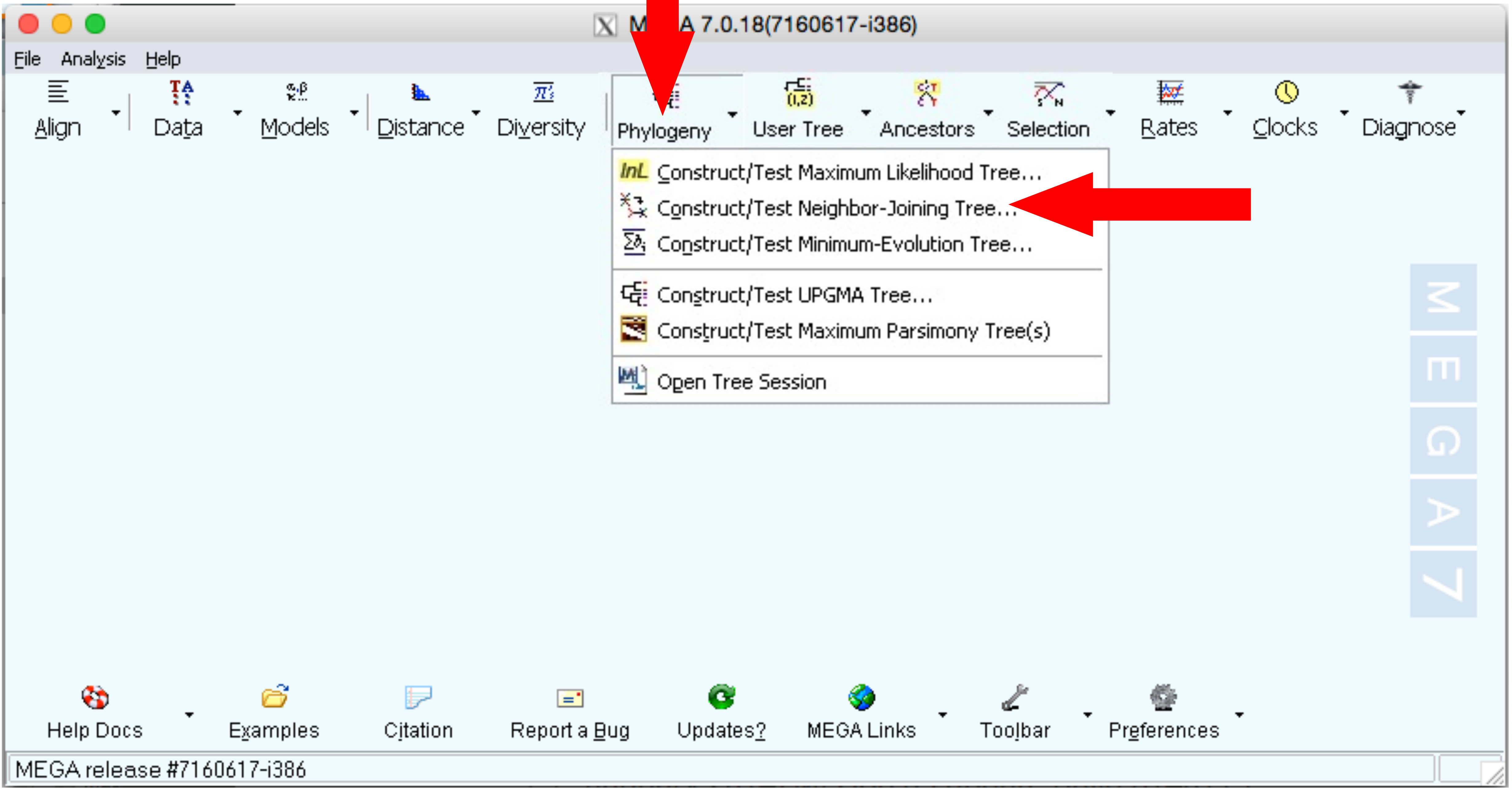
The algorithm approach: Neighbor-joining Calculation



National Institute of Allergy and Infectious Diseases

MEGA7

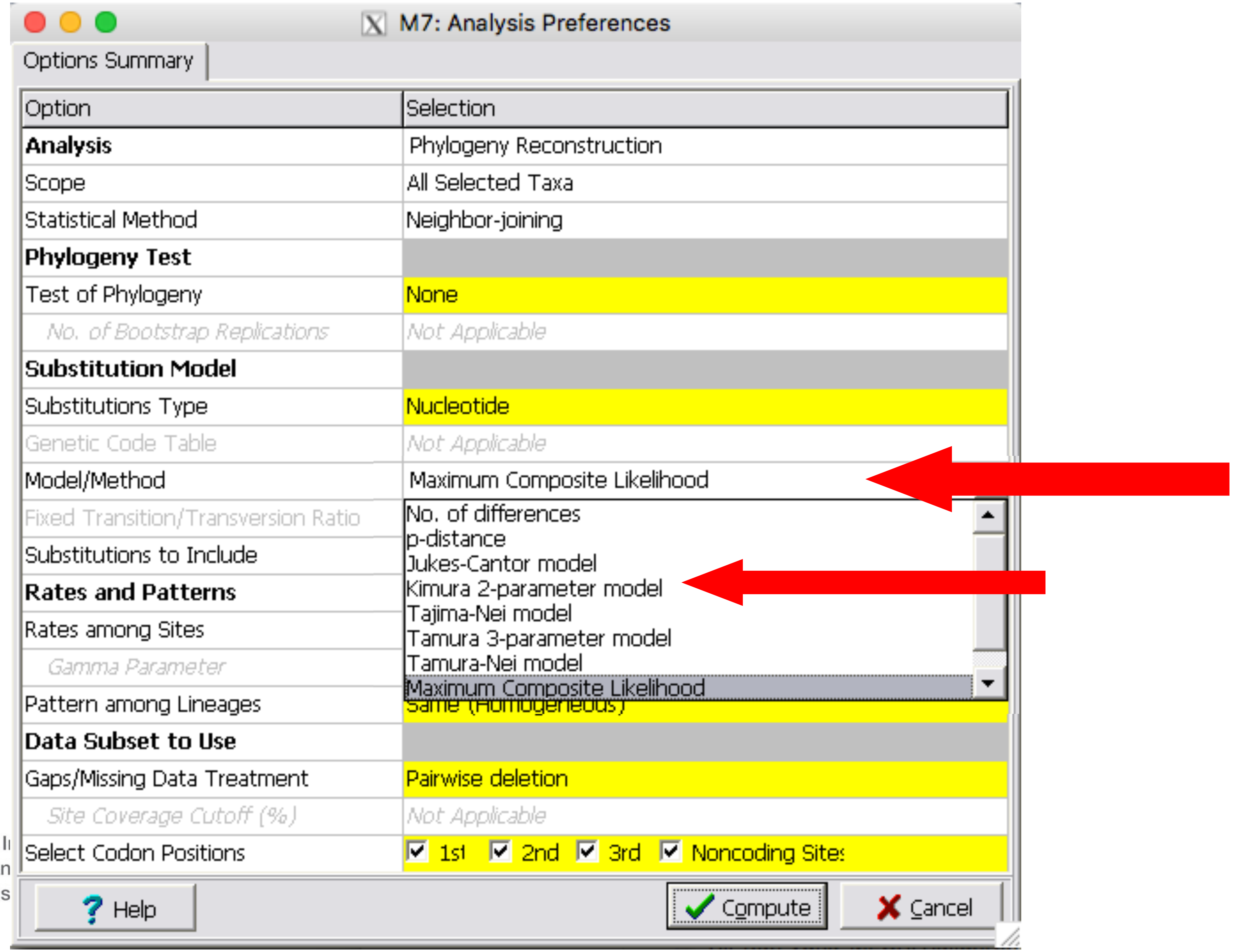
Phylogeny Construction: Neighbor-Joining


MEGA 7.0.18(7160617-i386)
File Analysis Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Construct/Test Maximum Likelihood Tree...
Construct/Test Neighbor-Joining Tree...
Construct/Test Minimum-Evolution Tree...
Construct/Test UPGMA Tree...
Construct/Test Maximum Parsimony Tree(s)
Open Tree Session
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7160617-i386
MEGA7

National Institute of Allergy and Infectious Diseases

MEGA7

Phylogeny Construction: Neighbor-Joining

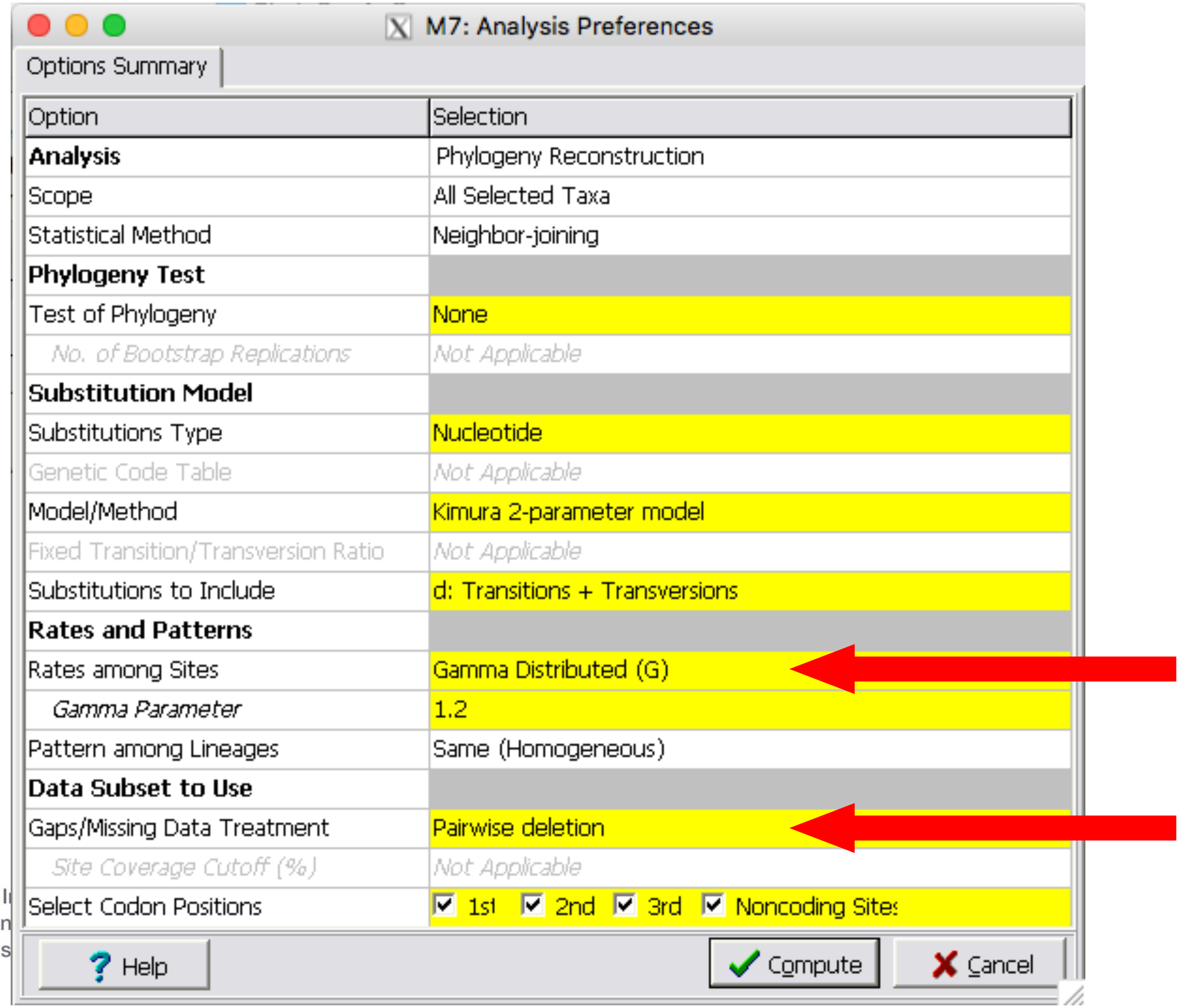


Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Maximum Composite Likelihood
Fixed Transition/Transversion Ratio	
Substitutions to Include	
Rates and Patterns	
Rates among Sites	
Gamma Parameter	
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Pairwise deletion
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites

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

MEGA7

Phylogeny Construction: Neighbor-Joining



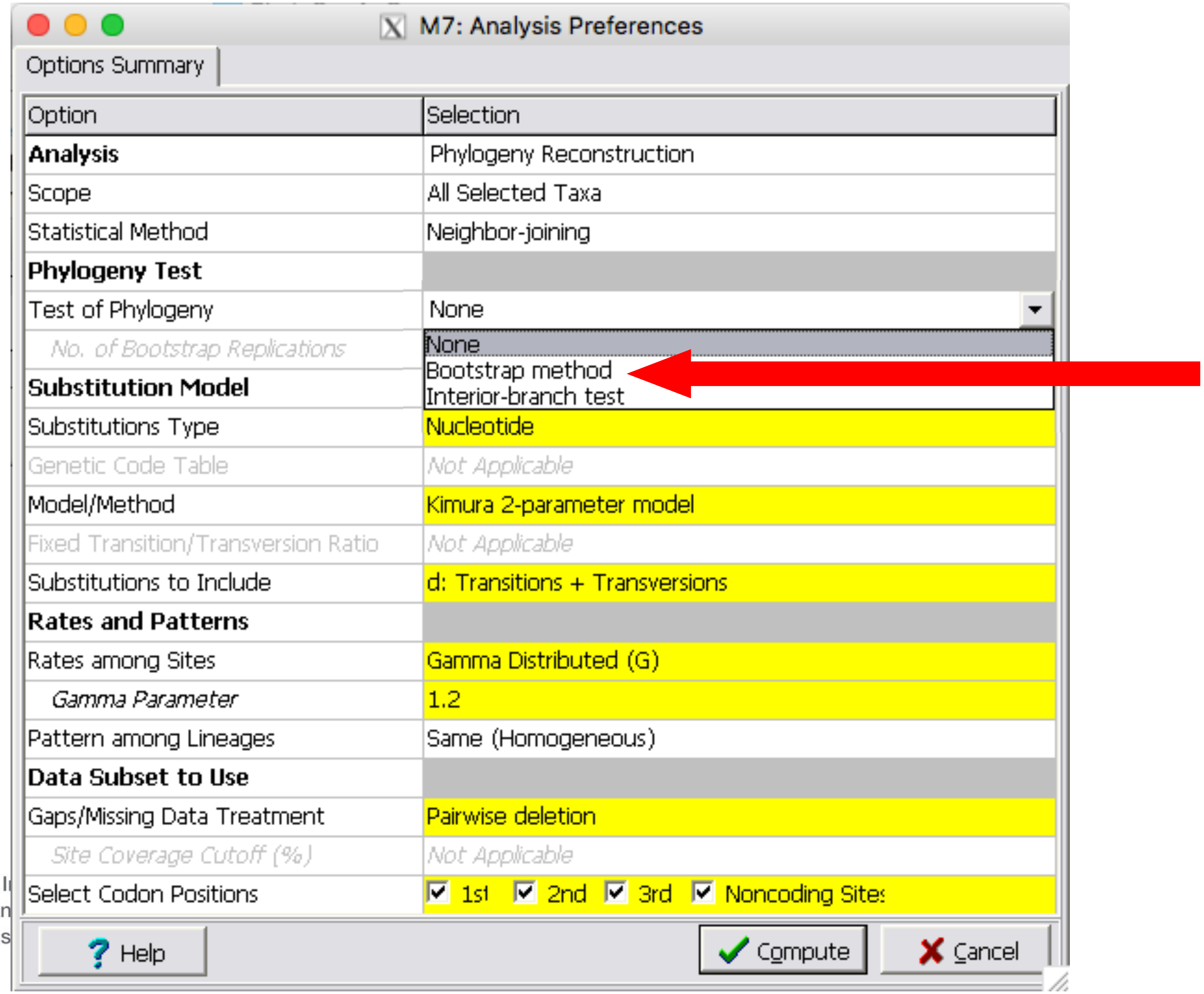
Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Model/Method	Kimura 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	Pairwise deletion
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites

NIH National I
Allergy an
Infectious

NIAID

MEGA7

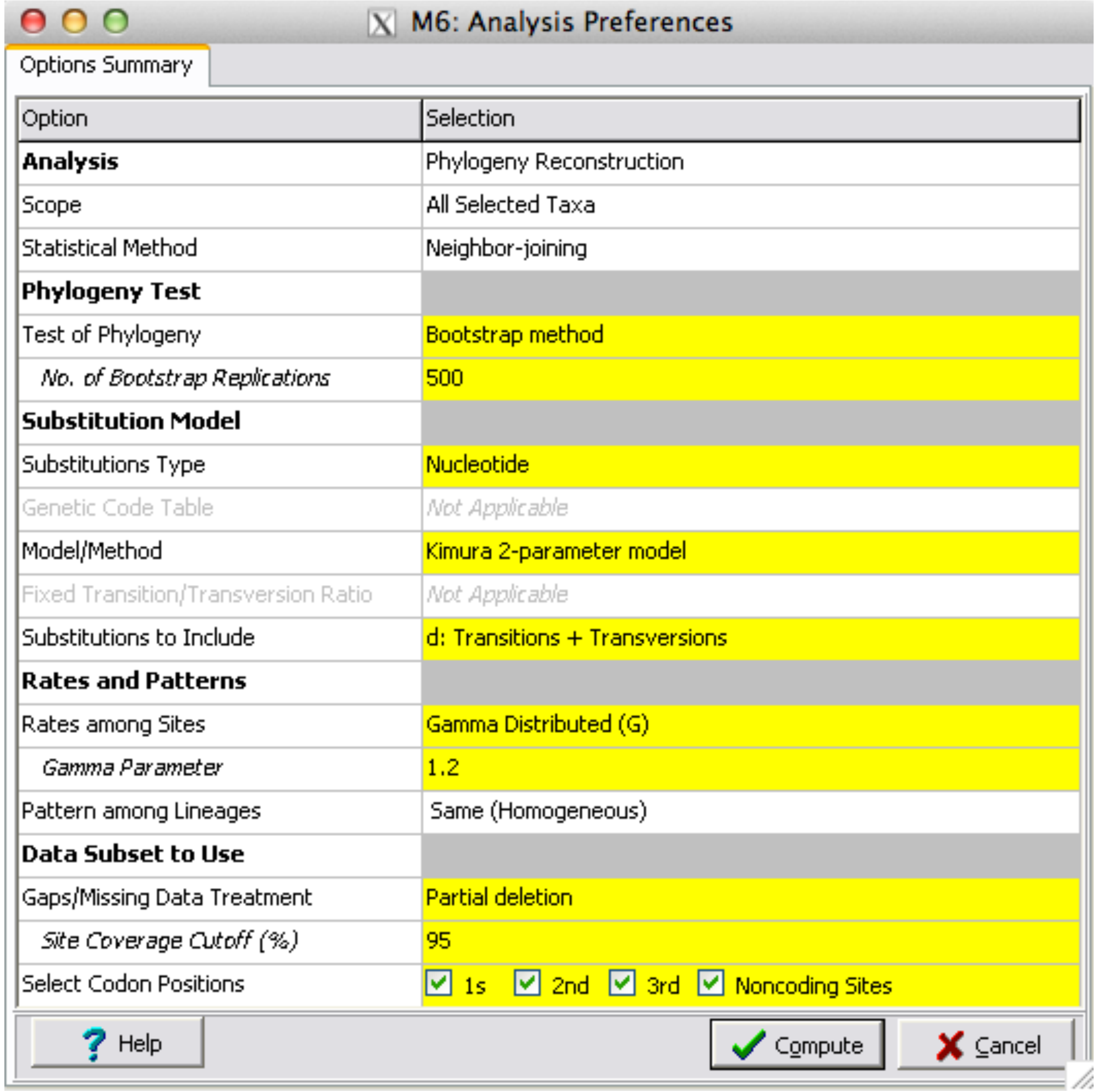
Phylogeny Construction: Neighbor-Joining



Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	None None Bootstrap method Interior-branch test
No. of Bootstrap Replications	
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	Pairwise deletion
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites

MEGA7

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

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)

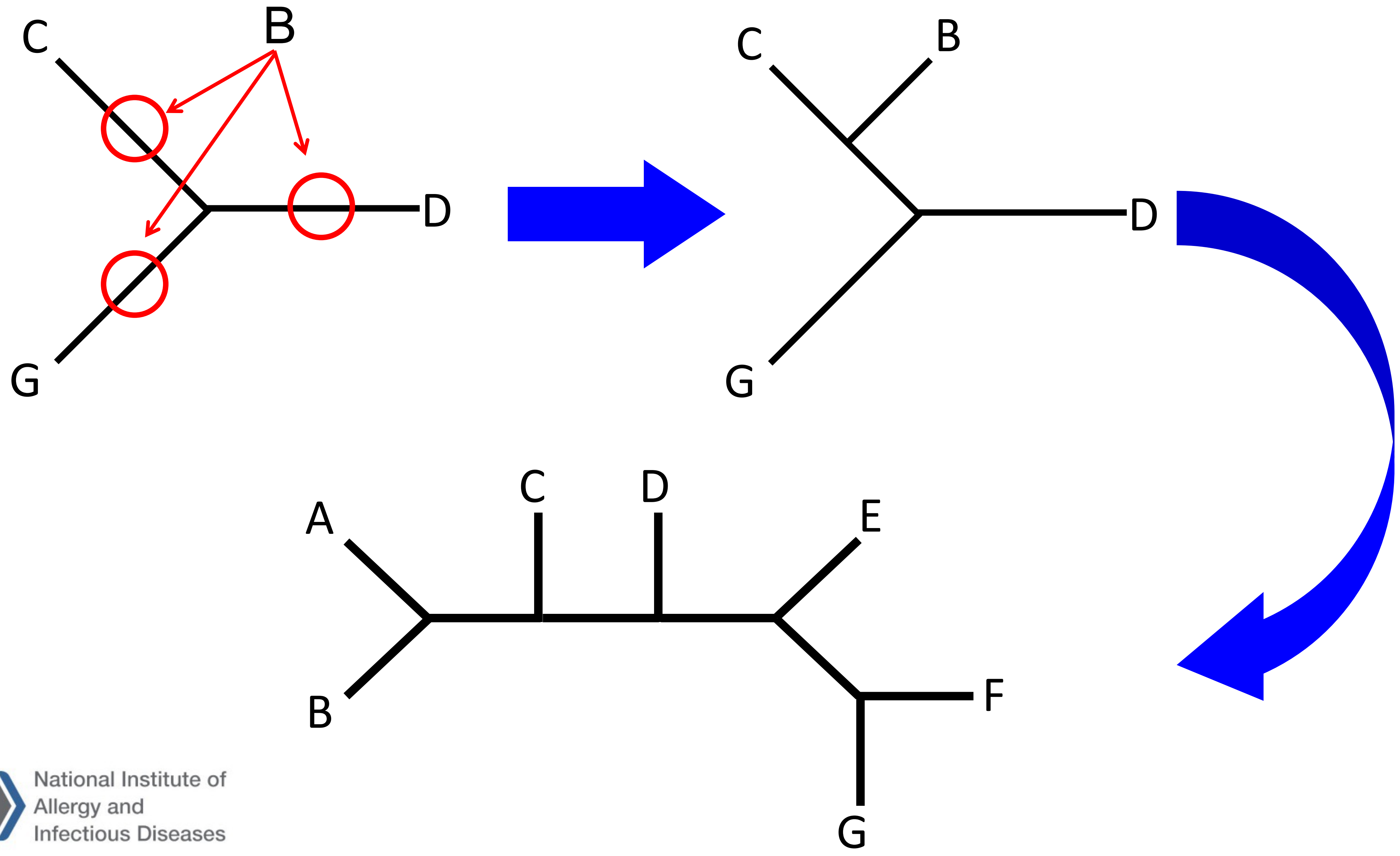
The optimality criterion approach

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



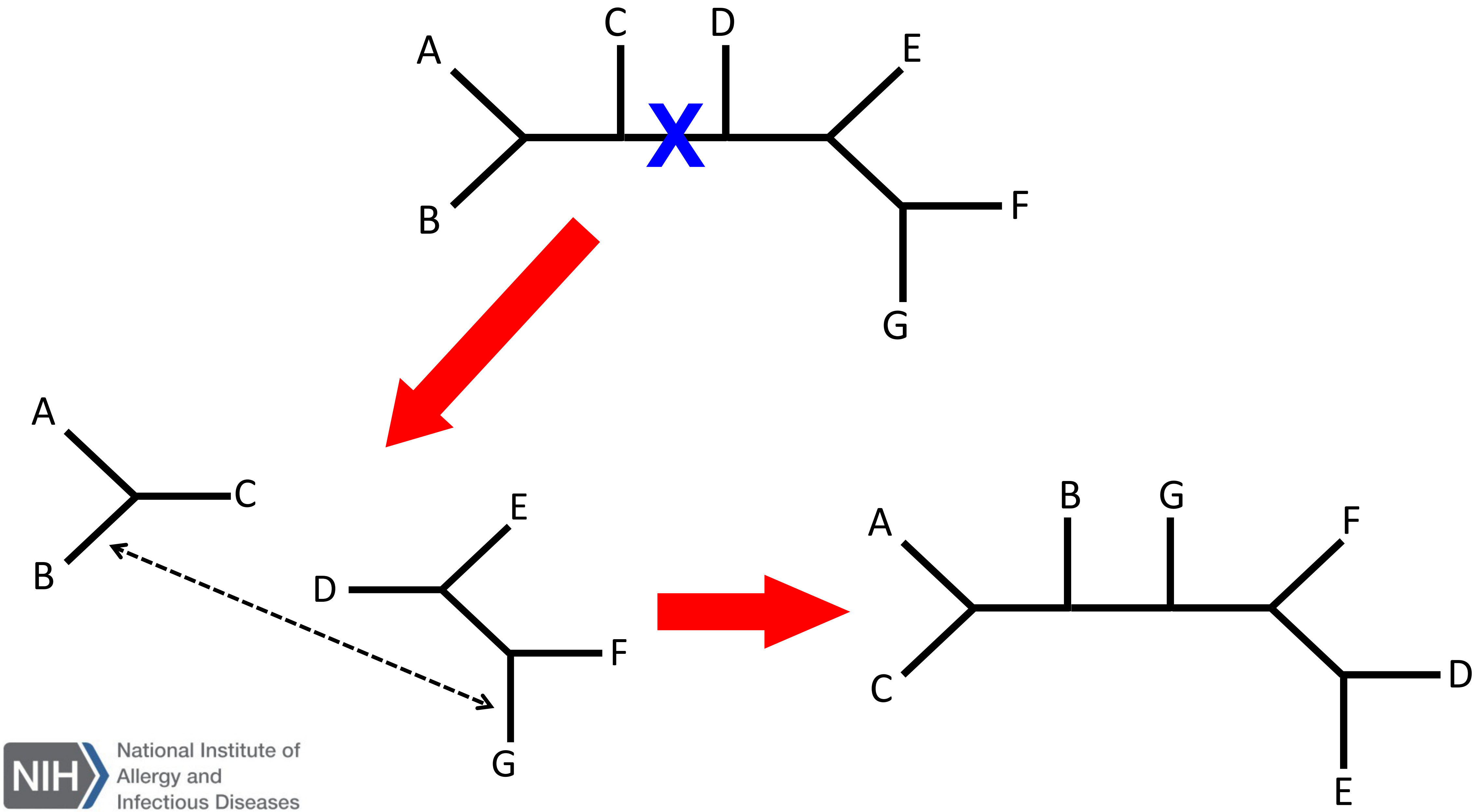
The optimality criterion approach

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



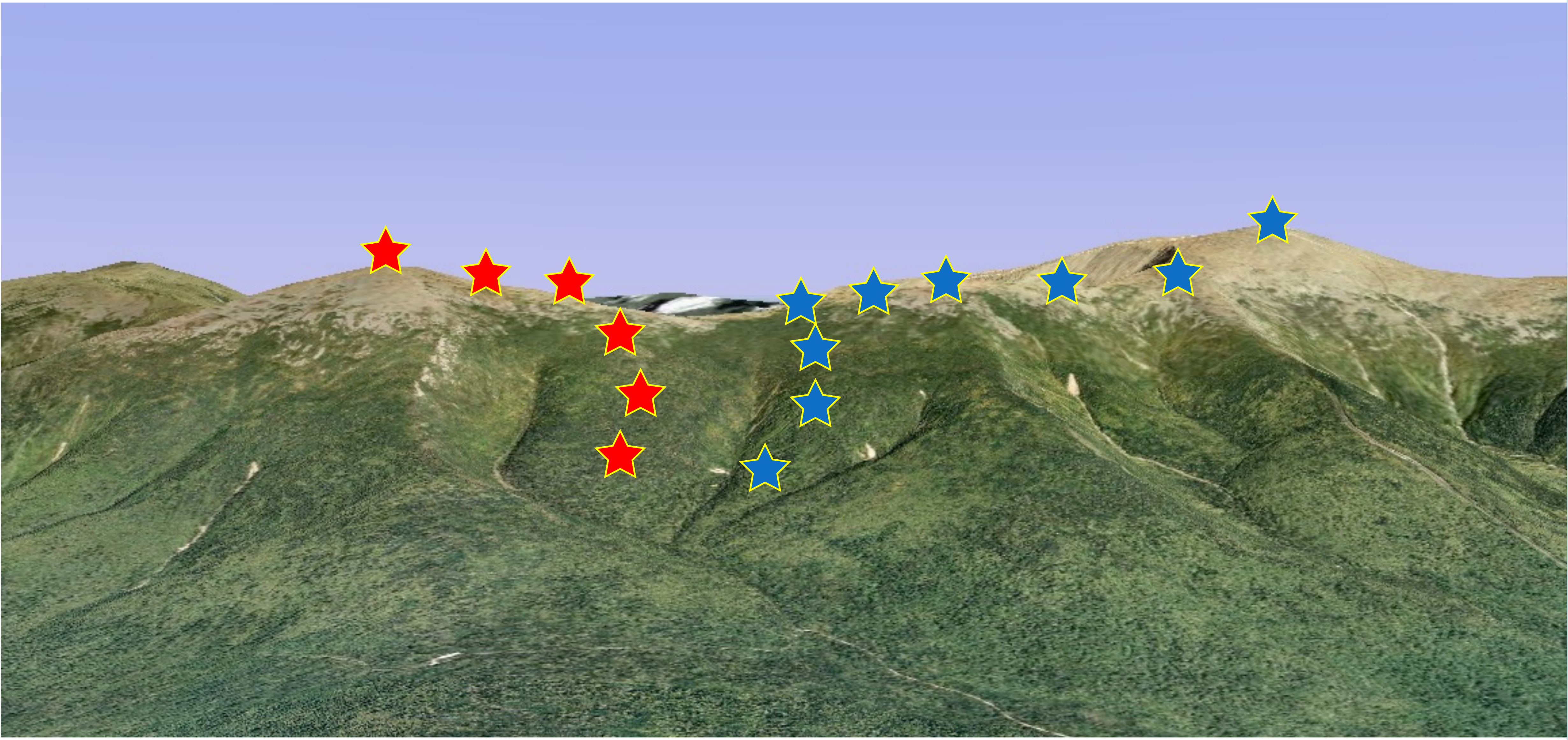
The optimality criterion approach

Exploring tree space

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

The optimality criterion approach

Exploring tree space



National Institute of Allergy and Infectious Diseases

NIAID

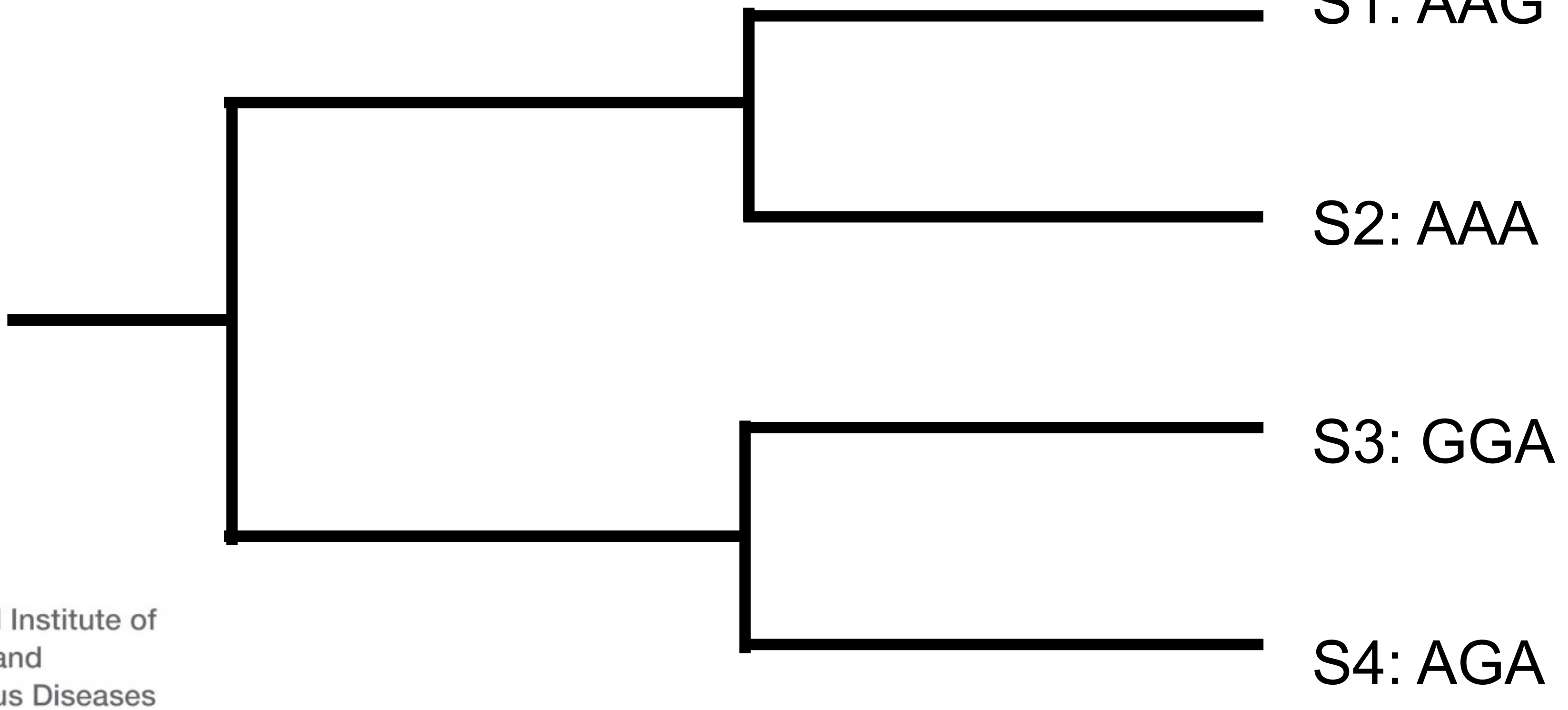
The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data



The optimality criterion approach

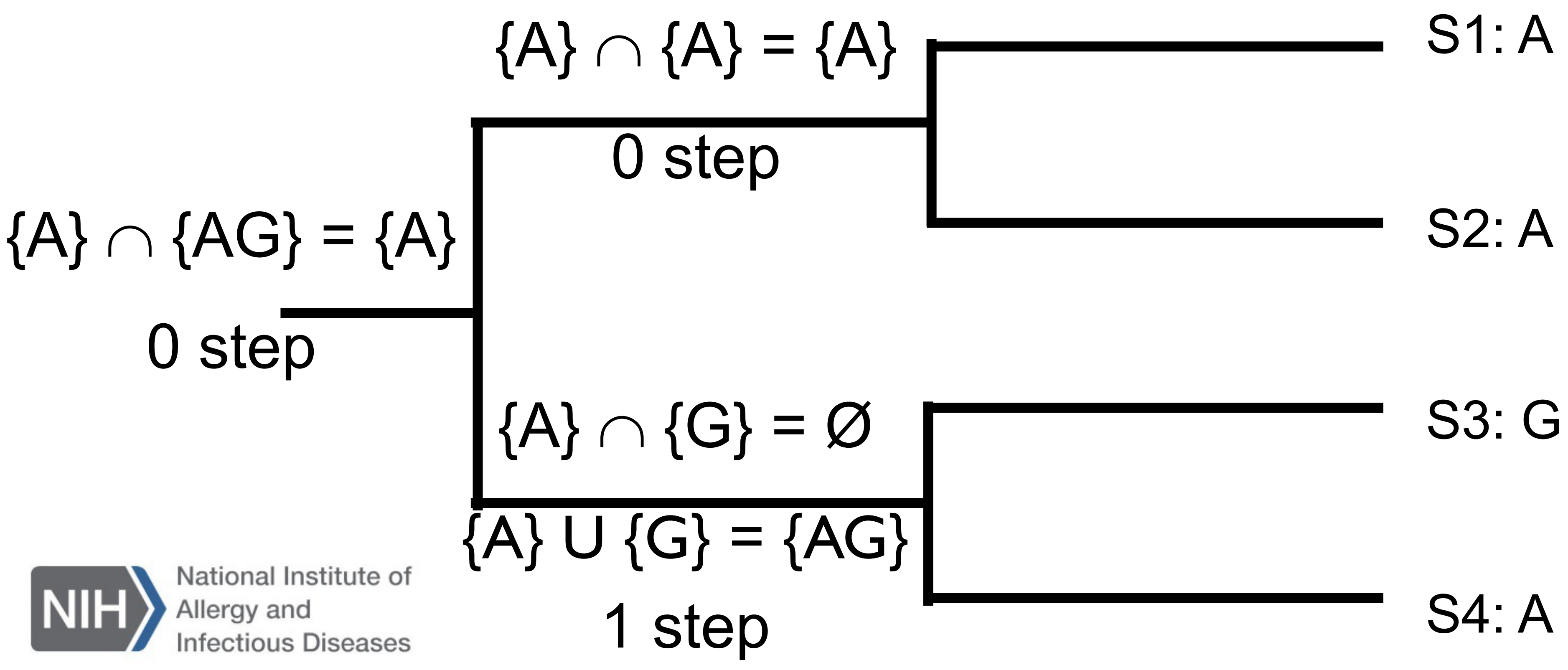
Is this tree optimal?

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

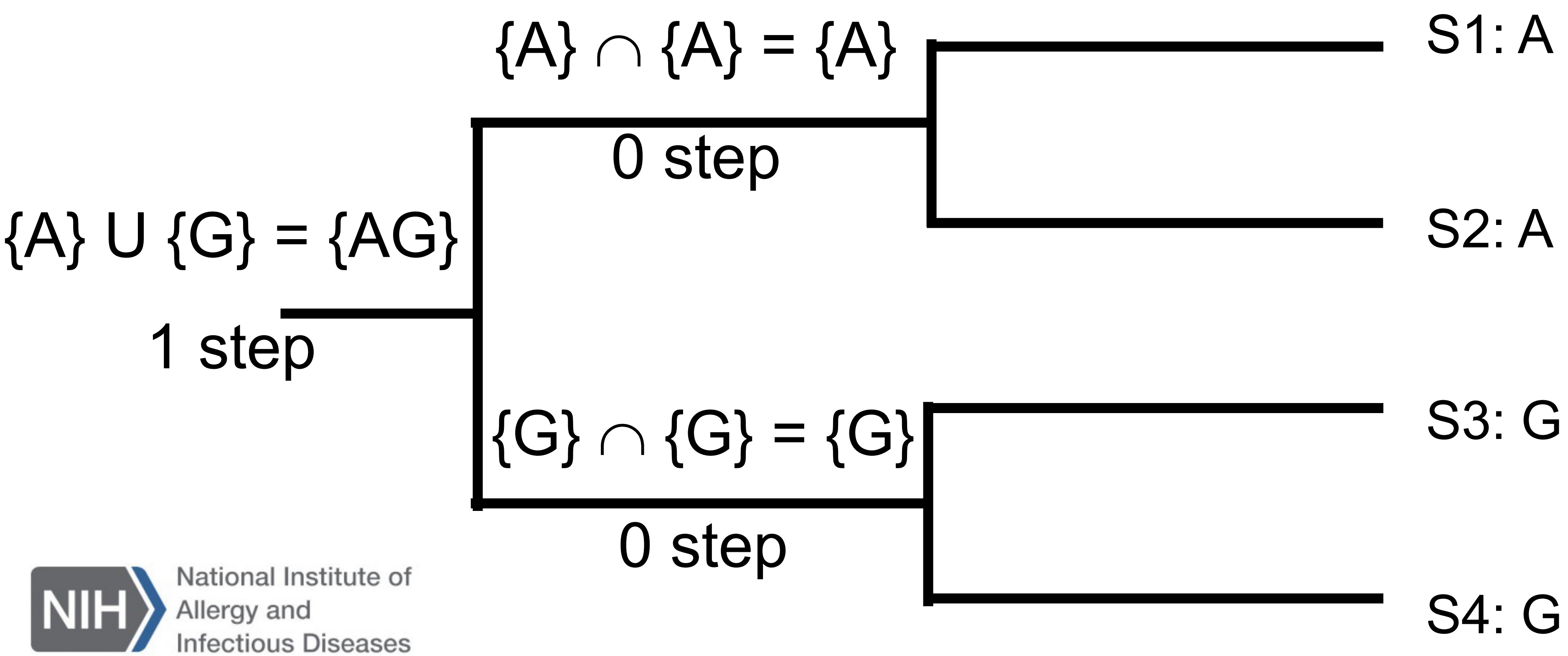
Is this tree optimal?

Counting changes (Fitch parsimony)

Position 2

The tree

The data



National Institute of Allergy and Infectious Diseases

The optimality criterion approach

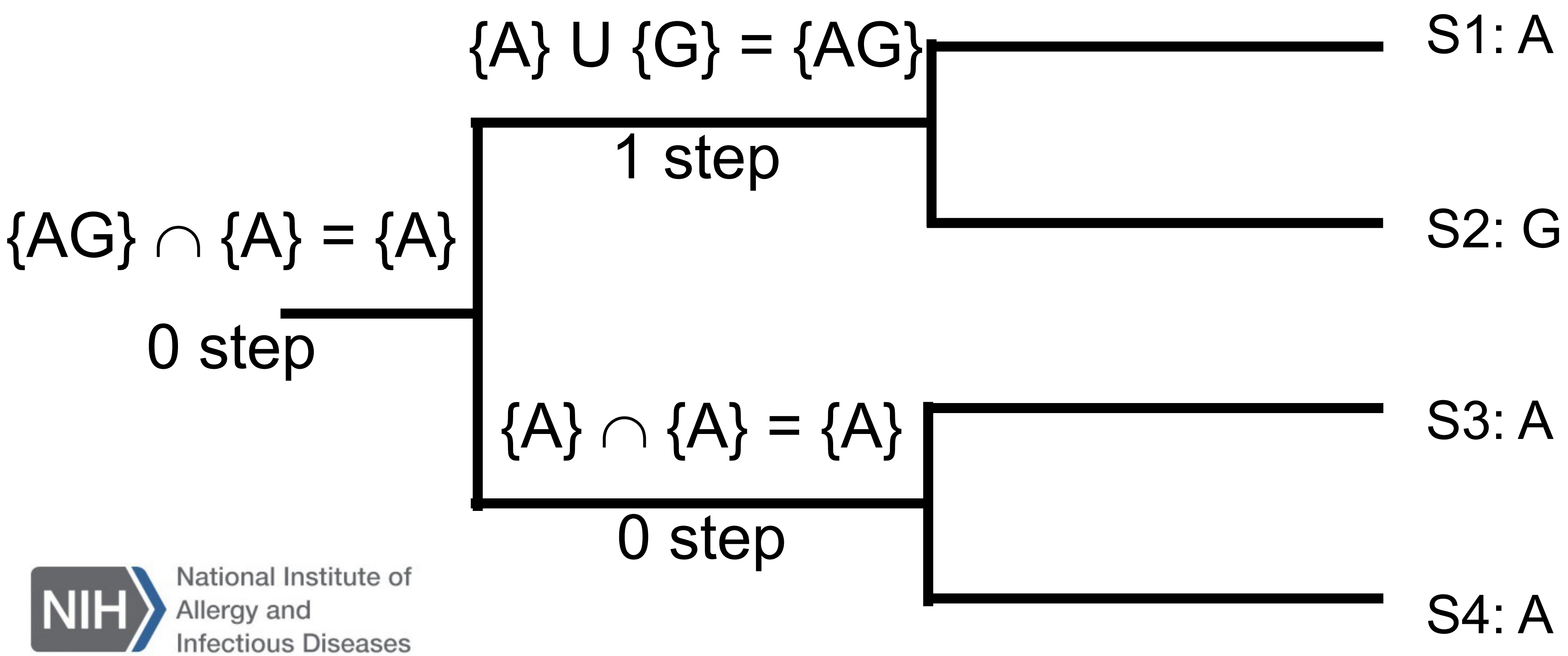
Is this tree optimal?

Counting changes (Fitch parsimony)

Position 3

The tree

The data



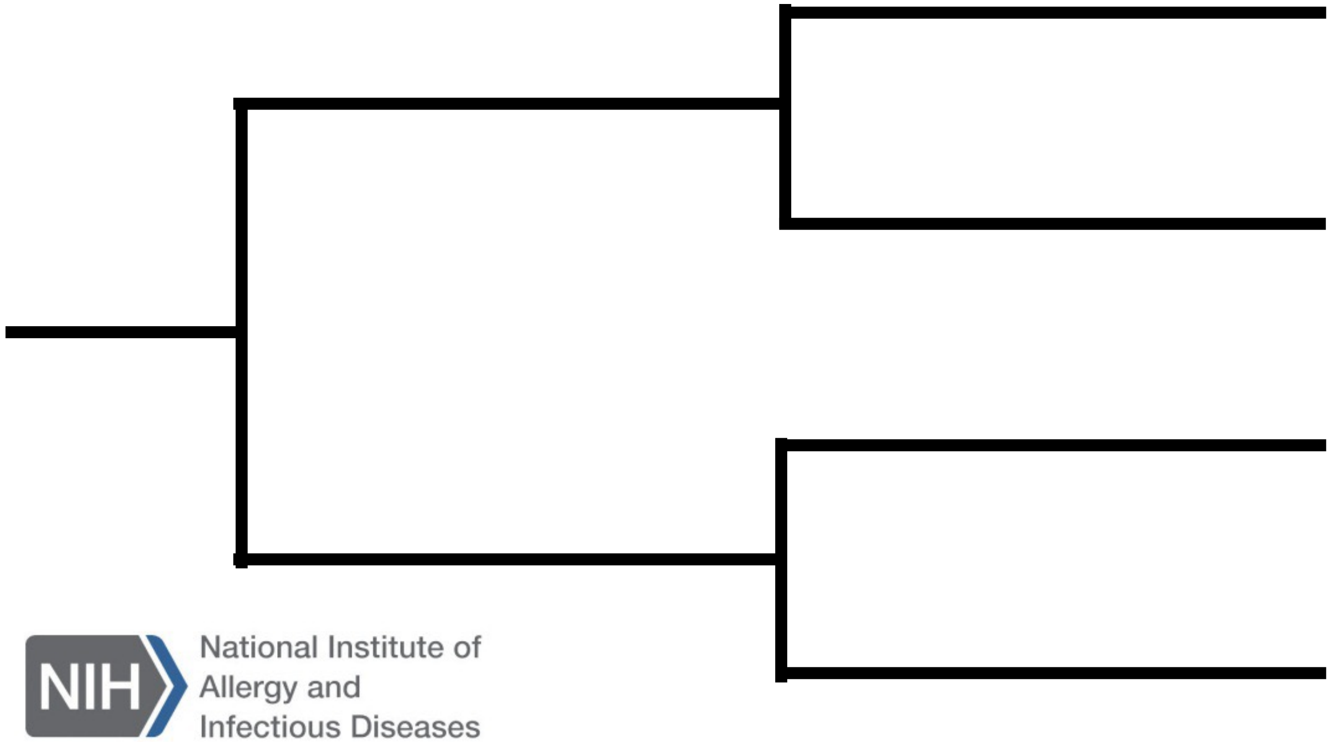
The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data



S1: AAG

S2: AAA

S3: GGA

S4: AGA

Total steps: 3

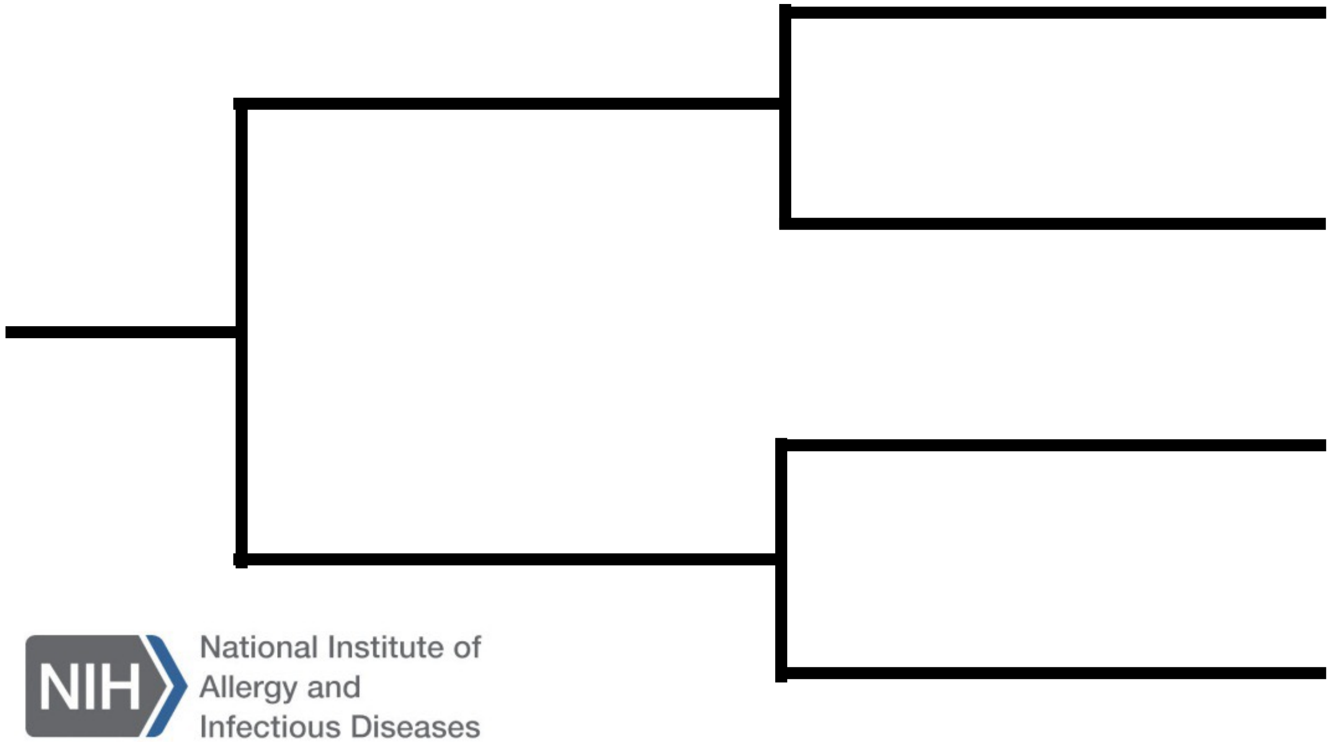
The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data



S1: AAG

S3: GGA

S2: AAA

S4: AGA

Total steps: 4

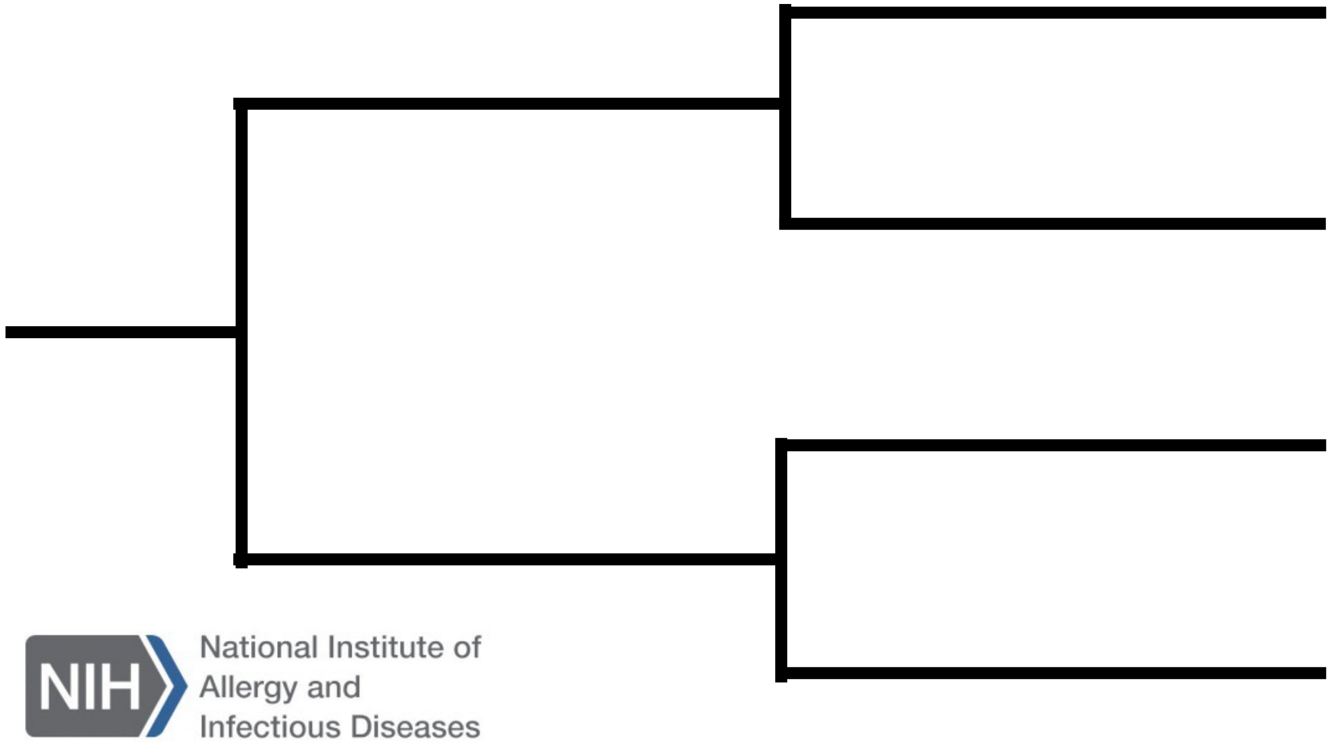
The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data



S1: AAG

S4: AGA

S2: AAA

S3: GGA

Total steps: 4

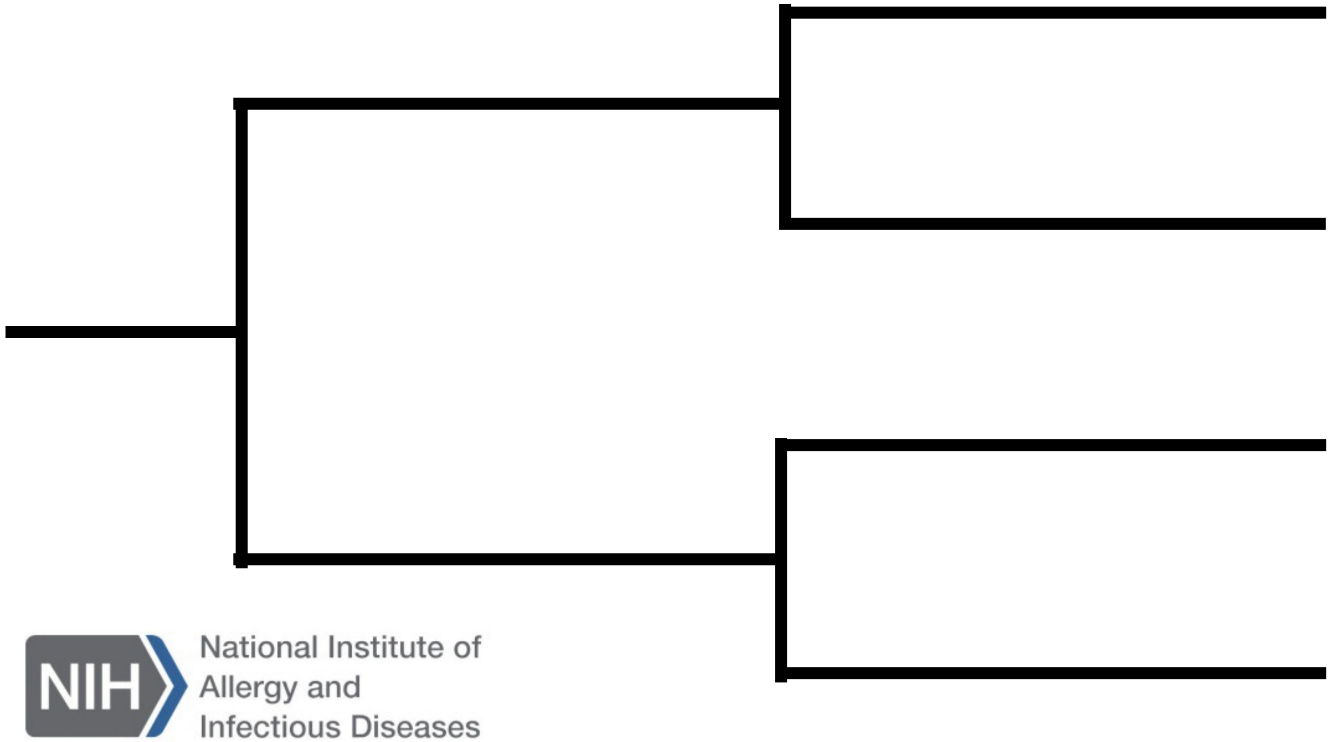
The optimality criterion approach

Is this tree optimal?

Counting changes (Fitch parsimony)

The tree

The data



S1: AAG

S2: AAA

S3: GGA

S4: AGA

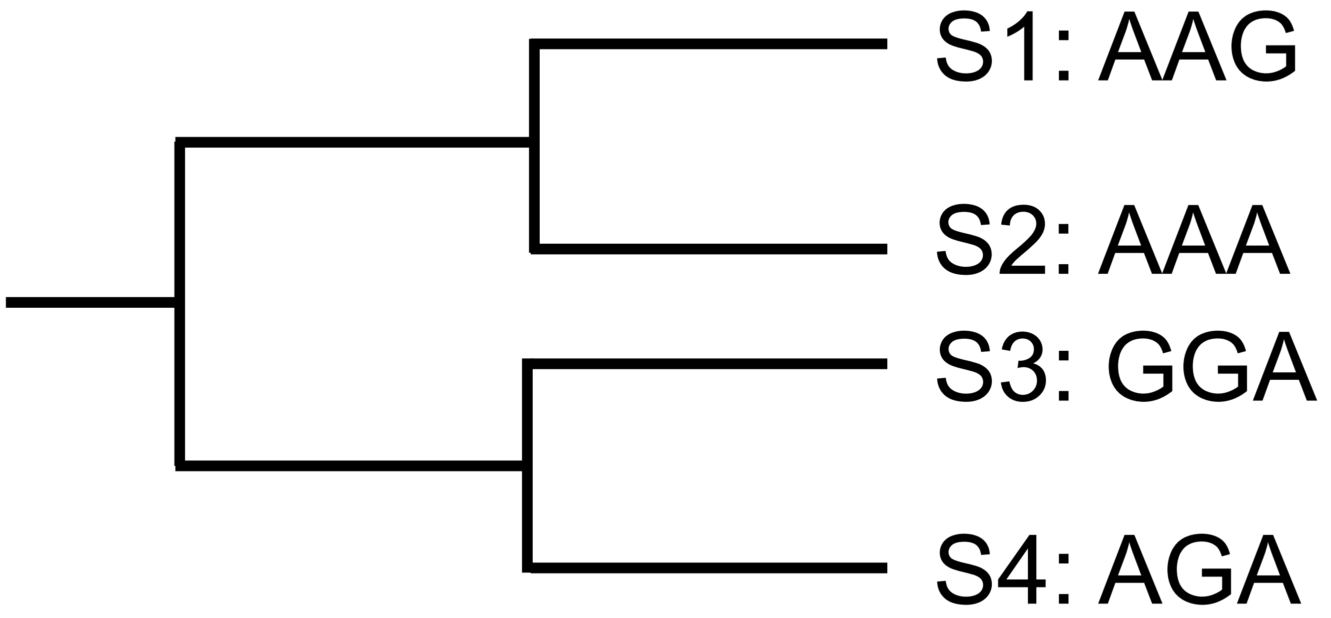
Optimal Tree!

Total steps: 3

Optimality Criterion: Likelihood

Calculating likelihood

The Tree The Data



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

National Institute of Allergy and Infectious Diseases

Optimality Criterion: Likelihood

Calculating likelihood: Setting parameters

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

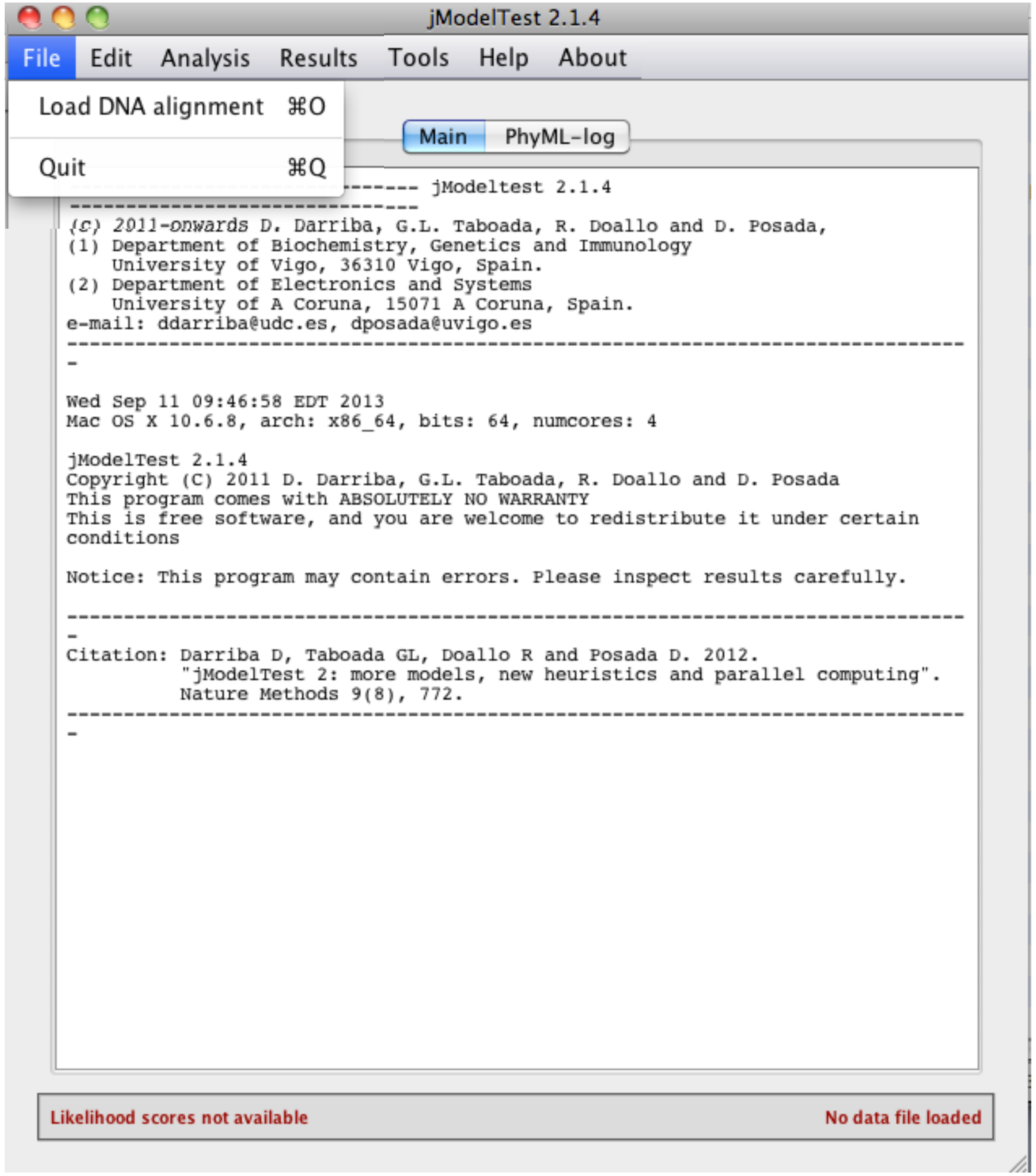
What values do you use for the substitution model?

Run jModelTest (or ProtTest for protein MSAs)

National Institute of Allergy and Infectious Diseases

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest



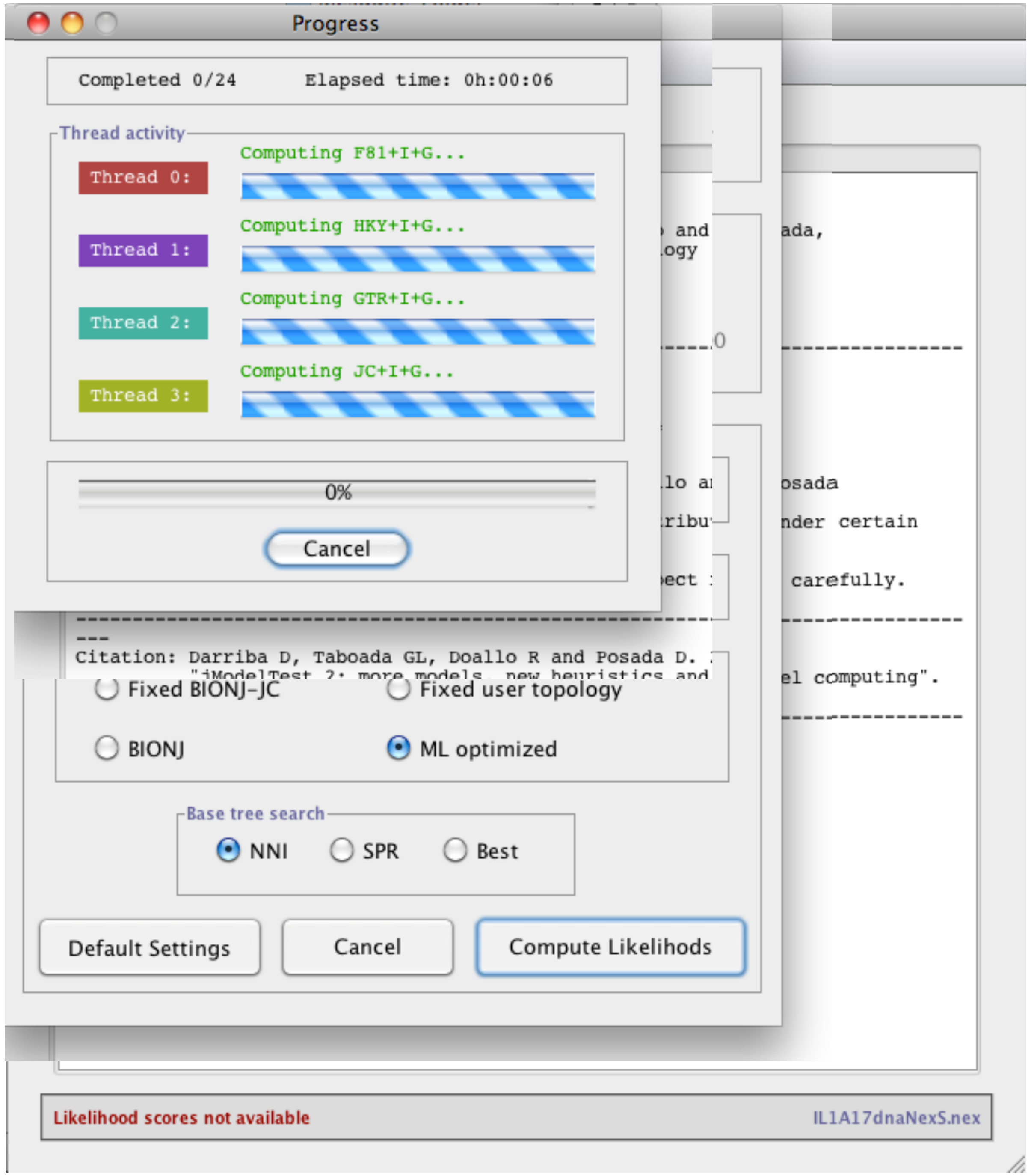
```
                              -------- jModeltest 2.1.4
------------------------------
(c) 2011-onwards D. Darriba, G.L. Taboada, R. Doallo and D. Posada,
(1) Department of Biochemistry, Genetics and Immunology
    University of Vigo, 36310 Vigo, Spain.
(2) Department of Electronics and Systems
    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

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest



Progress

Completed 0/24   Elapsed time: 0h:00:06

Thread activity

Thread 0: Computing F81+I+G...

Thread 1: Computing HKY+I+G...

Thread 2: Computing GTR+I+G...

Thread 3: Computing JC+I+G...

0%

Cancel

Citation: Darriba D, Taboada GL, Doallo R and Posada D.

○ Fixed BIONJ-JC ○ Fixed user topology

○ BIONJ ● ML optimized

Base tree search

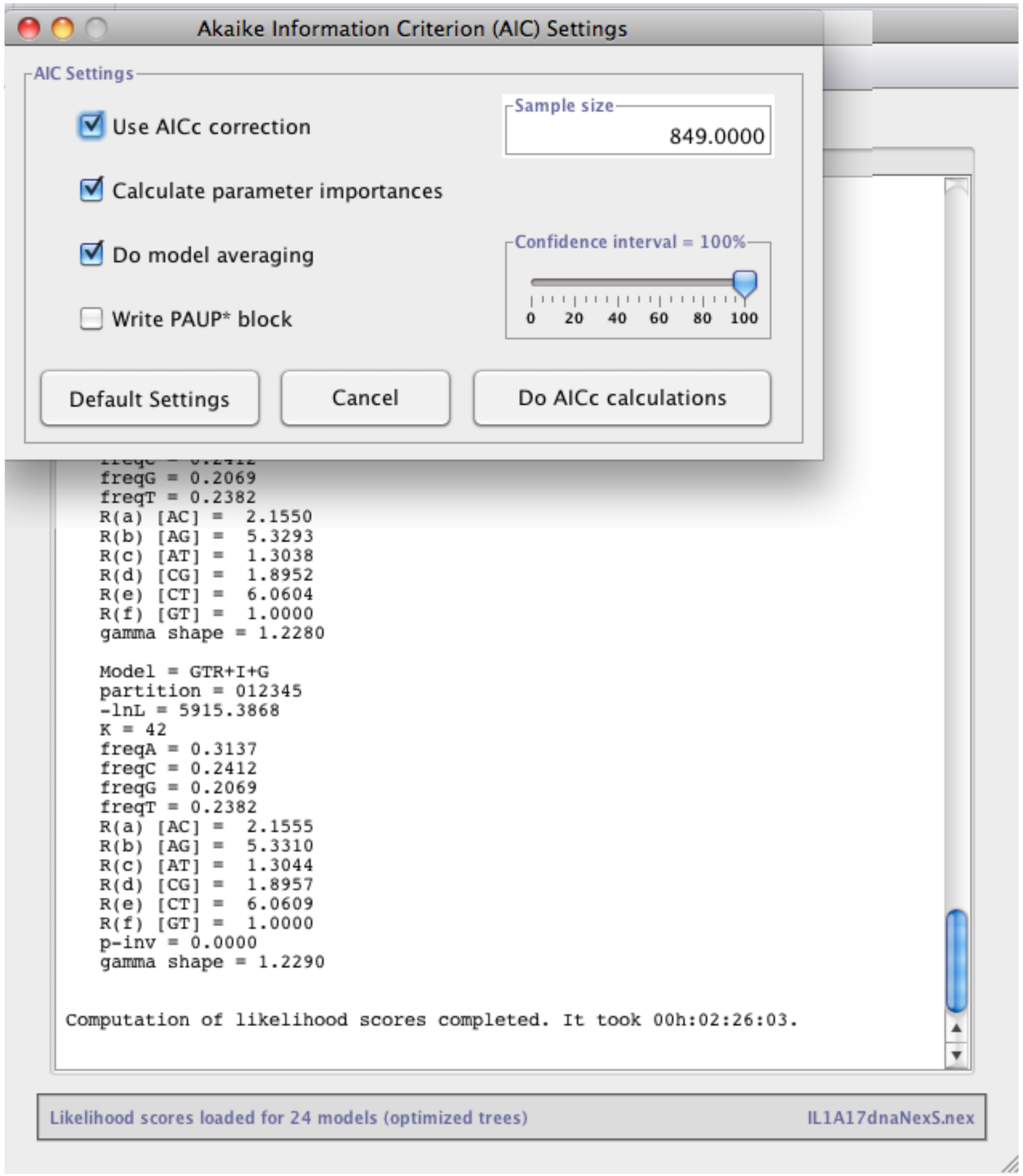
● NNI ○ SPR ○ Best

Default Settings | Cancel | Compute Likelihods

Likelihood scores not available   IL1A17dnaNexS.nex

Optimality Criterion: Likelihood

Calculating likelihood: jModelTest Results



Akaike Information Criterion (AIC) Settings

AIC Settings

☑ Use AICc correction

☑ Calculate parameter importances

☑ Do model averaging

□ Write PAUP* block

Sample size: 849.0000

Confidence interval = 100%

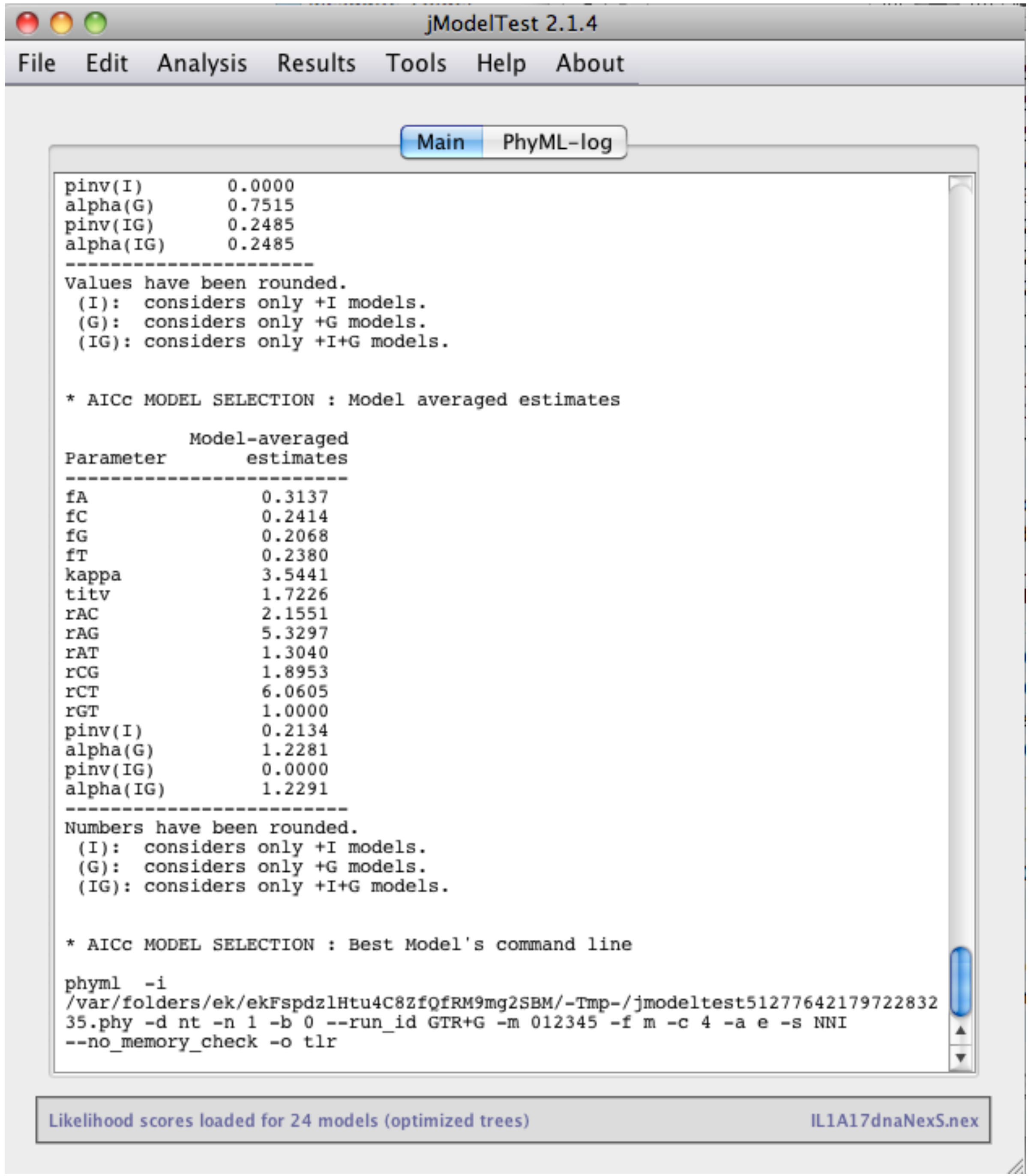
```
freqG = 0.2069
freqT = 0.2382
R(a) [AC] =  2.1550
R(b) [AG] =  5.3293
R(c) [AT] =  1.3038
R(d) [CG] =  1.8952
R(e) [CT] =  6.0604
R(f) [GT] =  1.0000
gamma shape = 1.2280

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] =  1.3044
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

National Institute of Allergy and Infectious Diseases

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/ekFspdzlHtu4C8ZfQfRM9mg2SBM/-Tmp-/jmodeltest512776421797228323
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
```

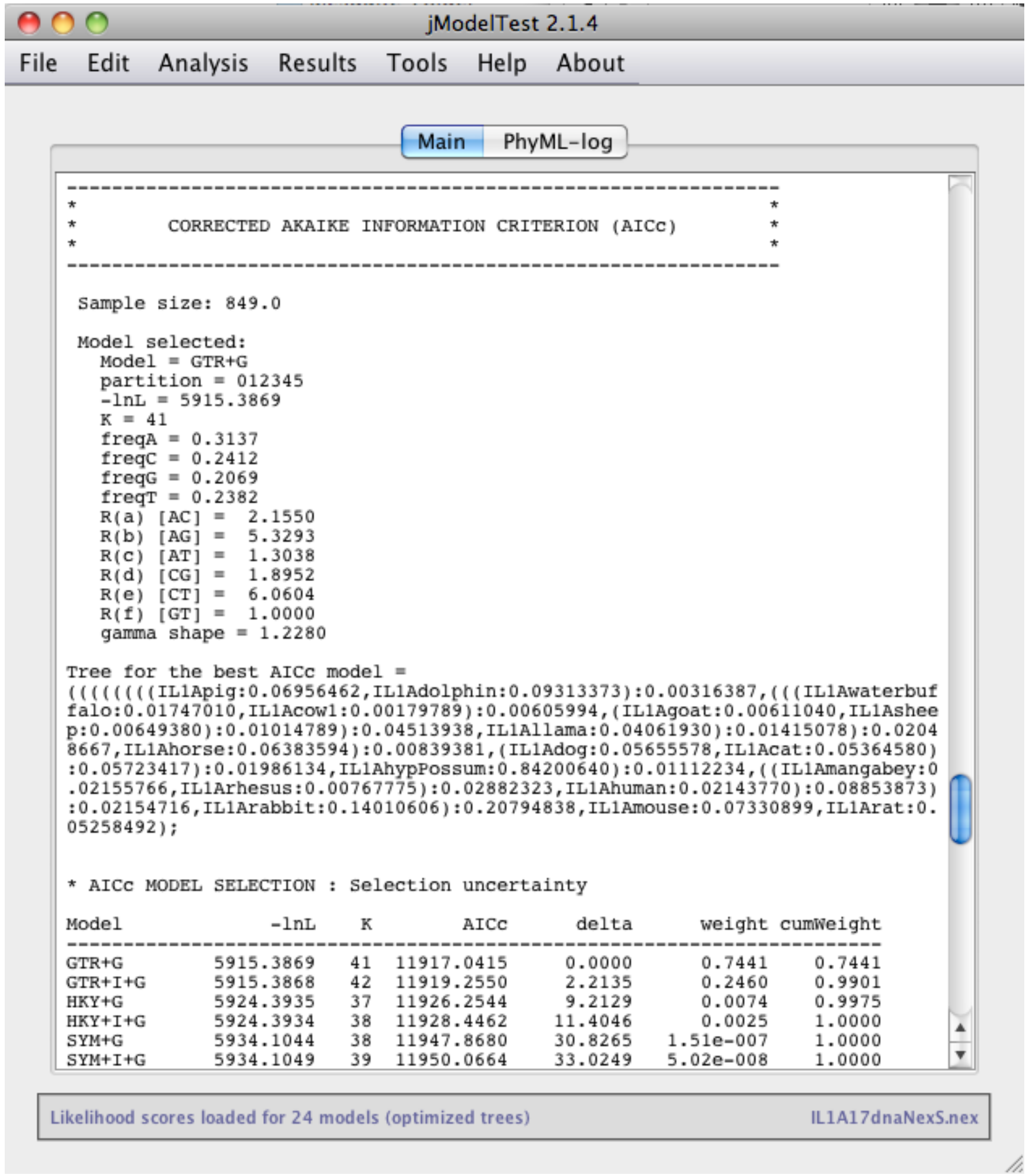
Likelihood scores loaded for 24 models (optimized trees) IL1A17dnaNexS.nex

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,IL1AhypPossum: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) IL1A17dnaNexS.nex

National Institute of Allergy and Infectious Diseases

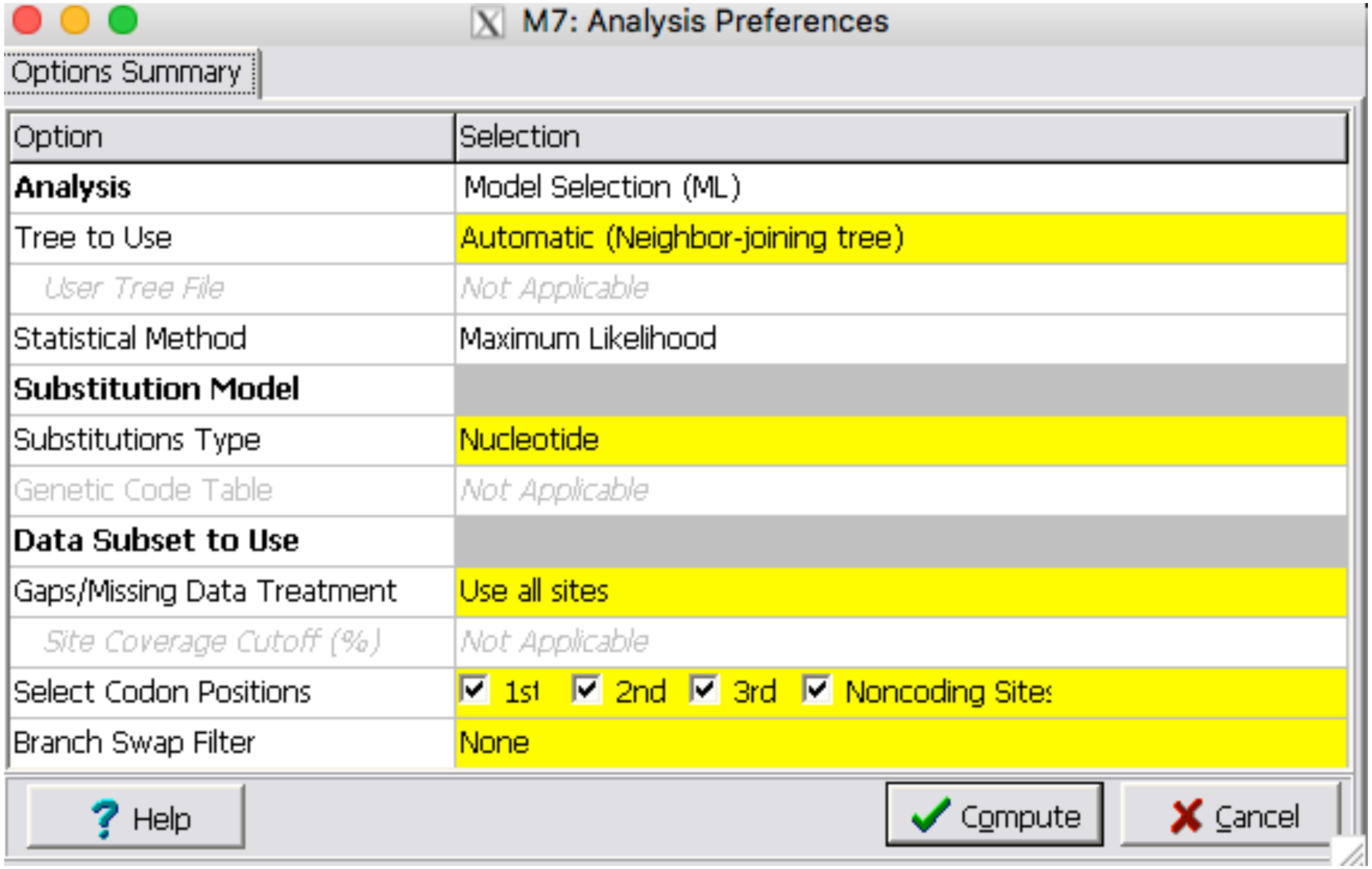
NIAID

Substitution models

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

Optimality Criterion: Likelihood

Calculating likelihood: MEGA7 Options



Option	Selection
Analysis	Model Selection (ML)
Tree to Use	Automatic (Neighbor-joining tree)
User Tree File	*Not Applicable*
Statistical Method	Maximum Likelihood
Substitution Model	
Substitutions Type	Nucleotide
Genetic Code Table	*Not Applicable*
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Site Coverage Cutoff (%)	*Not Applicable*
Select Codon Positions	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites
Branch Swap Filter	None

Optimality Criterion: Likelihood

Calculating likelihood: MEGA7 Results

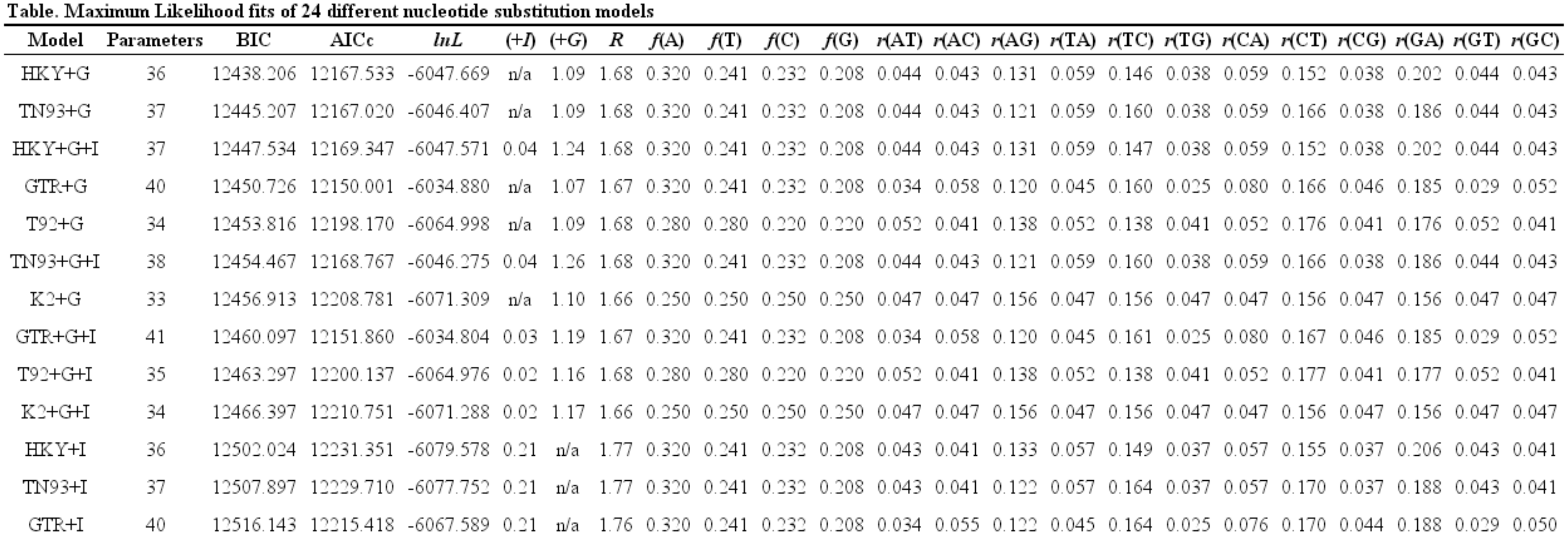
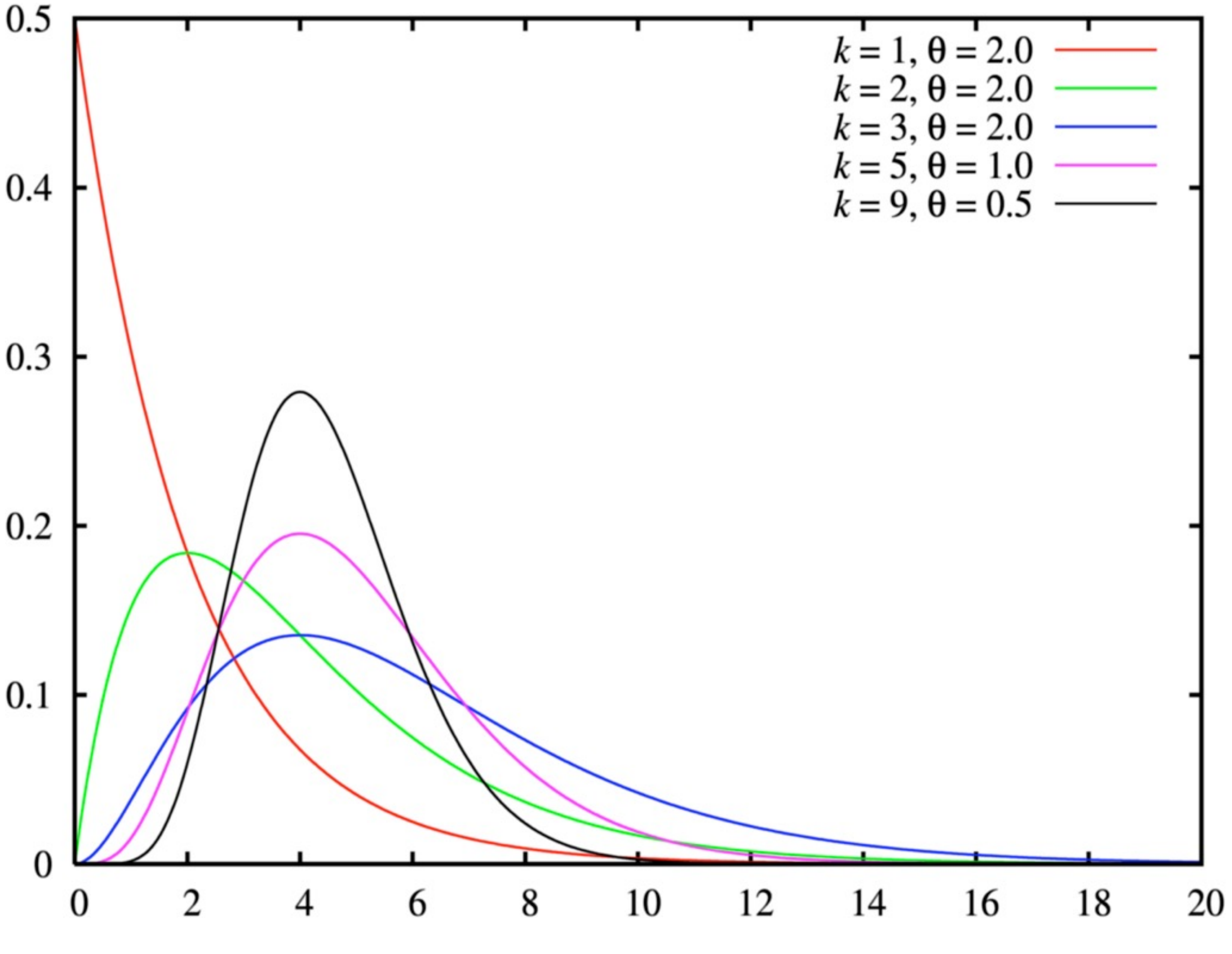


Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
HKY+G	36	12438.206	12167.533	-6047.669	n/a	1.09	1.68	0.320	0.241	0.232	0.208	0.044	0.043	0.131	0.059	0.146	0.038	0.059	0.152	0.038	0.202	0.044	0.043
TN93+G	37	12445.207	12167.020	-6046.407	n/a	1.09	1.68	0.320	0.241	0.232	0.208	0.044	0.043	0.121	0.059	0.160	0.038	0.059	0.166	0.038	0.186	0.044	0.043
HKY+G+I	37	12447.534	12169.347	-6047.571	0.04	1.24	1.68	0.320	0.241	0.232	0.208	0.044	0.043	0.131	0.059	0.147	0.038	0.059	0.152	0.038	0.202	0.044	0.043
GTR+G	40	12450.726	12150.001	-6034.880	n/a	1.07	1.67	0.320	0.241	0.232	0.208	0.034	0.058	0.120	0.045	0.160	0.025	0.080	0.166	0.046	0.185	0.029	0.052
T92+G	34	12453.816	12198.170	-6064.998	n/a	1.09	1.68	0.280	0.280	0.220	0.220	0.052	0.041	0.138	0.052	0.138	0.041	0.052	0.176	0.041	0.176	0.052	0.041
TN93+G+I	38	12454.467	12168.767	-6046.275	0.04	1.26	1.68	0.320	0.241	0.232	0.208	0.044	0.043	0.121	0.059	0.160	0.038	0.059	0.166	0.038	0.186	0.044	0.043
K2+G	33	12456.913	12208.781	-6071.309	n/a	1.10	1.66	0.250	0.250	0.250	0.250	0.047	0.047	0.156	0.047	0.156	0.047	0.047	0.156	0.047	0.156	0.047	0.047
GTR+G+I	41	12460.097	12151.860	-6034.804	0.03	1.19	1.67	0.320	0.241	0.232	0.208	0.034	0.058	0.120	0.045	0.161	0.025	0.080	0.167	0.046	0.185	0.029	0.052
T92+G+I	35	12463.297	12200.137	-6064.976	0.02	1.16	1.68	0.280	0.280	0.220	0.220	0.052	0.041	0.138	0.052	0.138	0.041	0.052	0.177	0.041	0.177	0.052	0.041
K2+G+I	34	12466.397	12210.751	-6071.288	0.02	1.17	1.66	0.250	0.250	0.250	0.250	0.047	0.047	0.156	0.047	0.156	0.047	0.047	0.156	0.047	0.156	0.047	0.047
HKY+I	36	12502.024	12231.351	-6079.578	0.21	n/a	1.77	0.320	0.241	0.232	0.208	0.043	0.041	0.133	0.057	0.149	0.037	0.057	0.155	0.037	0.206	0.043	0.041
TN93+I	37	12507.897	12229.710	-6077.752	0.21	n/a	1.77	0.320	0.241	0.232	0.208	0.043	0.041	0.122	0.057	0.164	0.037	0.057	0.170	0.037	0.188	0.043	0.041
GTR+I	40	12516.143	12215.418	-6067.589	0.21	n/a	1.76	0.320	0.241	0.232	0.208	0.034	0.055	0.122	0.045	0.164	0.025	0.076	0.170	0.044	0.188	0.029	0.050

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

The Gamma Distribution



Mean = kθ   Shape parameter = θ

Coefficient of Variation = 1/√θ

How reliable are my trees?

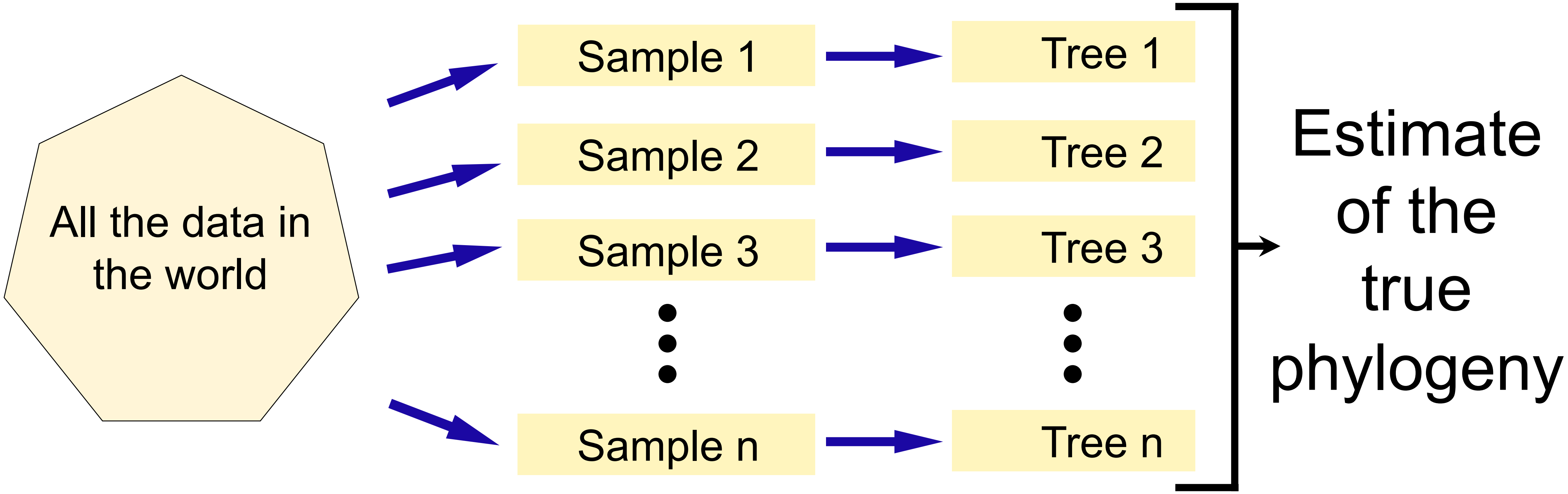
Bootstrapping (nonparametric)

NIH National Institute of Allergy and Infectious Diseases

NIAID

Bootstrapping

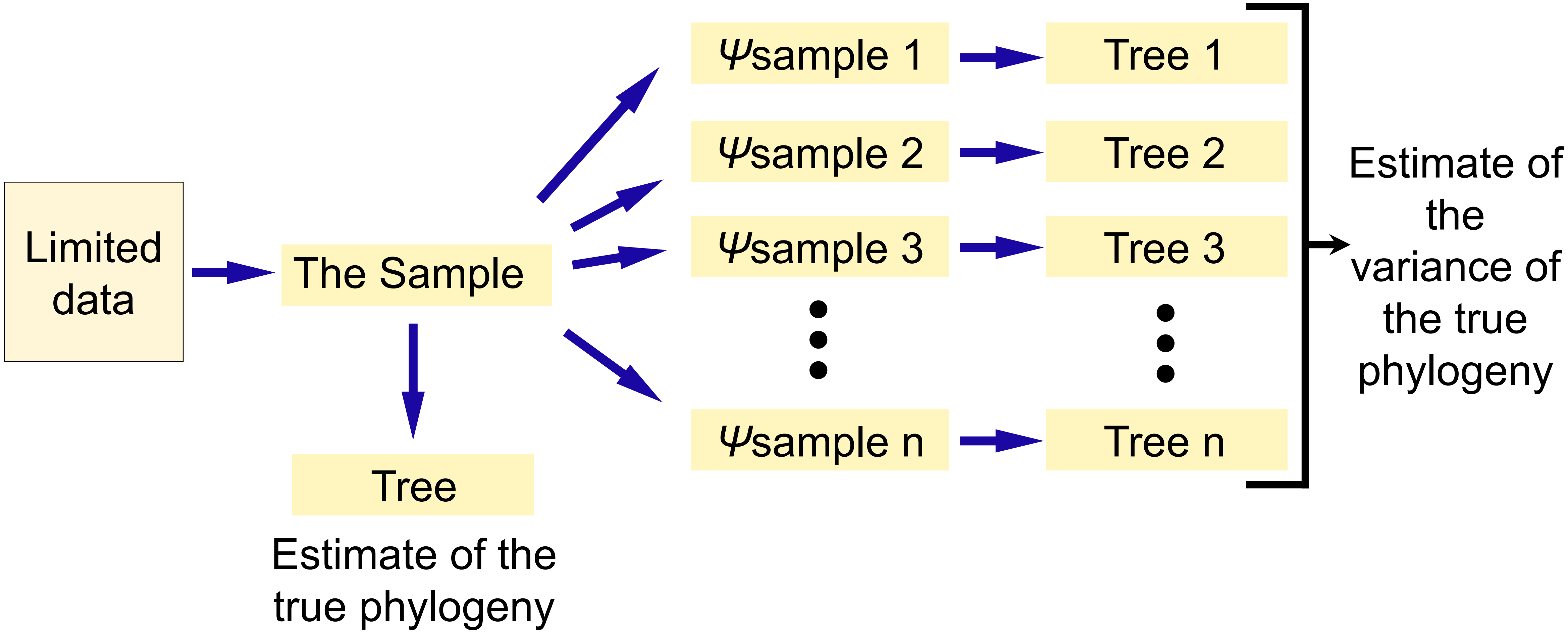
the ideal world



Build replicates by resampling from unlimited data

Bootstrapping

the real world



Limited data

The Sample

Tree

Estimate of the true phylogeny

Ψsample 1

Ψsample 2

Ψsample 3

Ψsample n

Tree 1

Tree 2

Tree 3

Tree n

Estimate of the variance of the true phylogeny

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

NIH National Institute of Allergy and Infectious Diseases

NIAID

Calculating likelihood: Programs

PAUP* – Commercial, NIH Biowulf, or NIAID HPC

DNA only

PHYLIP – Download, NIH Biowulf, or NIAID HPC

dnaml and proml programs

MEGA – Download for free from www.megasoftware.net

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

Generally the user has more flexibility with a local program.

But local programs can hog your computer.

Input File Formats

Phylogenetics program input file formats

FASTA

```
>MC1_01B4fs
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
>MC1_01A10
TGCACT---AATCTGACAAAGGCTATTAAGACCAATGGGAATGCTAATAATACCAGTACT
>MC1_01C1
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
>MC1_01A20
TGCACTAATAATCTGACAAAGGCTAGTAATGCCACTGAGAAGGCTAATAATACCATTACT
>MC1_01TA1
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
```

Input File Formats

Phylogenetics program input file formats

PHYLIP

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

```
5 60
MC1_01B4fsTGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
MC1_01A10 TGCACT---AATCTGACAAAGGCTATTAAGACCAATGGGAATGCTAATAATACCAGTACT
MC1_01C1  TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
MC1_01A20 TGCACTAATAATCTGACAAAGGCTAGTAATGCCACTGAGAAGGCTAATAATACCATTACT
MC1_01TA1 TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
```

Input File Formats

Phylogenetics program input file formats

NEXUS

```
#NEXUS
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;
```

Input File Formats

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

What’s next?

After the break

Building Bayesian trees with our MSA

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

National Institute of Allergy and Infectious Diseases

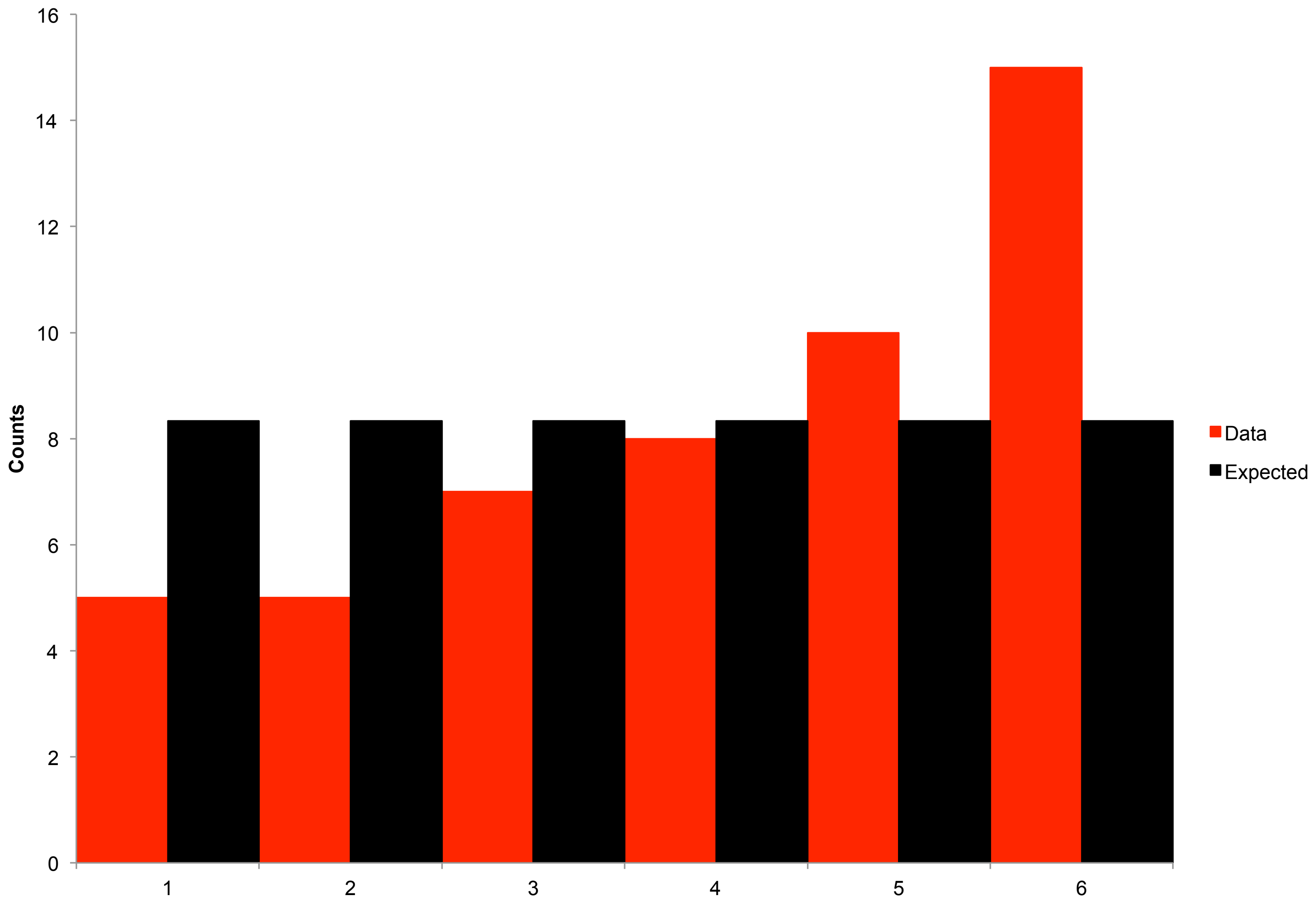
NIAID

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

What is Bayesian analysis?

Likelihood that this die is unbiased?



Bayesian Analysis

Exploring the posterior probability distribution

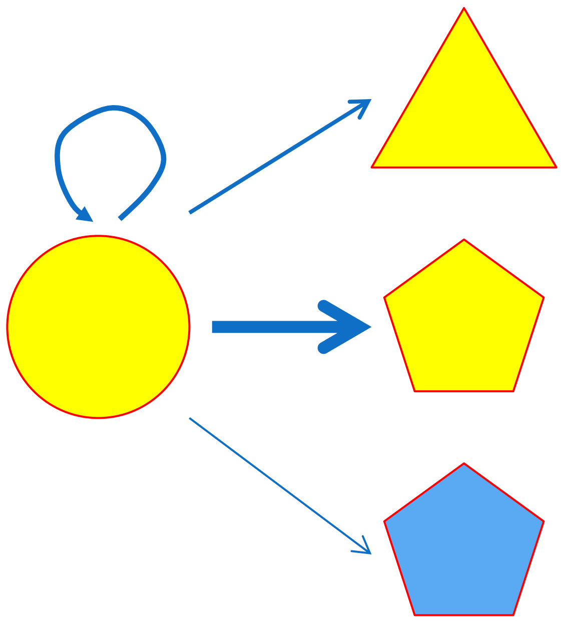
Posterior probabilities of trees and parameters are approximated using Markov Chain Monte Carlo (MCMC) sampling

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

What is MCMC?

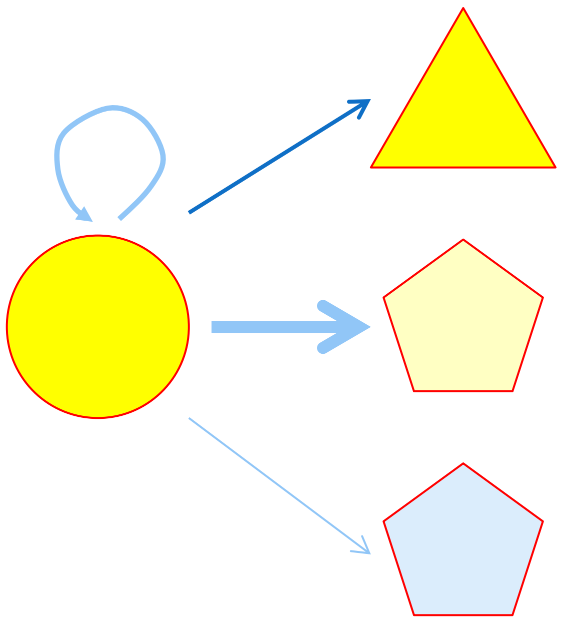
Markov Chain Monte Carlo

Markov chain



One link in the chain

Monte Carlo



Choosing a link

National Institute of Allergy and Infectious Diseases

Bayesian Analysis

Markov Chain example: Jukes-Cantor

$\mu t = 0.25$

	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

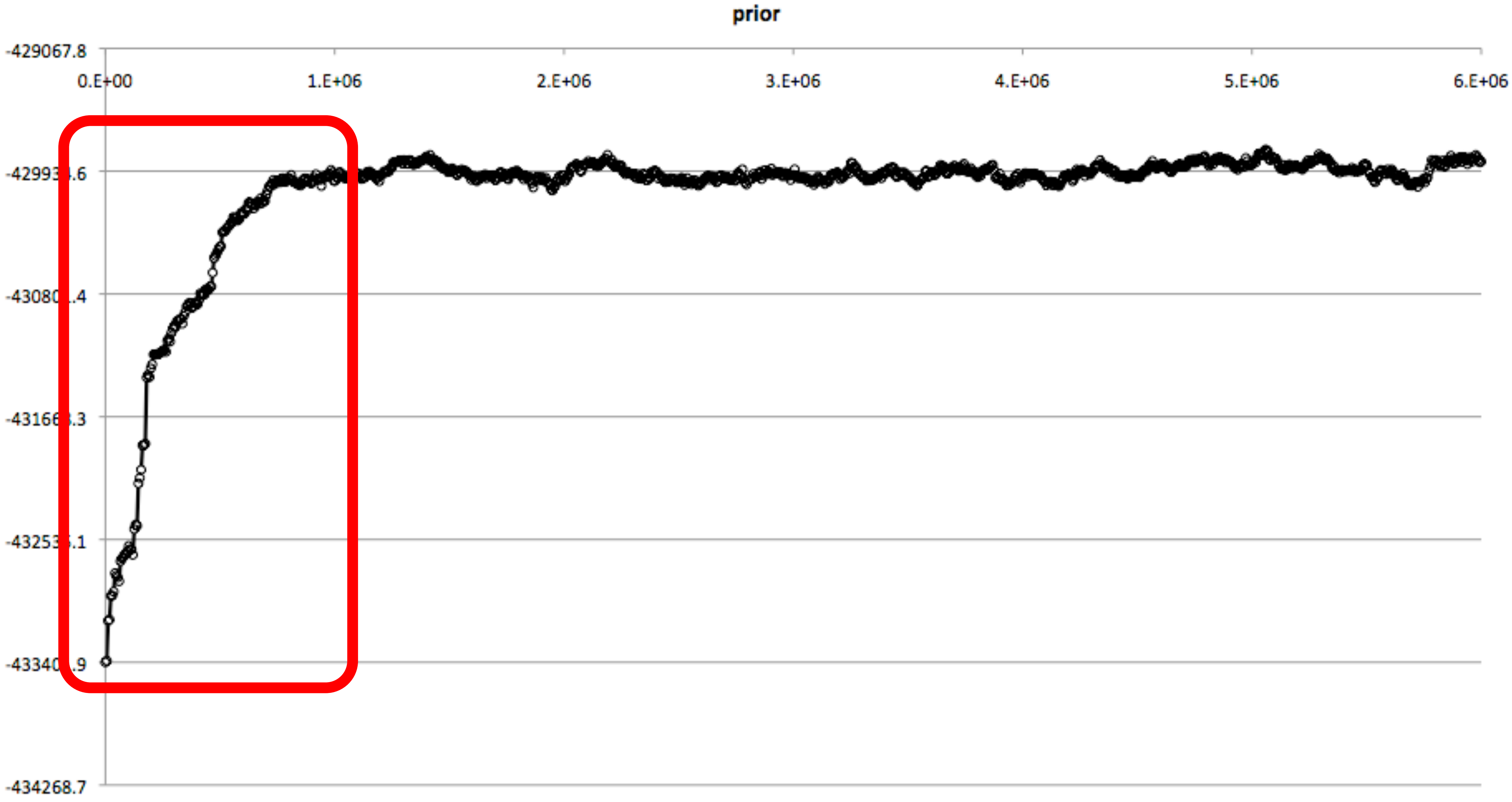
Running MrBayes: General

```
#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)
```

National Institute of
Allergy and
Infectious Diseases

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

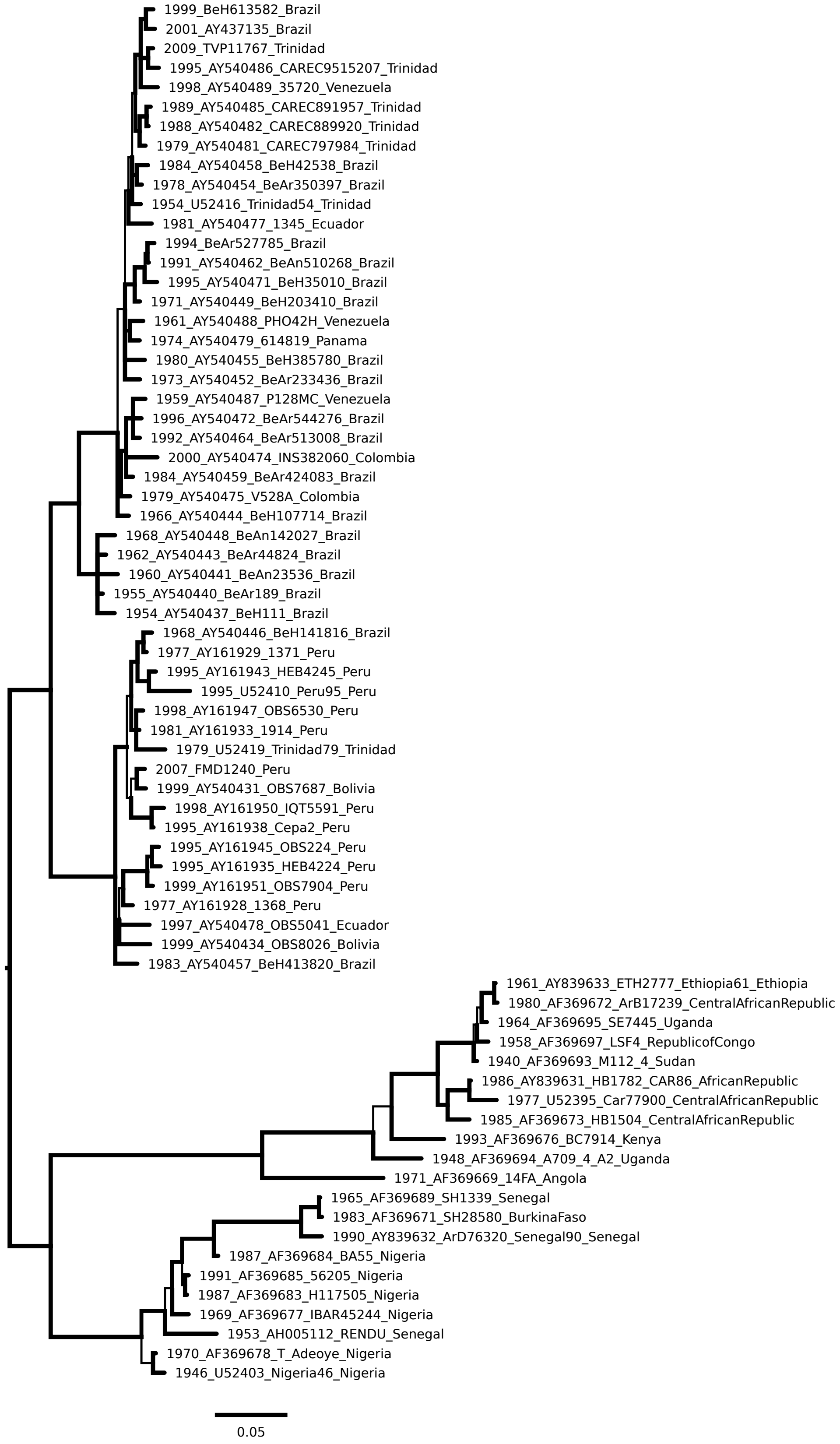
```
                                          95% HPD Interval
                                       --------------------
Parameter      Mean      Variance     Lower       Upper       Median    min ESS*  avg ESS    PSRF+
--------------------------------------------------------------------------------------------------
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
------------------------------------=-------------------------------------------------------------
```

National Institute of
Allergy and
Infectious Diseases

NIAID

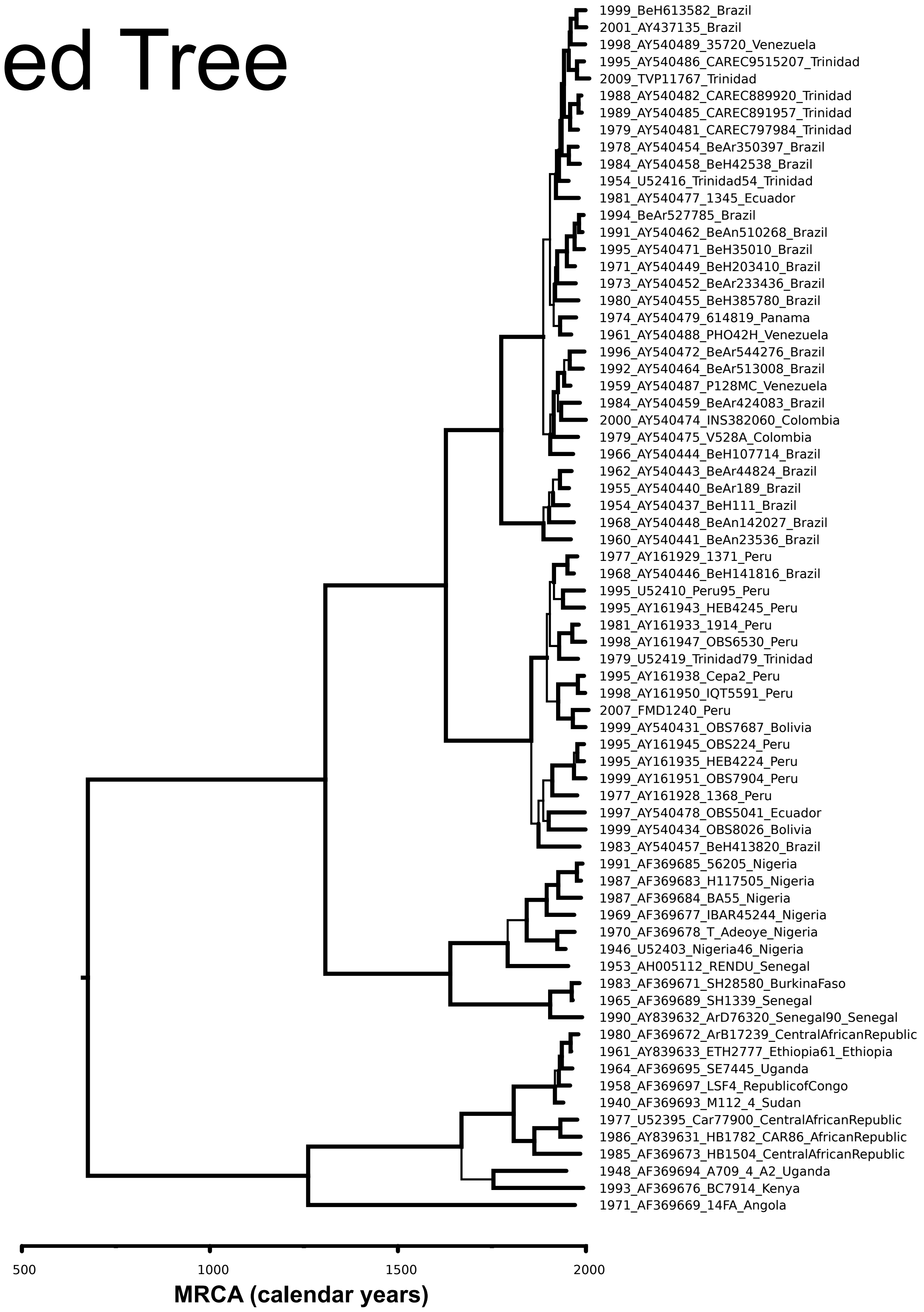
Bayesian Analysis

The Consensus Tree


1999_BeH613582_Brazil
2001_AY437135_Brazil
2009_TVP11767_Trinidad
1995_AY540486_CAREC9515207_Trinidad
1998_AY540489_35720_Venezuela
1989_AY540485_CAREC891957_Trinidad
1988_AY540482_CAREC889920_Trinidad
1979_AY540481_CAREC797984_Trinidad
1984_AY540458_BeH42538_Brazil
1978_AY540454_BeAr350397_Brazil
1954_U52416_Trinidad54_Trinidad
1981_AY540477_1345_Ecuador
1994_BeAr527785_Brazil
1991_AY540462_BeAn510268_Brazil
1995_AY540471_BeH35010_Brazil
1971_AY540449_BeH203410_Brazil
1961_AY540488_PHO42H_Venezuela
1974_AY540479_614819_Panama
1980_AY540455_BeH385780_Brazil
1973_AY540452_BeAr233436_Brazil
1959_AY540487_P128MC_Venezuela
1996_AY540472_BeAr544276_Brazil
1992_AY540464_BeAr513008_Brazil
2000_AY540474_INS382060_Colombia
1984_AY540459_BeAr424083_Brazil
1979_AY540475_V528A_Colombia
1966_AY540444_BeH107714_Brazil
1968_AY540448_BeAn142027_Brazil
1962_AY540443_BeAr44824_Brazil
1960_AY540441_BeAn23536_Brazil
1955_AY540440_BeAr189_Brazil
1954_AY540437_BeH111_Brazil
1968_AY540446_BeH141816_Brazil
1977_AY161929_1371_Peru
1995_AY161943_HEB4245_Peru
1995_U52410_Peru95_Peru
1998_AY161947_OBS6530_Peru
1981_AY161933_1914_Peru
1979_U52419_Trinidad79_Trinidad
2007_FMD1240_Peru
1999_AY540431_OBS7687_Bolivia
1998_AY161950_IQT5591_Peru
1995_AY161938_Cepa2_Peru
1995_AY161945_OBS224_Peru
1995_AY161935_HEB4224_Peru
1999_AY161951_OBS7904_Peru
1977_AY161928_1368_Peru
1997_AY540478_OBS5041_Ecuador
1999_AY540434_OBS8026_Bolivia
1983_AY540457_BeH413820_Brazil
1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1980_AF369672_ArB17239_CentralAfricanRepublic
1964_AF369695_SE7445_Uganda
1958_AF369697_LSF4_RepublicofCongo
1940_AF369693_M112_4_Sudan
1986_AY839631_HB1782_CAR86_AfricanRepublic
1977_U52395_Car77900_CentralAfricanRepublic
1985_AF369673_HB1504_CentralAfricanRepublic
1993_AF369676_BC7914_Kenya
1948_AF369694_A709_4_A2_Uganda
1971_AF369669_14FA_Angola
1965_AF369689_SH1339_Senegal
1983_AF369671_SH28580_BurkinaFaso
1990_AY839632_ArD76320_Senegal90_Senegal
1987_AF369684_BA55_Nigeria
1991_AF369685_56205_Nigeria
1987_AF369683_H117505_Nigeria
1969_AF369677_IBAR45244_Nigeria
1953_AH005112_RENDU_Senegal
1970_AF369678_T_Adeoye_Nigeria
1946_U52403_Nigeria46_Nigeria
0.05

Bayesian Analysis

The Time-Stamped Tree

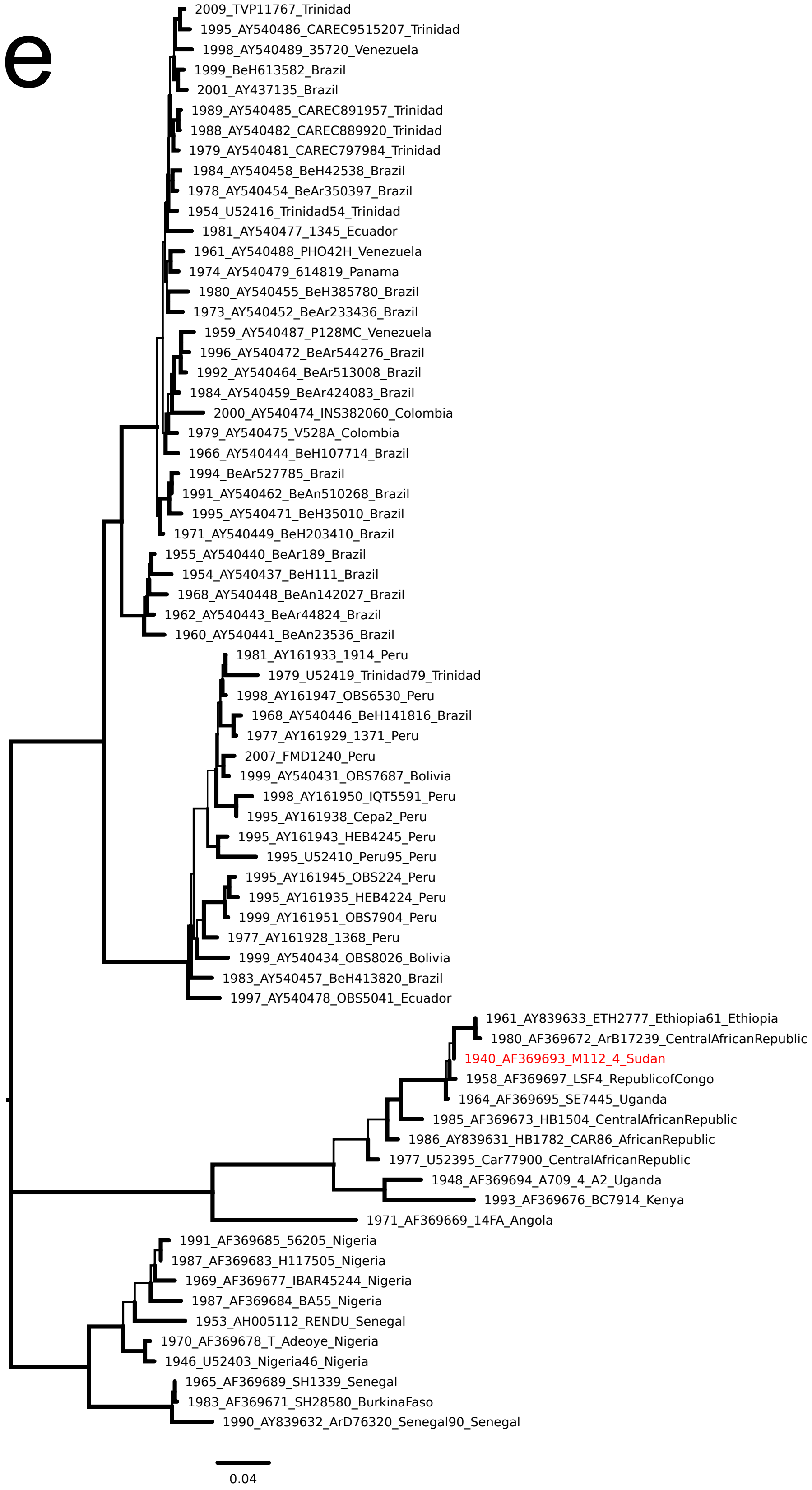


1999_BeH613582_Brazil
2001_AY437135_Brazil
1998_AY540489_35720_Venezuela
1995_AY540486_CAREC9515207_Trinidad
2009_TVP11767_Trinidad
1988_AY540482_CAREC889920_Trinidad
1989_AY540485_CAREC891957_Trinidad
1979_AY540481_CAREC797984_Trinidad
1978_AY540454_BeAr350397_Brazil
1984_AY540458_BeH42538_Brazil
1954_U52416_Trinidad54_Trinidad
1981_AY540477_1345_Ecuador
1994_BeAr527785_Brazil
1991_AY540462_BeAn510268_Brazil
1995_AY540471_BeH35010_Brazil
1971_AY540449_BeH203410_Brazil
1973_AY540452_BeAr233436_Brazil
1980_AY540455_BeH385780_Brazil
1974_AY540479_614819_Panama
1961_AY540488_PHO42H_Venezuela
1996_AY540472_BeAr544276_Brazil
1992_AY540464_BeAr513008_Brazil
1959_AY540487_P128MC_Venezuela
1984_AY540459_BeAr424083_Brazil
2000_AY540474_INS382060_Colombia
1979_AY540475_V528A_Colombia
1966_AY540444_BeH107714_Brazil
1962_AY540443_BeAr44824_Brazil
1955_AY540440_BeAr189_Brazil
1954_AY540437_BeH111_Brazil
1968_AY540448_BeAn142027_Brazil
1960_AY540441_BeAn23536_Brazil
1977_AY161929_1371_Peru
1968_AY540446_BeH141816_Brazil
1995_U52410_Peru95_Peru
1995_AY161943_HEB4245_Peru
1981_AY161933_1914_Peru
1998_AY161947_OBS6530_Peru
1979_U52419_Trinidad79_Trinidad
1995_AY161938_Cepa2_Peru
1998_AY161950_IQT5591_Peru
2007_FMD1240_Peru
1999_AY540431_OBS7687_Bolivia
1995_AY161945_OBS224_Peru
1995_AY161935_HEB4224_Peru
1999_AY161951_OBS7904_Peru
1977_AY161928_1368_Peru
1997_AY540478_OBS5041_Ecuador
1999_AY540434_OBS8026_Bolivia
1983_AY540457_BeH413820_Brazil
1991_AF369685_56205_Nigeria
1987_AF369683_H117505_Nigeria
1987_AF369684_BA55_Nigeria
1969_AF369677_IBAR45244_Nigeria
1970_AF369678_T_Adeoye_Nigeria
1946_U52403_Nigeria46_Nigeria
1953_AH005112_RENDU_Senegal
1983_AF369671_SH28580_BurkinaFaso
1965_AF369689_SH1339_Senegal
1990_AY839632_ArD76320_Senegal90_Senegal
1980_AF369672_ArB17239_CentralAfricanRepublic
1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1964_AF369695_SE7445_Uganda
1958_AF369697_LSF4_RepublicofCongo
1940_AF369693_M112_4_Sudan
1977_U52395_Car77900_CentralAfricanRepublic
1986_AY839631_HB1782_CAR86_AfricanRepublic
1985_AF369673_HB1504_CentralAfricanRepublic
1948_AF369694_A709_4_A2_Uganda
1993_AF369676_BC7914_Kenya
1971_AF369669_14FA_Angola
500
1000
1500
2000
MRCA (calendar years)

National Institute of Allergy and Infectious Diseases

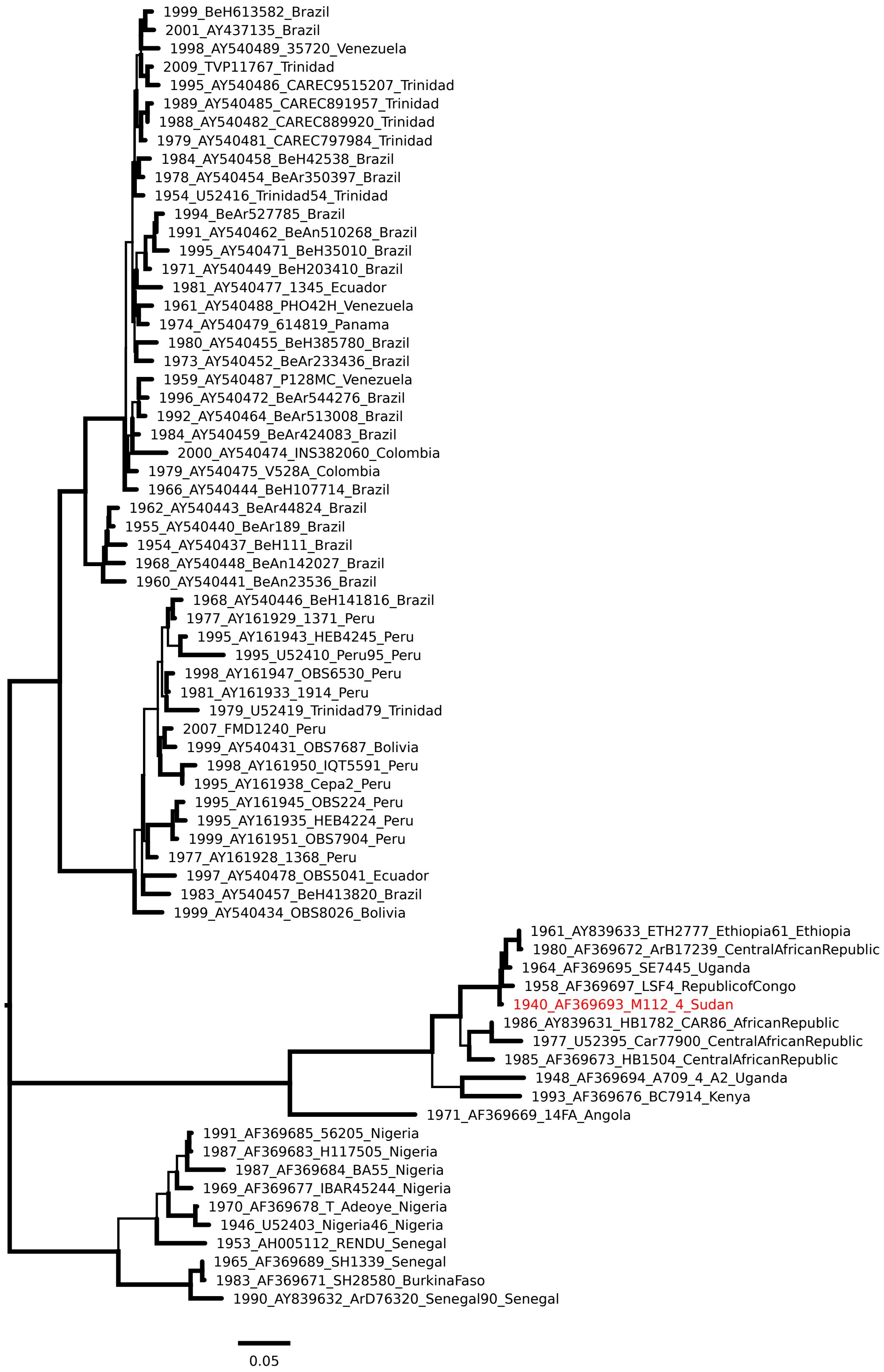
Distance Analysis

The Neighbor-Joining Tree


2009_TVP11767_Trinidad
1995_AY540486_CAREC9515207_Trinidad
1998_AY540489_35720_Venezuela
1999_BeH613582_Brazil
2001_AY437135_Brazil
1989_AY540485_CAREC891957_Trinidad
1988_AY540482_CAREC889920_Trinidad
1979_AY540481_CAREC797984_Trinidad
1984_AY540458_BeH42538_Brazil
1978_AY540454_BeAr350397_Brazil
1954_U52416_Trinidad54_Trinidad
1981_AY540477_1345_Ecuador
1961_AY540488_PHO42H_Venezuela
1974_AY540479_614819_Panama
1980_AY540455_BeH385780_Brazil
1973_AY540452_BeAr233436_Brazil
1959_AY540487_P128MC_Venezuela
1996_AY540472_BeAr544276_Brazil
1992_AY540464_BeAr513008_Brazil
1984_AY540459_BeAr424083_Brazil
2000_AY540474_INS382060_Colombia
1979_AY540475_V528A_Colombia
1966_AY540444_BeH107714_Brazil
1994_BeAr527785_Brazil
1991_AY540462_BeAn510268_Brazil
1995_AY540471_BeH35010_Brazil
1971_AY540449_BeH203410_Brazil
1955_AY540440_BeAr189_Brazil
1954_AY540437_BeH111_Brazil
1968_AY540448_BeAn142027_Brazil
1962_AY540443_BeAr44824_Brazil
1960_AY540441_BeAn23536_Brazil
1981_AY161933_1914_Peru
1979_U52419_Trinidad79_Trinidad
1998_AY161947_OBS6530_Peru
1968_AY540446_BeH141816_Brazil
1977_AY161929_1371_Peru
2007_FMD1240_Peru
1999_AY540431_OBS7687_Bolivia
1998_AY161950_IQT5591_Peru
1995_AY161938_Cepa2_Peru
1995_AY161943_HEB4245_Peru
1995_U52410_Peru95_Peru
1995_AY161945_OBS224_Peru
1995_AY161935_HEB4224_Peru
1999_AY161951_OBS7904_Peru
1977_AY161928_1368_Peru
1999_AY540434_OBS8026_Bolivia
1983_AY540457_BeH413820_Brazil
1997_AY540478_OBS5041_Ecuador
1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1980_AF369672_ArB17239_CentralAfricanRepublic
1940_AF369693_M112_4_Sudan
1958_AF369697_LSF4_RepublicofCongo
1964_AF369695_SE7445_Uganda
1985_AF369673_HB1504_CentralAfricanRepublic
1986_AY839631_HB1782_CAR86_AfricanRepublic
1977_U52395_Car77900_CentralAfricanRepublic
1948_AF369694_A709_4_A2_Uganda
1993_AF369676_BC7914_Kenya
1971_AF369669_14FA_Angola
1991_AF369685_56205_Nigeria
1987_AF369683_H117505_Nigeria
1969_AF369677_IBAR45244_Nigeria
1987_AF369684_BA55_Nigeria
1953_AH005112_RENDU_Senegal
1970_AF369678_T_Adeoye_Nigeria
1946_U52403_Nigeria46_Nigeria
1965_AF369689_SH1339_Senegal
1983_AF369671_SH28580_BurkinaFaso
1990_AY839632_ArD76320_Senegal90_Senegal
0.04

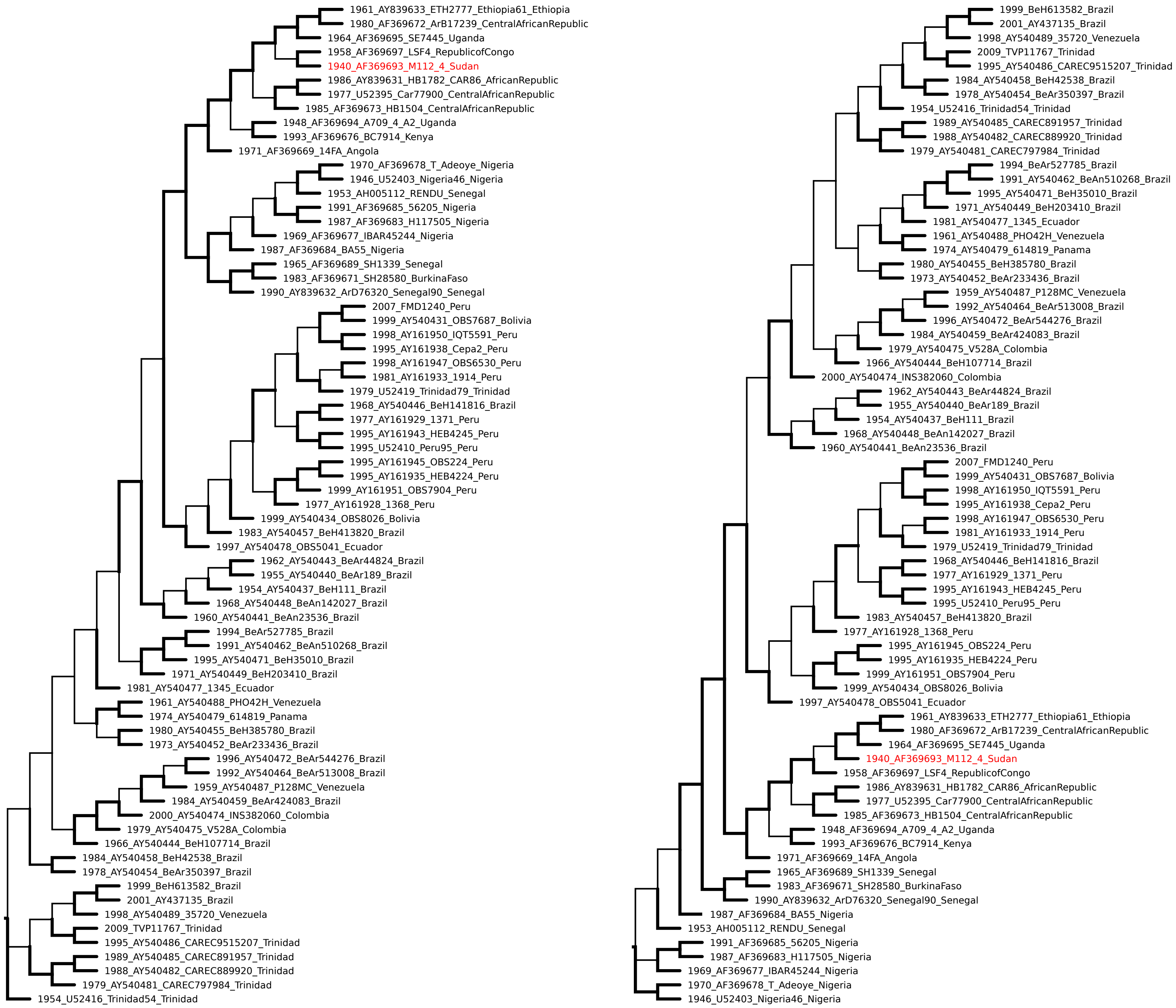
Likelihood Analysis

The Best Tree


1999_BeH613582_Brazil
2001_AY437135_Brazil
1998_AY540489_35720_Venezuela
2009_TVP11767_Trinidad
1995_AY540486_CAREC9515207_Trinidad
1989_AY540485_CAREC891957_Trinidad
1988_AY540482_CAREC889920_Trinidad
1979_AY540481_CAREC797984_Trinidad
1984_AY540458_BeH42538_Brazil
1978_AY540454_BeAr350397_Brazil
1954_U52416_Trinidad54_Trinidad
1994_BeAr527785_Brazil
1991_AY540462_BeAn510268_Brazil
1995_AY540471_BeH35010_Brazil
1971_AY540449_BeH203410_Brazil
1981_AY540477_1345_Ecuador
1961_AY540488_PHO42H_Venezuela
1974_AY540479_614819_Panama
1980_AY540455_BeH385780_Brazil
1973_AY540452_BeAr233436_Brazil
1959_AY540487_P128MC_Venezuela
1996_AY540472_BeAr544276_Brazil
1992_AY540464_BeAr513008_Brazil
1984_AY540459_BeAr424083_Brazil
2000_AY540474_INS382060_Colombia
1979_AY540475_V528A_Colombia
1966_AY540444_BeH107714_Brazil
1962_AY540443_BeAr44824_Brazil
1955_AY540440_BeAr189_Brazil
1954_AY540437_BeH111_Brazil
1968_AY540448_BeAn142027_Brazil
1960_AY540441_BeAn23536_Brazil
1968_AY540446_BeH141816_Brazil
1977_AY161929_1371_Peru
1995_AY161943_HEB4245_Peru
1995_U52410_Peru95_Peru
1998_AY161947_OBS6530_Peru
1981_AY161933_1914_Peru
1979_U52419_Trinidad79_Trinidad
2007_FMD1240_Peru
1999_AY540431_OBS7687_Bolivia
1998_AY161950_IQT5591_Peru
1995_AY161938_Cepa2_Peru
1995_AY161945_OBS224_Peru
1995_AY161935_HEB4224_Peru
1999_AY161951_OBS7904_Peru
1977_AY161928_1368_Peru
1997_AY540478_OBS5041_Ecuador
1983_AY540457_BeH413820_Brazil
1999_AY540434_OBS8026_Bolivia
1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1980_AF369672_ArB17239_CentralAfricanRepublic
1964_AF369695_SE7445_Uganda
1958_AF369697_LSF4_RepublicofCongo
1940_AF369693_M112_4_Sudan
1986_AY839631_HB1782_CAR86_AfricanRepublic
1977_U52395_Car77900_CentralAfricanRepublic
1985_AF369673_HB1504_CentralAfricanRepublic
1948_AF369694_A709_4_A2_Uganda
1993_AF369676_BC7914_Kenya
1971_AF369669_14FA_Angola
1991_AF369685_56205_Nigeria
1987_AF369683_H117505_Nigeria
1987_AF369684_BA55_Nigeria
1969_AF369677_IBAR45244_Nigeria
1970_AF369678_T_Adeoye_Nigeria
1946_U52403_Nigeria46_Nigeria
1953_AH005112_RENDU_Senegal
1965_AF369689_SH1339_Senegal
1983_AF369671_SH28580_BurkinaFaso
1990_AY839632_ArD76320_Senegal90_Senegal
0.05

Parsimony Analysis

Two Equally Parsimonious Trees


1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1980_AF369672_ArB17239_CentralAfricanRepublic
1964_AF369695_SE7445_Uganda
1958_AF369697_LSF4_RepublicofCongo
1940_AF369693_M112_4_Sudan
1986_AY839631_HB1782_CAR86_AfricanRepublic
1977_U52395_Car77900_CentralAfricanRepublic
1985_AF369673_HB1504_CentralAfricanRepublic
1948_AF369694_A709_4_A2_Uganda
1993_AF369676_BC7914_Kenya
1971_AF369669_14FA_Angola
1970_AF369678_T_Adeoye_Nigeria
1946_U52403_Nigeria46_Nigeria
1953_AH005112_RENDU_Senegal
1991_AF369685_56205_Nigeria
1987_AF369683_H117505_Nigeria
1969_AF369677_IBAR45244_Nigeria
1987_AF369684_BA55_Nigeria
1965_AF369689_SH1339_Senegal
1983_AF369671_SH28580_BurkinaFaso
1990_AY839632_ArD76320_Senegal90_Senegal
2007_FMD1240_Peru
1999_AY540431_OBS7687_Bolivia
1998_AY161950_IQT5591_Peru
1995_AY161938_Cepa2_Peru
1998_AY161947_OBS6530_Peru
1981_AY161933_1914_Peru
1979_U52419_Trinidad79_Trinidad
1968_AY540446_BeH141816_Brazil
1977_AY161929_1371_Peru
1995_AY161943_HEB4245_Peru
1995_U52410_Peru95_Peru
1995_AY161945_OBS224_Peru
1995_AY161935_HEB4224_Peru
1999_AY161951_OBS7904_Peru
1977_AY161928_1368_Peru
1999_AY540434_OBS8026_Bolivia
1983_AY540457_BeH413820_Brazil
1997_AY540478_OBS5041_Ecuador
1962_AY540443_BeAr44824_Brazil
1955_AY540440_BeAr189_Brazil
1954_AY540437_BeH111_Brazil
1968_AY540448_BeAn142027_Brazil
1960_AY540441_BeAn23536_Brazil
1994_BeAr527785_Brazil
1991_AY540462_BeAn510268_Brazil
1995_AY540471_BeH35010_Brazil
1971_AY540449_BeH203410_Brazil
1981_AY540477_1345_Ecuador
1961_AY540488_PHO42H_Venezuela
1974_AY540479_614819_Panama
1980_AY540455_BeH385780_Brazil
1973_AY540452_BeAr233436_Brazil
1996_AY540472_BeAr544276_Brazil
1992_AY540464_BeAr513008_Brazil
1959_AY540487_P128MC_Venezuela
1984_AY540459_BeAr424083_Brazil
2000_AY540474_INS382060_Colombia
1979_AY540475_V528A_Colombia
1966_AY540444_BeH107714_Brazil
1984_AY540458_BeH42538_Brazil
1978_AY540454_BeAr350397_Brazil
1999_BeH613582_Brazil
2001_AY437135_Brazil
1998_AY540489_35720_Venezuela
2009_TVP11767_Trinidad
1995_AY540486_CAREC9515207_Trinidad
1989_AY540485_CAREC891957_Trinidad
1988_AY540482_CAREC889920_Trinidad
1979_AY540481_CAREC797984_Trinidad
1954_U52416_Trinidad54_Trinidad
1999_BeH613582_Brazil
2001_AY437135_Brazil
1998_AY540489_35720_Venezuela
2009_TVP11767_Trinidad
1995_AY540486_CAREC9515207_Trinidad
1984_AY540458_BeH42538_Brazil
1978_AY540454_BeAr350397_Brazil
1954_U52416_Trinidad54_Trinidad
1989_AY540485_CAREC891957_Trinidad
1988_AY540482_CAREC889920_Trinidad
1979_AY540481_CAREC797984_Trinidad
1994_BeAr527785_Brazil
1991_AY540462_BeAn510268_Brazil
1995_AY540471_BeH35010_Brazil
1971_AY540449_BeH203410_Brazil
1981_AY540477_1345_Ecuador
1961_AY540488_PHO42H_Venezuela
1974_AY540479_614819_Panama
1980_AY540455_BeH385780_Brazil
1973_AY540452_BeAr233436_Brazil
1959_AY540487_P128MC_Venezuela
1992_AY540464_BeAr513008_Brazil
1996_AY540472_BeAr544276_Brazil
1984_AY540459_BeAr424083_Brazil
1979_AY540475_V528A_Colombia
1966_AY540444_BeH107714_Brazil
2000_AY540474_INS382060_Colombia
1962_AY540443_BeAr44824_Brazil
1955_AY540440_BeAr189_Brazil
1954_AY540437_BeH111_Brazil
1968_AY540448_BeAn142027_Brazil
1960_AY540441_BeAn23536_Brazil
2007_FMD1240_Peru
1999_AY540431_OBS7687_Bolivia
1998_AY161950_IQT5591_Peru
1995_AY161938_Cepa2_Peru
1998_AY161947_OBS6530_Peru
1981_AY161933_1914_Peru
1979_U52419_Trinidad79_Trinidad
1968_AY540446_BeH141816_Brazil
1977_AY161929_1371_Peru
1995_AY161943_HEB4245_Peru
1995_U52410_Peru95_Peru
1983_AY540457_BeH413820_Brazil
1977_AY161928_1368_Peru
1995_AY161945_OBS224_Peru
1995_AY161935_HEB4224_Peru
1999_AY161951_OBS7904_Peru
1999_AY540434_OBS8026_Bolivia
1997_AY540478_OBS5041_Ecuador
1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1980_AF369672_ArB17239_CentralAfricanRepublic
1964_AF369695_SE7445_Uganda
1940_AF369693_M112_4_Sudan
1958_AF369697_LSF4_RepublicofCongo
1986_AY839631_HB1782_CAR86_AfricanRepublic
1977_U52395_Car77900_CentralAfricanRepublic
1985_AF369673_HB1504_CentralAfricanRepublic
1948_AF369694_A709_4_A2_Uganda
1993_AF369676_BC7914_Kenya
1971_AF369669_14FA_Angola
1965_AF369689_SH1339_Senegal
1983_AF369671_SH28580_BurkinaFaso
1990_AY839632_ArD76320_Senegal90_Senegal
1987_AF369684_BA55_Nigeria
1953_AH005112_RENDU_Senegal
1991_AF369685_56205_Nigeria
1987_AF369683_H117505_Nigeria
1969_AF369677_IBAR45244_Nigeria
1970_AF369678_T_Adeoye_Nigeria
1946_U52403_Nigeria46_Nigeria

National Institute of Allergy and Infectious Diseases

Making a Tree Figure: FigTree

http://tree.bio.ed.ac.uk/software/figtree/

In Conclusion

Where have we been? What have we done?

- Why are virus' biological sequences special?
- Calculating a multiple sequence alignment.
- Calculating 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

Questions?

Email us!

bioinformatics@niaid.nih.gov

NIH National Institute of Allergy and Infectious Diseases

NIAID