Introduction, Sequence Alignment, and BLAST

Kurt Wollenberg, PhD

Phylogenetics Specialist

Bioinformatics and Computational Biosciences Branch

Office of Cyber Infrastructure and Computational Biology

February 2024

We Are BCBB!

- Bioinformatics Software Developers
- Computational Biologists
- Project Management & Analysis Professionals

NIH National Institute of Allergy and Infectious Diseases

NIAID

Contact BCBB…

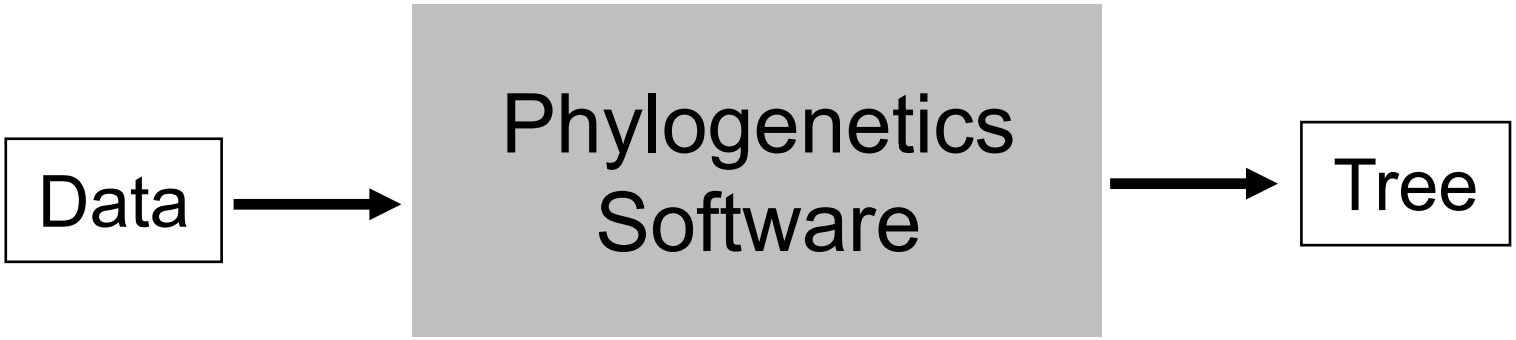
- NIH Users: Access a menu of BCBB services on the NIAID Intranet:
  - http://bioinformatics.niaid.nih.gov/
  - Requires the use of your NIH login credentials
- Outside of NIH –
  - search “BCBB” on the NIAID Public Internet Page: www.niaid.nih.gov

– or –

- Email us at:
  - bioinformatics@niaid.nih.gov

The Goal


Data
Phylogenetics Software
Tree

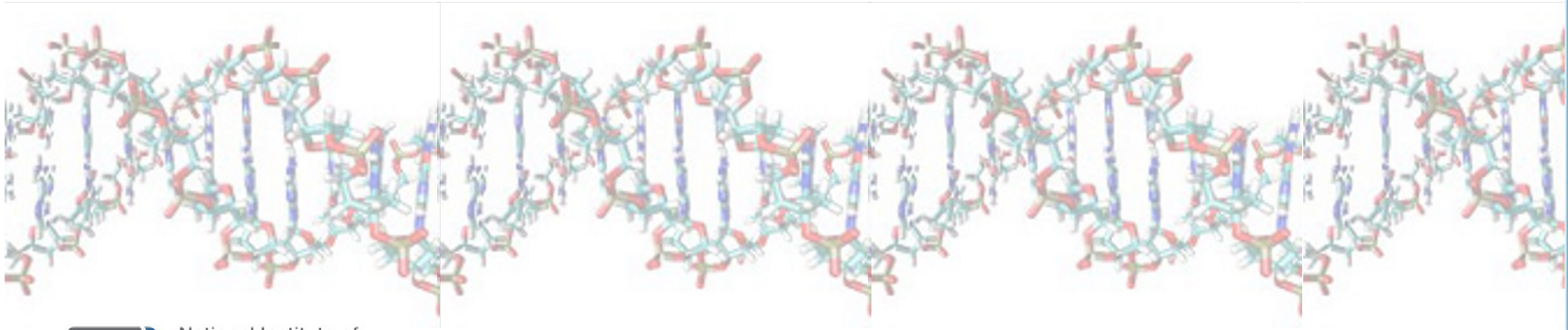

Data
Phylogenetics Software
Tree

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



NIH National Institute of Allergy and Infectious Diseases

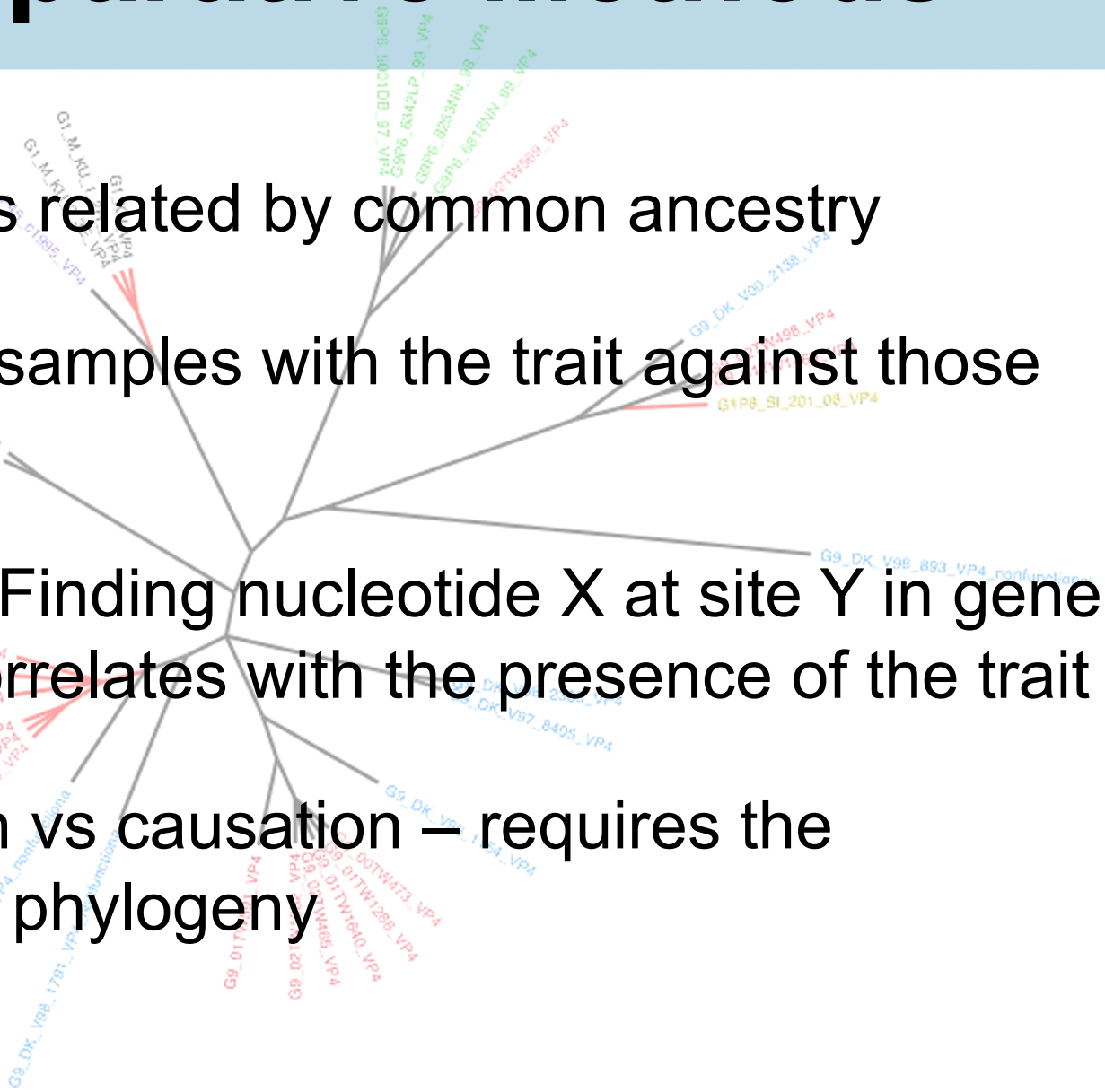
NIAID

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 – requires the underlying phylogeny



NIH National Institute of Allergy and Infectious Diseases

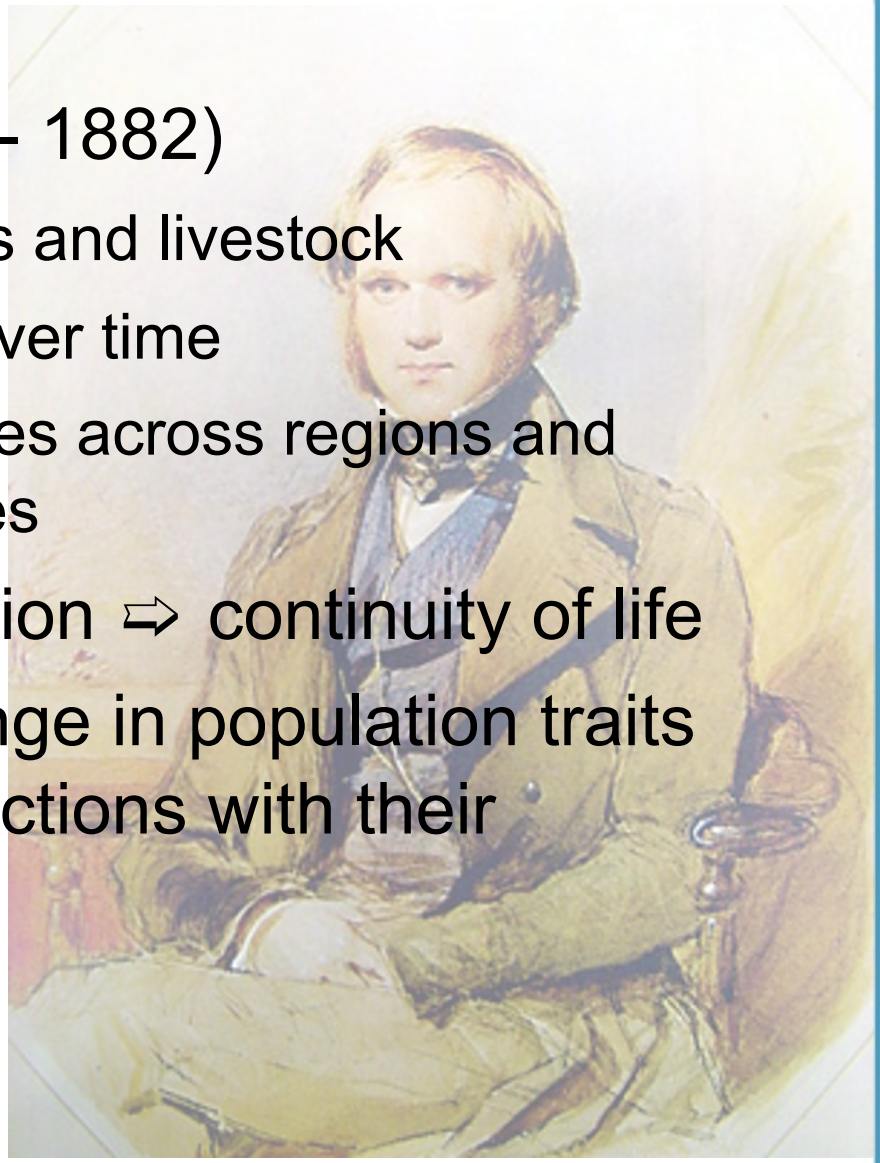
NIAID

Hierarchy of Life

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

Common Ancestry

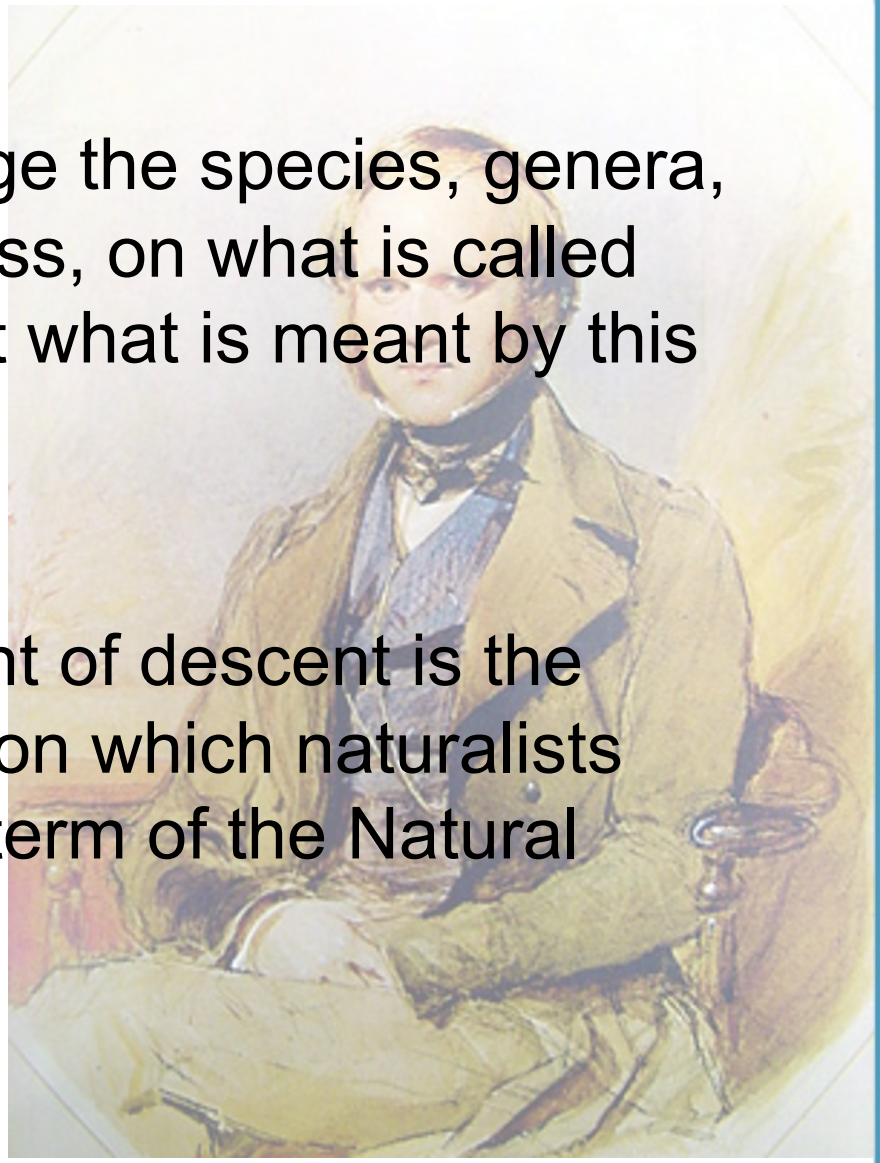
- Charles Darwin (1809 - 1882)
  - Artificial selection: crops and livestock
  - Fossil record: change over time
  - Biogeography: similarities across regions and differences within locales
- Descent with modification ⇨ continuity of life
- Natural selection: change in population traits due to individual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



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



Artificial Selection

Artificial Selection

National Institute of Allergy and Infectious Diseases

NIAID

Today…

Pairwise sequence alignment

- How does it work?

BLAST

- How does it work?
- The many flavors of BLAST

Multiple Sequence Alignment

- How does it work?
- Inspect and correct your MSA

BLAST and Sequence Alignment Demo

PAIRWISE ALIGNMENT

and **BLAST**: **B**asic **L**ocal **A**lignment **S**earch **T**ool

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

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, Geneious
  - Alignment of genomic reads
    - BWA, Bowtie, SOAP, minimap, canu

PAIRWISE ALIGNMENT

- Single locus

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

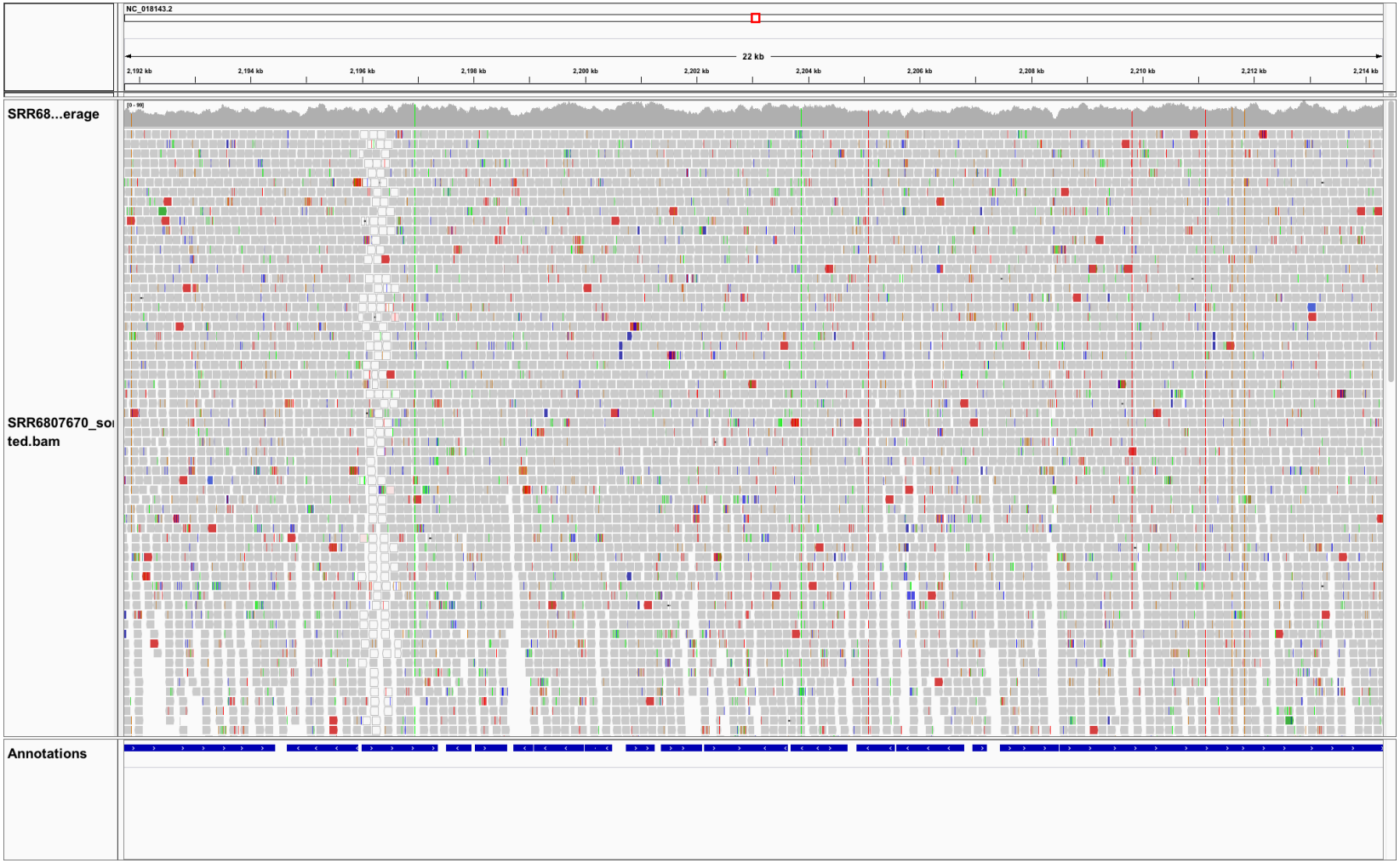
PAIRWISE ALIGNMENT

- Overlapping contigs

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

PAIRWISE ALIGNMENT

- Genomic reads (short)



HOMOLOGY vs. ANALOGY

common ancestry

convergence

PAIRWISE ALIGNMENT

Pairing of sites based on an assessment of homology

Homology assessed using Substitution Matrices

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
```

PAIRWISE ALIGNMENT

Substitution Matrices

- Derived mathematically
- Derived from data

"A substitution matrix (even one derived by arbitrarily assigning probabilities to pairs) is a statement of the probability of observing these pairs in real alignment."

PAIRWISE ALIGNMENT

DNA Substitution Matrices

- Single parameter - Jukes-Cantor
  - Equal base frequencies
  - Uniform rates of change
- Two parameter - Kimura
  - Equal base probabilities
  - Two rates of change

PAIRWISE ALIGNMENT

DNA Substitution Matrices

- More (5) parameters - HKY
  - Unequal base frequencies
  - Two rates of change
- Fully (9) parameterized - GTR
  - Unequal base probabilities
  - Six rates of change

PAIRWISE ALIGNMENT

Jukes-Cantor Substitution Probabilities

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-4\mu t} & i = j \\ \frac{1}{4} - \frac{1}{4} e^{-4\mu t} & i \neq j \end{cases}$$

PAIRWISE ALIGNMENT

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

PAIRWISE ALIGNMENT

Kimura Two-Parameter Substitution Model

If the probability of transitions (A ⇔ G, C ⇔ T) is different from the probability of transversions (A ⇔T, G ⇔T, A ⇔C, G ⇔C), then there are two relative rate parameters expressed as the transition/transversion rate ratio κ

PAIRWISE ALIGNMENT

Kimura Two-Parameter Substitution Probabilities

$$P_{ij}(t) = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-4\mu t} & i \neq j, transversion \\ \frac{1}{4} + \frac{1}{4}e^{-4\mu t} - \frac{1}{2}e^{-2(\kappa+1)\mu t} & i \neq j, transition \\ \frac{1}{4} + \frac{1}{4}e^{-4\mu t} + \frac{1}{2}e^{-2(\kappa+1)\mu t} & i = j \end{cases}$$

PAIRWISE ALIGNMENT

Kimura Two-Parameter Substitution Probabilities

$\mu t = 0.25\ \kappa = 2.0$

	A	C	G	T
A	0.4535	0.1580	0.2304	0.1580
C	0.1580	0.4535	0.1580	0.2304
G	0.2304	0.1580	0.4535	0.1580
T	0.1580	0.2304	0.1580	0.4535

PAIRWISE ALIGNMENT

HKY Substitution Probabilities

$$P_{ij}(t) = \begin{cases} \pi_j + \pi_j\left(\frac{1}{\Pi_j} - 1\right)e^{-\mu t} + \left(\frac{\Pi_j - \pi_j}{\Pi_j}\right)e^{-\mu t A} & (i = j) \\ \pi_j + \pi_j\left(\frac{1}{\Pi_j} - 1\right)e^{-\mu t} + \left(\frac{\pi_j}{\Pi_j}\right)e^{-\mu t A} & (i \neq j, transition) \\ \pi_j\left(1 - e^{-\mu t}\right) & (i \neq j, transversion) \end{cases}$$

PAIRWISE ALIGNMENT

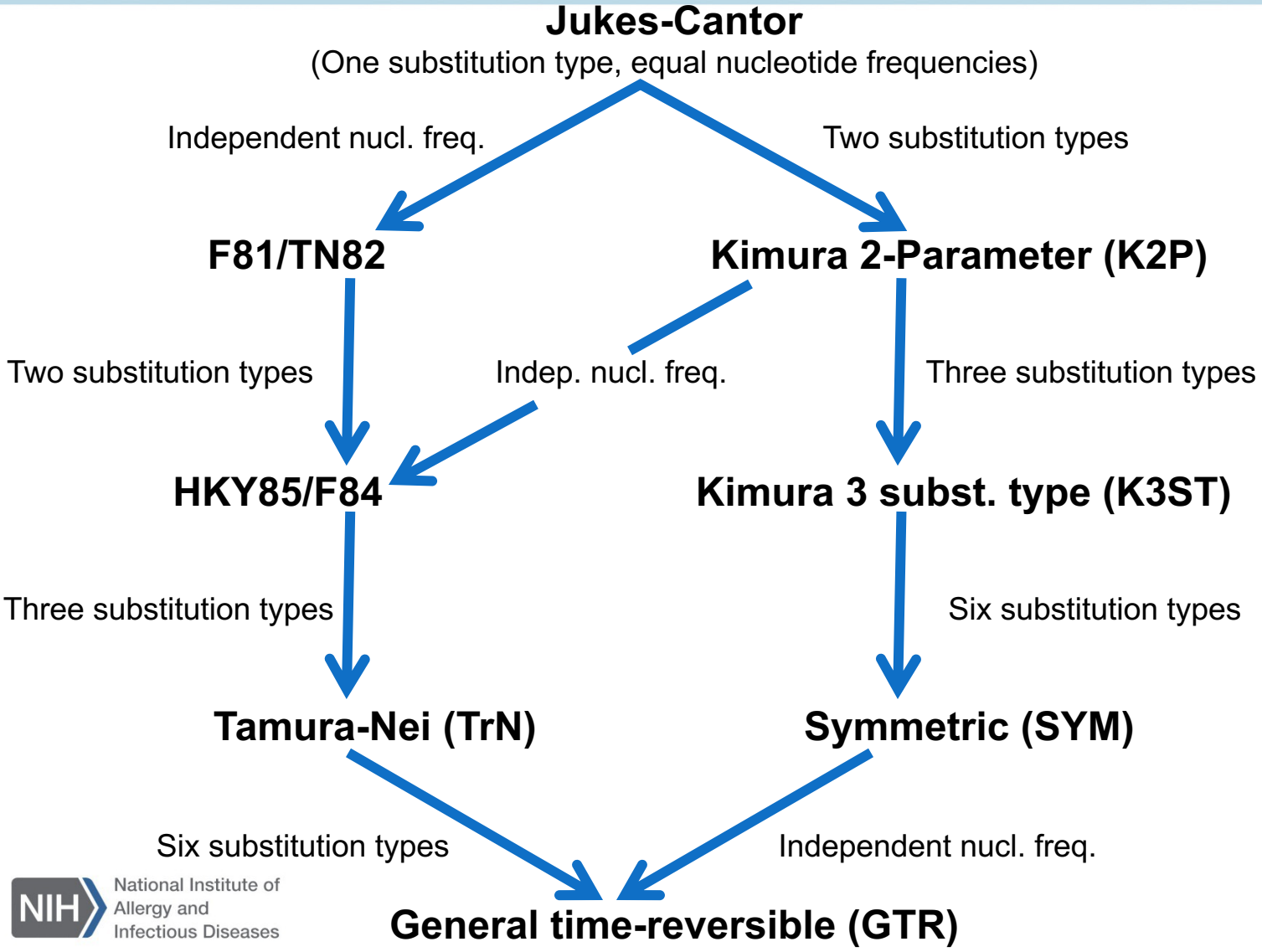
HKY Substitution Probabilities

$$\Pi_j = \pi_A + \pi_G \quad \text{if } j \text{ is a purine}$$

$$\Pi_j = \pi_C + \pi_T \quad \text{if } j \text{ is a pyrimidine}$$

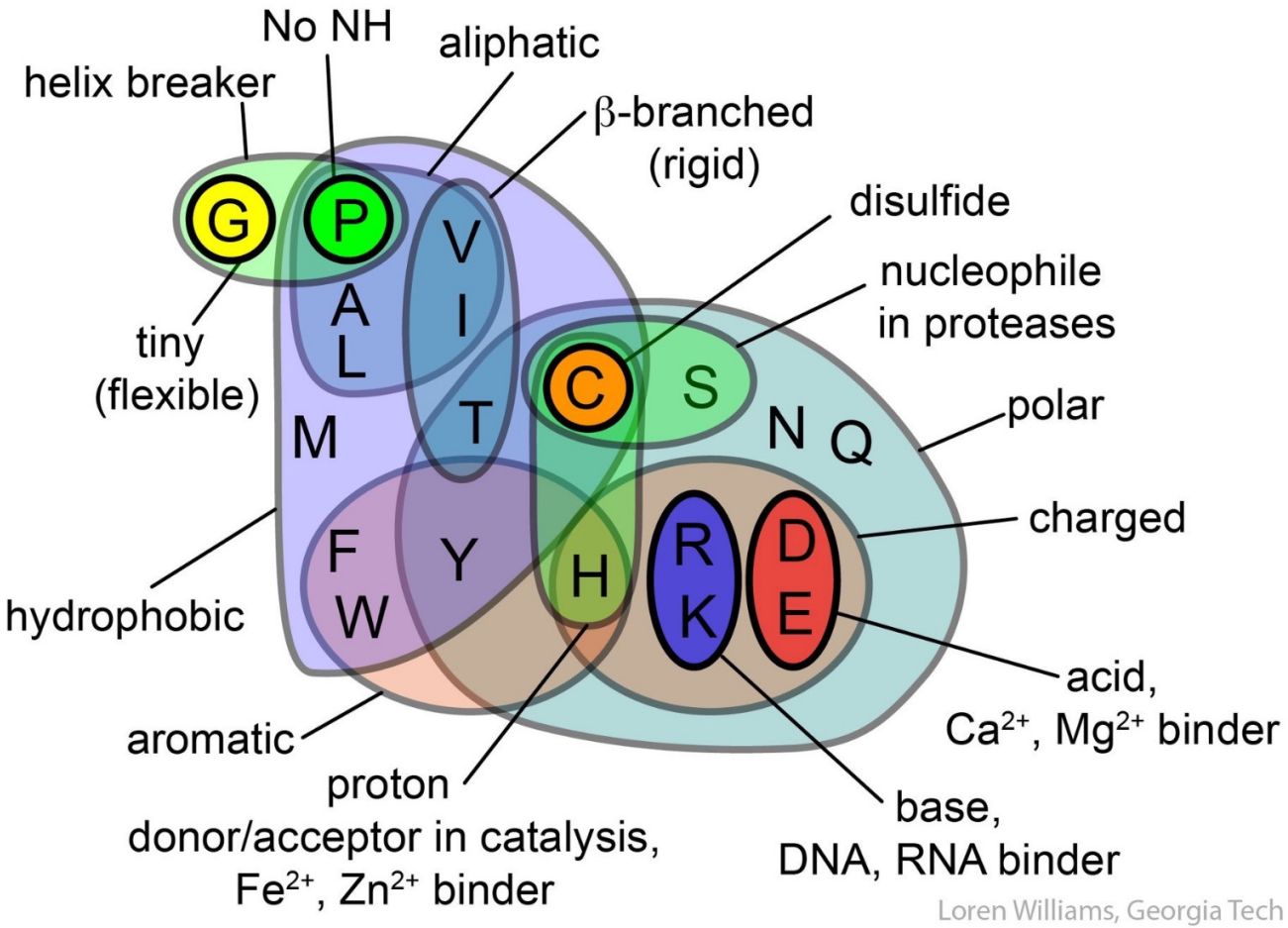
$$A = 1 + \Pi_j(\kappa - 1)$$

Substitution Models



PAIRWISE ALIGNMENT

Protein Score Matrices
Similarity of Amino Acids



Loren Williams, Georgia Tech

NIAID

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 Score (Substitution) Matrices

The log-odds ratio:
$s(a,b) = \log(p_{ab}/q_a q_b)$

q_a = frequency of residue a in the data

p_{ab} = probability that residues a and b have been derived from a common ancestor

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

PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	−2	1	3																	
P	−3	1	0	6																
A	−2	1	1	1	2															
G	−3	1	0	−1	1	5														
N	−4	1	0	−1	0	0	2													
D	−5	0	0	−1	0	1	2	4												
E	−5	0	0	−1	0	0	1	3	4											
Q	−5	−1	−1	0	0	−1	1	2	2	4										
H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−6	−3	2	−3	−4	−5	−2	−6	0	0	17

National Institute of Allergy and Infectious Diseases

NIAID

Protein Substitution Matrices

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
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

NIH National Institute of Allergy and Infectious Diseases

NIAID

Protein Substitution Matrices

W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−6	−3	2	−3	−4	−5	−2	−6	0	0	17	*P250*
W	−2	−3	2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11	*B62*
	C	*S*	*T*	*P*	*A*	*G*	*N*	*D*	*E*	*Q*	*H*	*R*	*K*	*M*	*I*	*L*	*V*	*F*	*Y*	*W*	

BLAST and Sequence Alignment

How do two sequences get “aligned”?

- Global alignment (Needleman-Wunsch)
  - Assign homology across the entire sequence
  - Clustal
- Local alignment (Smith-Waterman)
  - Assign homology for subsequences
  - MUSCLE and BLAST
  - MAFFT is also a local algorithm
  - Good for aligning very divergent sequences

SEQUENCE ALIGNMENT

HEAGAWGHEE ⇔ PAWHEAE

Build a matrix of score values for all site pairs

PAM250

	H	E	A	G	A	W	G	H	E	E
P	0	-1	1	0	1	-6	0	0	-1	-1
A	-1	0	2	1	2	-6	1	-1	0	0
W	-3	-7	-6	-7	-6	17	-7	-3	-7	-7
H	6	1	-1	-2	-1	-3	-2	6	1	1
E	1	4	0	0	0	-7	0	1	4	4
A	-1	0	2	1	2	-6	1	-1	0	0
E	1	4	0	0	0	-7	0	1	4	4

BLOSUM62

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	4	0	4	-3	0	-2	-1	-1
W	-2	-3	-3	-2	-3	11	-2	-2	-3	-3
H	8	0	-2	-2	-2	-2	-2	8	0	0
E	0	5	-1	-2	-1	-3	-2	0	5	5
A	-2	-1	4	0	4	-3	0	-2	-1	-1
E	0	5	-1	-2	-1	-3	-2	0	5	5

SEQUENCE ALIGNMENT

What about gaps?

- Score penalty for opening
- Score penalty for extending

Penalties are log probabilities of a gap of a specific length

SEQUENCE ALIGNMENT

Standard gap costs

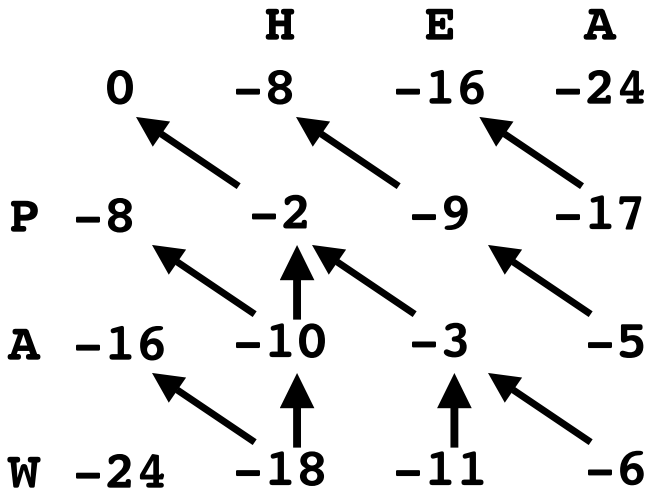
Substitution Matrix	Gap Costs (Open, Extend)
PAM30	(9,1)
PAM70	(10,1)
BLOSUM80	(10,1)
BLOSUM62	(10,1)
BLOSUM45	(15,2)

SEQUENCE ALIGNMENT

Dynamic Programming:
Calculate a matrix of alignment scores

BLOSUM62

	H	E	A
P	-2	-1	-1
A	-2	-1	4
W	-2	-3	-3



SEQUENCE ALIGNMENT

Dynamic Programming

1) Calculate a full matrix
2) Traceback to get the Global Alignment

		H	E	A	G	A	W	G	H	E	E
	0	**-8**	**-16**	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	**-17**	**-25**	-33	-41	-49	-57	-65	-73
A	-16	-10	-3	-5	-13	**-21**	-29	-37	-45	-53	-61
W	-24	-18	-11	-6	-7	-15	**-10**	-18	-26	-34	-41
H	-32	-16	-18	-13	-8	-9	-17	**-12**	-10	-18	-26
E	-40	-24	-11	-19	-15	-9	-12	-19	**-12**	-5	-13
A	-48	-32	-19	-7	-15	-11	-12	-12	-20	**-13**	-6
E	-58	-40	-27	-15	-9	-16	-14	-14	-12	-15	**-8**

H E A G A W G H E E

\- - P - A W H E A E

SEQUENCE ALIGNMENT

Local Alignment

- Alignment of subsequences
- Good for aligning very divergent sequences

Score Calculation

- Minimum score is zero
- Traceback begins at the highest score
- Score = 0 ➔ End of subsequence

SEQUENCE ALIGNMENT

Local Alignment

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	4	0	**4**	0	0	0	0	0
W	0	0	0	0	0	0	**15**	**7**	0	0	0
H	0	**8**	0	0	0	0	7	13	**15**	7	0
E	0	0	**13**	5	0	0	0	5	13	**20**	12
A	0	0	5	**17**	9	4	0	0	5	12	17
E	0	0	5	9	15	8	0	0	0	10	17

A W G H E

A W - H E

Overlap Match

H E A G A W G H E e

p A W - H E a e

Repeat Match

H E A G A W G H E e

p a w H E A e

p A W - H E a e

SEQUENCE ALIGNMENT

Scoring alignments and expect values

Score := Value in the dynamic programming matrix where the traceback began.

Expect (**E**) value := Number of matches expected due to chance, with a score greater than **S**, based on a stochastic sequence model.

P value := Probability of finding at least one match with score ≥ **S**

$$\mathbf{P} = 1\text{-}e^{-\mathbf{E}(\mathbf{S})}$$

BLAST
(Basic Local Alignment Search Tool)

How does BLAST work?

- Create a list of query sequence "words"
  - Word lengths: 11 nucleotides, 3 amino acids
- Create a list of neighborhood words
  - Similar to query words and above a score threshold
- Search for matches in the database
- Extend matches
  - Below threshold? Discard!
  - Above threshold? Keep it!
- Format and output maximally extended matches

BLAST
(Basic Local Alignment Search Tool)

How does BLAST work?

How does BLAST evaluate matches?

It uses (local) alignment scores

BLAST

The Many Flavors of BLAST

- BLASTn and BLASTp
- short, nearly-exact match BLAST
- Translated BLAST
  - BLASTx nt → aa ⇨ protein db
  - tBLASTn aa ⇨ protein db ← DNA db
  - tBLASTx nt → aa ⇨ protein db ← DNA db
- PSI-BLAST (Position-Specific Iterated BLAST)
- bl2seq

BLAST

short, nearly-exact match BLAST

- Increase Expect threshold
- Reduce word size (7 for nt, 2 for aa)
- Turn off low complexity filter
- Protein: Use a more stringent substitution matrix

BLAST

PSI-BLAST
(Position-Specific Iterated BLAST)

- Perform initial BLASTp search
- Generate a Position Specific Score Matrix (PSSM) from results
- BLASTp using the PSSM
- Iterate until no new sequences are found
- Convergence

BLAST

Position Specific Score Matrix

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

Next Iteration

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

Next Iteration

NIAID

BLAST

Sequence Profile

`[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]`

[ ] = Any of the residues within the brackets

\- = spacer separating sites in the profile

x = Any residue

x(a,b) = Any residues a to b in length

VGERGLEEDKRKRSAWMQC

MGETALRRRKKEDEERTANVYT

FGEAAMPGGPHQSRSAFAWV

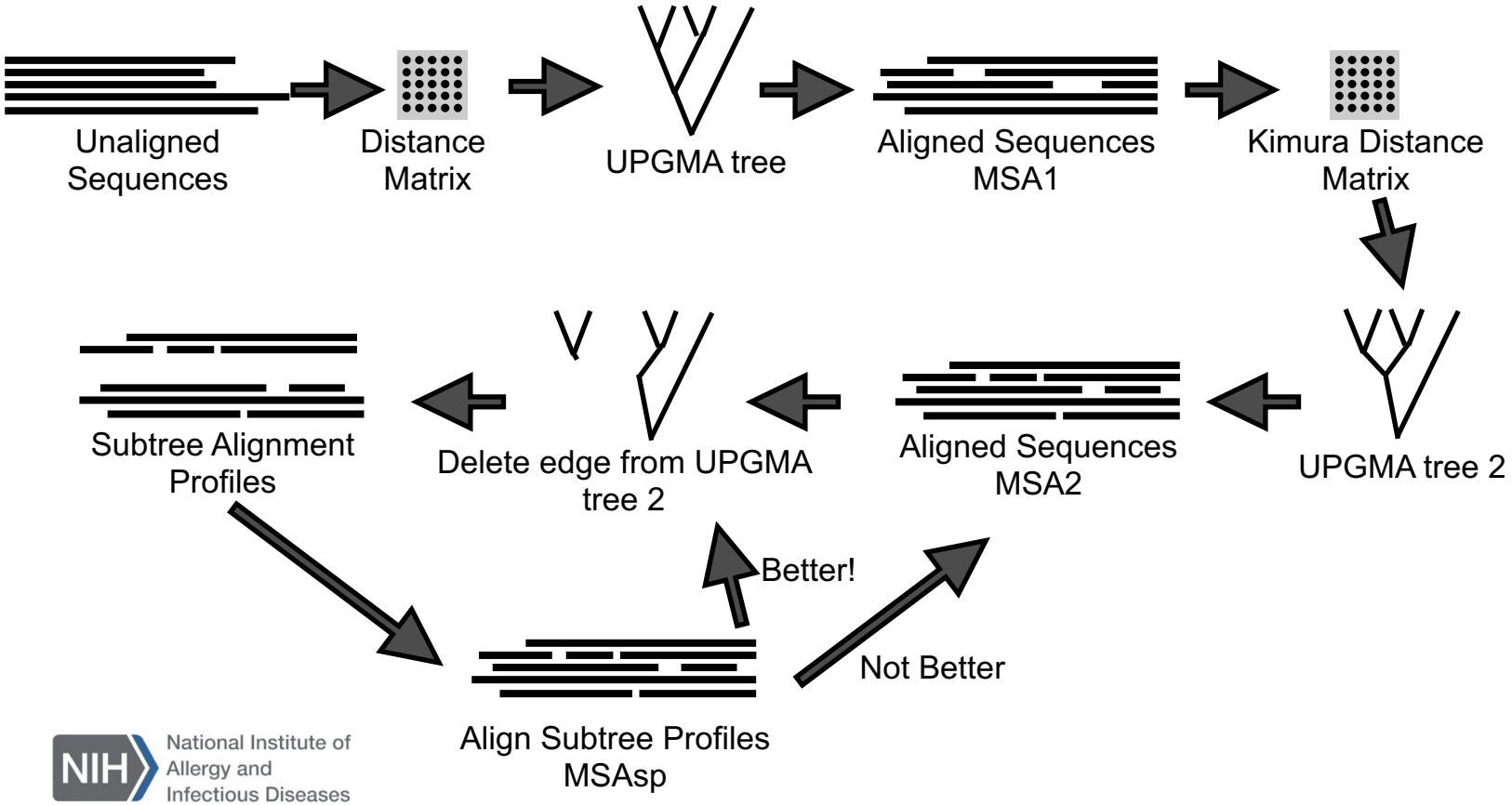
BLAST

Access to BLAST

- NCBI web page
  - https://blast.ncbi.nlm.nih.gov/Blast.cgi
- Your own computer
- NIAID HPC cluster – Locus/Skyline
- NIH HPC cluster - Biowulf

Multiple Sequence Alignment

Multiple Sequence Alignment
The Progressive Alignment Algorithm



From RC Edgar 2004, NAR 32: 1792-1797

Multiple Sequence Alignment

Programs

- Clustal
  - Your own computer
  - Web Server
  - HPC clusters (Biowulf, Locus)
- MUSCLE
  - Your own computer
  - Web Server
  - HPC clusters
- MAFFT
  - Web Server
  - HPC clusters

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

Multiple Sequence Alignment Editors

Commercial Software

- Geneious
- MacVector
- MegAlign (Lasergene)

Public Domain Software

- MEGA
- AliView
- GeneDoc
- BioEdit

Web Resources

ClustalW2

http://www.clustal.org/

Muscle

http://www.drive5.com/muscle/index.htm

MAFFT

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

AliView

https://github.com/AliView/AliView

GeneDoc

http://nrbsc.org/gfx/genedoc - Last update 2007

UGENE

http://ugene.net/

MEGA

https://megasoftware.net

DEMO – Sequence alignment

Multiple sequence alignment using MEGA11

1. Under “Align” choose “Do BLAST search”
2. Use query sequence NM_000575
3. In the “Organism” field limit results to Mammals
4. Under “Algorithm Parameters” change “Max target sequences” to 250
5. Run the search
6. Unselect “All” results and choose specific sequences
7. Change view to “Genbank”
8. In Genbank view, change format to “FASTA(text)”
9. Add to Alignment (at top of MEGA11 window)

DEMO – Sequence alignment

Multiple sequence alignment using MEGA11

1. Under “Edit” menu choose “Select All”
2. Click on the icon to run a Muscle alignment
3. These sequences include more than the coding sequence, so let’s edit them
4. Search for motif ATGGCCAAA
5. Select and delete the block of sequence before the ATG
6. Search for motif TAGGTCT
7. Select and delete the block of sequence after TAG

DEMO – Sequence alignment

Multiple sequence alignment using MEGA11

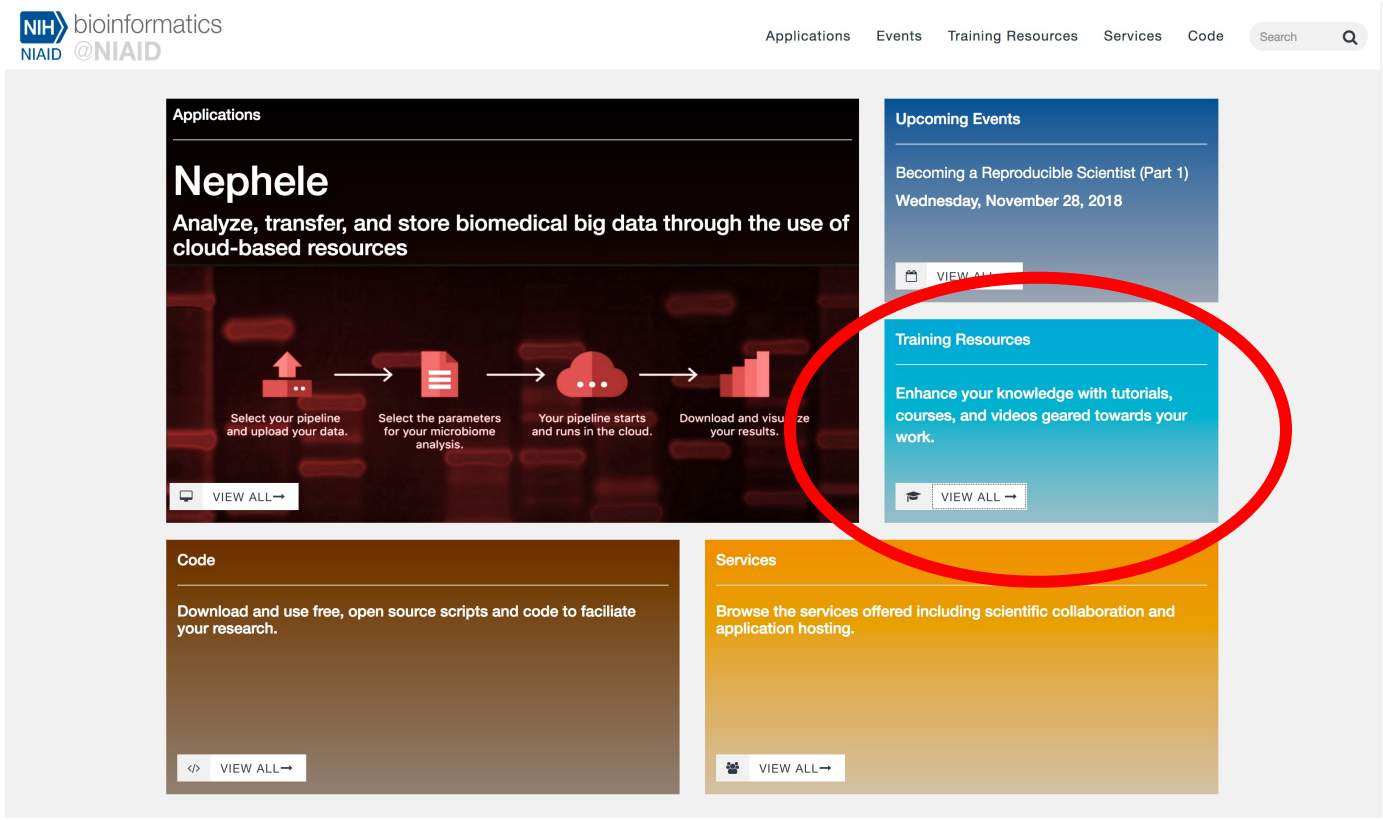
1. Click on “Translated Protein Sequences”
2. Accept the standard code
3. Look for “?” sites
4. Select site this site and click on “DNA Sequences”
5. Correct the split ATG codon
6. Continue and correct remaining misaligned codons
7. From the “Data” menu export the alignment as a fasta formatted file
8. To make the alignment the active data for further analysis, choose “Phylogenetic Analysis” from the “Data” menu

Recapitulation

- BLAST and sequence alignment are two applications of the same process.
- Sequence alignment can be global or local.
- Alignment scores are cumulative, so maximum value will depend on sequence length
- Alignment algorithms are not perfect, and generally do not respect the reading frame, so always inspect the alignment if possible.

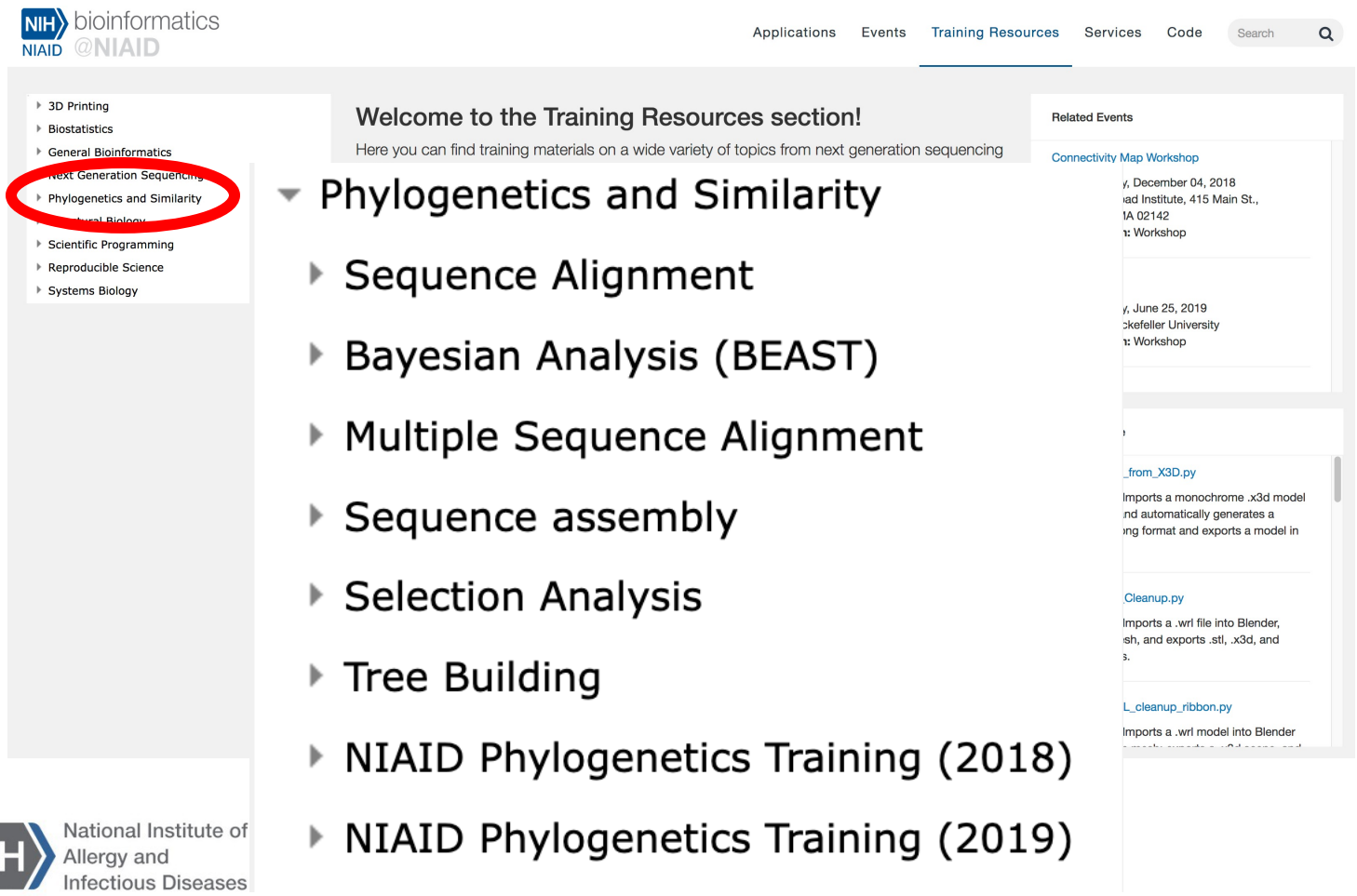
Seminar Follow-Up Site

https://bioinformatics.niaid.nih.gov



Seminar Follow-Up Site

https://bioinformatics.niaid.nih.gov



NIH National Institute of Allergy and Infectious Diseases

NIAID

Questions?

Email us!

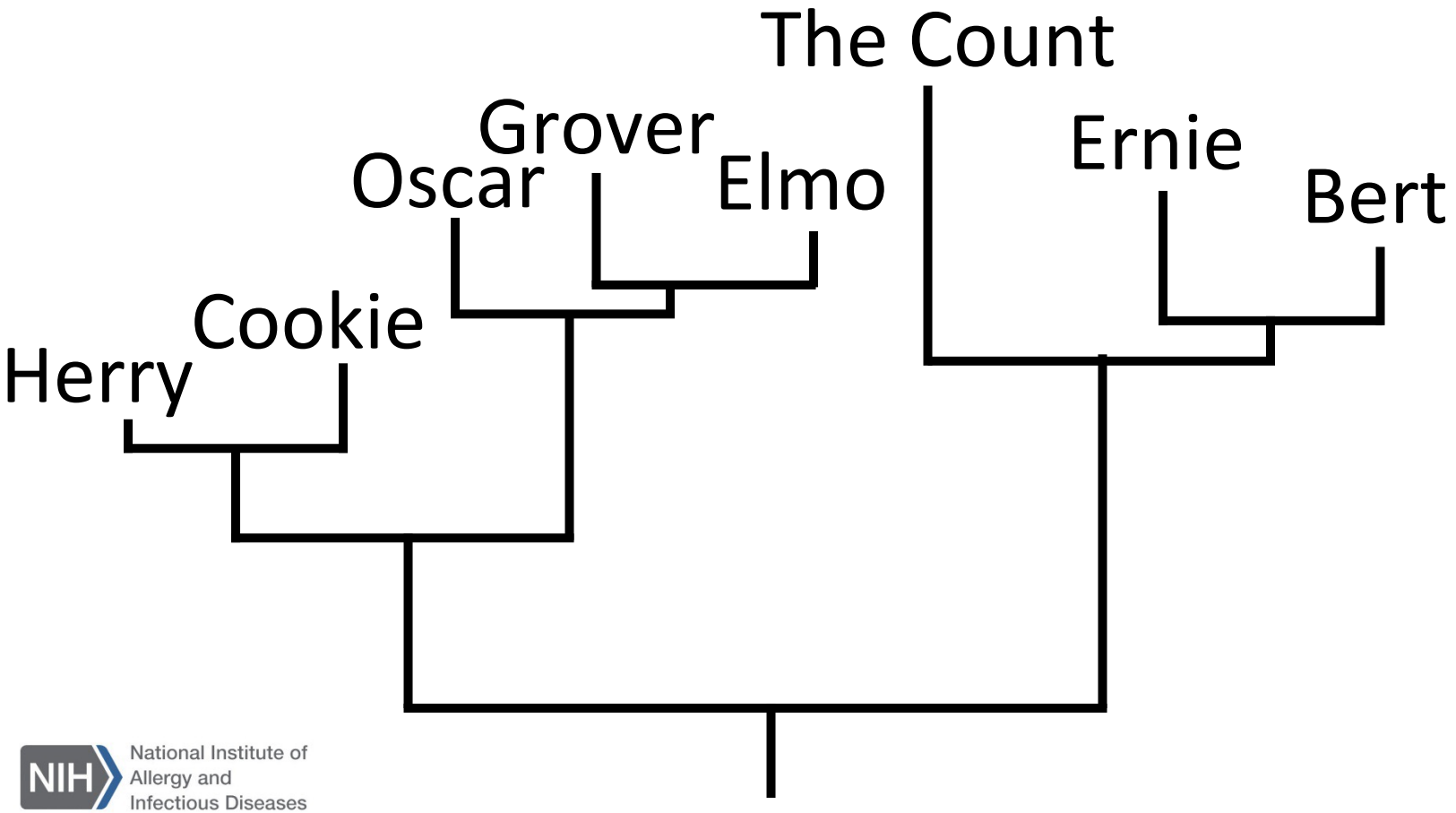
bioinformatics@niaid.nih.gov

NIH National Institute of Allergy and Infectious Diseases

NIAID

Next Lecture

Wednesday, 06 March



NIH National Institute of Allergy and Infectious Diseases

NIAID