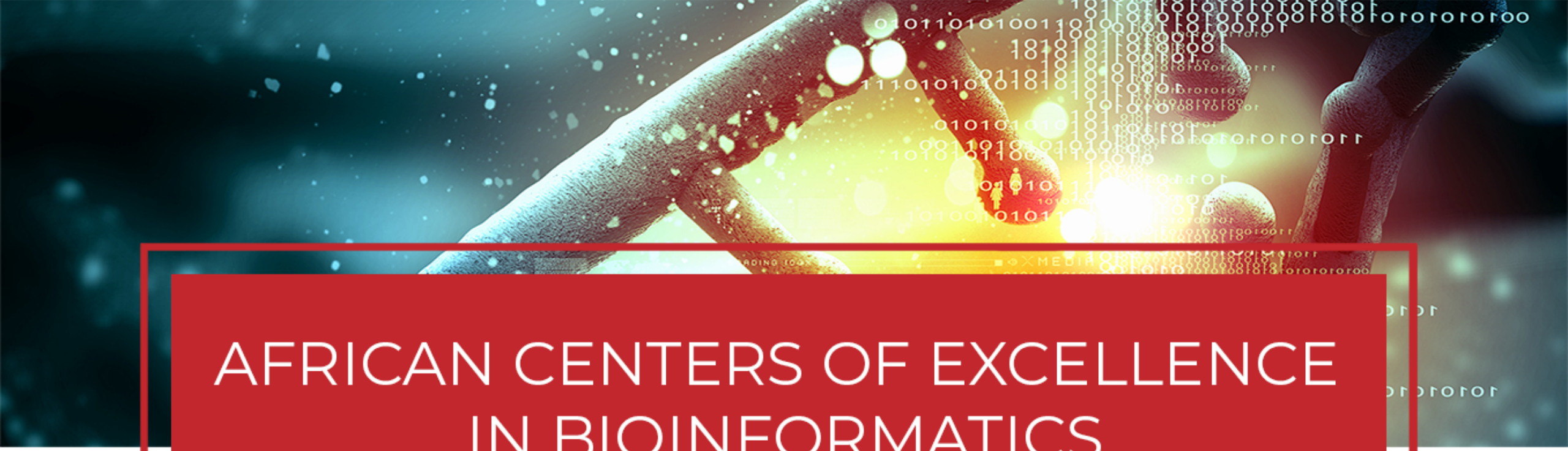
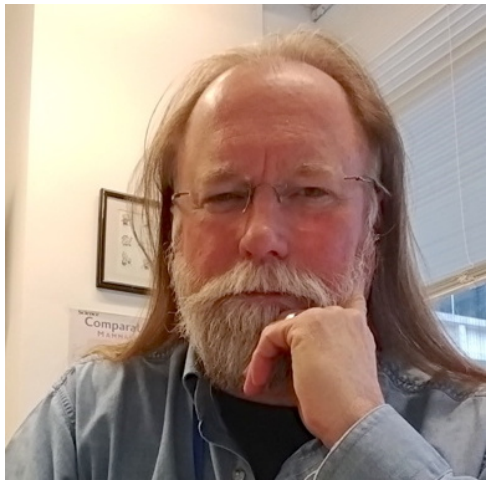


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

DYNAMIC PROGRAMMING AND ALIGNMENT ALGORITHMS

Today's Instructor



Dr. Kurt Wollenberg, Ph.D. in Genetics

Ongoing Computational Biology projects:

- Hepatitis B molecular evolution
- CLAG protein family evolution

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: kurt.wollenberg@nih.gov

Class Materials

- Directory on Uganda ACE server:
  - File directory: user@kla-ac-bio-03:/home/bcbb_teaching_files
  - Large data files
- NIAID github repository:
  - https://github.com/niaid/ACE-2020
  - Code
  - Data files
  - Copies of lecture slides

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
  - 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
```

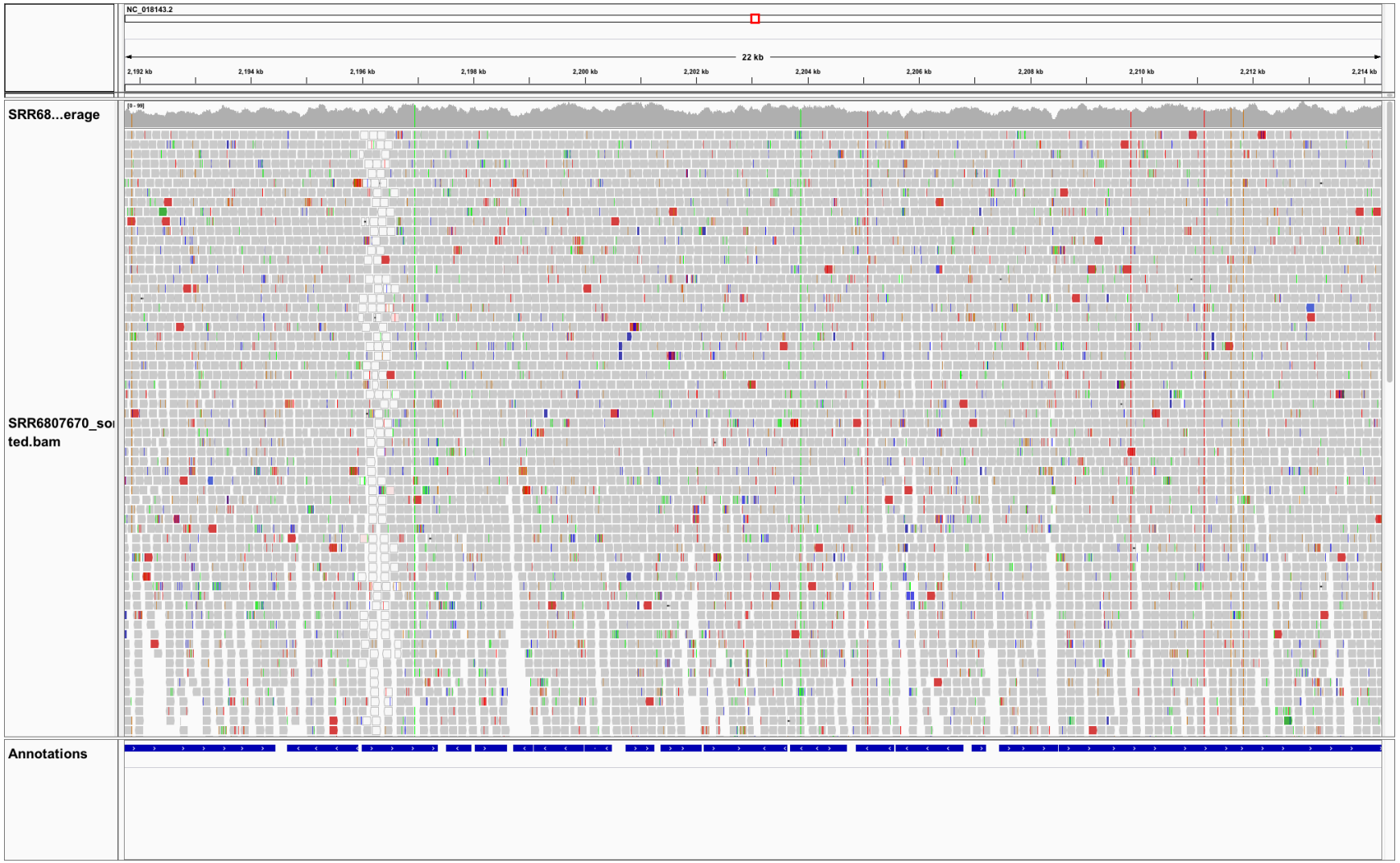
PAIRWISE ALIGNMENT

- Overlapping contigs

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

PAIRWISE ALIGNMENT

- Short reads



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 parameters - HKY
  - Unequal base frequencies
  - Two rates of change
- Fully 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 Probabilites

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

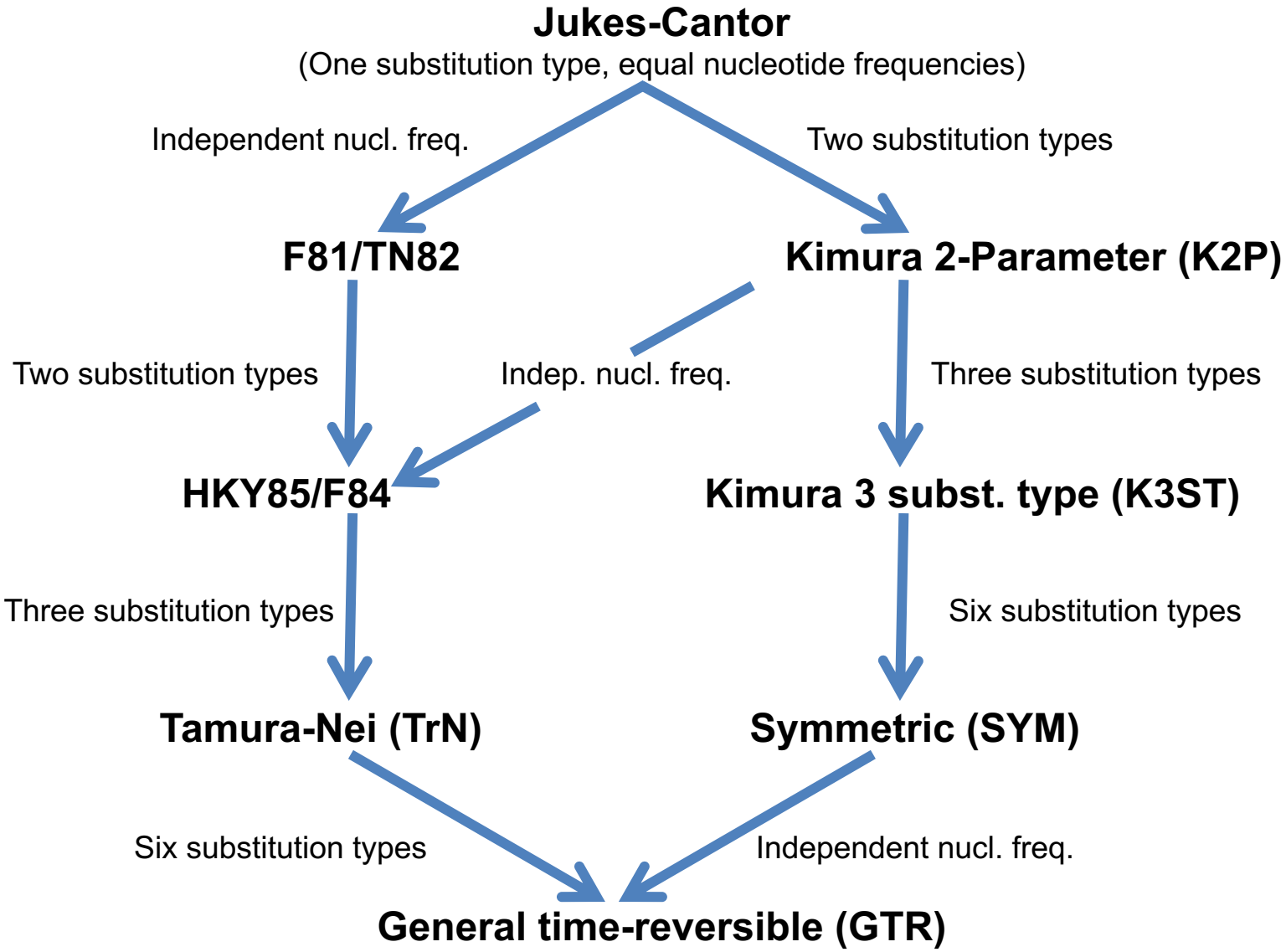
PAIRWISE ALIGNMENT

Kimura Two-Parameter Substitution Probabilites

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

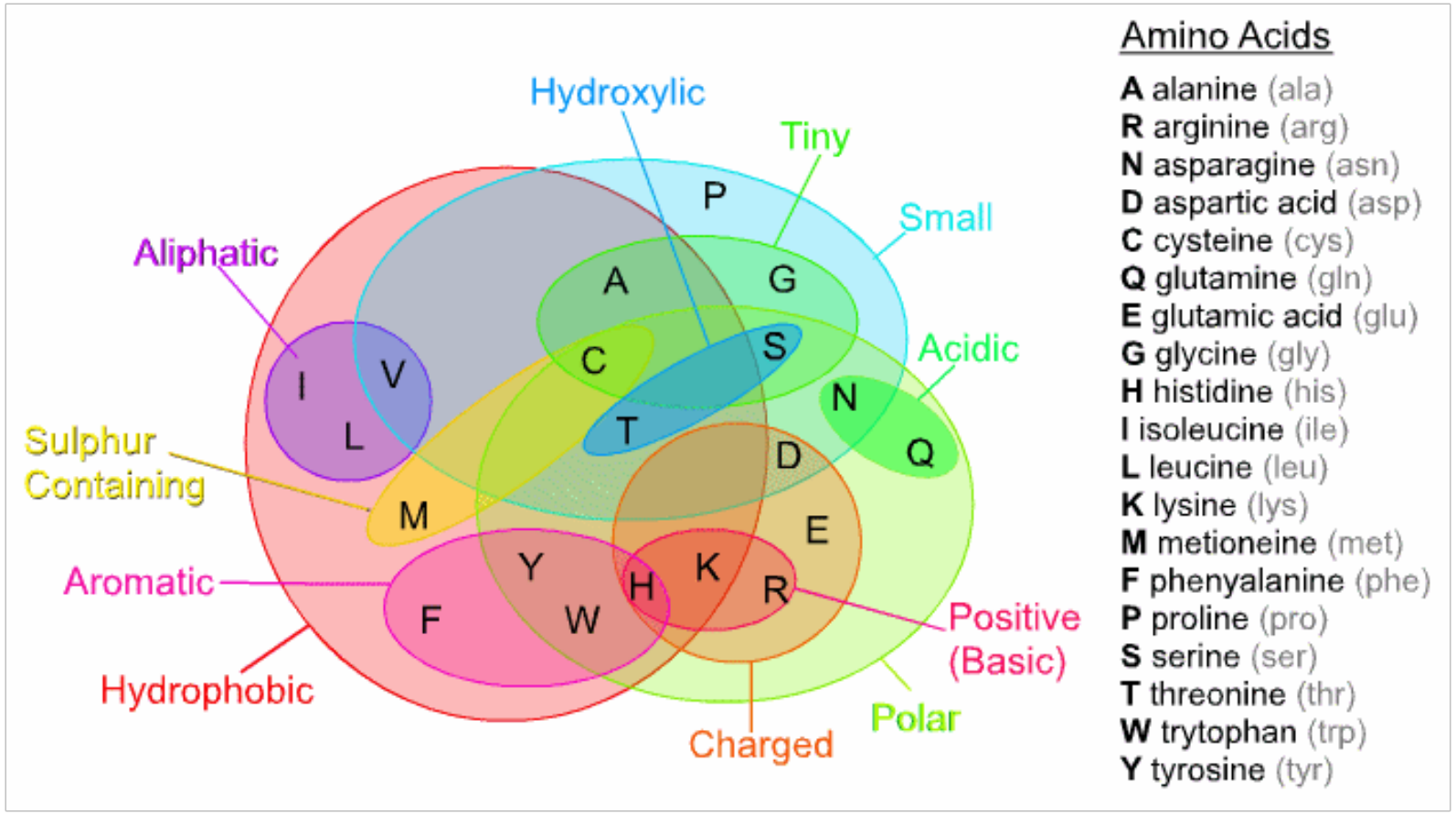
	A	C	G	T
A	0.4535	0.1580	0.2304	0.1580
C	0.1580	0.4535	0.1580	0.2304
G	0.2304	0.1580	0.4535	0.1580
T	0.1580	0.2304	0.1580	0.4535

SUBSTITUTION MODELS



PAIRWISE ALIGNMENT

Protein Score Matrices
Similarity of Amino Acids



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 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 Score (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.
- Both matrices: log odds values are scaled and rounded to the nearest integer values.

PAIRWISE ALIGNMENT

Protein Score (Substitution) Matrices

```
W -8 -2 -5 -6 -6 -7 -4 -7 -7 -6 -3  2 -3 -4 -5 -2 -6  0  0 17 PAM250
W -5 -4 -3 -4 -3 -3 -4 -5 -3 -1 -3 -3 -3 -1 -3 -2 -3  1  2 15 BLOSUM50
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

PAIRWISE 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
  - Good for aligning very divergent sequences

PAIRWISE 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

PAIRWISE ALIGNMENT

What about gaps?

- Score penalty for opening
- Score penalty for extending

Penalties are log probabilities of a gap of a specific length

PAIRWISE ALIGNMENT

Standard gap costs

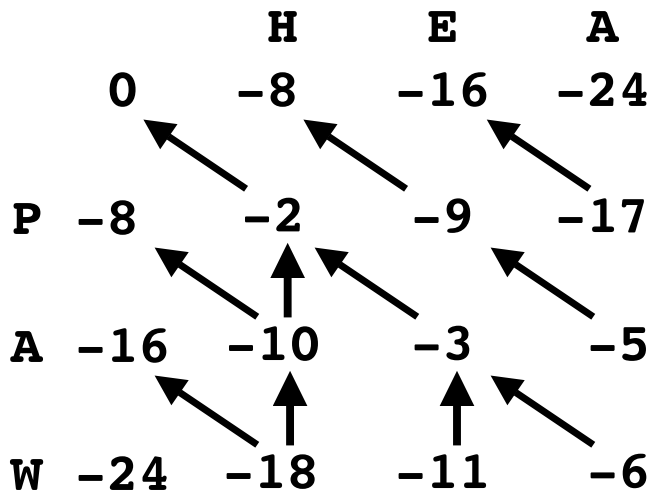
Substitution Matrix	Gap Costs (Open, Extend)
PAM30	(9,1)
PAM70	(10,1)
BLOSUM80	(10,1)
BLOSUM62	(10,1)
BLOSUM45	(15,2)

PAIRWISE 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



PAIRWISE 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

PAIRWISE 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

PAIRWISE 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

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

Overlap Match

H E A G A W G H E e

p A W - H E a e

PAIRWISE ALIGNMENT

Scoring alignments and expect values

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

Scores are a function of length of the sequences

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 sequence profile from results
- BLASTp using the profile
- Iterate until no new sequences are found
- Convergence

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

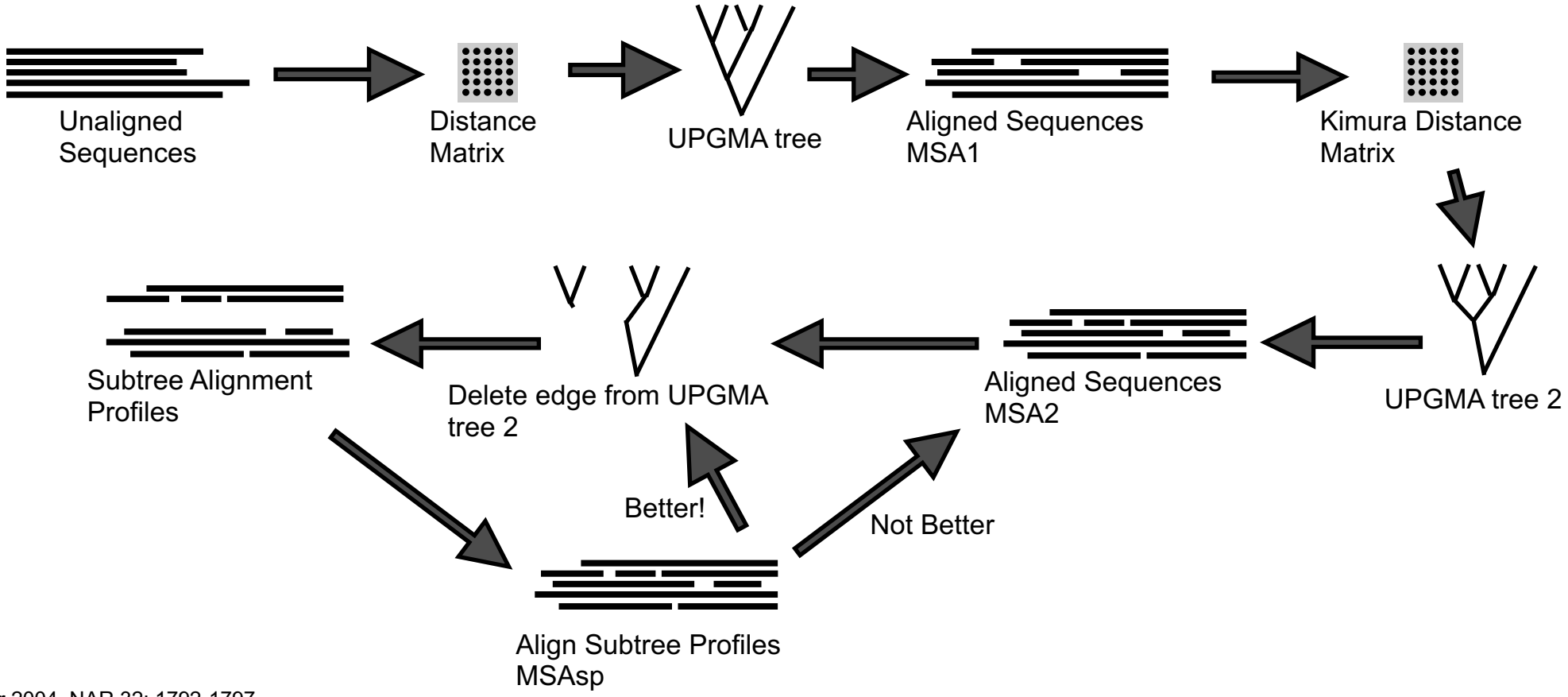
VGERGLEEDKRKRSAWMQC

MGETALRRRKKEDEERTANVYT

FGEAAMPGGPHQSRSAFAWV

MULTIPLE SEQUENCE ALIGNMENT

The Progressive Alignment Algorithm



From RC Edgar 2004, NAR 32: 1792-1797

MULTIPLE SEQUENCE ALIGNMENT

Programs

- ClustalW2
  - Your own computer
  - Web Server
  - Other sequence analysis packages
- MUSCLE
  - Your own computer
  - Web Server
  - Other sequence analysis packages
- MAFFT
  - Your own computer
  - Web Server
  - Other sequence analysis packages

MULTIPLE SEQUENCE ALIGNMENT

NEVER

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

ALWAYS

inspect the alignment to correct or improve it.

MULTIPLE SEQUENCE ALIGNMENT

Multiple Sequence Alignment Editors

Commercial Software

- Geneious
- MacVector
- MegAlign (Lasergene)

Public Domain Software

- AliView
- Seaview
- GeneDoc
- BioEdit
- MEGA

MULTIPLE SEQUENCE ALIGNMENT

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

MEGA

https://megasoftware.net

PAIRWISE ALIGNMENT

MEGAX

1. Under “Align” choose “Perform BLAST search”
2. Use query sequence NM_000575
3. In the “Organism” field limit results to Mammals
4. Under “General Paramters” 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 MEGAX window)

PAIRWISE ALIGNMENT

MEGAX

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 TAGGCTC
7. Select and delete the block of sequence after TAG

PAIRWISE ALIGNMENT

MEGAX

1. Click on “Translated Protein Sequences”
2. Accept the standard code
3. Look for “?” sites
4. Select site 48 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