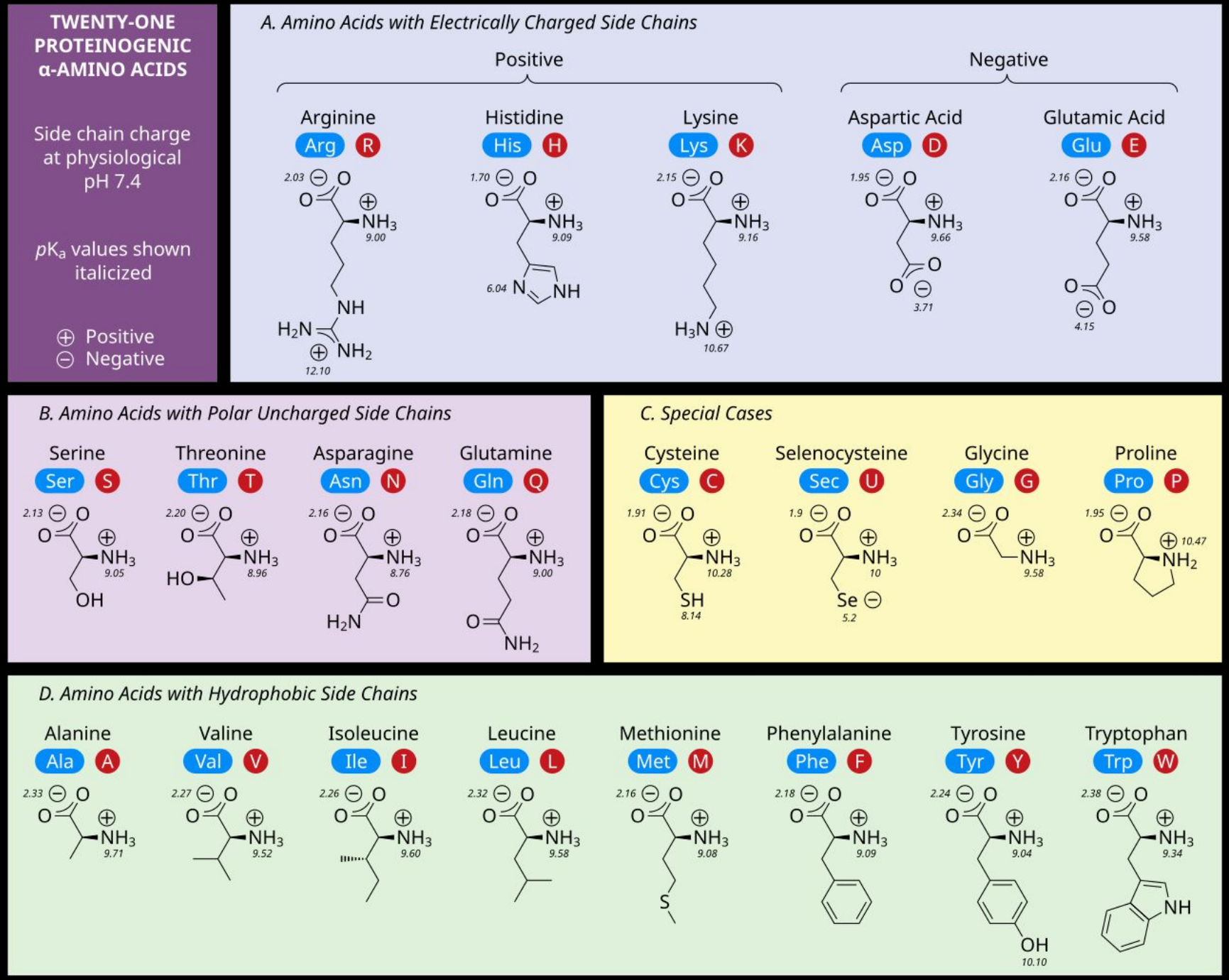

Sequence Alignment
S-qua-ce Am--xnedt
-

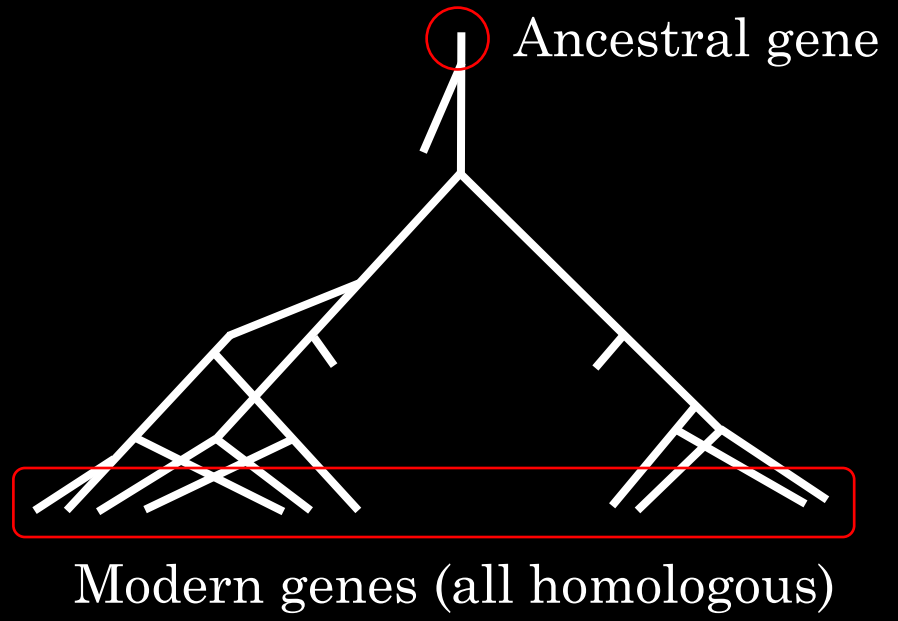
Amino acid code reminder

https://en.wikipedia.org/wiki/Amino_acid#/media/File:ProteinogenicAminoAcids.svg



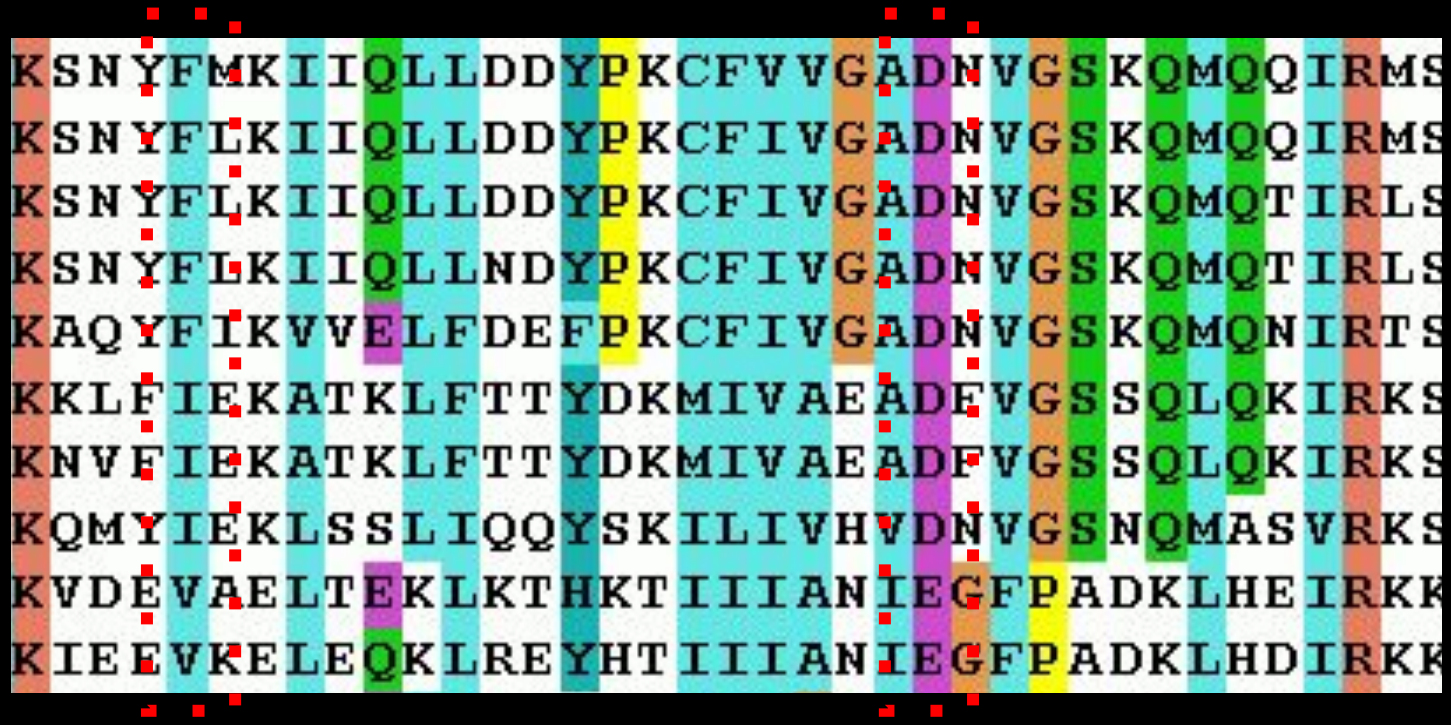
Homology:
More than just genes!

HOMOLOGOUS genes
share a common ancestor



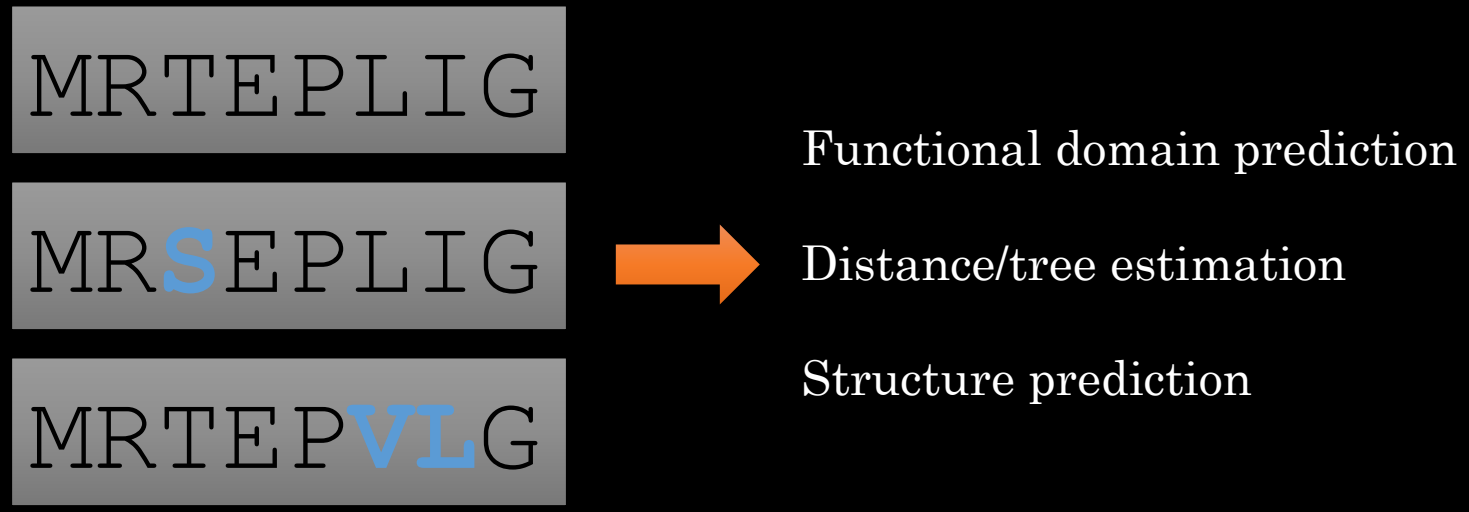
Homology:
More than just genes!

DNA / protein “residues”
(nucleotides and amino acids)
can also be homologous

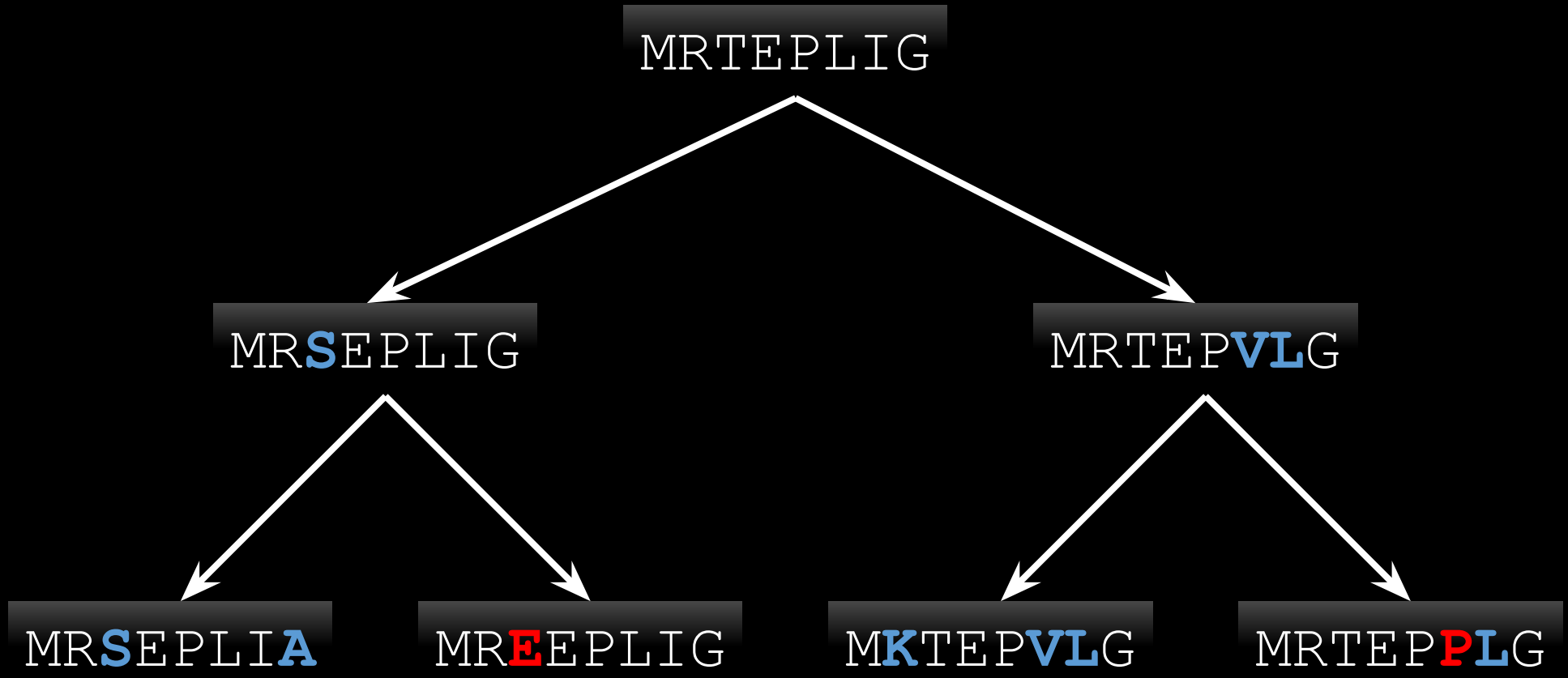


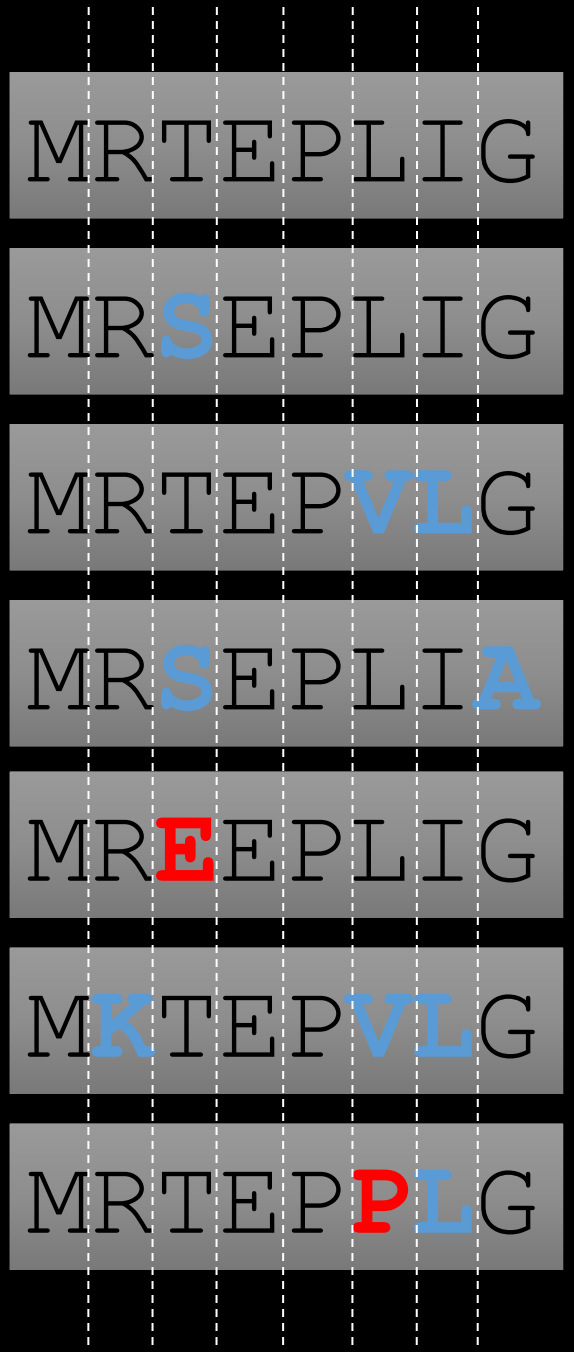
Each column is a homologous position *within* the proteins

For many applications of sequence analysis, we would like to know which residues (nucleotides, amino acids) are homologous between sequences



In a world where substitutions were the only type of mutation, the homology of residues would be obvious



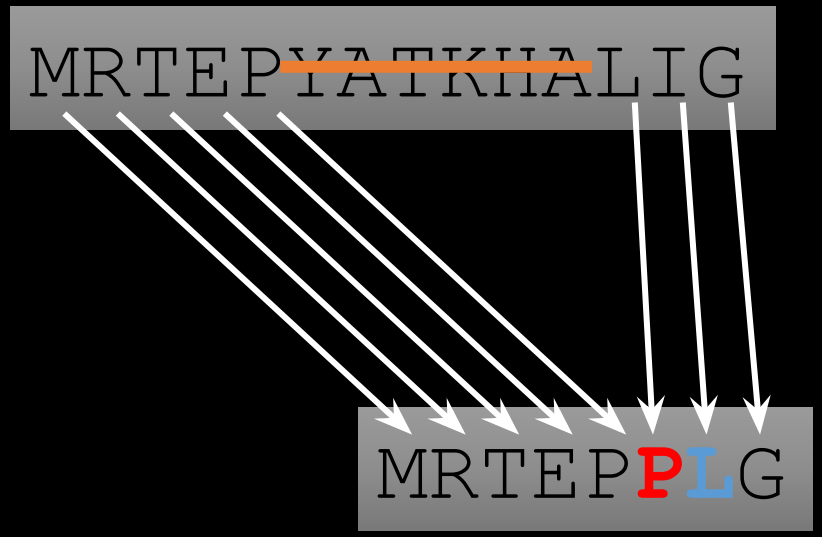


Each column contains a set of residues that are **homologous**

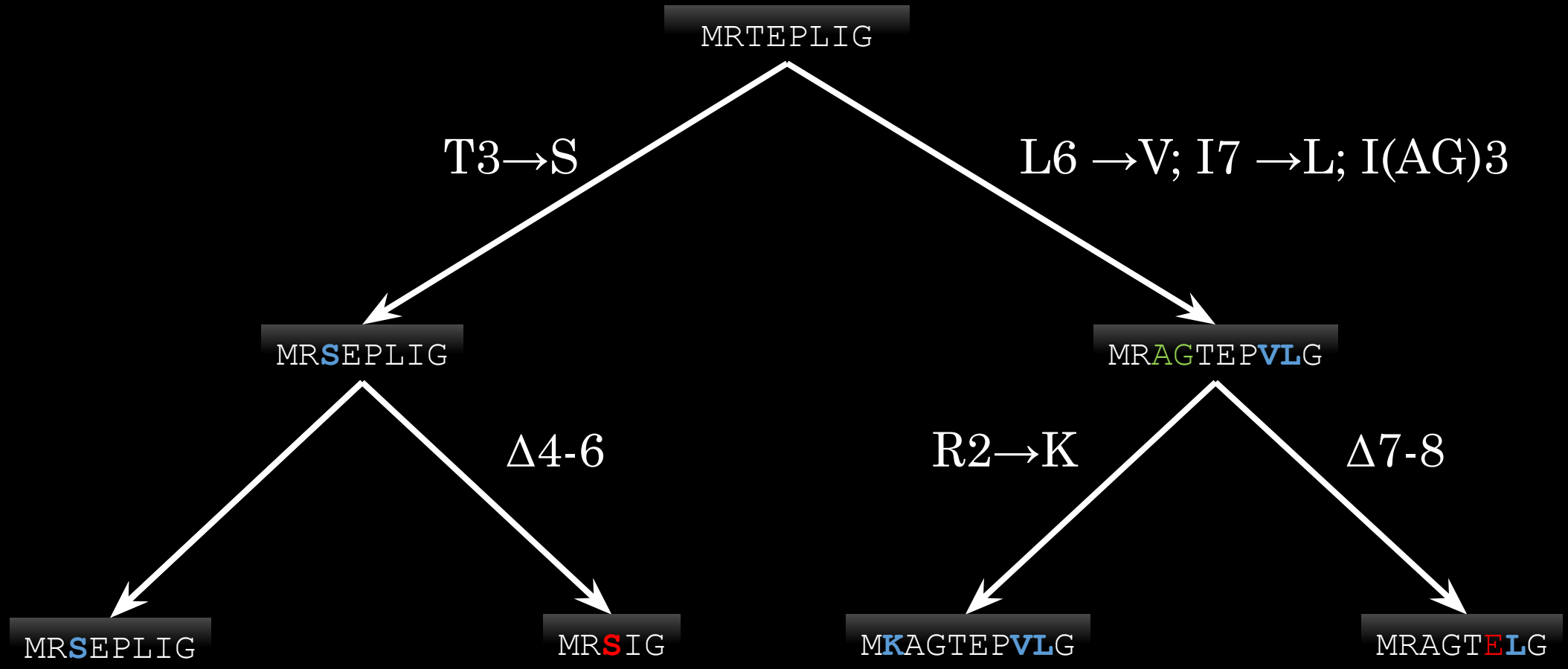
This is a **sequence alignment** (albeit a trivial one!)

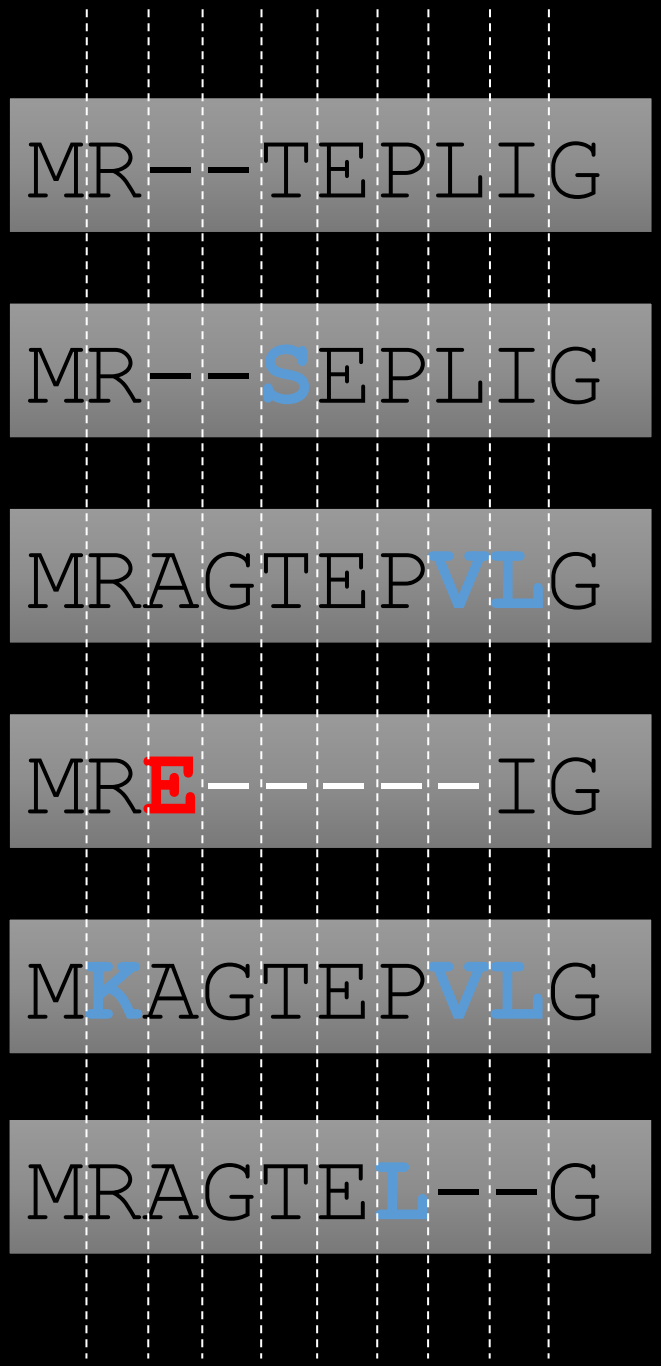
But Life is Not so Easy…

Insertions and deletions (and more complex changes) can complicate the process



The process


MRTEPLIG
T3→S
L6 →V; I7 →L; I(AG)3
MRSEPLIG
MRAGTEPVLG
Δ4-6
R2→K
Δ7-8
MRSEPLIG
MRSIG
MKAGTEPVLG
MRAGTELG



To bring homologous residues together, we need to perform a SEQUENCE ALIGNMENT by introducing gap characters

But how do we get to an alignment, and how do we decide which is best?

Keys to sequence alignment

1. A SCORING SYSTEM for an alignment of two or more sequences
   - Is the alignment any good?
   - Is the similarity between the two sequences better than random?

2. An ALGORITHM to find the best alignment, or a set of highly probable alignments
   - What is the complexity of finding the optimal solution?
   - To what extent can we trade away optimality for efficiency?

Elements of a scoring system

1. Residue **frequencies** $f(x_i)$ in the set of sequences
2. Transition **probabilities** $p(x_i,x_j)$ between residues
3. A scheme G for penalizing **gaps**
4. A formula for computing the **score**, given F, P, and G

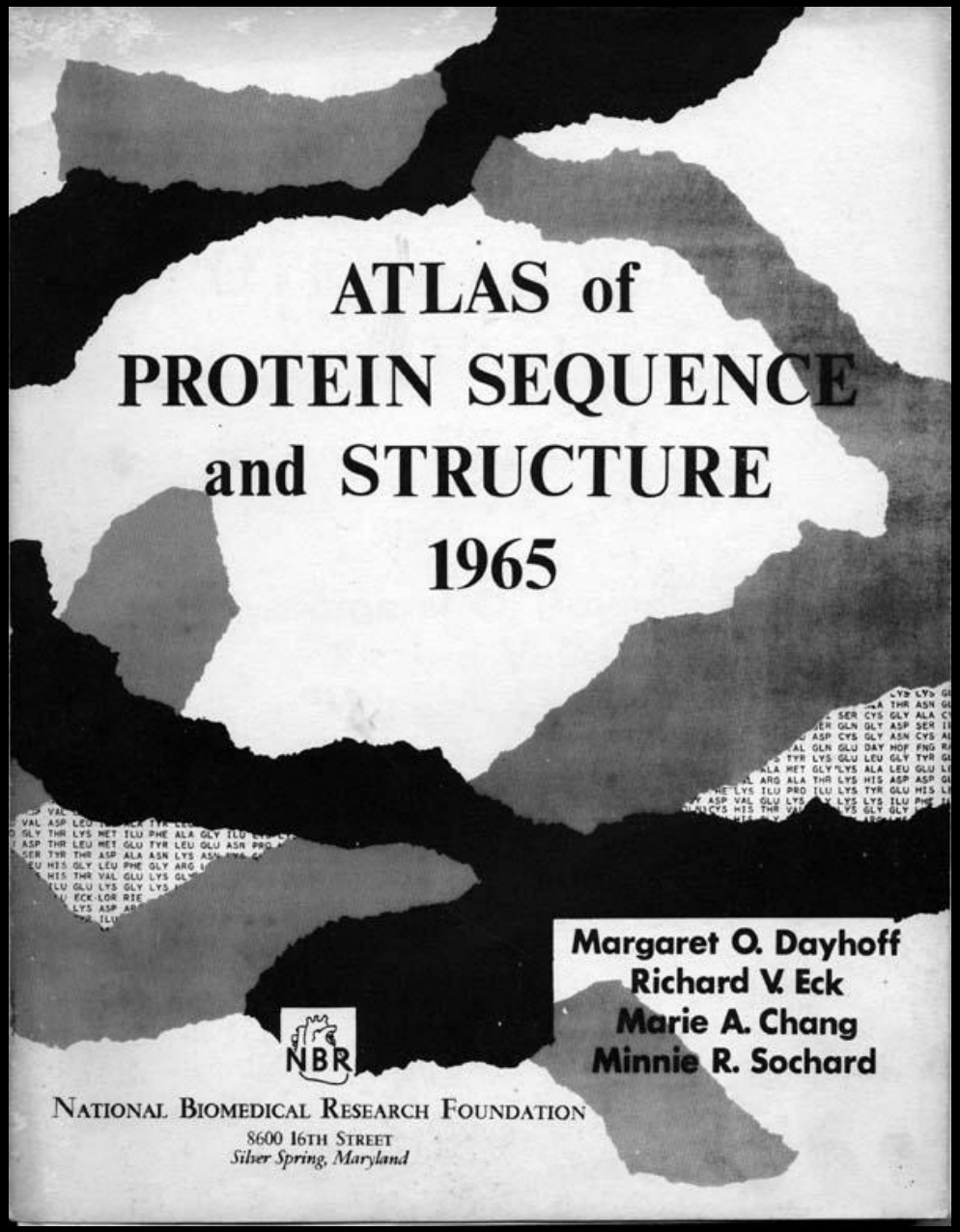
Part the first: substitution probabilities

1. Build a **reference dataset** with certain desirable properties

2. Construct **alignments** (?!) of the sequences within this dataset

3. Compute the probabilities of different substitutions based on observed **frequencies**

Margaret Dayhoff and PAM



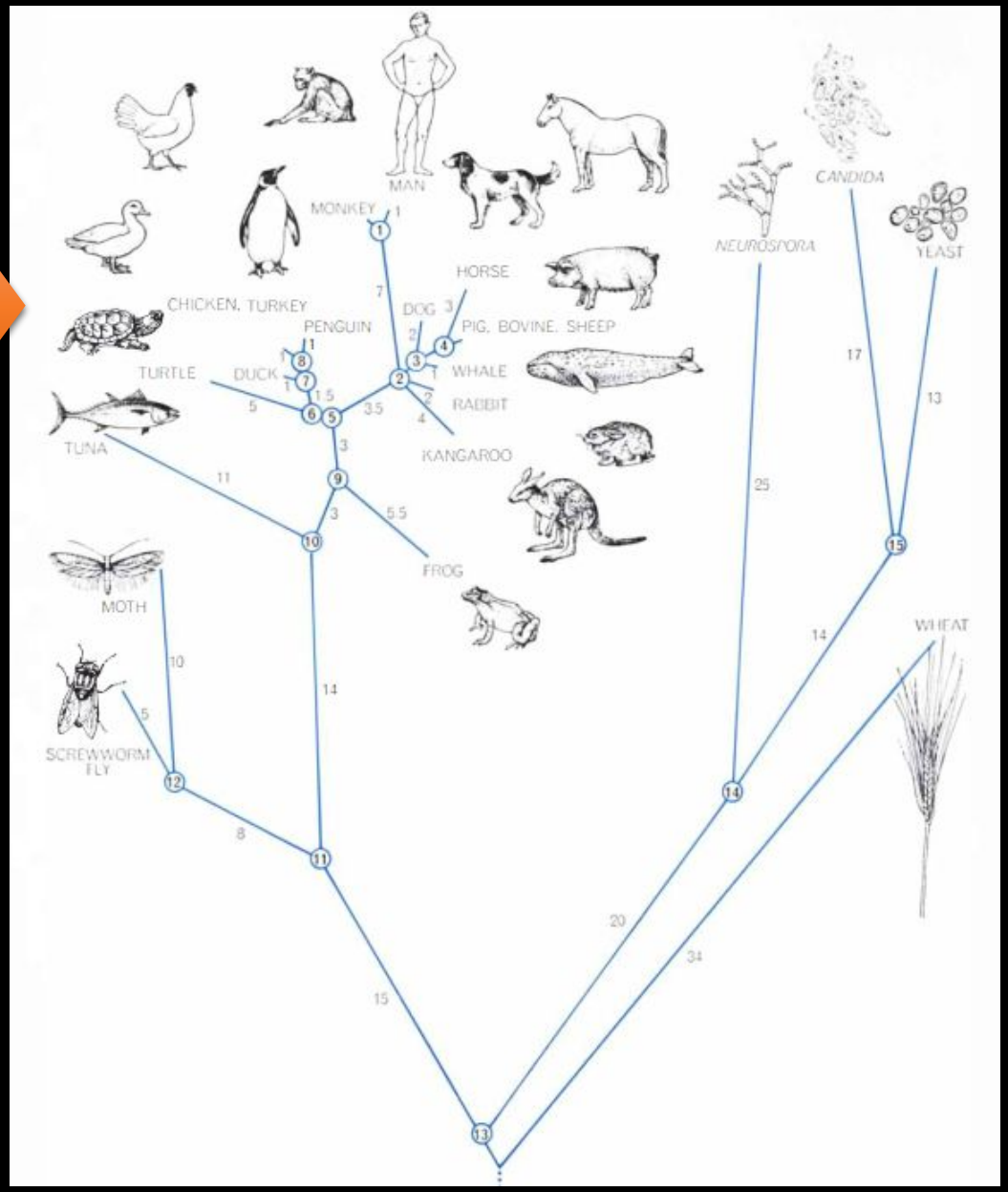
65 protein sequences

“Responding to the sudden increase in the rate of nucleic acid sequencing, **Dr. Dayhoff established an online computer database and a sophisticated retrieval system, accessible by phone to outside users, in September 1980**. A home computer system had been used to prove the feasibility of this approach. This nucleic acid sequence database is currently one of the largest in the world, containing over 2 000 000 sequenced nucleotides with references and annotations. Since September 1981, the Protein Sequence Database has also been available on-line as well as on magnetic tape.”

http://www.dayhoff.cc/MODBiography.html

Other Dayhoff

- First phylogenetic tree calculated using a computer
- Origins of life / Early planetary evolution
- Cancer biology
- Protein families and superfamilies

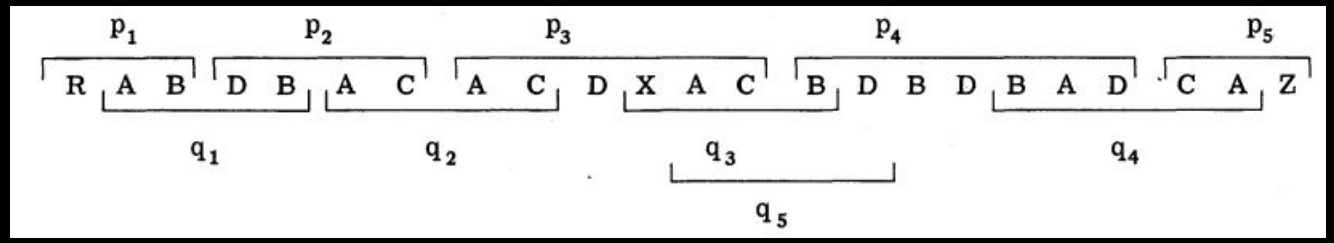


Dayhoff (1969) *Sci Am*

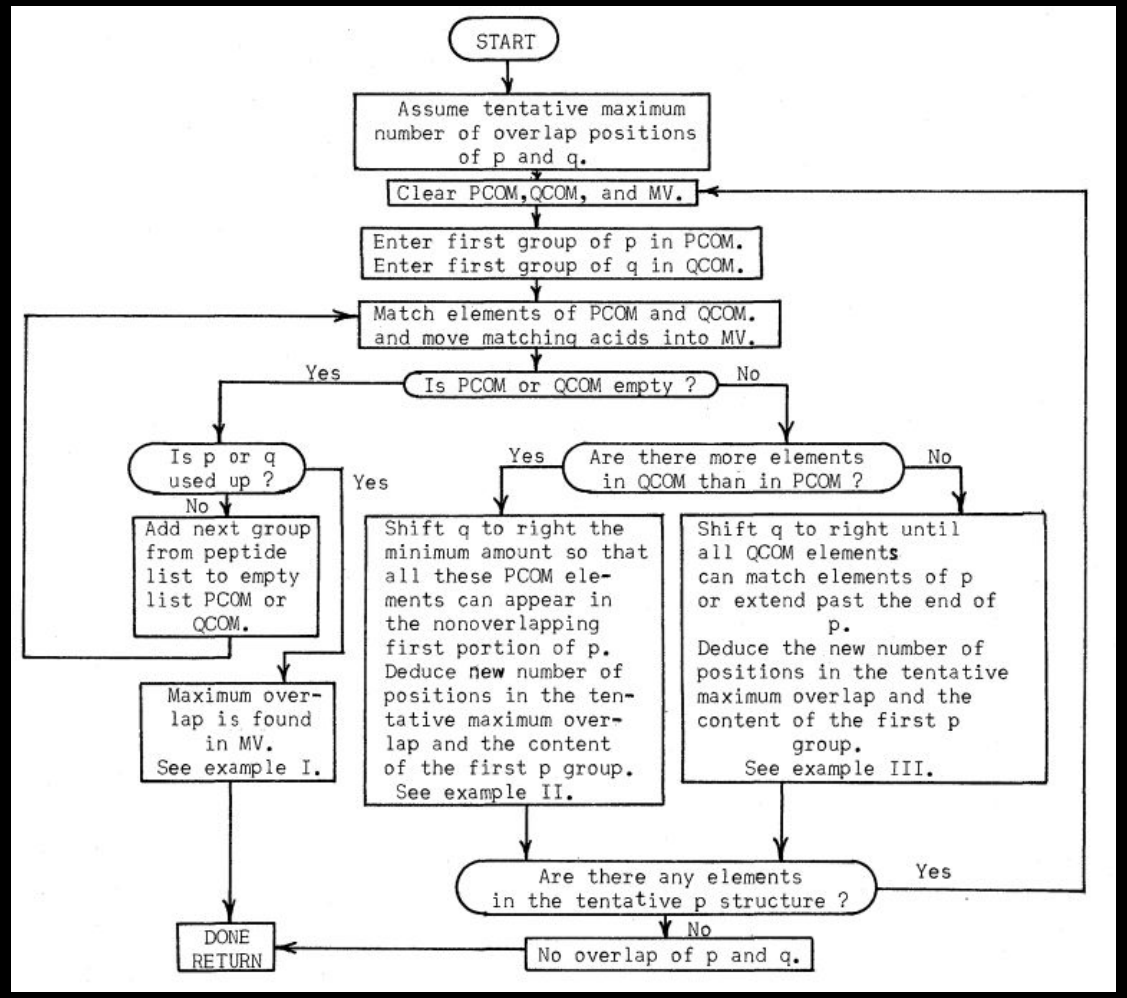
The first bioinformatics tool

COMPROTEIN: A COMPUTER PROGRAM TO AID PRIMARY PROTEIN STRUCTURE DETERMINATION*

Margaret Oakley Dayhoff and Robert S. Ledley
National Biomedical Research Foundation
Silver Spring, Maryland



IBM 7090



Proceedings of the Fall Joint Computer Conference, 1962

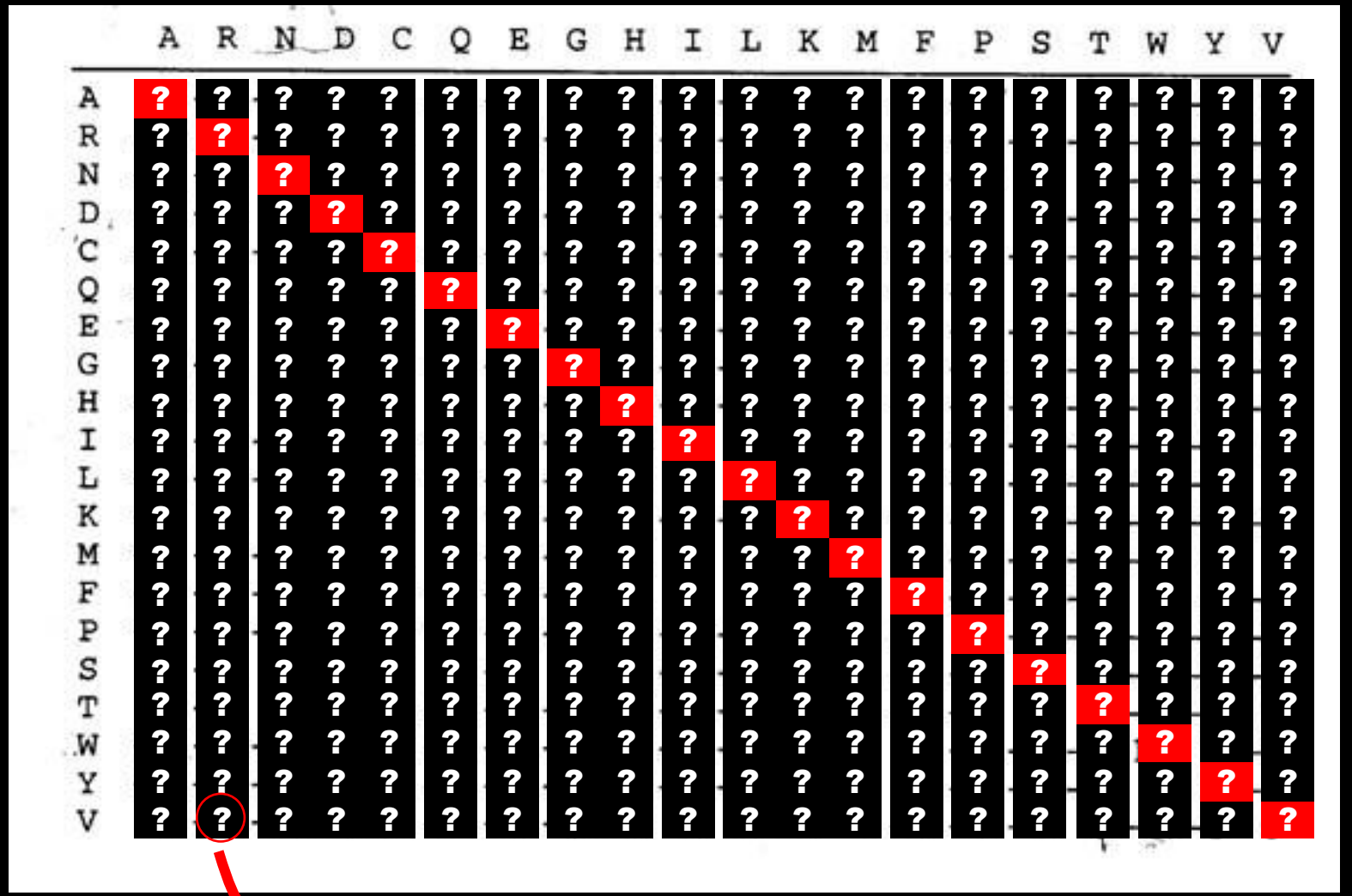
And…

The amino acid alphabet

ACDEFGHIKLMNPQRSTVWY

https://www.whatisbiotechnology.org/index.php/people/summary/Dayhoff

Building a Substitution Matrix



Some measure of change from V to R

Building a Substitution Matrix

- One way is to define amino acids based on their chemical and/or structural properties, and build a matrix based on their similarity

	Isoleucine	Leucine	Tryptophan
Isoleucine		↑	↓
Leucine	↑		↓
Tryptophan	↓	↓	

- e.g. Grantham matrix (1974). Doesn't reflect the evolutionary process – why not?

Percent Accepted Mutation (PAM)

- An ‘accepted’ mutation changes one or more amino acids and doesn’t lead to insta-death or selective costs

- PAM*n* matrix – *n* substitutions per 100 sites

  PAM1: Sequences with 1 substitution / 100 sites

  PAM250: Sequences with 250 substitutions / 100 sites

Um, what?

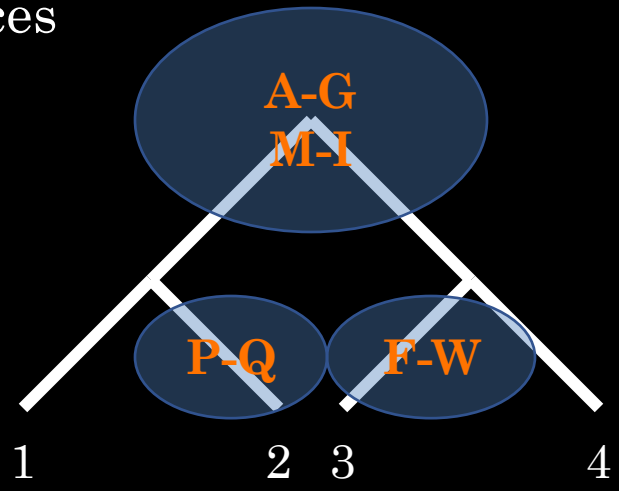
Building the PAM1 matrix

- Assume that amino acid substitution is a Markovian process (?)
- Reference data set (1978): set of protein alignments, 71 families in total
- Consider only **blocks** – ungapped alignment regions ≥ 85% identical (minimize double substitutions!)

Map onto a PHYLOGENETIC TREE that shows the history of the sequences

```
1  AAAILGMVFP
2  AAAILGMVFQ
3  AAGILGIVWP
4  AAGILGIVFP
```

Count this change only once!



Treat substitutions as REVERSIBLE (so our matrix will be symmetric)

$M \leftrightarrow I$

Also compute the vector of frequencies:

$f(A) = 10/40 = 0.25$

$f(F) = 3/40 = 0.075$

etc…

(1) Matrix of Counts

A

	A	C	D	…
A	9981	15	31	…
C	15	6744	12	…
D	31	12	8330	…
…	…	…	…	…

DIAGONALS (no change) dominate
in closely related sequences

(2) Matrix of Probabilities

Normalize by row, all row sums == 1

B

	A	C	D	…	Sum
A	0.97	0.0002	0.005	…	1.0
C	0.0002	0.995	0.0003	…	1.0
D	0.005	0.0003	0.982	…	1.0
…	…	…	…	…	

What is the relative rate of change of A ↔ C, or "change" between A ↔ A?

(3) Matrix of Scaled Probabilities (1 PAM)

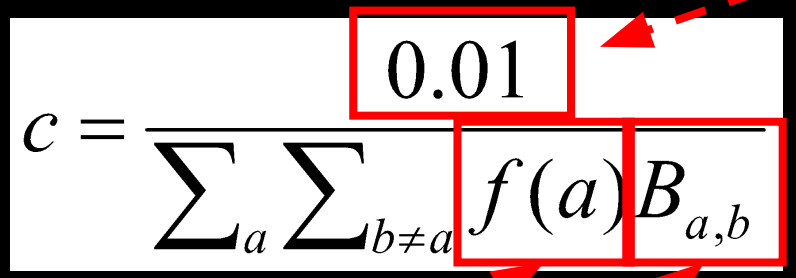
The amount of sequence change in **B** is dependent on whatever sequences we used to create our dataset

More-distant sequences: diagonals smaller, off-diagonals larger

We want to rescale the matrix based on frequencies so the expected number of amino-acid substitutions per site is equal to 0.01

(3) Matrix of Scaled Probabilities (1 PAM)

Each off-diagonal element is multiplied by c, where



Change diagonals so each row sums to 1.0, and the rest of the matrix sums to 1 PAM

Total amount of change = ???

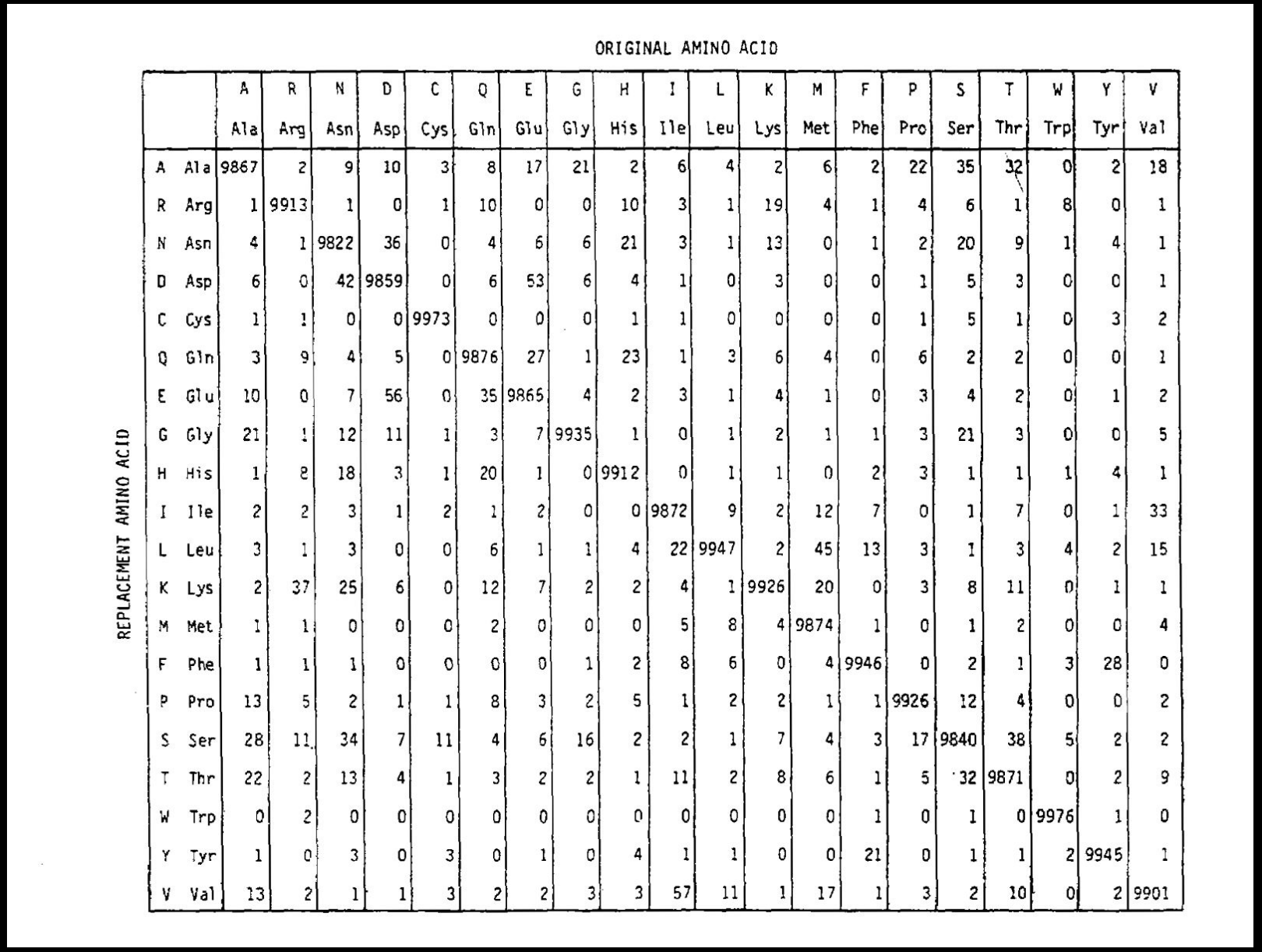
B

	A	C	D	…	Sum
A	0.97	0.0002	0.005	…	1.0
C	0.0002	0.995	0.0003	…	1.0
D	0.005	0.0003	0.982	…	1.0
…	…	…	…	…	

Total amount of change = 0.01 substitutions per site

C

	A	C	D	…
A	0.9994	0.00002	0.0005	…
C	0.00002	0.9985	0.00003	…
D	0.0005	0.00003	0.9911	…
…	…	…	…	…



ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82(!): PAM1 probability matrix
(divide by 10,000 to get probabilities)

https://cs.rice.edu/~ogilvie/comp571/pam/

For higher-order PAM matrices:

$$\mathrm{PAM}n = (\mathrm{PAM1})^n$$

For higher-order PAM matrices, values on the diagonal will decrease, while off-diagonals will increase

(greater evolutionary distance)

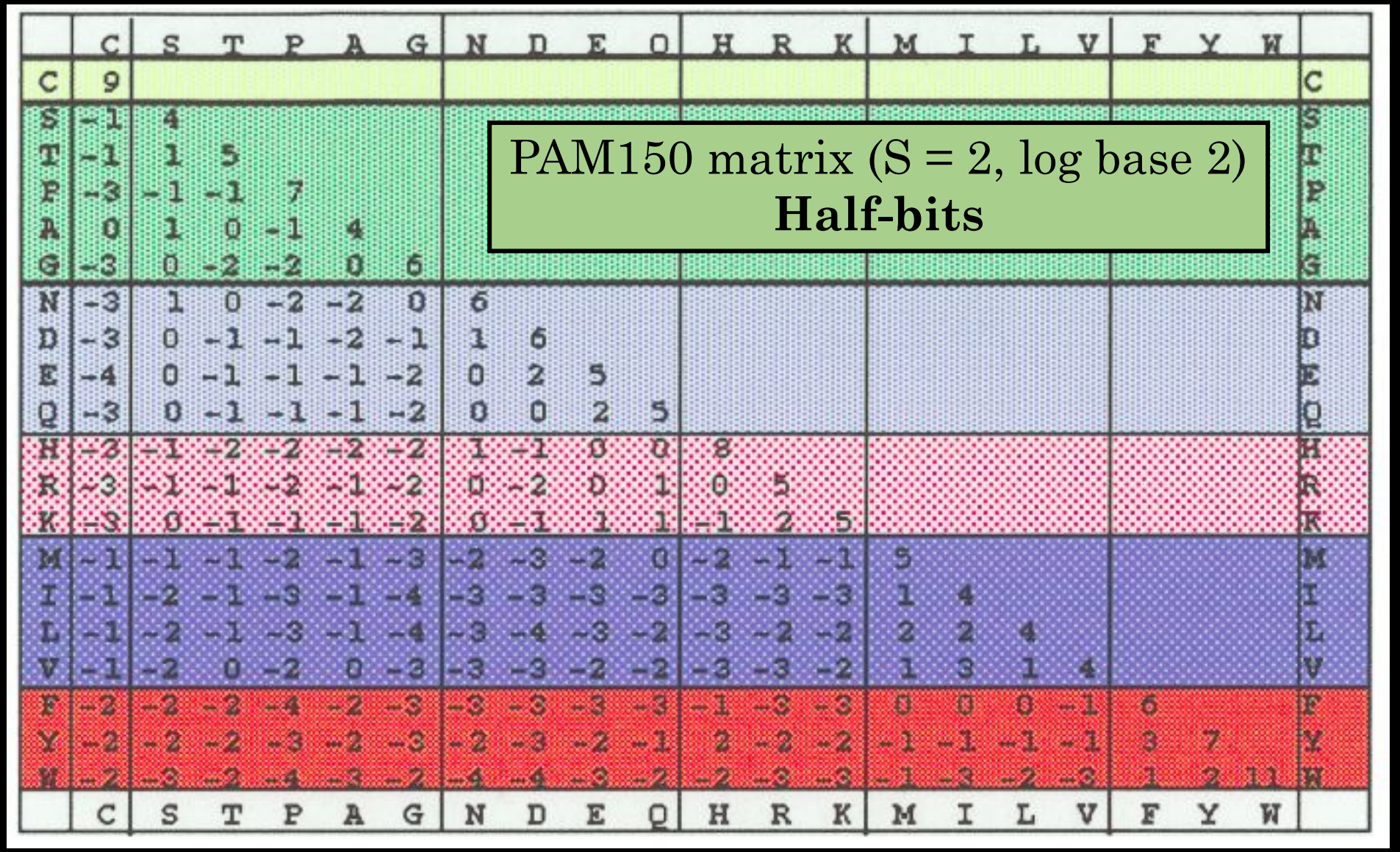
Exponentiation (rather than changing the scaling constant) is necessary to properly account for multiple substitutions

(4) The last step

- We need to generate a matrix that captures the probability of seeing residues *i* and *j* together due to homology, relative to a random expectation

$$\mathbf{D}_{a,b} = S \cdot \log\left(\frac{C_{a,b}}{f(a)f(b)}\right)$$

More frequent than random: D > 0
Random: D = 0
Less frequent than random: D < 0



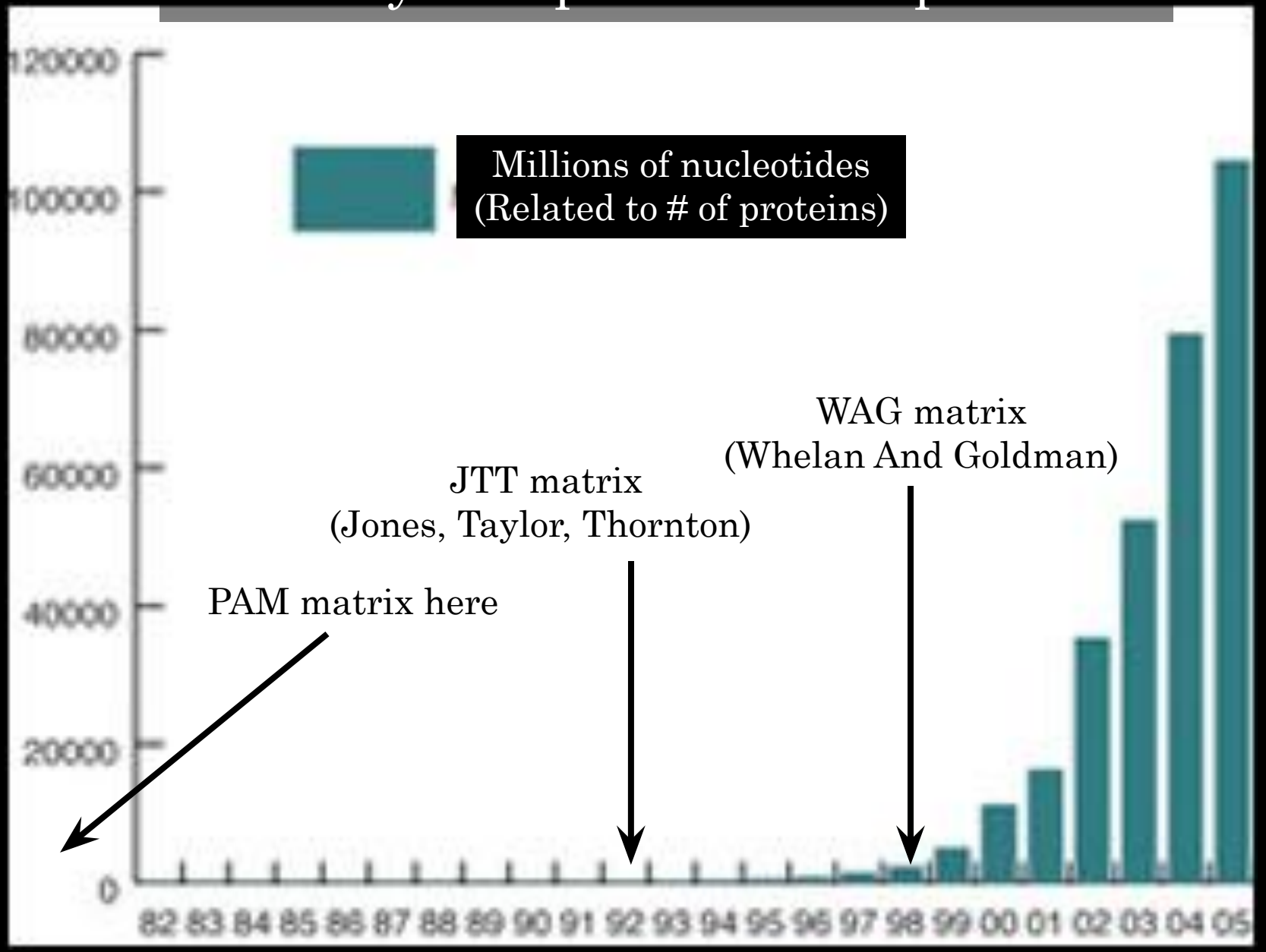
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

https://bioinformaticamente.com/2020/10/22/alignment-algorithms/

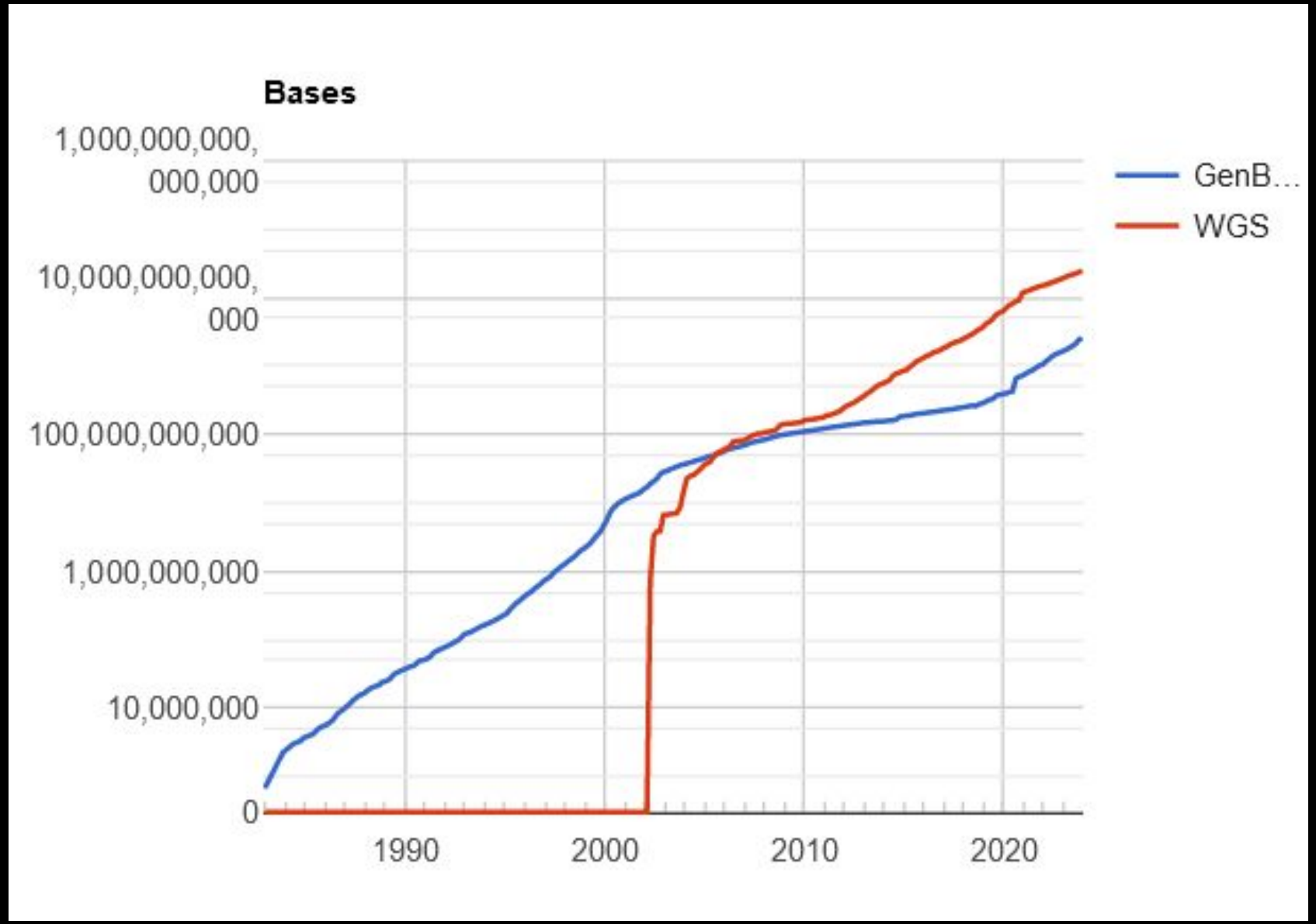
Thoughts on PAM

Limitations?

Accuracy is dependent on input data!



http://www.innovations-report.com/bilder_neu/48071_data.jpg

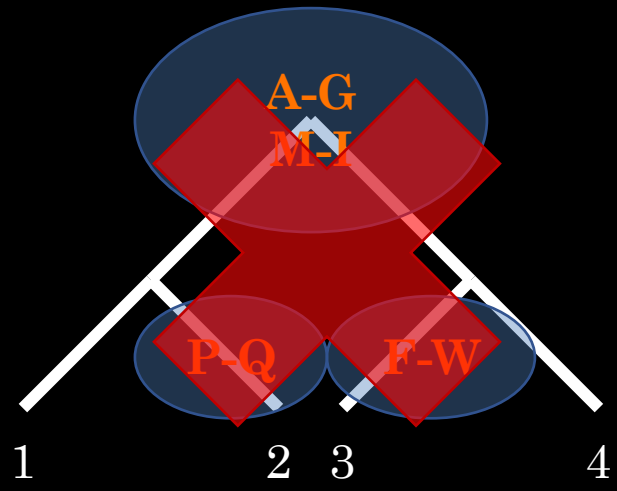


https://www.ncbi.nlm.nih.gov/genbank/statistics/

$$\mathrm{PAM}n = (\mathrm{PAM1})^n$$

Extrapolation!!!

What if the tree is wrong?



The BLOSUM matrix – clusters instead of trees

Subdivide homologous sequences into CLUSTERS with at least L% identity
Count substitutions between clusters only

```
1  AAAILGMVFP
2  AAAILGMVFQ

3  AAGILGIVWP
4  AAGILGIVFP
```

	P	Q
P	2	-
Q	2	0

A

	P	Q
P	2	-
Q	2	0

C

$$C_{P,Q} = \frac{A_{P,Q}}{\sum_{c,d} A_{c,d}} = 1/2$$

Log-odds matrix: as before!

$$\mathbf{D} = S \cdot \log\left(\frac{C_{a,b}}{f(a)f(b)}\right)$$

Better than random: D > 0
Random: D = 0
Worse than random: D < 0

BLOSUM62

D

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

https://en.wikipedia.org/wiki/BLOSUM

Why BLOSUM?

- No reliance on an inferred tree
- No extrapolation; differences are observed directly from alignments with at least L% divergence
- Higher accuracy relative to PAM in detecting remote homologs

BLOSUM x Matrices

x = the % identity within blocks

BLOSUM 62 is based on more-similar sequences than BLOSUM 45

(opposite of PAM!)

Choosing a matrix

- Matrices can be compared based on their relative entropy

$$H = \sum_{i,j} q_{i,j} \log_2 \frac{q_{i,j}}{p_i p_j}$$

- Very similar sequences: high entropy
- As distance approaches ∞, H approaches 0

Altschul (1991) *J Mol Biol*

Difference between BLOSUM62 and PAM160

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

FIG. 2. BLOSUM 62 substitution matrix (*Lower*) and difference matrix (*Upper*) obtained by subtracting the PAM 160 matrix position by position. These matrices have identical relative entropies (0.70); the expected value of BLOSUM 62 is −0.52; that for PAM 160 is −0.57.

Henikoff and Henikoff (1992) *Proc Natl Acad Sci USA*

Table 1

The relative entropy H *of PAM matrices*

PAM distance	*H* (bits)	Min. significant length (30 bits)	PAM distance	*H* (bits)	Min. significant length (30 bits)
0	4·17	8	180	0·60	51
10	3·43	9	190	0·55	55
20	2·95	11	200	0·51	59
30	2·57	12	210	0·48	63
40	2·26	14	220	0·45	68
50	2·00	15	230	0·42	73
60	1·79	17	240	0·39	78
70	1·60	19	250	0·36	83
80	1·44	21	260	0·34	89
90	1·30	24	270	0·32	94
100	1·18	26	280	0·30	100
110	1·08	28	290	0·28	107
120	0·98	31	300	0·27	113
130					120
140					127
150					134
160					141
170					149

age (Fig. 1). Based on relative entropy, the PAM 250 matrix is comparable to BLOSUM 45 with relative entropy of ≈0.4 bit, while PAM 120 is comparable to BLOSUM 80 with relative entropy of ≈1 bit. BLOSUM 62 (Fig. 2 *Lower*) is intermediate in both clustering percentage and relative entropy (0.7 bit) and is comparable to PAM 160. Matrices with comparable relative entropies also have similar expected scores.

Altschul (1991) *J Mol Biol*
Henikoff and Henikoff (1992) *Proc Natl Acad Sci USA*

There's more than ± 10 ways to do it

RAxML	Inference	JC, K80, HKY, GTR	Blosum62, CpRev, Dayhoff, DUMMY, FLU, HIVb, HIVw, JTT, JonesDCMUT, LG, Mtart, Mtmam, Mtrev, Mtzoa, PMB, RtRev, STMREV, VT, WAG +F	Partitioned models can be specified	+I +G

Tree inference software (coming later!)

Models:

- Different originating datasets (HIVb)
- Larger datasets (JTT)
- Fancy likelihoods (WAG, LG)

Arenas (2015) *Front Genet*

Great. We can score alignments.

But what about gaps??

```
QVKQIYKTPPIKYFGGFNFSQILPDPSKPSKRSPIEDLLF---------
QVKQIYKTPPIK----------D--------------FGGFNFSQIL
```

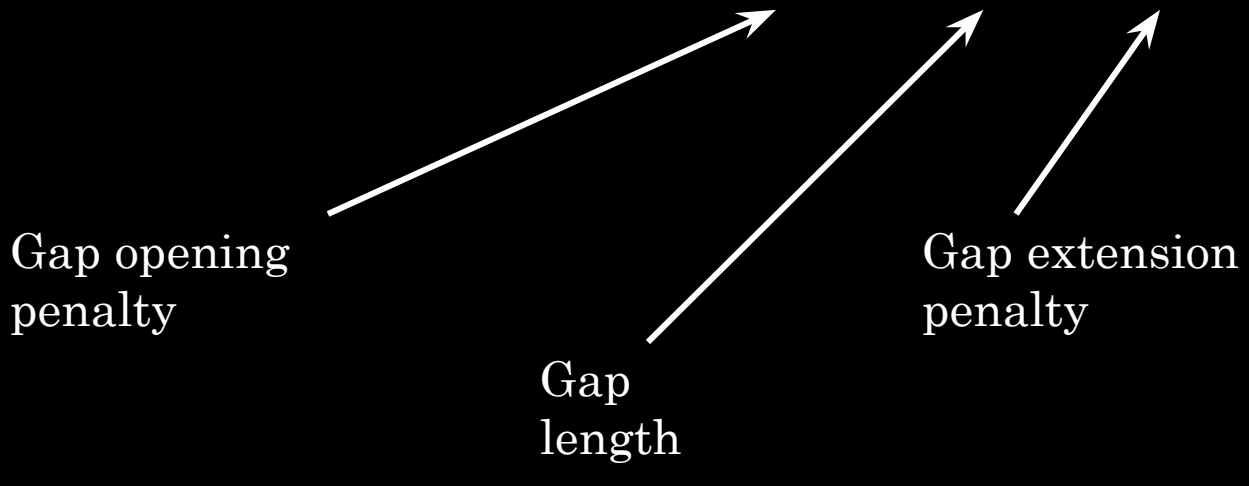
It's a lot more difficult to build rigorous statistics for gaps

GAP Penalties!

- Two types:

LINEAR: $\gamma(g) = -gd$

AFFINE: $\gamma(g) = -d - (g-1)e$



Computing an Alignment Score

X =

```
MKAGTEPVLG
MRAGTEL--G
```

$$S(X) = D_{M,M} + D_{K,R} + D_{A,A} + D_{G,G} + D_{T,T} + D_{E,E} + D_{P,L} + \gamma(g = 2) + D_{G,G}$$

Using PAM250, a gap opening penalty of 5 and a gap extension penalty of 2,

$$S(X) = 6 + 3 + 2 + 5 + 3 + 4 + (-3) + (-7) + 5$$

$$= 18$$

A_1 =

```
MKAGTEPVLG
MRAGTEL--G
```

S(X) = 18

$$S(X) = 6 + 3 + 2 + 5 + 3 + 4 + (-3) + (-7) + 5$$

Contrast with alignment A_2:

A_2 =

```
MKAGTEPVLG
MRA--GTELG
```

S(Y) = 13

$$S(Y) = 6 + 3 + 2 + (-7) + 0 + 0 + (-2) + 6 + 5$$

um, DNA?

Something like this usually works pretty well:

	A	G	C	T
A	1	-1	-1	-1
G	-1	1	-1	-1
C	-1	-1	1	-1
T	-1	-1	-1	1

Or possibly this:

	A	G	C	T
A	1	0.5	-1	-1
G	0.5	1	-1	-1
C	-1	-1	1	0.5
T	-1	-1	0.5	1

For protein-coding sequences, it is most common to align the amino acid sequences, then match the corresponding DNA codons against this sequence

¿Why?

The goal of sequence alignment is (usually) to find the best alignment score – maximize the probability of observing aligned residues, relative to the null model

But optimal methods are slow – as you will see!