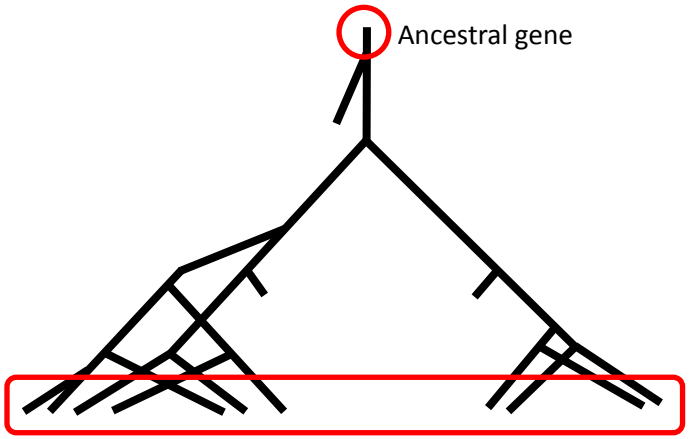
Sequence Alignment

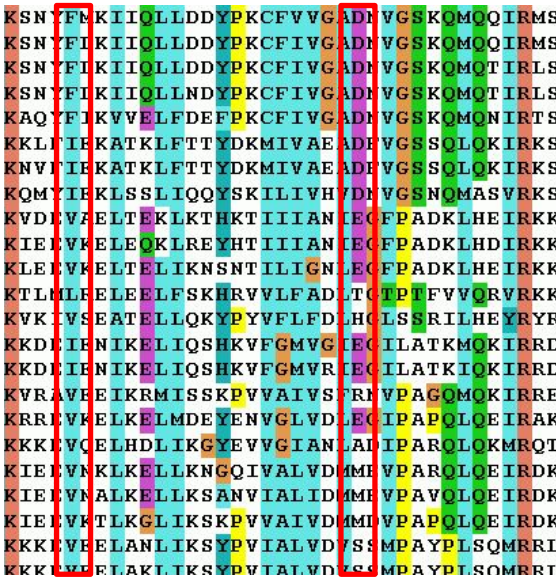
S-qua-ce Am--xnedt

Homology:
More than just genes!

HOMOLOGOUS genes
share a common ancestor

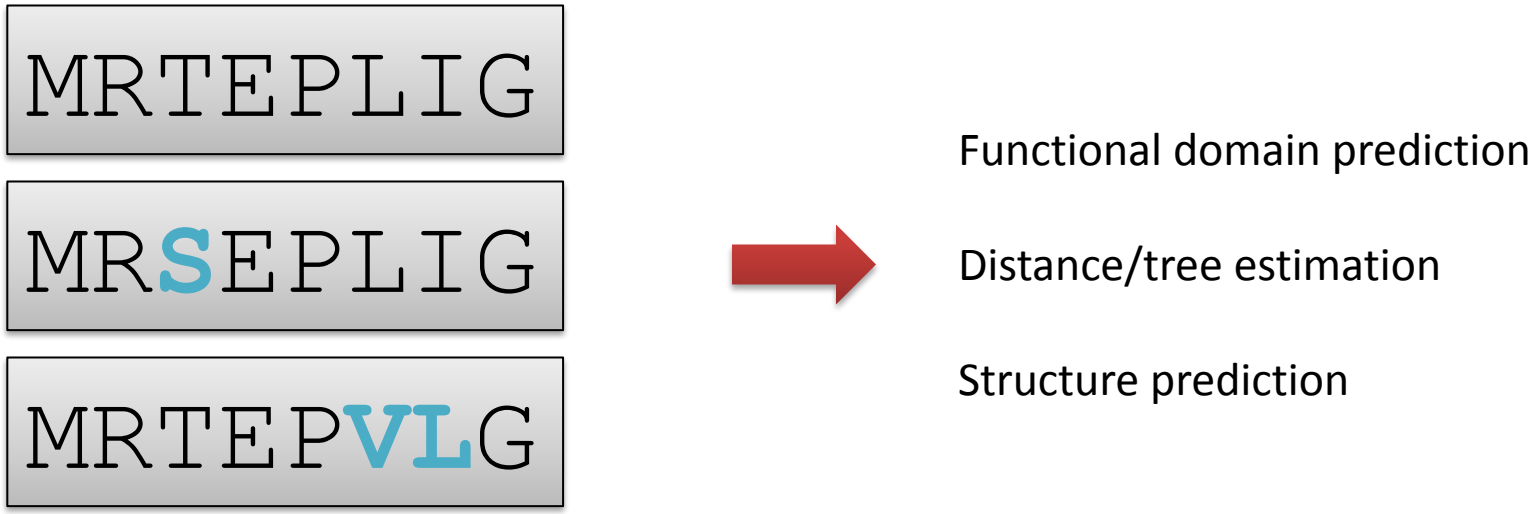


DNA / protein “residues”
(nucleotides and amino acids)
can also be homologous

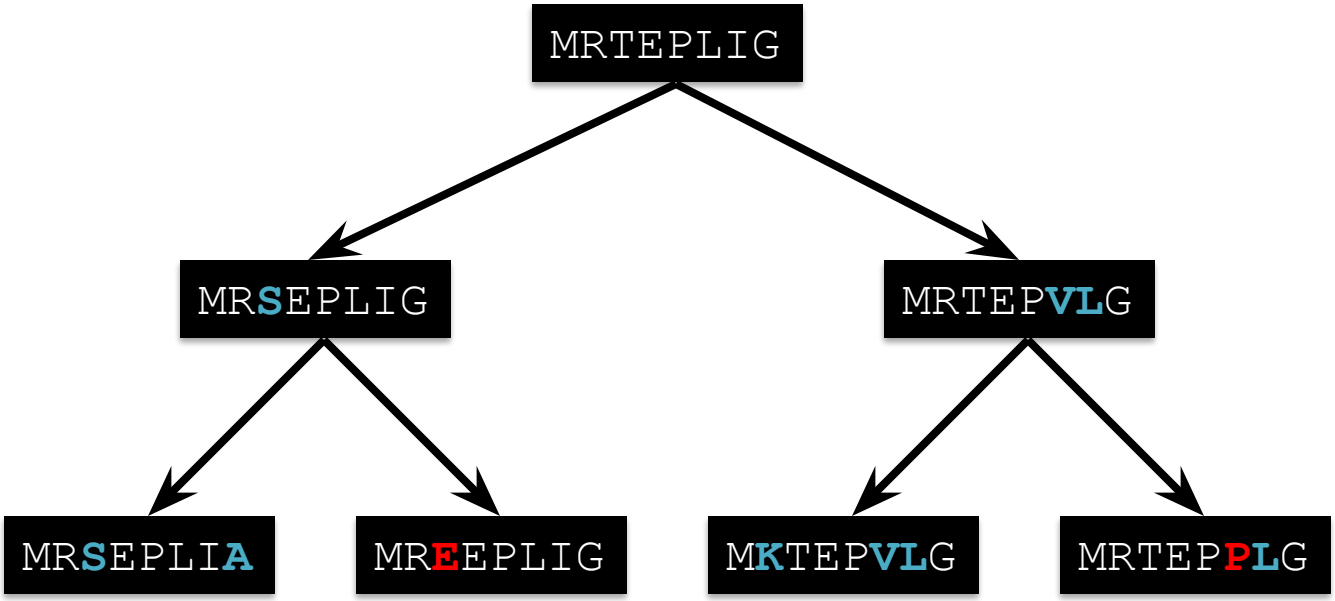


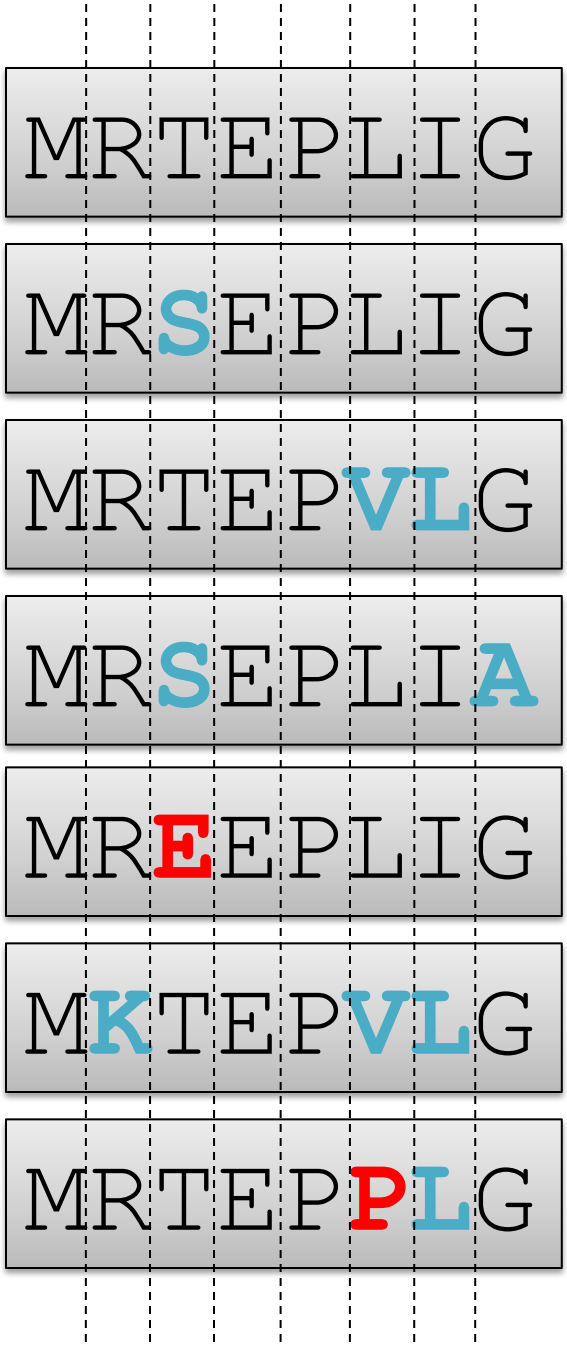
Each column is a homologous position
within the proteins

For many applications of sequence analysis, we would like to know which residues are **homologous** between sequences



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



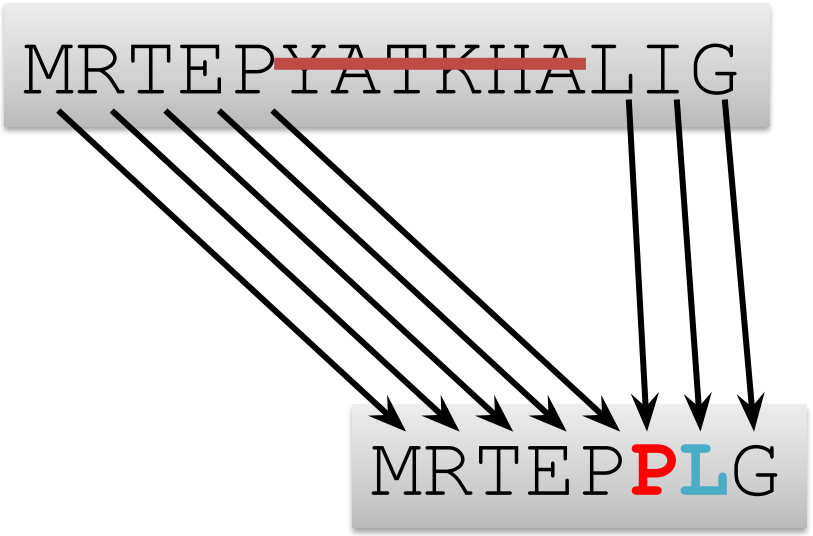


Each column contains a set of residues that are **homologous**

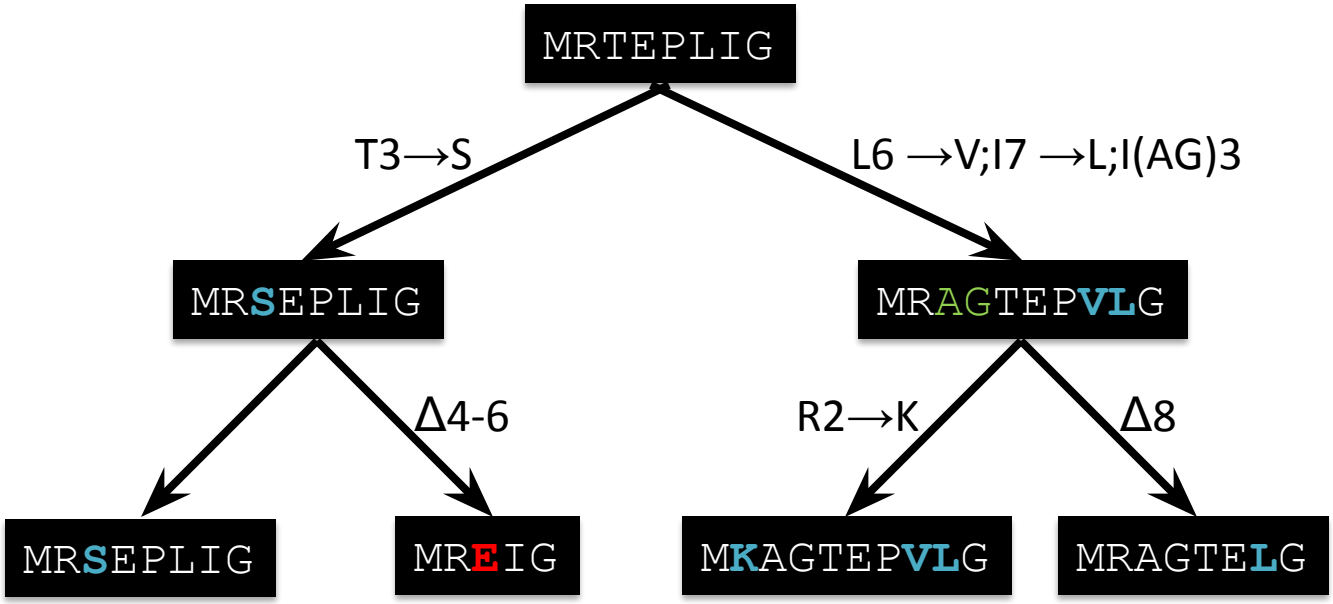
This is a **sequence alignment** (albeit a trivial one!)

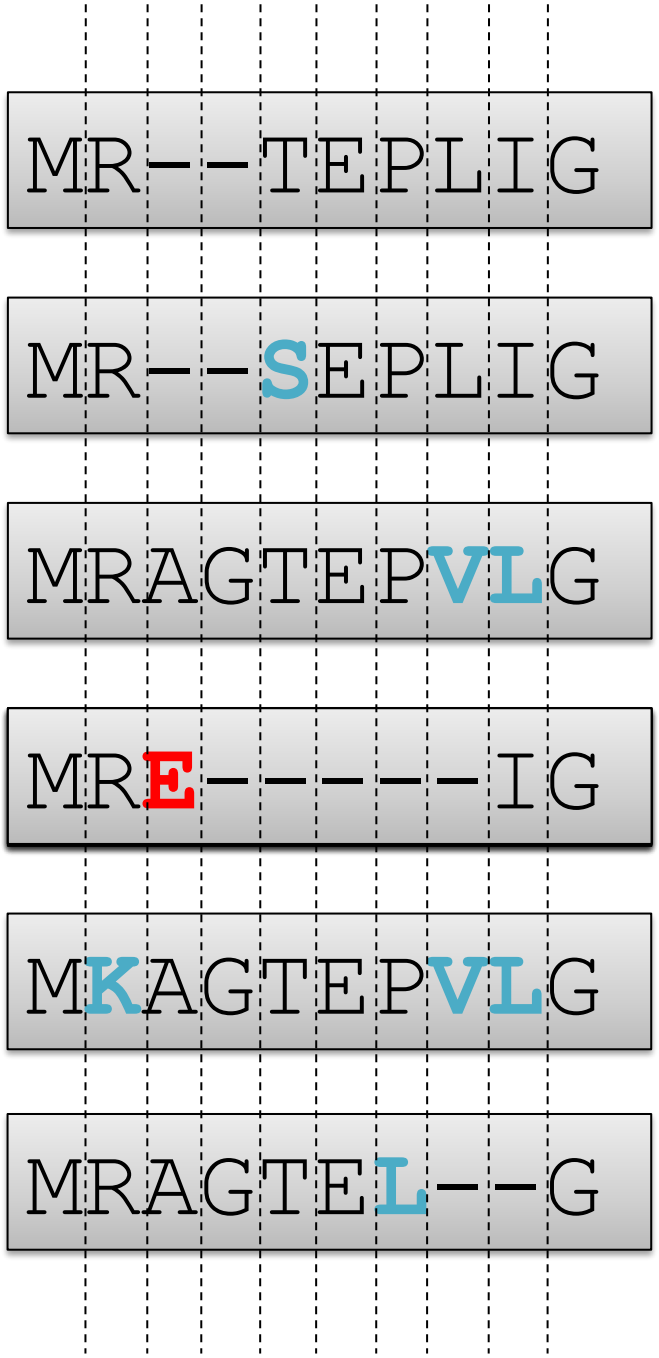
But Life is Not so Easy…

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
Δ8
MRSEPLIG
MREIG
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. We need 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. We also need 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

- Residue **frequencies** $f(x_i)$
  and **transition probabilities** $p(x_i,x_j)$

- A scheme G for penalizing **gaps**

- A formula for computing the **score**, given F, P, and G

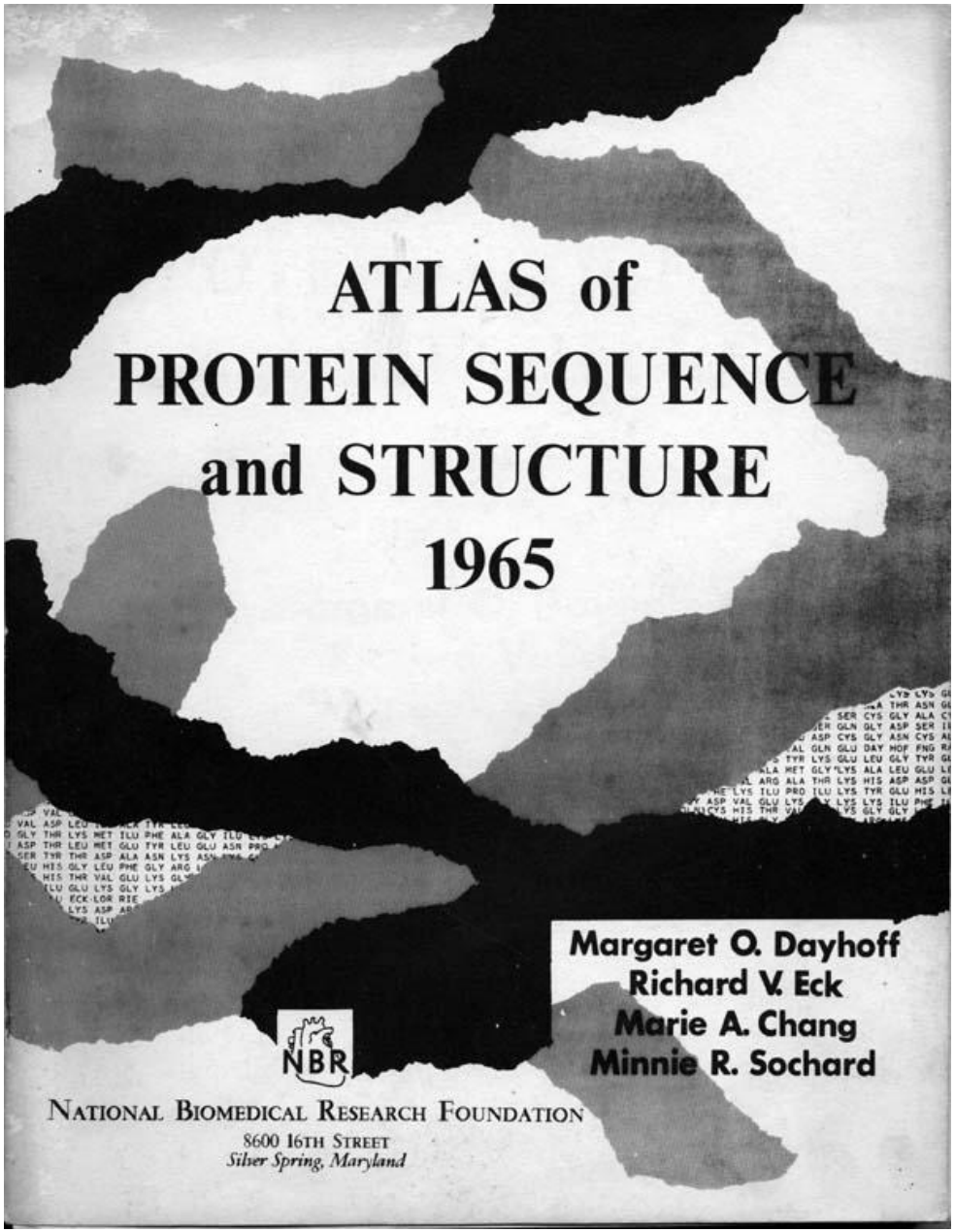
Part the first: substitution probabilities

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

First DNA gene sequence was 1972

“Responding to the sudden increase in the rate of nucleic acid sequencing, **Dr. Dayhoff established an on-line computer database and a sophisticated retrieval system, accessable 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.”

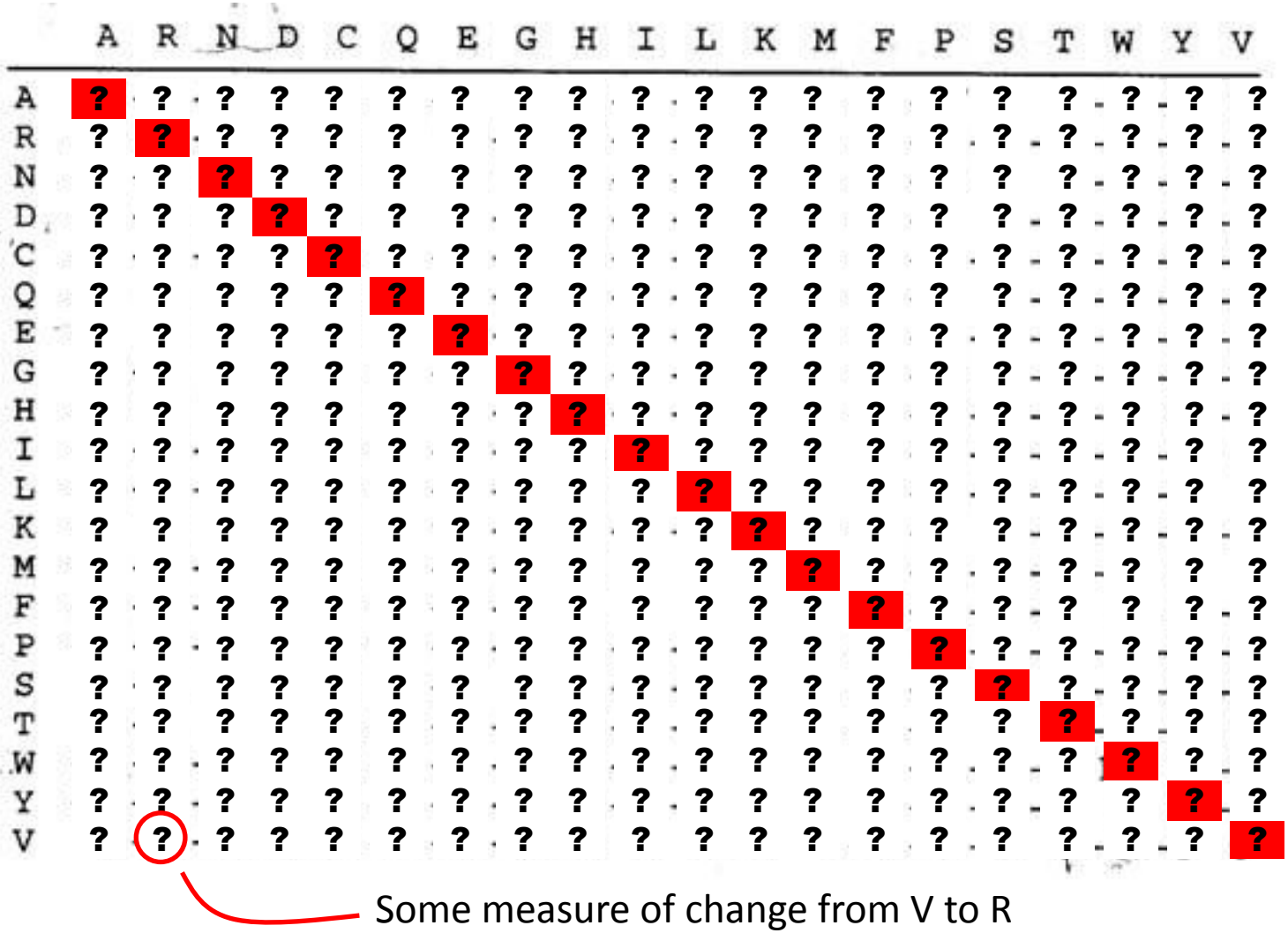
http://www.dayhoff.cc/MODBiography.html

Other Dayhoff

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

ACDEFGHKLMNPQRSTVWY

Building a Substitution Matrix



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

Huh?

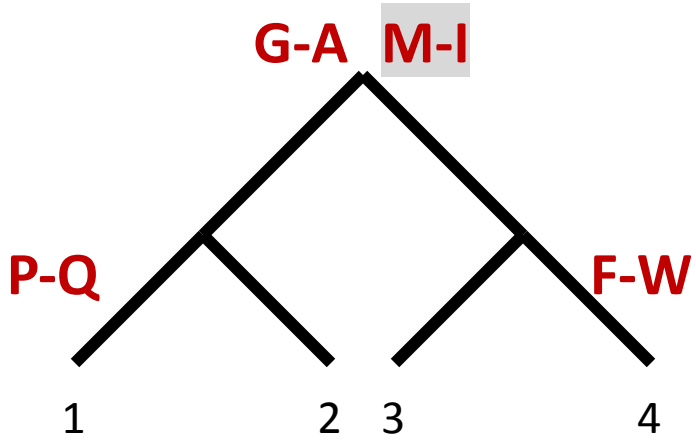
Building the PAM1 matrix

- 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  AAAILGMVFQ
2  AAAILGMVFP
3  AAGILGIVFP
4  AAGILGIVWP
```

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…

Matrix of Counts

$$A_{a,b} = s(a \rightarrow b) + s(b \rightarrow a)$$

A

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

DIAGONALS (no change) dominate
in closely related sequences

Matrix of Probabilities

Normalize by **row**, all row sums == 1

$$B_{a,b} = \frac{A_{a,b}}{\sum_c A_{a,c}}$$

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

Matrix of Scaled Probabilities (1 PAM)

The **amount** of evolution in **B** is arbitrary, based on whatever sequences we used to create our dataset

Rescale the matrix based on **frequencies** so the expected number of substitutions per site is equal to 0.01

Each off-diagonal element is multiplied by *c*, where

$$c = \frac{0.01}{\sum_a \sum_{b \neq a} f(a) B_{a,b}}$$

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

B

Total amount of change = ???

	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	…
…	…	…	…	…

	A	R	N	D	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9973

Upper left-hand corner of PAM1 probability matrix
(divide by 10,000 to get probabilities)

http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec03/node9.html

For higher-order PAM matrices:

$$PAMn = (PAM1)^n$$

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

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

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

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

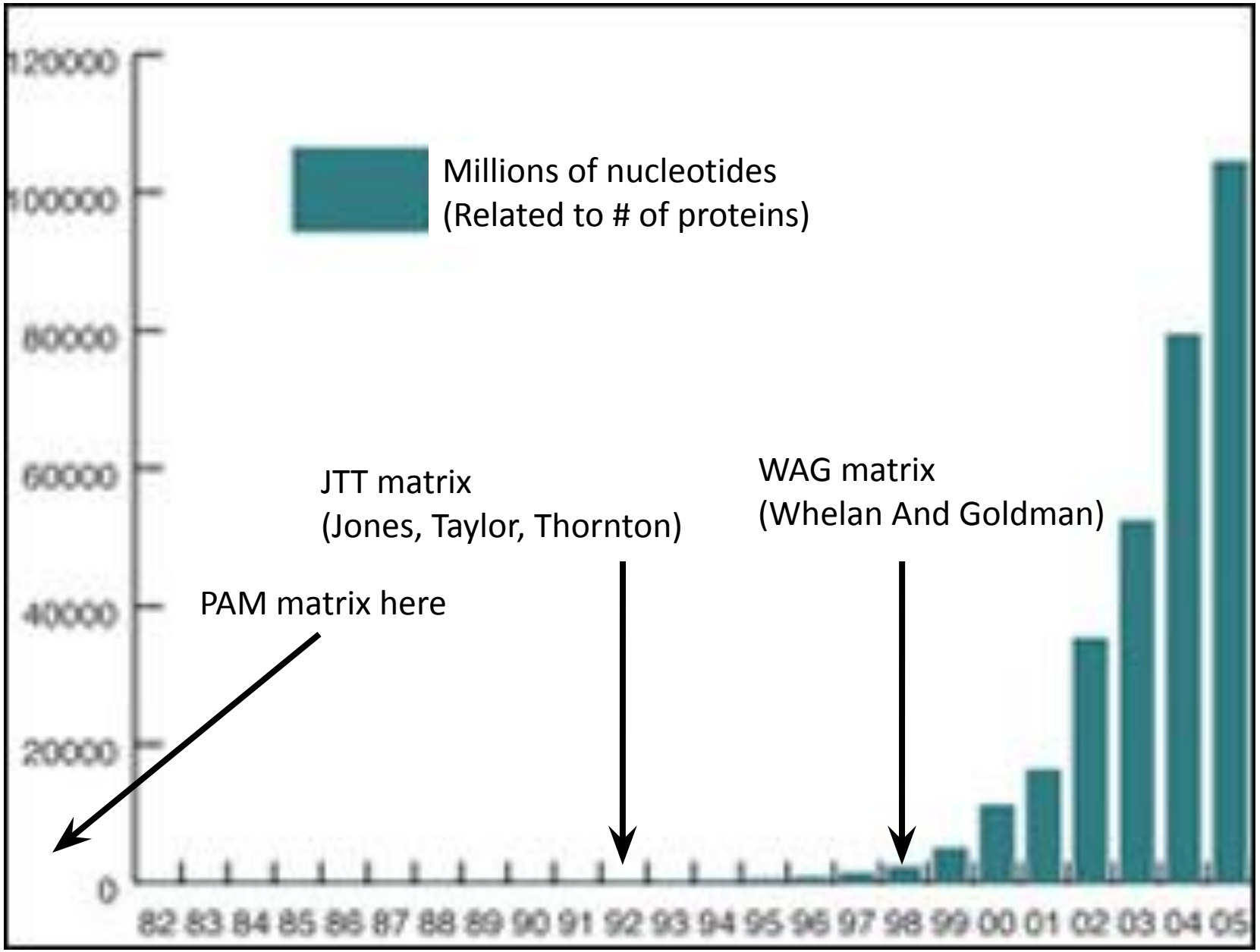
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

PAM250 matrix (S = 2, log base 2)
Half-bits

http://www.acm.org/crossroads/xrds13-1/gfx/13-1-11/Image7-BLOSUM62.jpg

Thoughts on PAM

Limitations?



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

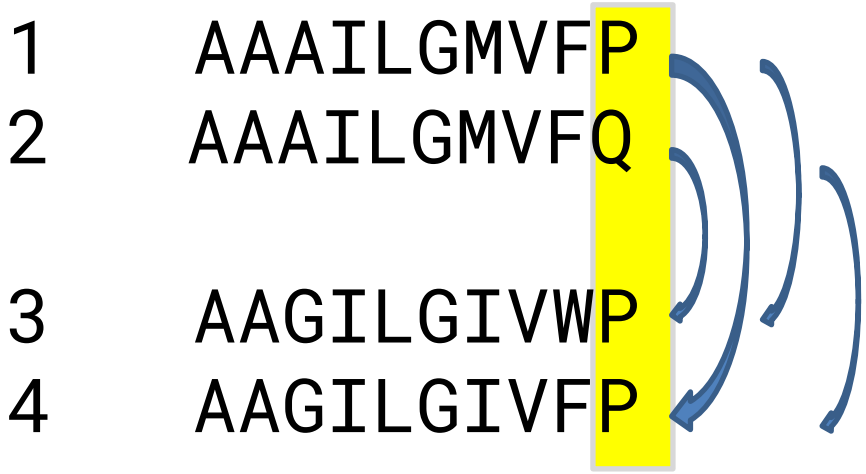
$$PAMn = (PAM1)^n$$

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



A

	P	Q
P	2	-
Q	2	0

A

	P	Q
P	2	-
Q	2	0

$$f(P) = \frac{A_{P,P} + \sum_{b \neq P} \frac{1}{2} A_{P,b}}{\sum_{c,d} A_{c,d}}$$

= 3/4

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

= 1/2

Why BLOSUM?

- No reliance on an inferred tree
- No extrapolation; differences are observed directly from alignments with at least L% divergence
- Choose matrix that matches alignment similarity

BLOSUM x Matrices

x = the % identity within blocks

BLOSUM 62 is based on more similar sequences than BLOSUM 50

(opposite of PAM!)

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 in a future module!)

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??

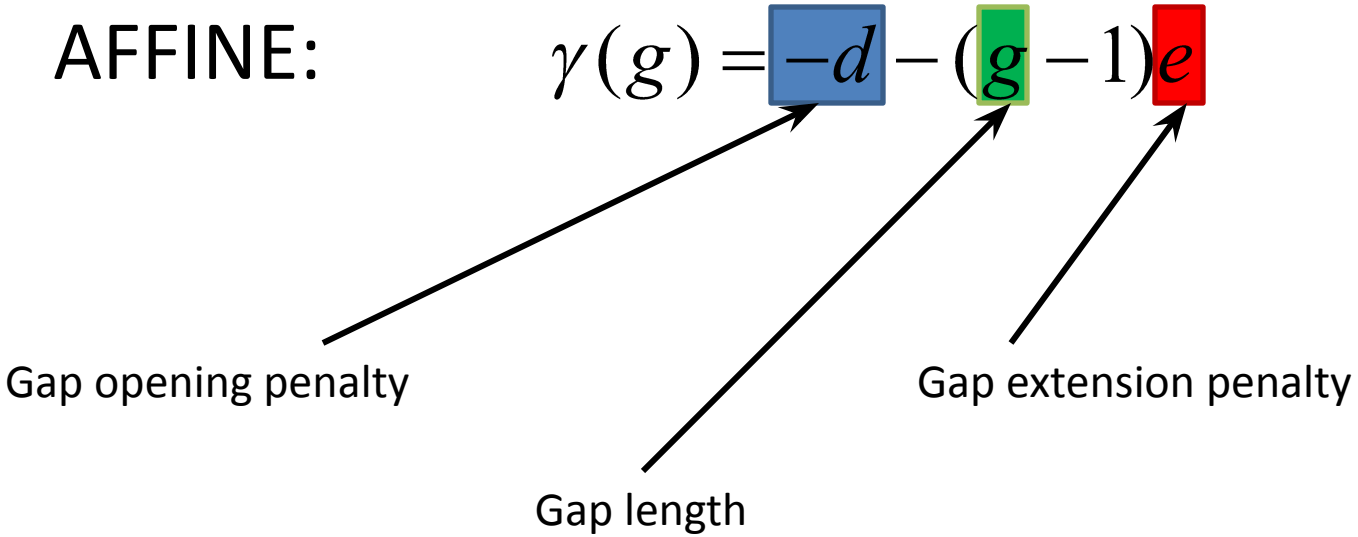
```
QVKQIYKTPPIKYFGGFNFSQILPDPSKPSKRSPIEDLLF---------
QVKQIYKTPPIK----------D--------------FGGFNFSQIL
```

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

X =

```
MKAGTEPVLG
MRAGTEL--G
```

S(X) = 18

Contrast with alignment Y:

Y =

```
MKAGTEPVLG
MRA--GTELG
```

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

S(Y) = 13

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

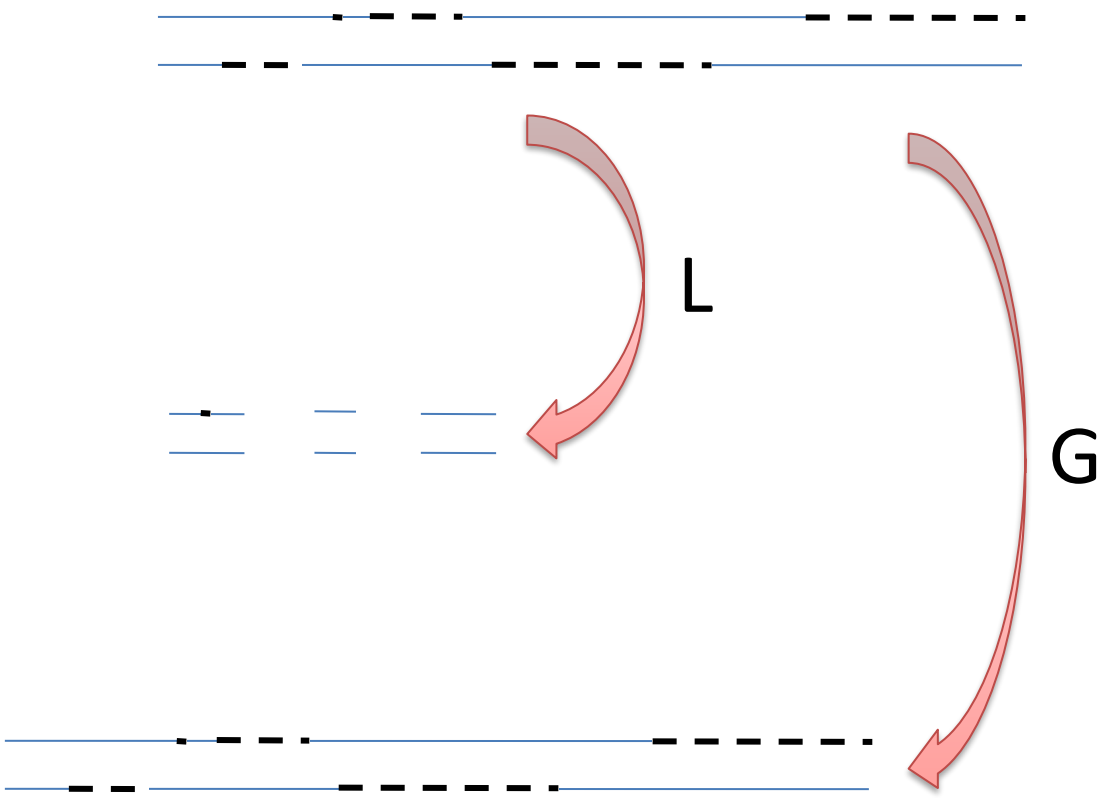
For protein-coding sequences, it is most common to align the amino acid sequences, then match the corresponding DNA codons against this sequence

¿Why?

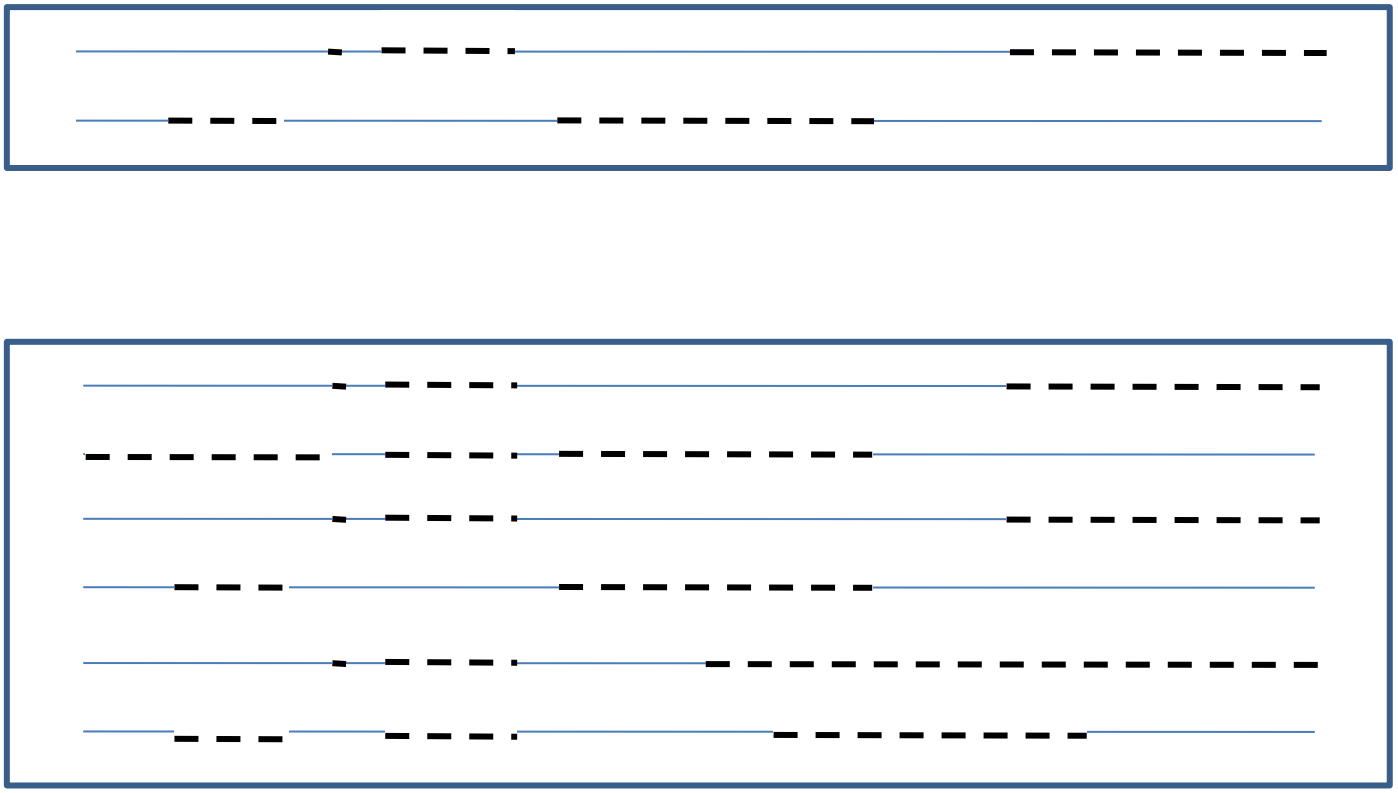
The goal of sequence alignment is (usually) to find the best alignment score – **maximize** the probability of observing aligned residues, relative to the **null model**

But optimal methods are slow – as you will see!

Global vs. Local alignment



Pairwise vs. Multiple alignment



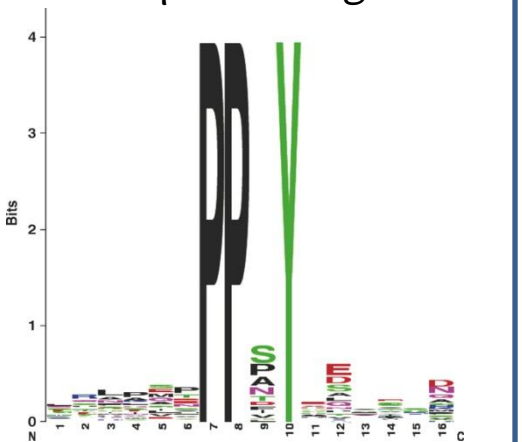
Alignment Representations

PWMs/PSSMs

PO	A	C	G	T
01	0.435	0.317	-0.128	-1.037
02	1.320	-3.121	0.349	-3.121
03	1.065	-3.121	0.301	-0.834
04	-3.121	-3.121	-3.121	1.870
05	-3.121	-3.121	1.870	-3.121
06	1.870	-3.121	-3.121	-3.121
07	-3.119	1.527	-3.119	-0.171
08	-3.121	-3.121	-3.121	1.870
09	-3.121	1.870	-3.121	-3.121
10	0.881	-0.061	-2.987	0.104

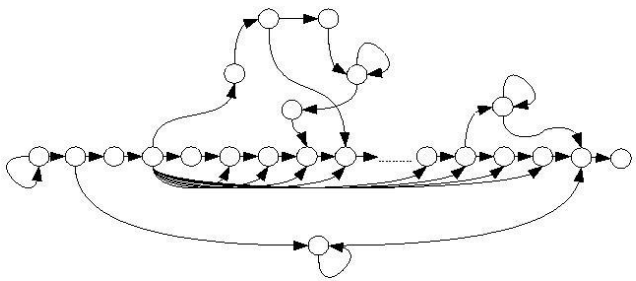
http://www.cbs.dtu.dk/courses/27619/project09.php

Sequence Logos



http://www.nature.com/msb/journal/v3/n1/images/msb4100159-f3.jpg

Hidden Markov Models



http://www.pdc.kth.se/~hakanv/modhmm/modhmm_web_pic.jpg

Overview

1. We need a SCORING SYSTEM for an alignment of two or more sequences
   - frequencies + substitutions + gaps = score
   - PAM/BLOSUM matrices capture the first two
   - Gap penalties can be linear or affine

2. But we still need algorithms that make use of our scoring system