Practical considerations of working with sequencing data

File Types

- Fastq ->aligner -> reference(genome) coordinates
- Coordinate files
    - SAM/BAM – most complete, contains all of the info in fastq and more!
    - Bedgraph – read density along the genome
    - Bed file –Read density reported in large continuous intervals
        - Genes/transcript and transcript structure
        - Transcription factor binding regions
- If someone does a sequencing experiment usually one of these is available and deposited in a public database

SAM/BAM

(1) The query name of the read is given (`M01121...`)

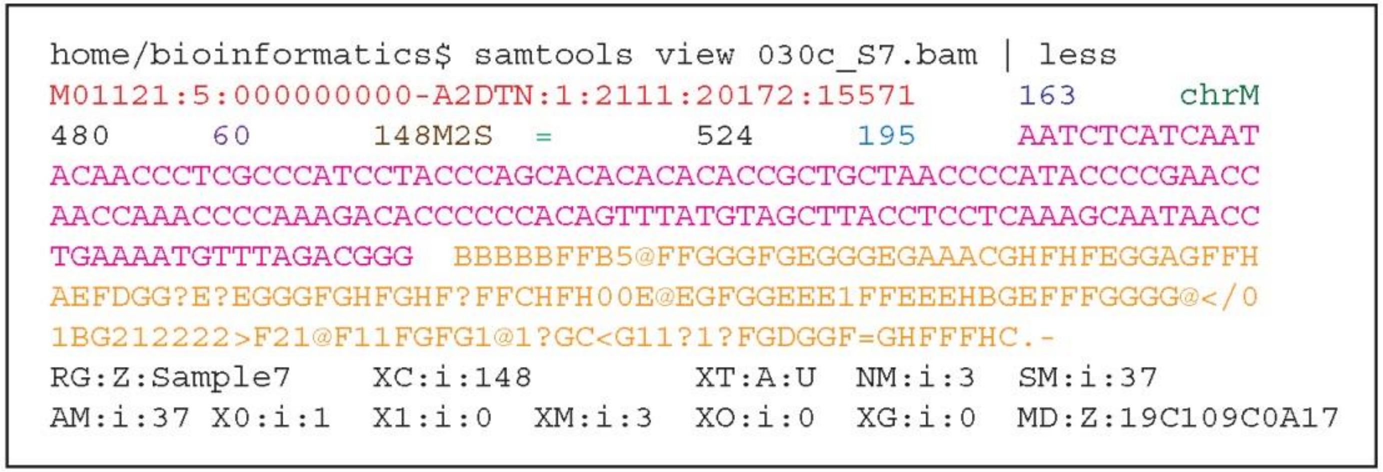
(2) The flag value is `163` (this equals 1+2+32+128)

(3) The reference sequence name, `chrM`, refers to the mitochondrial genome

(4) Position `480` is the left-most coordinate position of this read

(5) The Phred-scaled mapping quality is `60` (an error rate of 1 in 10^6)

(6) The CIGAR string (`148M2S`) shows 148 matches and 2 soft-clipped (unaligned) bases



```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571     163     chrM
480      60       148M2S   =         524      195      AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG   BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFH00E@EGFGGEEE1FFEEEHBGEFFFGGGG@</0
1BG212222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-
RG:Z:Sample7     XC:i:148         XT:A:U   NM:i:3   SM:i:37
AM:i:37 X0:i:1   X1:i:0   XM:i:3   XO:i:0   XG:i:0   MD:Z:19C109C0A17
```

(7) An = sign shows that the mate reference matches the reference name

(8) The 1-based left position is `524`

(9) The insert size is `195` bases

(10) The sequence begins `AATCT` and ends `ACGGG` (its length is 150 bases)

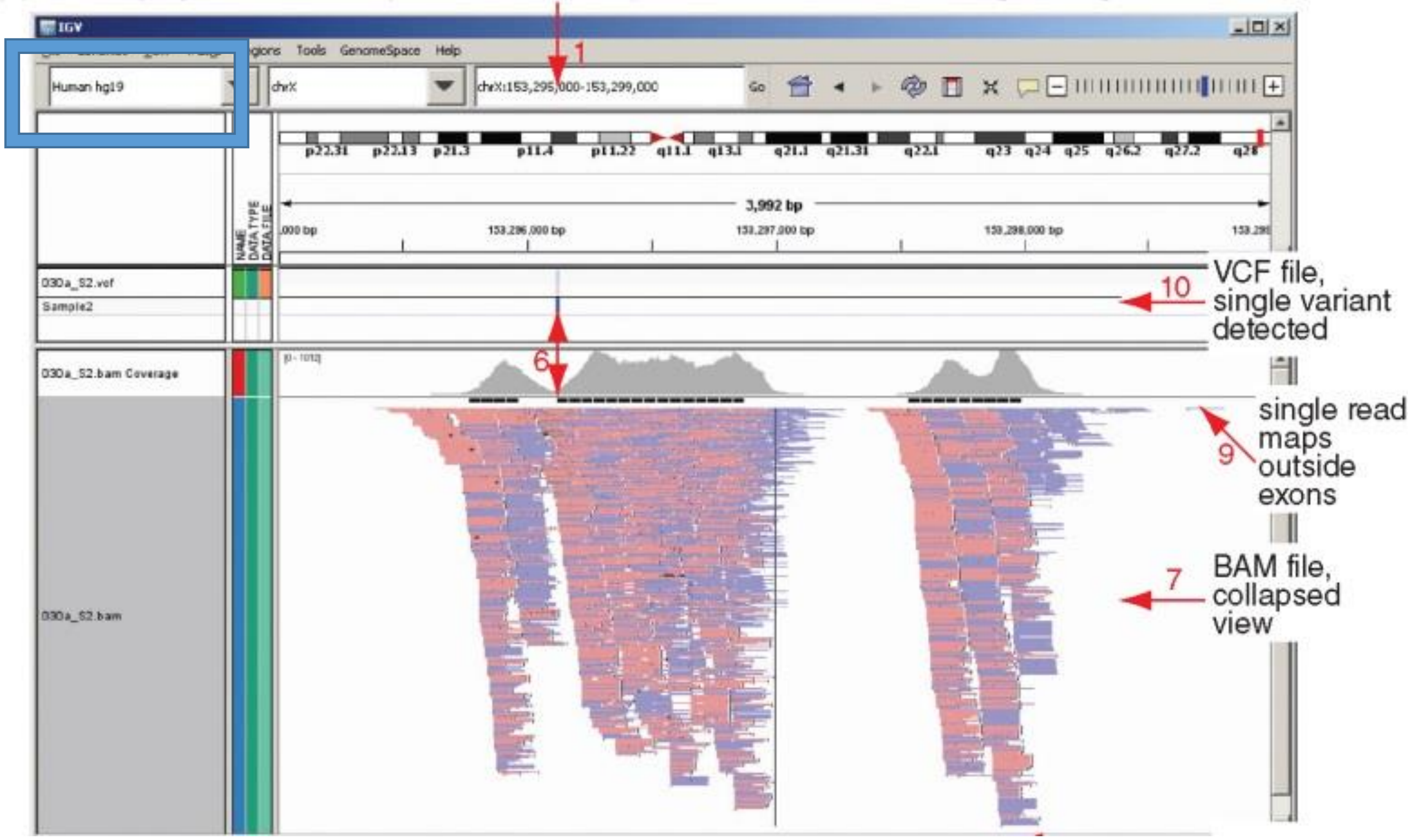
(11) Each base is assigned a quality score (from BBBBB ending `FHC.-`)

(12) This read has additional, optional fields that accompany the MiSeq analysis

Viewing genome coordinate files with IGV

- Integrated Genome Browser
- Cross-platform application
- Knows about common genomes
- Genome version is important!

(a) IGV display of a BAM file (at two resolutions) and a VCF in the *MECP2* gene region



Human hg19

chrX

chrX:153,295,000-153,299,000

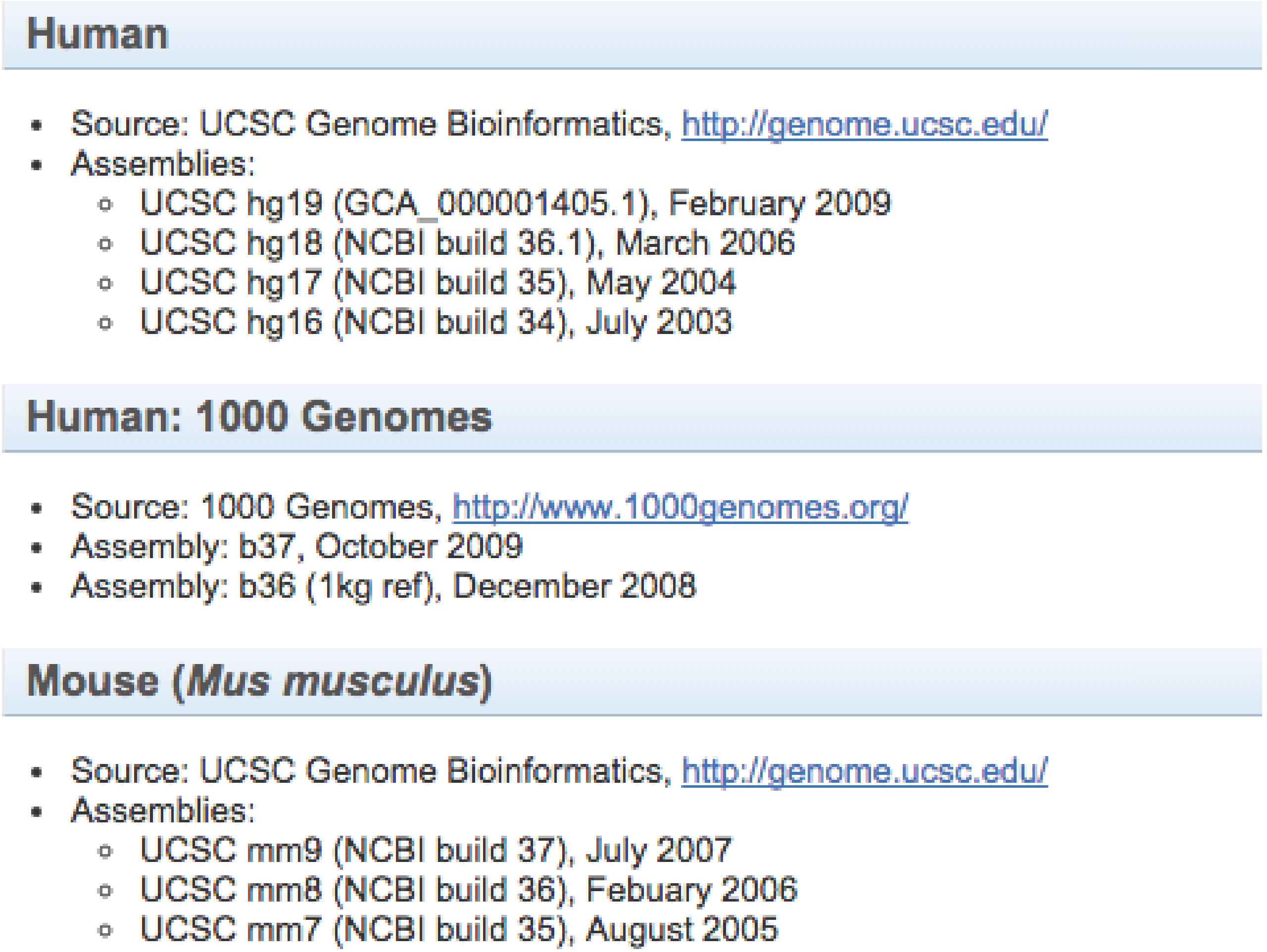
VCF file, single variant detected

single read maps outside exons

BAM file, collapsed view

Different assemblies

- Genome coordinates different between genome assemblies
  - Differences accumulate over chromosome length
- You have to know which assembly was used
- Sequencing files are non-randomly distributed relative to genes
  - RNAseq—should align with exons
  - TF binding sites—biased towards promoter regions



Human

- Source: UCSC Genome Bioinformatics, http://genome.ucsc.edu/
- Assemblies:
  - UCSC hg19 (GCA_000001405.1), February 2009
  - UCSC hg18 (NCBI build 36.1), March 2006
  - UCSC hg17 (NCBI build 35), May 2004
  - UCSC hg16 (NCBI build 34), July 2003

Human: 1000 Genomes

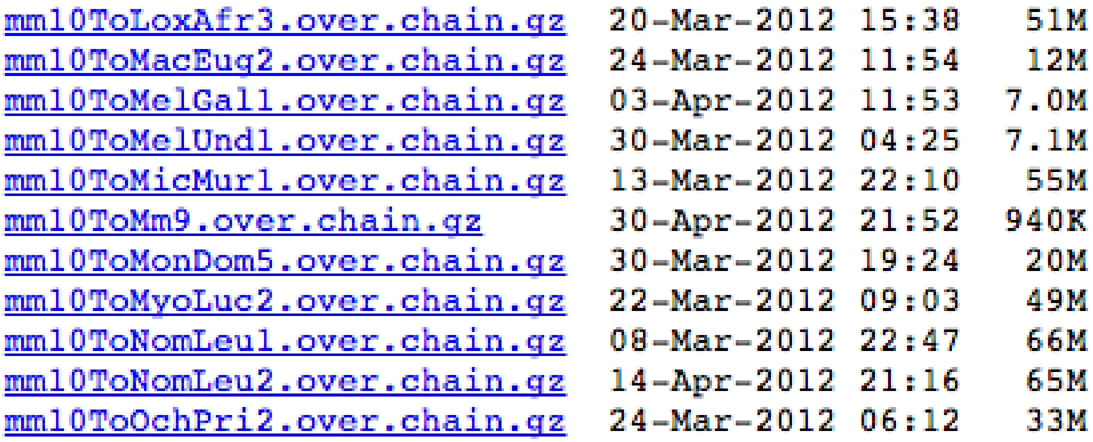
- Source: 1000 Genomes, http://www.1000genomes.org/
- Assembly: b37, October 2009
- Assembly: b36 (1kg ref), December 2008

Mouse (*Mus musculus*)

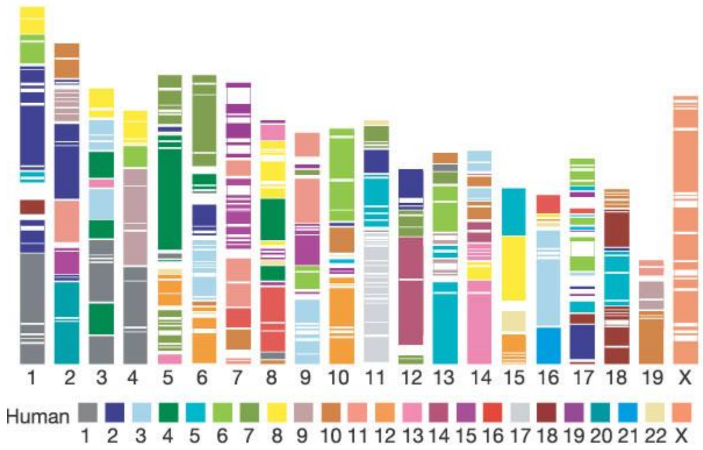
- Source: UCSC Genome Bioinformatics, http://genome.ucsc.edu/
- Assemblies:
  - UCSC mm9 (NCBI build 37), July 2007
  - UCSC mm8 (NCBI build 36), Febuary 2006
  - UCSC mm7 (NCBI build 35), August 2005

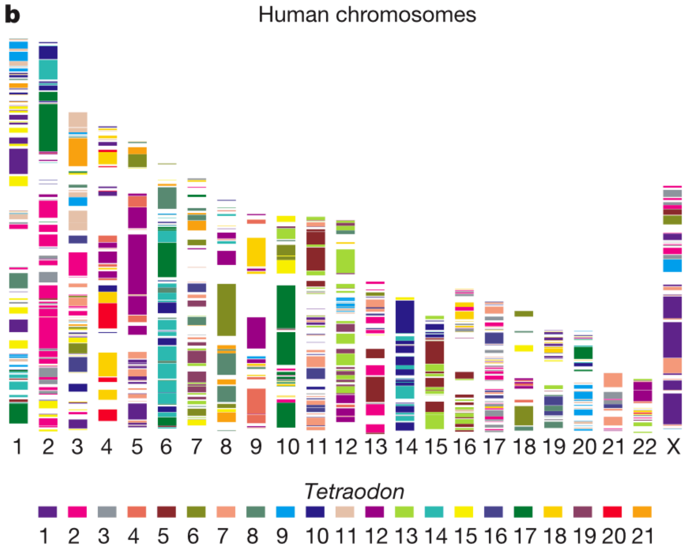
Converting coordinates

- UCSC liftOver -- converts genome coordinates
- Convert from one assembly to another
- Cross organism conversion
    - Mammals/vertebrates



```
mm10ToLoxAfr3.over.chain.gz     20-Mar-2012 15:38    51M
mm10ToMacEug2.over.chain.gz     24-Mar-2012 11:54    12M
mm10ToMelGal1.over.chain.gz     03-Apr-2012 11:53   7.0M
mm10ToMelUnd1.over.chain.gz     30-Mar-2012 04:25   7.1M
mm10ToMicMur1.over.chain.gz     13-Mar-2012 22:10    55M
mm10ToMm9.over.chain.gz         30-Apr-2012 21:52   940K
mm10ToMonDom5.over.chain.gz     30-Mar-2012 19:24    20M
mm10ToMyoLuc2.over.chain.gz     22-Mar-2012 09:03    49M
mm10ToNomLeu1.over.chain.gz     08-Mar-2012 22:47    66M
mm10ToNomLeu2.over.chain.gz     14-Apr-2012 21:16    65M
mm10ToOchPri2.over.chain.gz     24-Mar-2012 06:12    33M
```





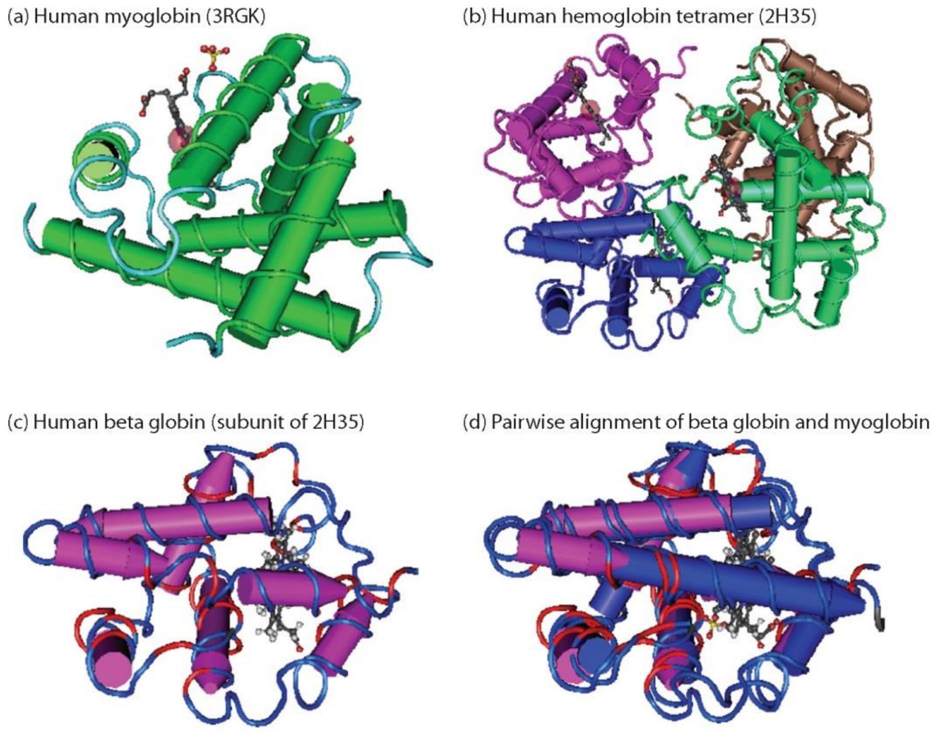
Sequence Alignment

To do:

- Global alignment
- Local alignment
- Scoring
  - Gaps
  - Scoring matrices
- Database Search
  - Statistical Significance
- Multiple Sequence alignment

Why compare sequences

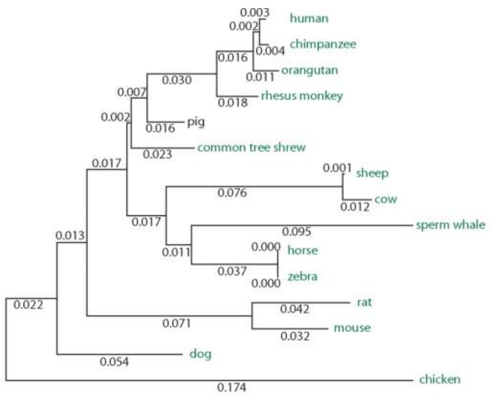
- Given a new sequence, infer its function based on similarity to another sequence
- Find important molecular regions – conserved across species
- Determine 3d structure with homology modeling
- **Homologs-**sequences that descended from a common ancestral sequence
  - **Orthologs**- separated by speciation
  - **Paralogs** separated by duplication in a single genome
- Basic unit of protein homology is a sufficient functional unit—typically much smaller than a whole gene

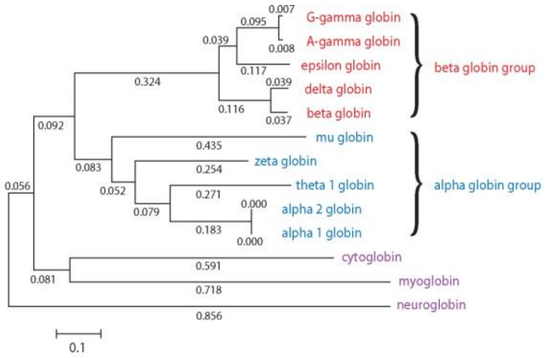

(a) Human myoglobin (3RGK)

(b) Human hemoglobin tetramer (2H35)

(c) Human beta globin (subunit of 2H35)

(d) Pairwise alignment of beta globin and myoglobin



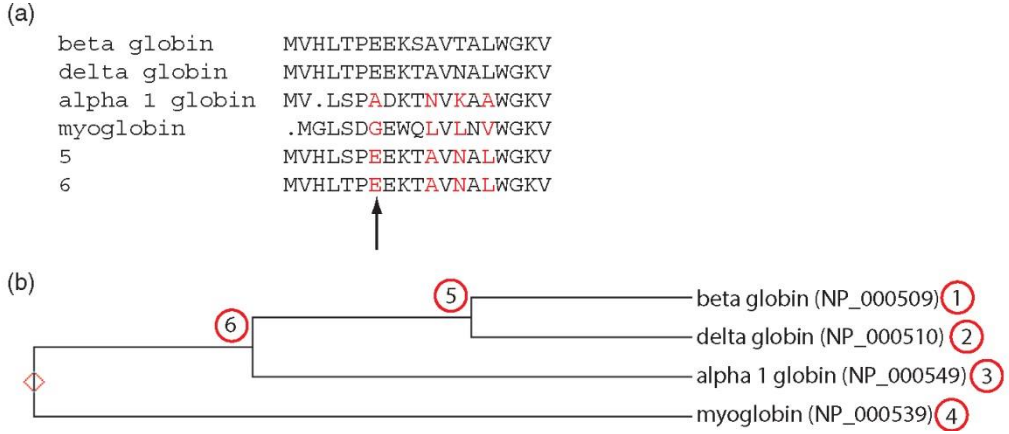


DNA vs Protein alignments

- Protein coding
  - Typically compared in amino acid space
  - Amino acid change slower than nucleotides
    - Some nucleotides can change without any change to a.a. sequence
  - Different levels of amino acid similarity can be accounted for
    - Not all a.a. changes are equally disruptive
  - Can detect very remote homology
- Non-coding regions
  - Smaller alphabet requires more matches to achieve significance
  - No notion of similarity—match or nor match
  - Diverge more rapidly though some are very conserved at short evolutionary distances

What is a good sequence alignment

- Theory: If two sequences are homologous we want to match up the residues such that each residue is descendant from a common ancestral residue
- Practice: approximate string matching
  - introduce gaps and padding to find best matching between two strings



(a)

```
beta globin      MVHLTPEEKSAVTALWGKV
delta globin     MVHLTPEEKTAVNALWGKV
alpha 1 globin   MV.LSPADKTNVKAAWGKV
myoglobin        .MGLSDGEWQLVLNVWGKV
5                MVHLSPEEKTAVNALWGKV
6                MVHLTPEEKTAVNALWGKV
```

(b)

5
6
beta globin (NP_000509) 1
delta globin (NP_000510) 2
alpha 1 globin (NP_000549) 3
myoglobin (NP_000539) 4

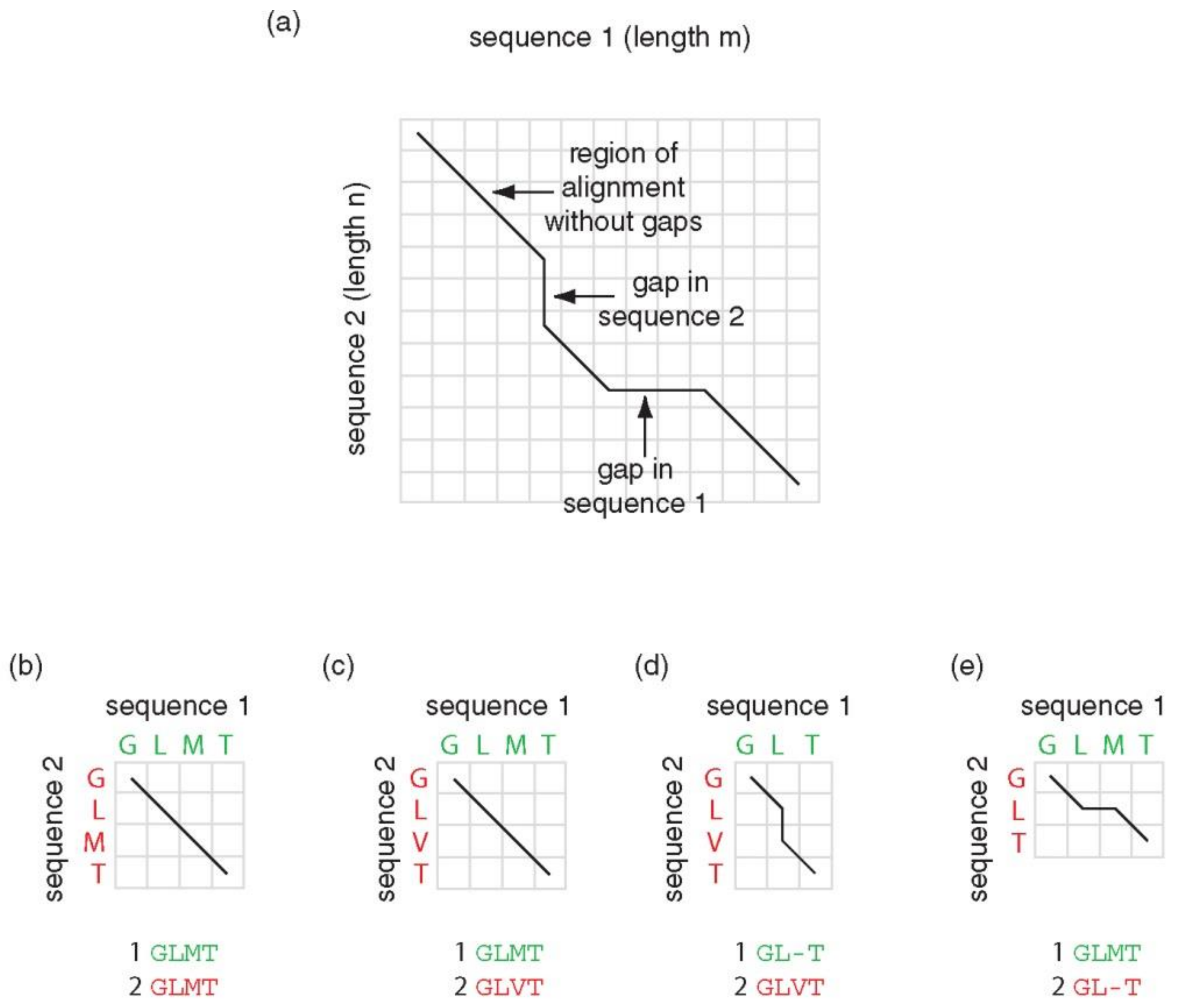
```
AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

Efficient alignment

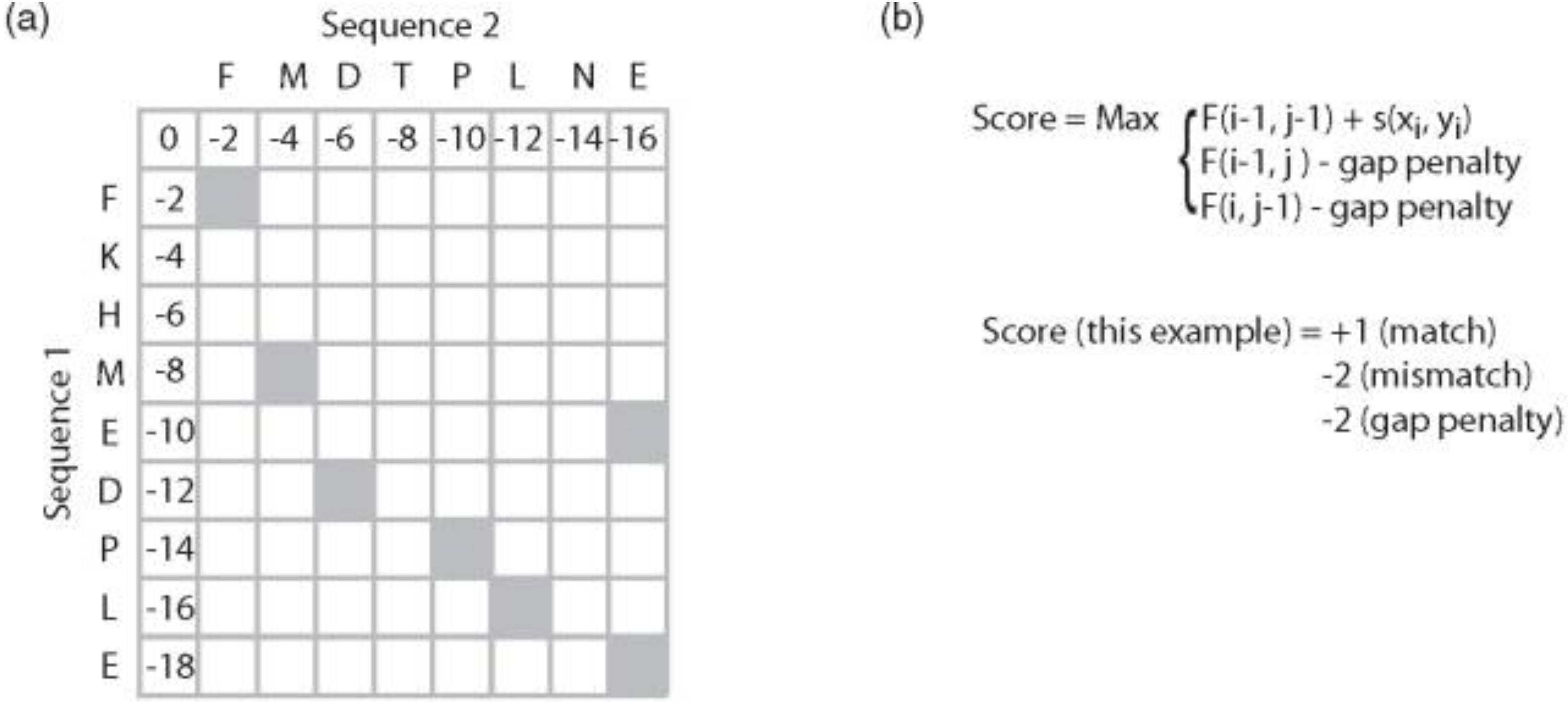
- What is the best alignment? – we need a scoring metric
  - Basic scoring metric (1 for matching, 0 for mismatching, 0 for a gap)
- Number of possible alignments is exponential in string length
- Scoring is local
- we apply dynamic programming
- dynamic programming –solve a large problem in terms of smaller subproblems
- Requirements
  - There is only a polynomial number of subproblems
    - Align $x_1 \ldots x_i$ to $y_1 \ldots y_j$
  - Original problem is one of the subproblems
    - Align $x_1 \ldots x_M$ to $y_1 \ldots y_N$
  - Each subproblem is easily solved from smaller subproblems

Matrix representation of an alignment



Dynamical programming approach

- Score the optimal alignment up to every (i,j) F(i,j)
- Scoring is local so F(i,j) depends only 3 other values



(a)

Sequence 2

Sequence 1

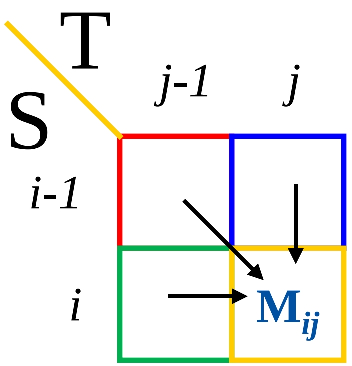
		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	▨							
K	-4								
H	-6								
M	-8		▨						
E	-10								▨
D	-12			▨					
P	-14					▨			
L	-16						▨		
E	-18								▨

(b)

$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j\,) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
-2 (mismatch)
-2 (gap penalty)

Global alignment



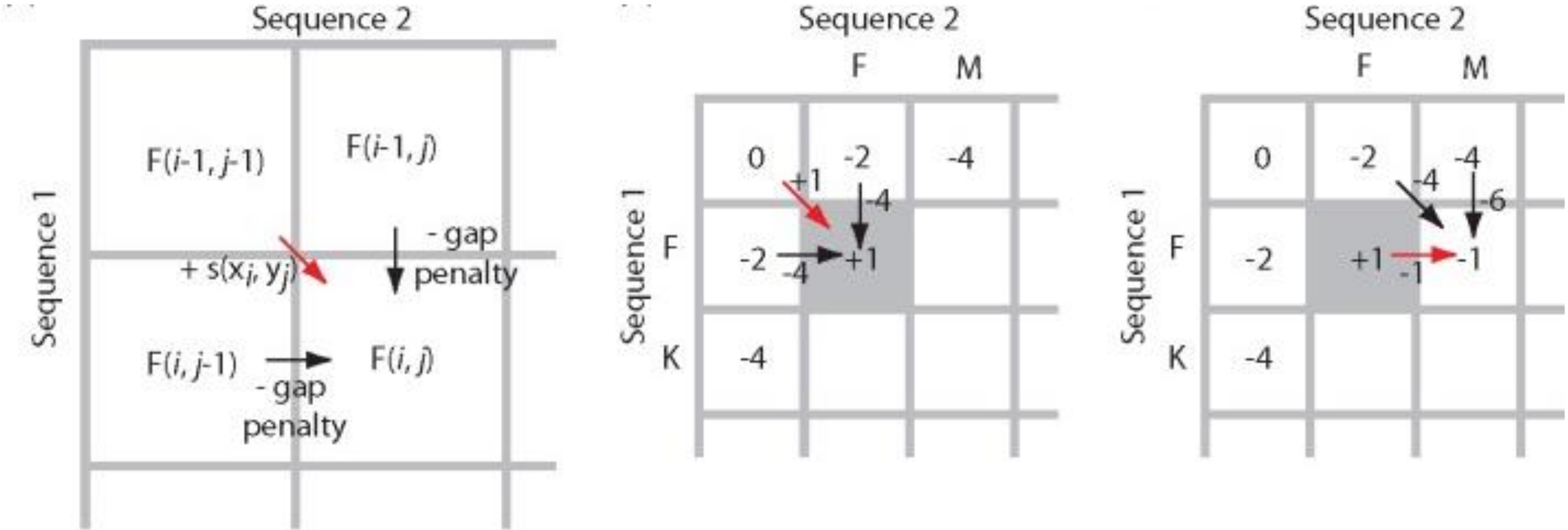
$$F_{i,j} = \text{MAX} \begin{cases} F_{i-1,\, j-1} + \text{Score}(S_i, T_j) \\ F_{i,j-1} + gp \\ F_{i-1,j} + gp \end{cases}$$

Gap penalty

Needleman & Wunsch, 1970

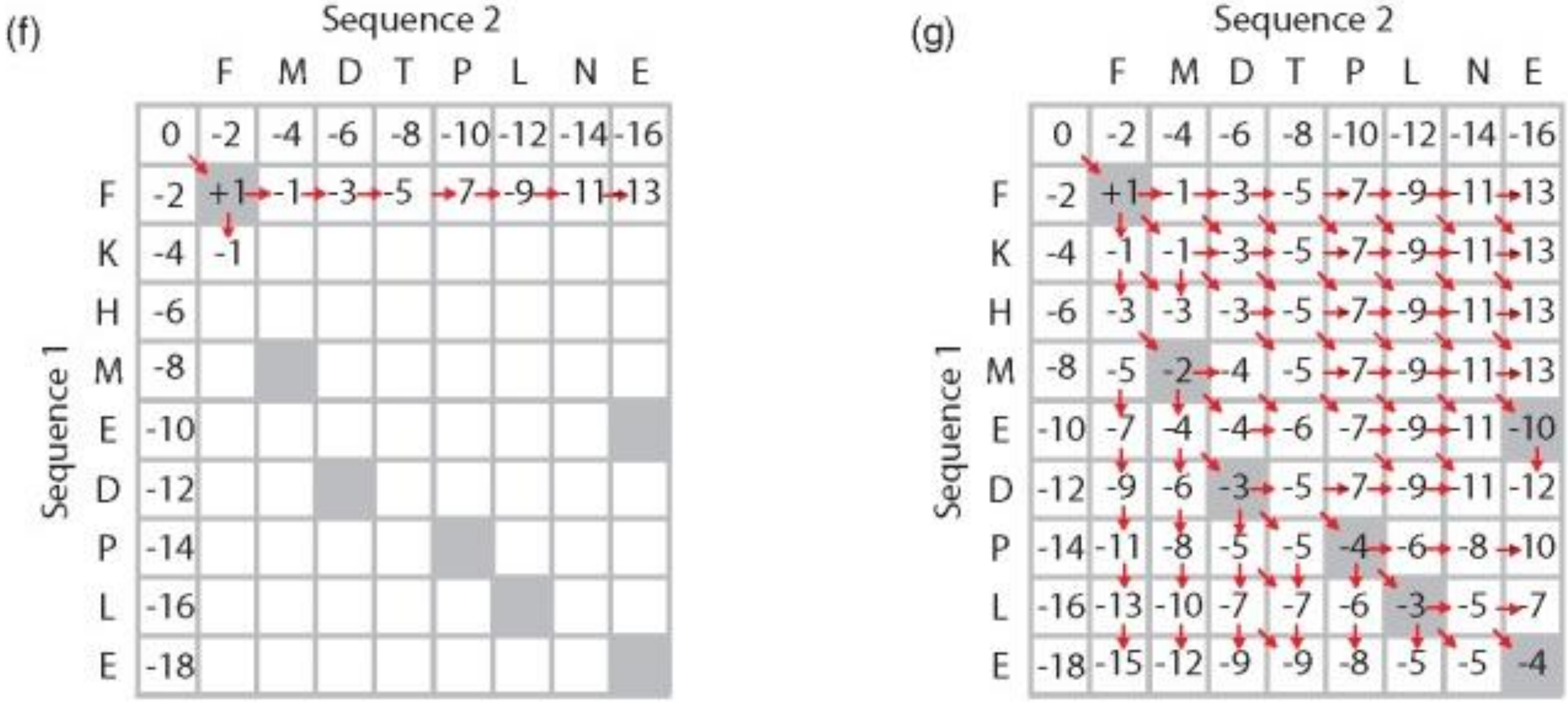
Example

Score (this example) = +1 (match)
-2 (mismatch)
-2 (gap penalty)



Keep track of the argmax!

Matrix filled out



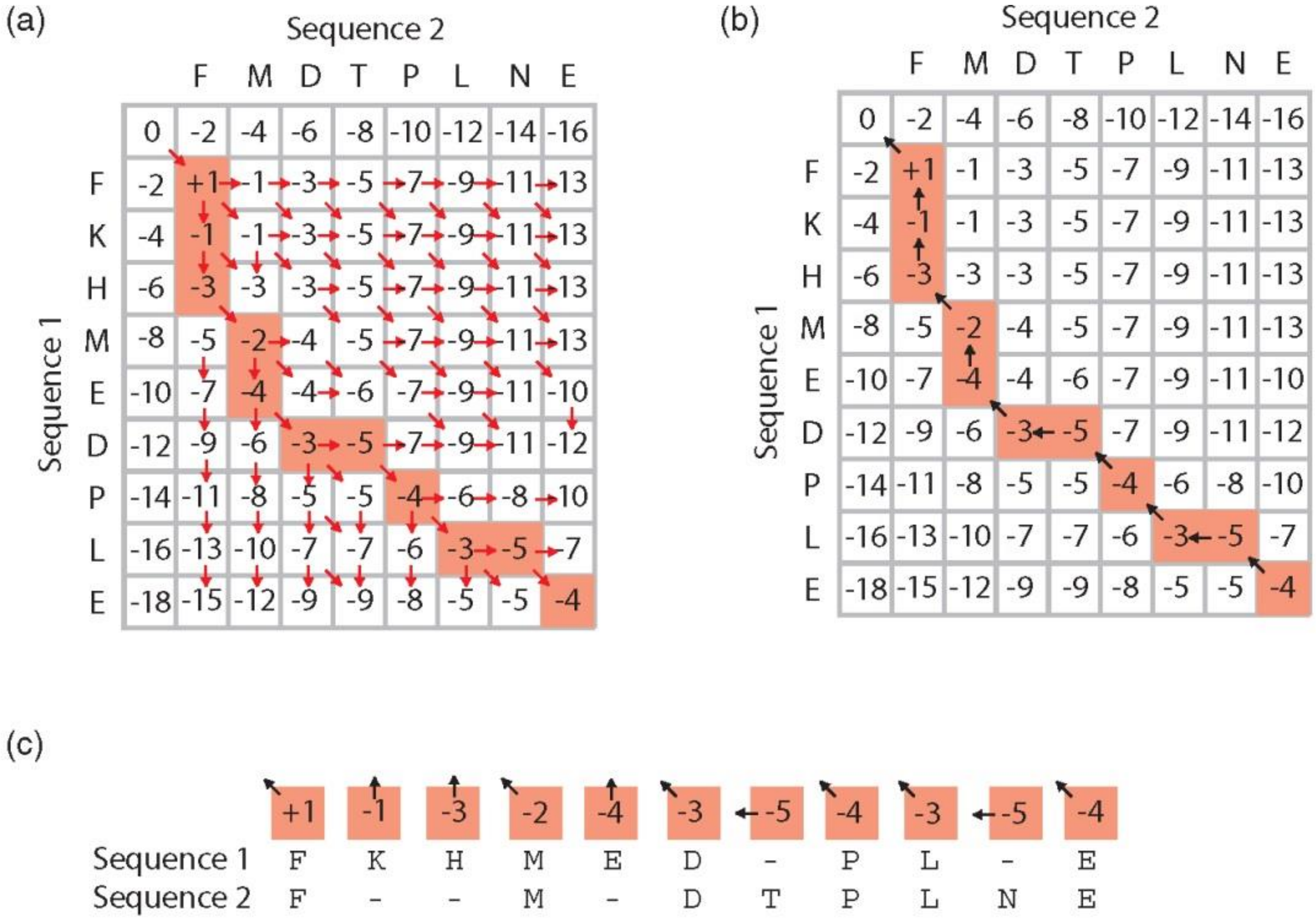
(f)

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1							
H	-6								
M	-8								
E	-10								
D	-12								
P	-14								
L	-16								
E	-18								

(g)

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

Finding the optimal alignment



(a)

Sequence 2

Sequence 1

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

(b)

Sequence 2

Sequence 1

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

(c)

	↖ +1	↑ -1	↑ -3	↖ -2	↑ -4	↖ -3	← -5	↖ -4	↖ -3	← -5	↖ -4
Sequence 1	F	K	H	M	E	D	-	P	L	-	E
Sequence 2	F	-	-	M	-	D	T	P	L	N	E

Complete Algorithm

- Initialization.

```
F(0,0) =0
F(0, j) = - j × go
F(i, 0) = - i × go
```

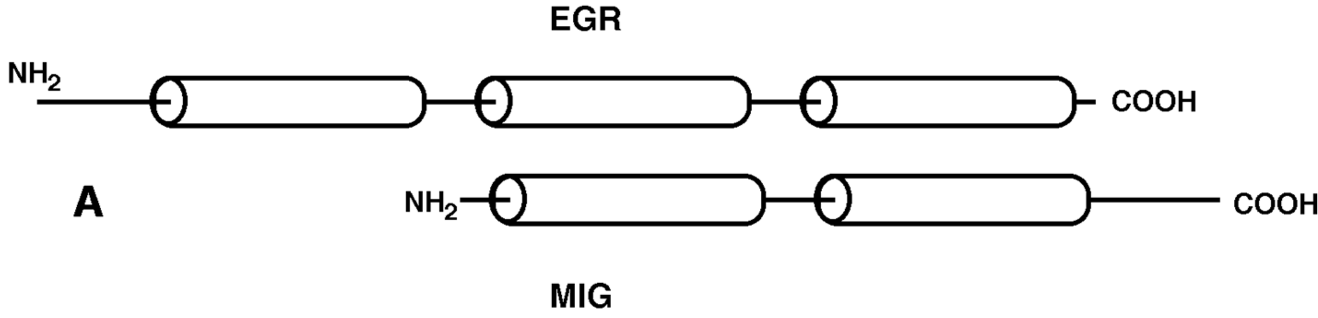
- Main Iteration. Filling-in partial alignments

```
For each i=1......M
    For each j = 1......N
        F(i, j) = max(F(i-1,j-1)+s(xi, yj)…
                                      F(i-1, j) – gp,…
                                                        F(i, j-1) – gp)

        Ptr(i,j) =DIAG                LEFT              UP
```

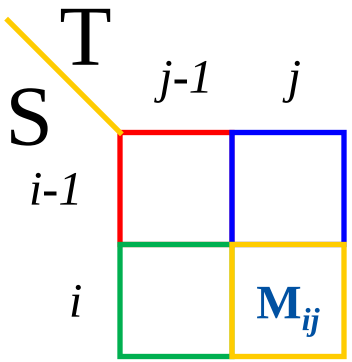
Local alignment

- Given two sequences, S and T, find two subsequences, s and t, whose alignment has the highest “score” amongst all subsequence pairs.
- Two genes in different species may be similar over short conserved regions and dissimilar over remaining regions.
- Example:
  - Homeobox genes have a short region called the *homeodomain* that is highly conserved between species.
  - A global alignment would not find the homeodomain because it would try to align the ENTIRE sequence
- Genes can have local similarity because of variable domain composition



```
EGR4_HUMAN   KA [FACPVESCVRSFARSDELNRHLRIH] TGHKP [FQCRICLRNFSRSDHLTSHVRTH] TGEKP [FACDV--CGRRFARSDEKKRHSKVH]
EGR4_RAT     KA [FACPVESCVRTFARSDELNRHLRIH] TGHKP [FQCRICLRNFSRSDHLTTHVRTH] TGEKP [FACDV--CGRRFARSDEKKRHSKVH]
EGR3_HUMAN   RP [HACPAEGCDRRFSRSDELTRHLRIH] TGHKP [FQCRICMRSFSRSDHLTTHIRTH] TGEKP [FACEF--CGRKFARSDERKRHAKIH]
EGR3_RAT     RP [HACPAEGCDRRFSRSDELTRHLRIH] TGHKP [FQCRICMRSFSRSDHLTTHIRTH] TGEKP [FACEF--CGRKFARSDERKRHAKIH]
EGR1_HUMAN   RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDI--CGRKFARSDERKRHTKIH]
EGR1_MOUSE   RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDI--CGRKFARSDERKRHTKIH]
EGR1_RAT     RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDI--CGRKFARSDERKRHTKIH]
EGR1_BRARE   RP [YACPVETCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACEI--CGRKFARSDERKRHTKIH]
EGR2_RAT     RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_XENLA   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_MOUSE   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_HUMAN   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_BRARE   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDF--CGRKFARSDERKRHTKIH]
MIG1_KLULA   -- [-------------------------] ---RP [YVCPICQRGFHRLEHQTRHIRTH] TGERP [HACDFPGCSKRFSRSDELTRHRRIH]
MIG1_KLUMA   -- [-------------------------] ---RP [YMCPICHRGFHRLEHQTRHIRTH] TGERP [HACDFPGCAKRFSRSDELTRHRRIH]
MIG1_YEAST   -- [-------------------------] ---RP [HACPICHRAFHRLEHQTRHMRIH] TGEKP [HACDFPGCVKRFSRSDELTRHRRIH]
MIG2_YEAST   -- [-------------------------] ---RP [FRCDTCHRGFHRLEHKKRHLRTH] TGEKP [HHCAFPGCGKSFSRSDELKRHMRTH]
                [                         ]    :* [. *  * * * * :* . *:* *] ***:* [. *    * : *:**** .** : *]
```

Local alignment



$$F_{i,j} = \text{MAX} \begin{cases} \boxed{0} & \text{No pointer, we start over} \\ F_{i-1,\, j-1} + \text{Score}(S_i,T_j\,) & \\ F_{i,j-1} + go & \\ F_{i-1,j} + go & \end{cases}$$

Gap penalty

Smith & Waterman, 1981

Similarity Scoring Expected value:
negative for random alignments
positive for highly similar sequences

Local alignment

- Initialization

  F(0,0) = F(0,j) = F(i,0) = 0

- Iteration

```
for i=1,…,M
    for j=1,…,N
        - calculate optimal F(i,j)
        - store Ptr(i,j) if score is positive
```

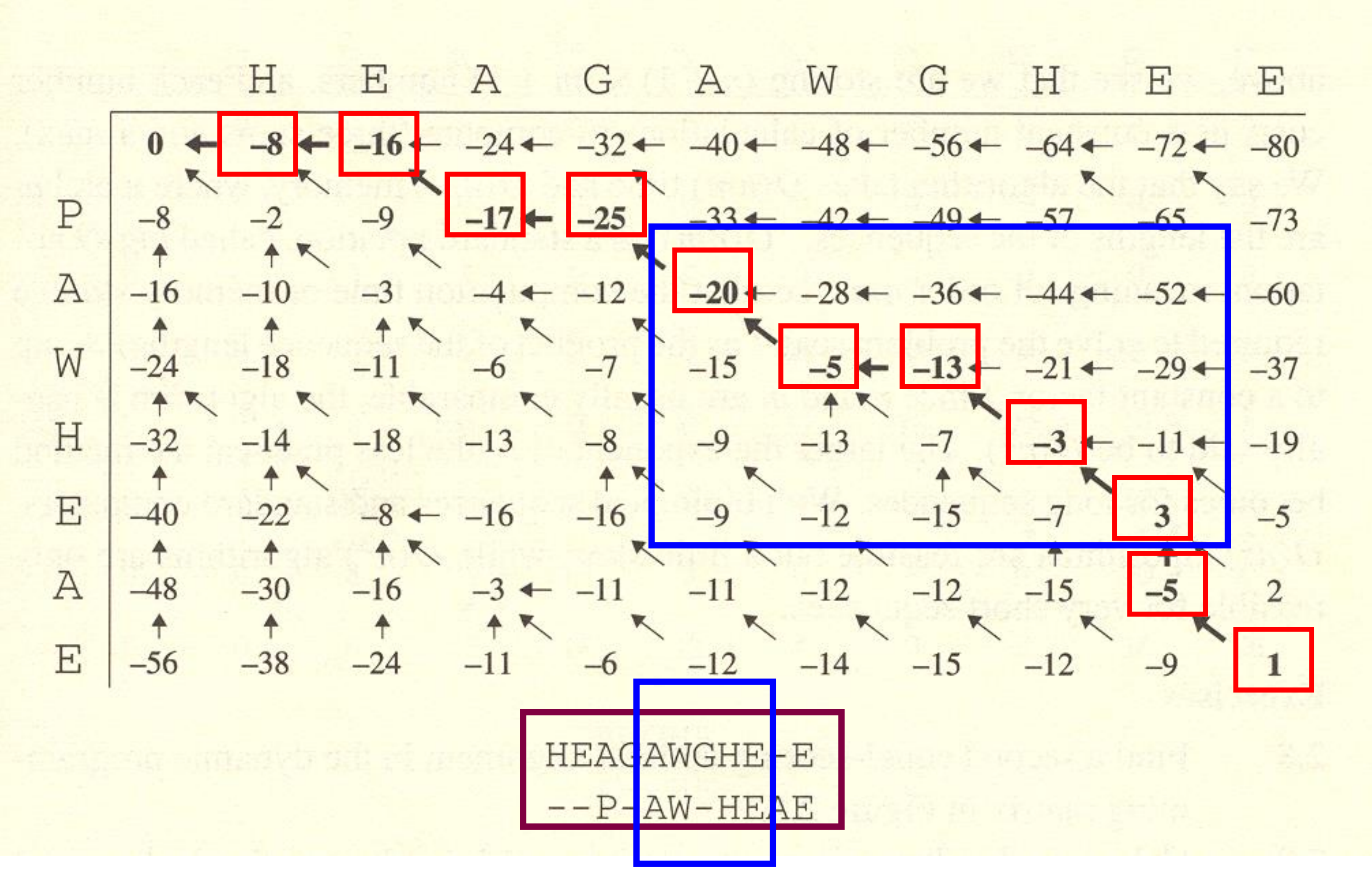
- Termination

  Find the end of the best alignment with FOPT = max{i,j} F(i,j) and trace back

OR

  Find all alignments with F(i,j) > threshold and trace back

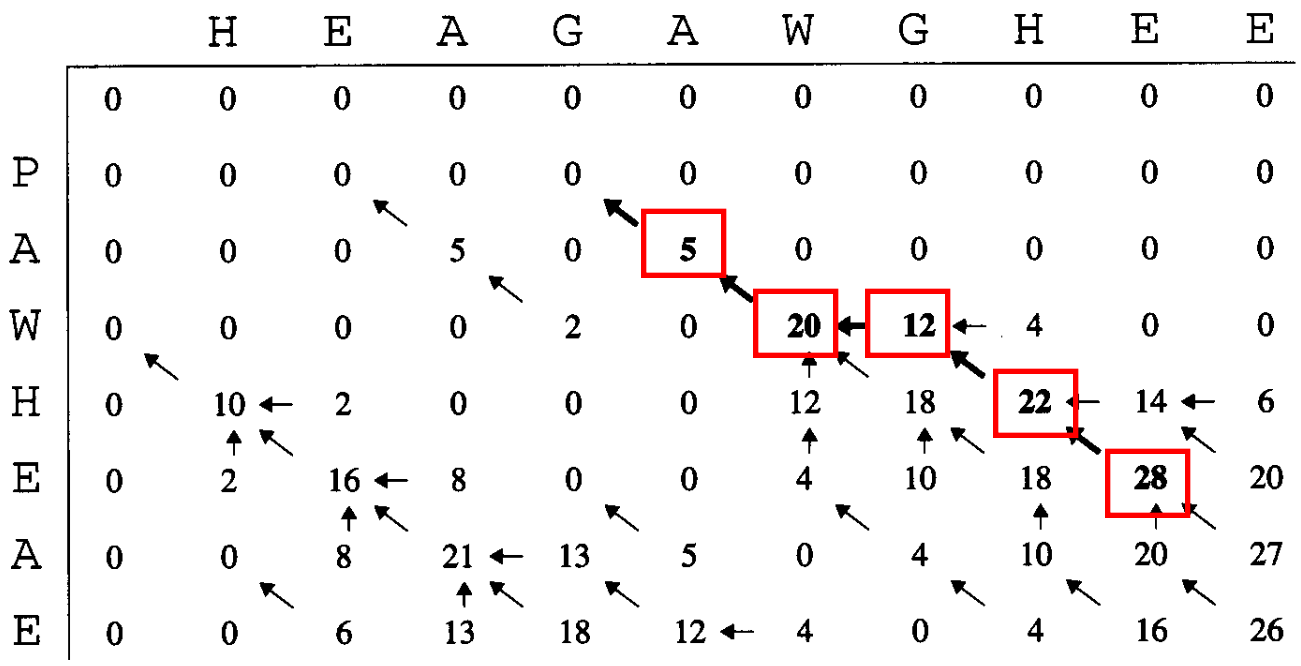
Local *vs.* global alignment



		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	–42	–49	–57	–65	–73
A	–16	–10	–3	–4	–12	–20	–28	–36	–44	–52	–60
W	–24	–18	–11	–6	–7	–15	–5	–13	–21	–29	–37
H	–32	–14	–18	–13	–8	–9	–13	–7	–3	–11	–19
E	–40	–22	–8	–16	–16	–9	–12	–15	–7	3	–5
A	–48	–30	–16	–3	–11	–11	–12	–12	–15	–5	2
E	–56	–38	–24	–11	–6	–12	–14	–15	–12	–9	1

```
HEAGAWGHE-E
--P-AW-HEAE
```

Local *vs.* global alignment (cntd)

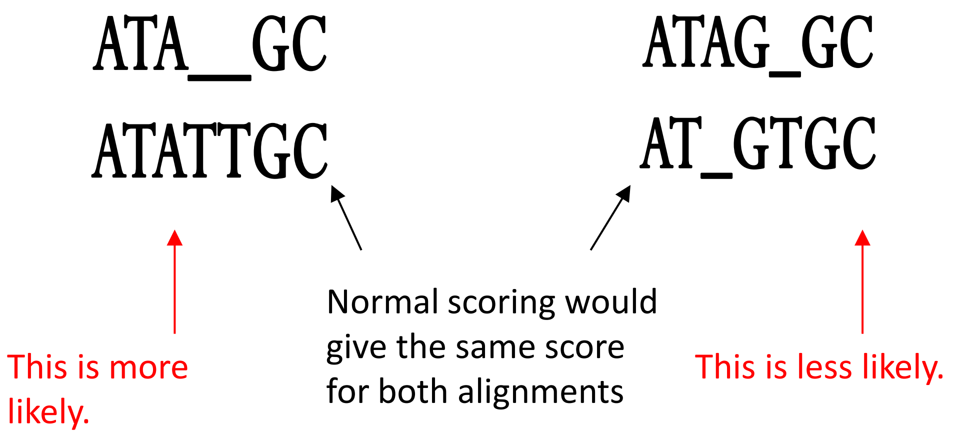


		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

```
AWGHE
AW-HE
```

More accurate gap model

- In nature, gaps often come as a single event rather than a series of single gaps
- Linear gap penalty is too stringent
- Convex gap penalty is expensive
  - Have to keep track of the length of gaps
- Compromise – Affine gap penalty
  - $\gamma(n) = -d - e * (n-1)$
  - d: gap initiation penalty
  - e: gap extension penalty



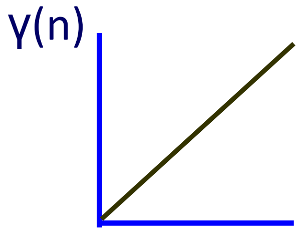
ATA__GC
ATATTGC

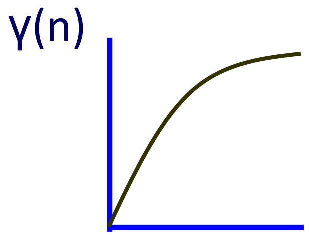
ATAG_GC
AT_GTGC

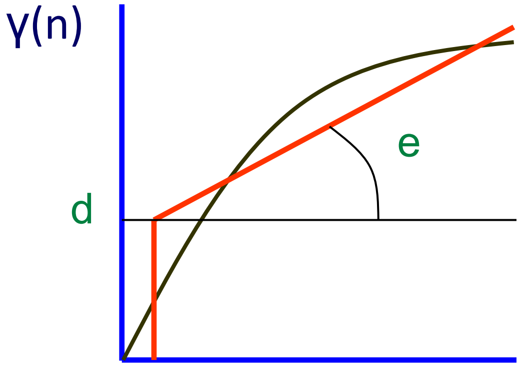
This is more likely.

Normal scoring would give the same score for both alignments

This is less likely.



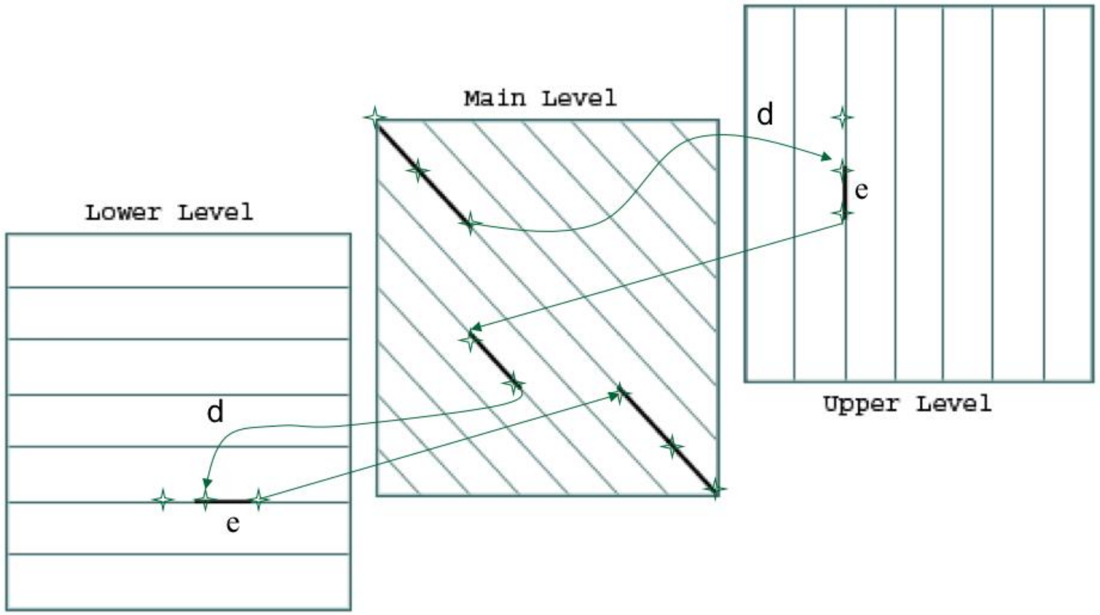




Affine gap algorithm

- Dynamical programming in 3 layers
  - The three recurrences for the scoring algorithm creates a 3-layered graph.
  - The top level creates/extends gaps in the sequence *w.*
  - The bottom level creates/extends gaps in sequence *v.*
  - The middle level extends matches and mismatches.

- Keep track of 3 matrices



Affine Gap Update rule

$$\overset{\downarrow}{F}_{i,j} = \max\begin{cases} \overset{\downarrow}{F}_{i-1,j} - e & \text{Continue Gap in } s \text{ (deletion)} \\ F_{i-1,j} - (d+e) & \text{Start Gap in } s \text{ (deletion): from middle} \end{cases}$$

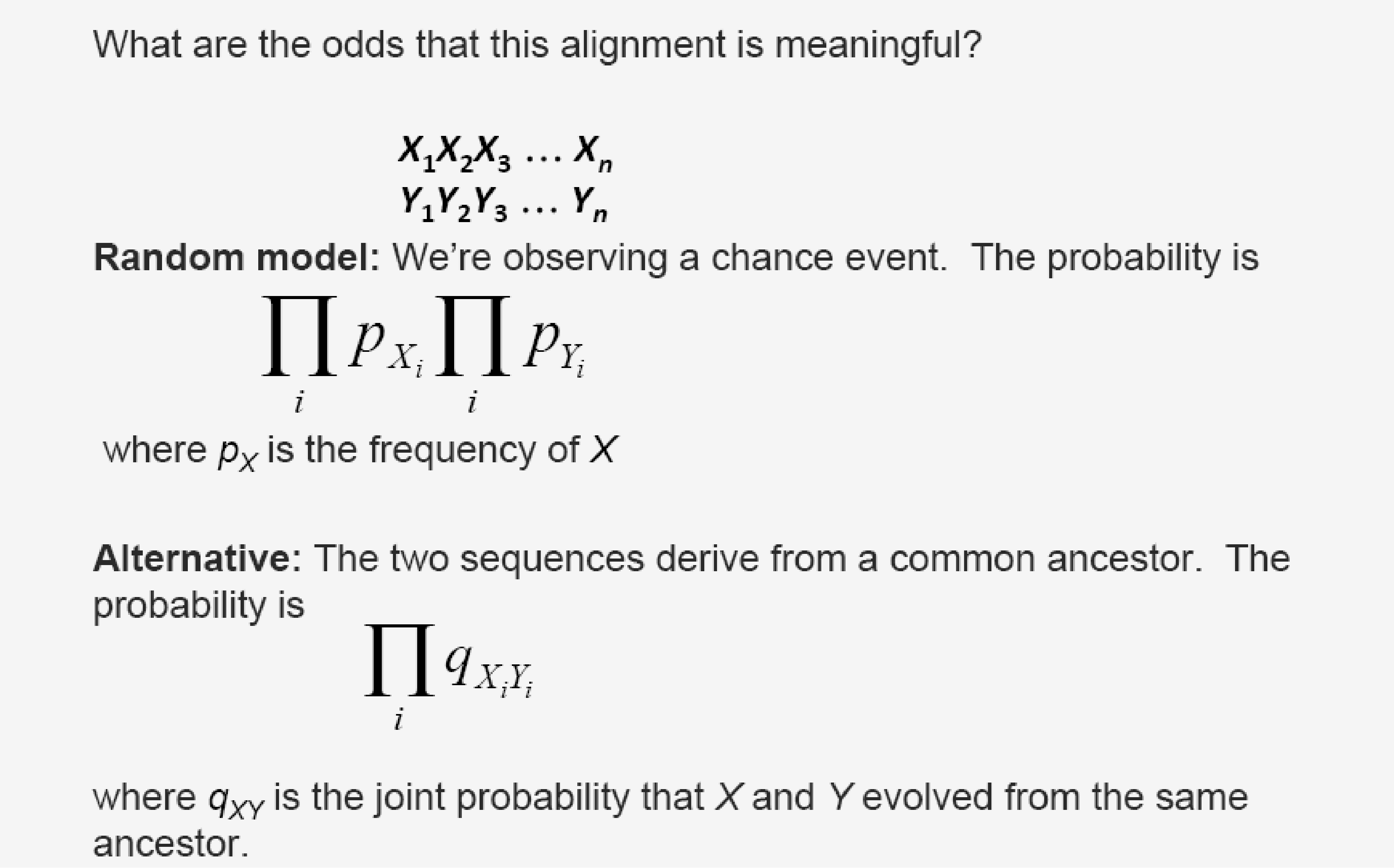
$$\overrightarrow{F}_{i,j} = \max\begin{cases} \overrightarrow{F}_{i,j-1} - e & \text{Continue Gap in } t \text{ (insertion)} \\ F_{i,j-1} - (d+e) & \text{Start Gap in } t \text{ (insertion):from middle} \end{cases}$$

$$F_{i,j} = \max\begin{cases} F_{i-1,j-1} + s(v_i, w_j) & \text{Match or Mismatch} \\ \overset{\downarrow}{F}_{i,j} & \text{End deletion: from top} \\ \overrightarrow{F}_{i,j} & \text{End insertion: from bottom} \end{cases}$$

How to decide on the correct scoring metric

- Scoring metrics should reflect the evolutionary process
- What are the odds that an alignment is biologically meaningful – the proteins are homologous
- Random model: product of chance events
- Non-random model: two sequences derived from a common ancestor
- Things to consider
    - What is the frequency of different mutations
    - Over what time scale?

Log-odds scoring



What are the odds that this alignment is meaningful?

$$\boldsymbol{X_1 X_2 X_3 \ldots X_n}$$
$$\boldsymbol{Y_1 Y_2 Y_3 \ldots Y_n}$$

Random model: We're observing a chance event. The probability is

$$\prod_i p_{X_i} \prod_i p_{Y_i}$$

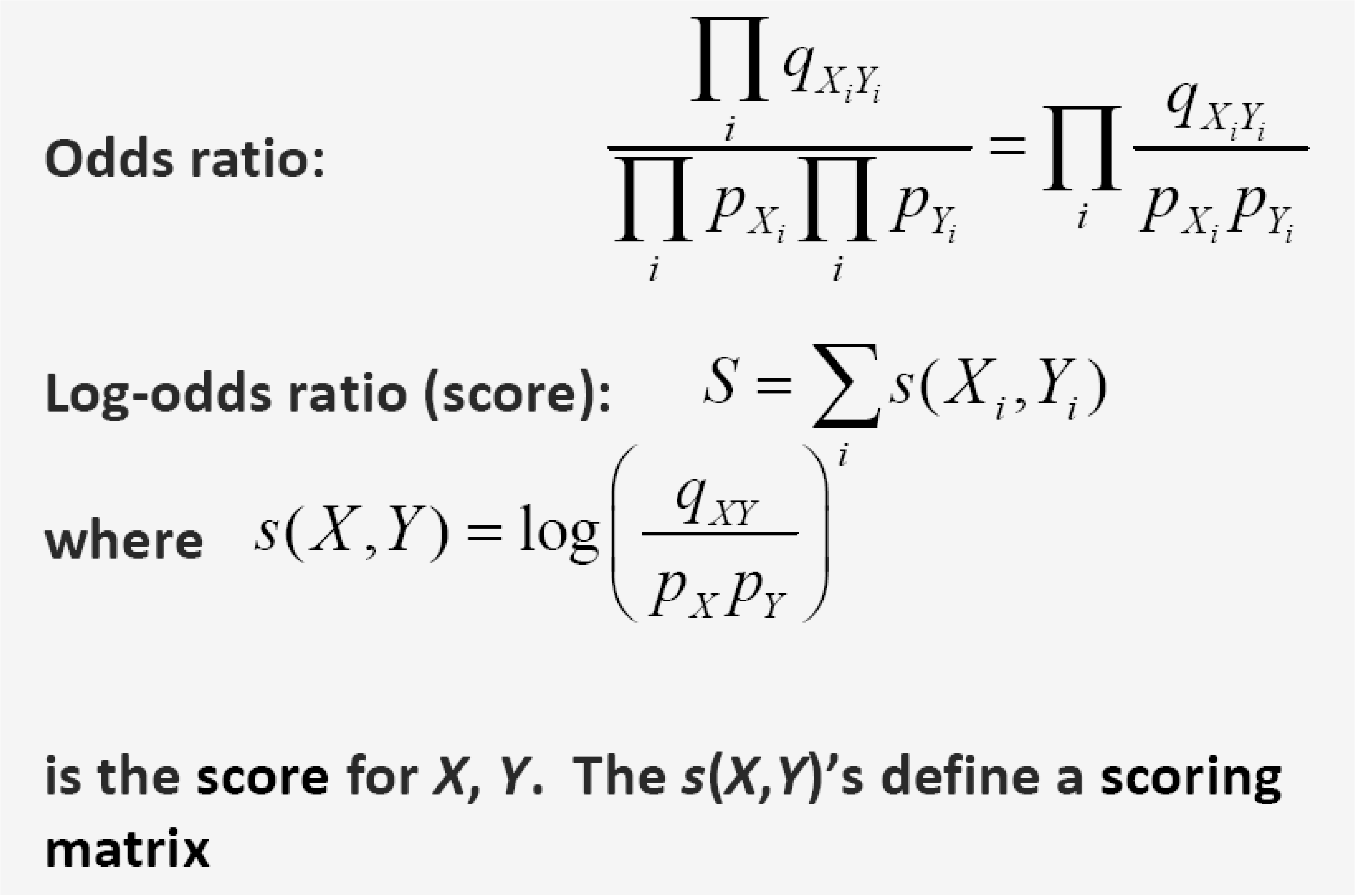
where p_X is the frequency of X

Alternative: The two sequences derive from a common ancestor. The probability is

$$\prod_i q_{X_i Y_i}$$

where q_{XY} is the joint probability that X and Y evolved from the same ancestor.

Log-odds scoring



Odds ratio:

$$\frac{\prod_i q_{X_iY_i}}{\prod_i p_{X_i}\prod_i p_{Y_i}} = \prod_i \frac{q_{X_iY_i}}{p_{X_i}p_{Y_i}}$$

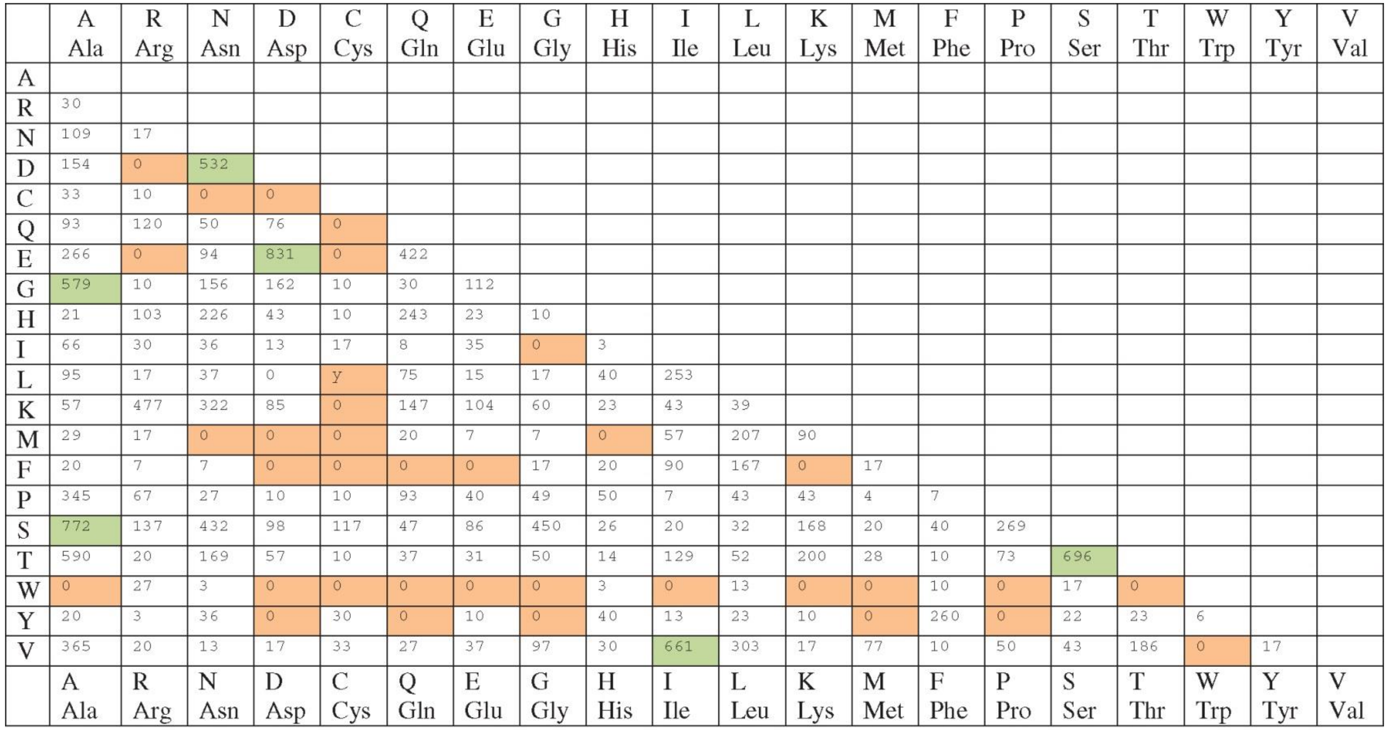
Log-odds ratio (score): $S = \sum_i s(X_i, Y_i)$

where $s(X,Y) = \log\left(\frac{q_{XY}}{p_X p_Y}\right)$

is the score for *X*, *Y*. The *s*(*X*,*Y*)'s define a scoring matrix

Accepted Point Mutation (PAM) model

- Where do we get q_{XY}
- Compare closely related proteins
- Find substitutions that are “accepted” to natural selection
- Very likely mutations E to D
- Very unlikely: involve C and W



	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																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	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

Conservative substitutions

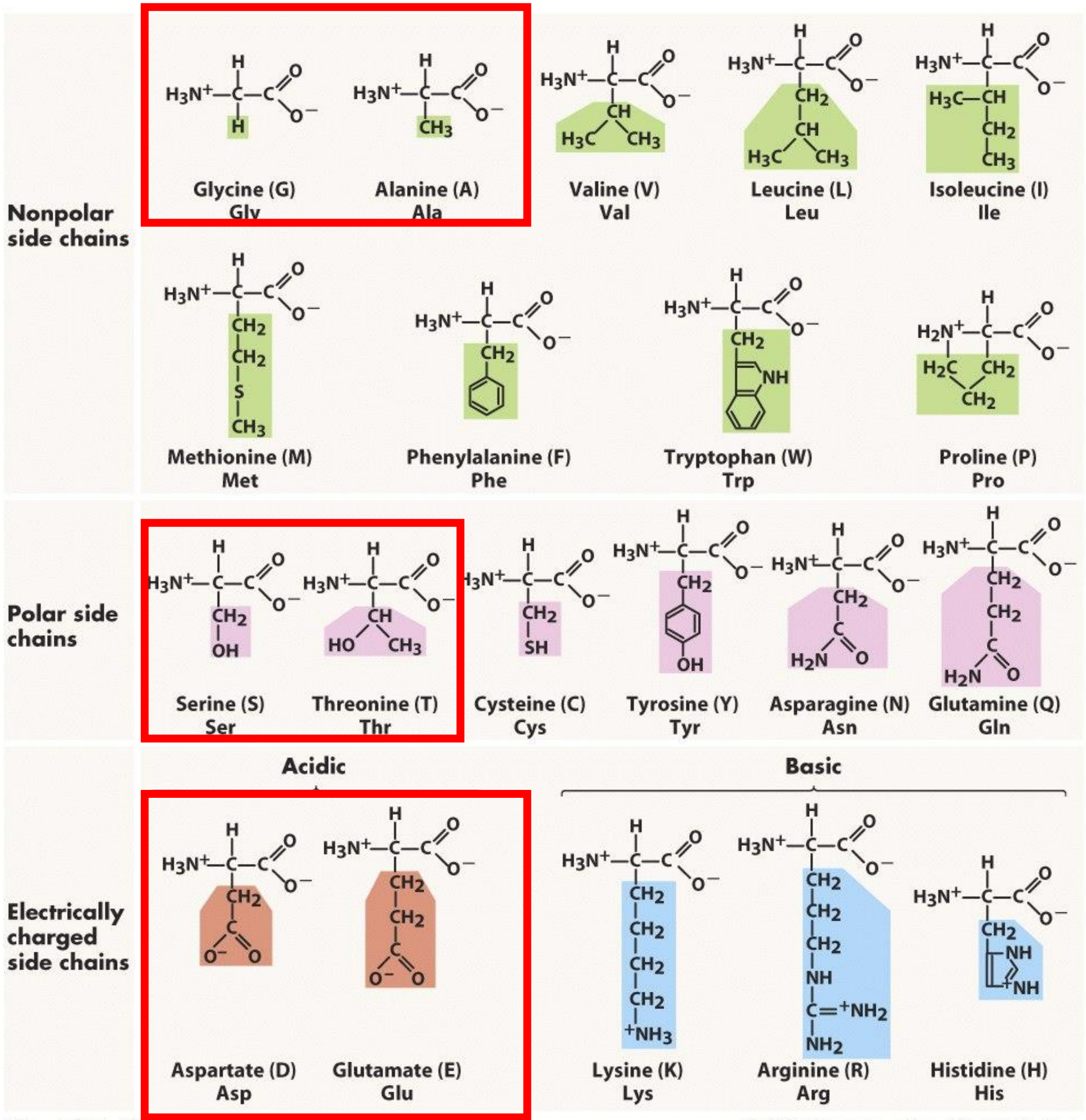
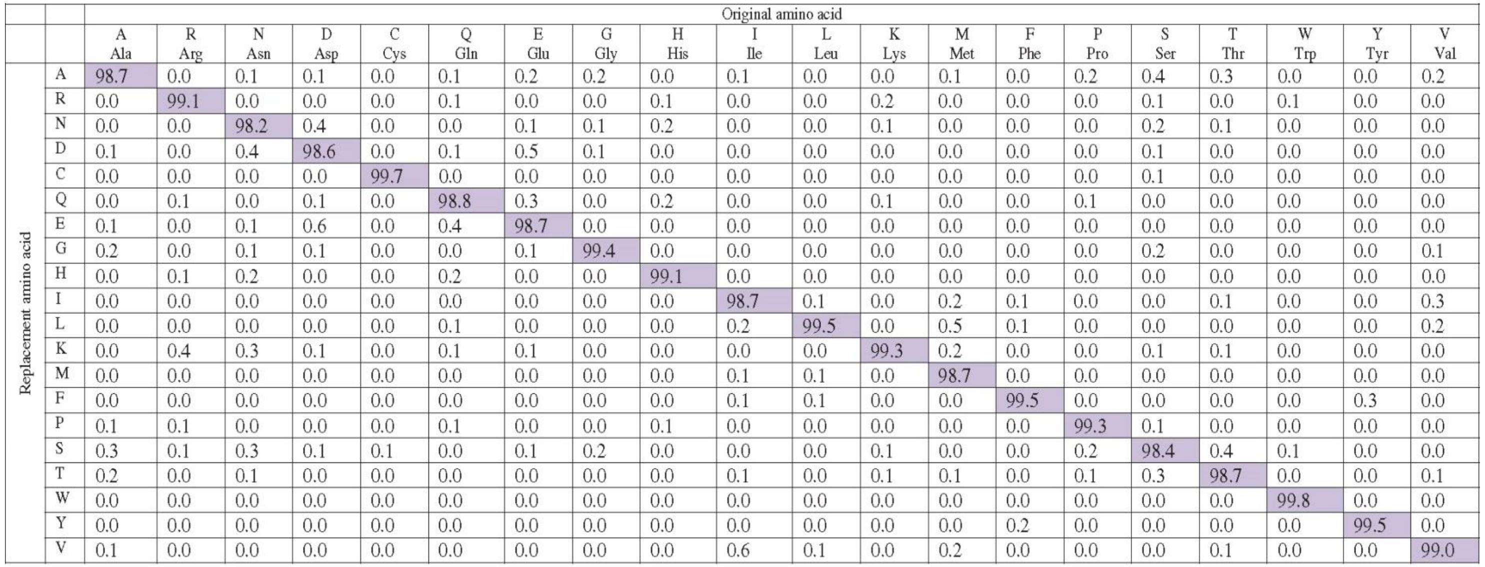


Figure 3-5 Biological Science, 2/e

© 2005 Pearson Prentice Hall, Inc.

PAM1 probability matrix

- PAM1 probability matrix
- Dayhoff et al (1978) estimated probability of one-step transitions
- Used a family of very closely related proteins
- Corresponds to 1 change per 100 a.a.



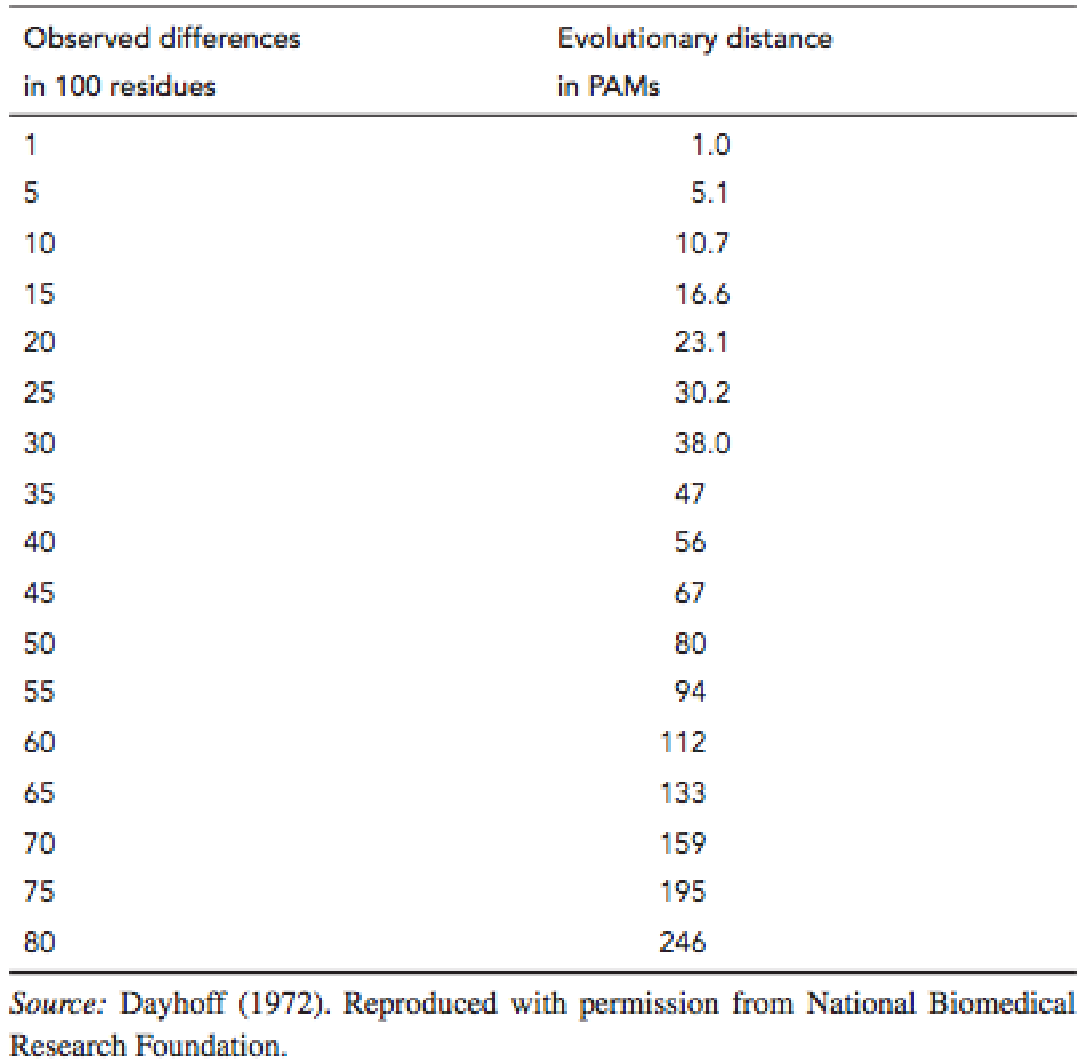
		Original amino acid																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.2
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0
	Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	99.5	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	99.0

FIGURE 3.9 The PAM1 mutation probability matrix. The original amino acid *j* is arranged in columns (across the top), while the replacement amino acid *i* is arranged in rows. Dayhoff et al. multiplied values by 10,000 (offering added precision) while here we multiply by 100 so that, for example, the first cell's value of 98.7 corresponds to 98.7% occurrence of ala remaining ala over this evolutionary interval.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics

PAM1 through PAM250

- We can multiply PAM1 by itself to get a probability matrix for longer time scales
- PAM is measured in number of changes not time
- Number of changes that occurred is not the same as number of observed changes

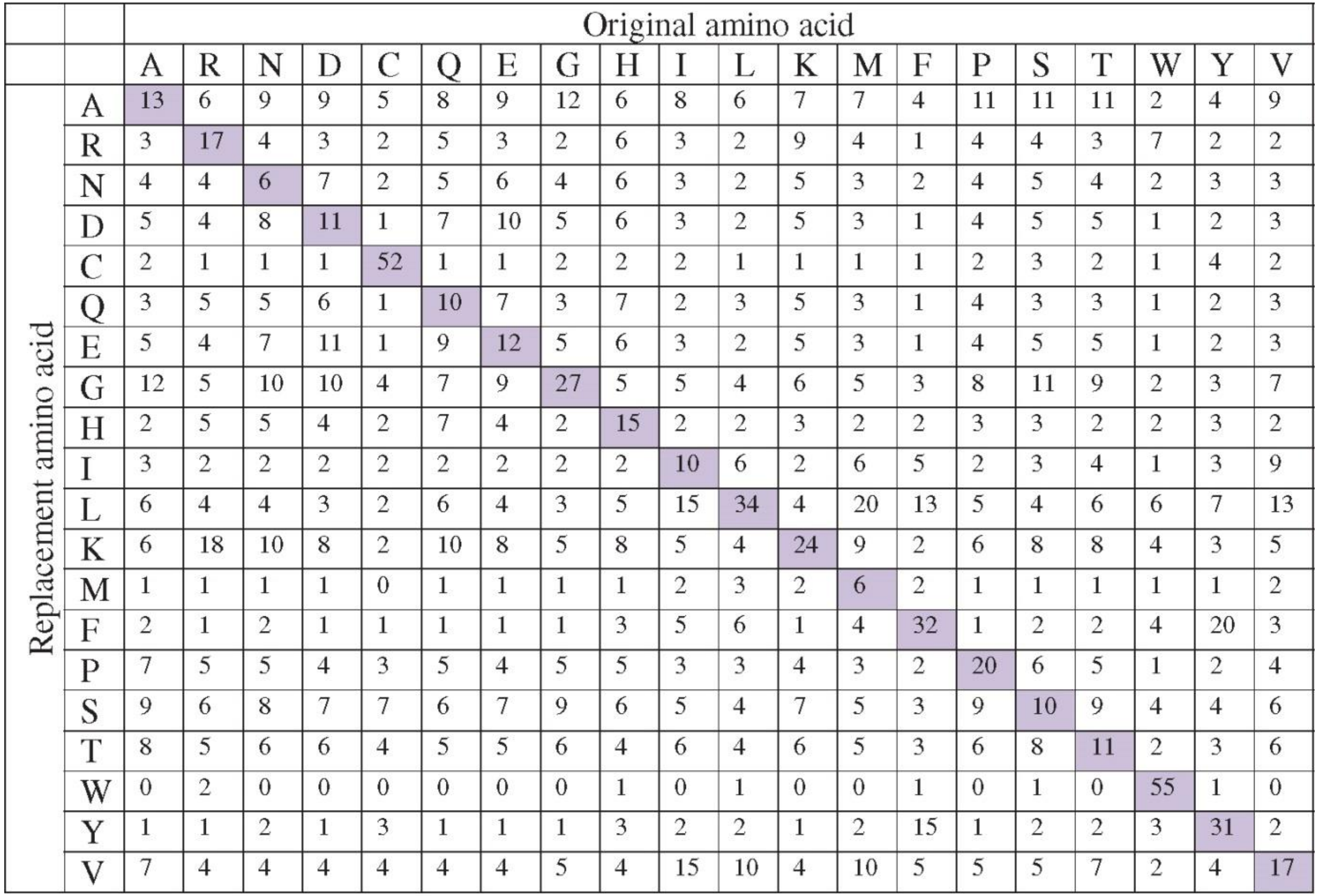


Observed differences in 100 residues	Evolutionary distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

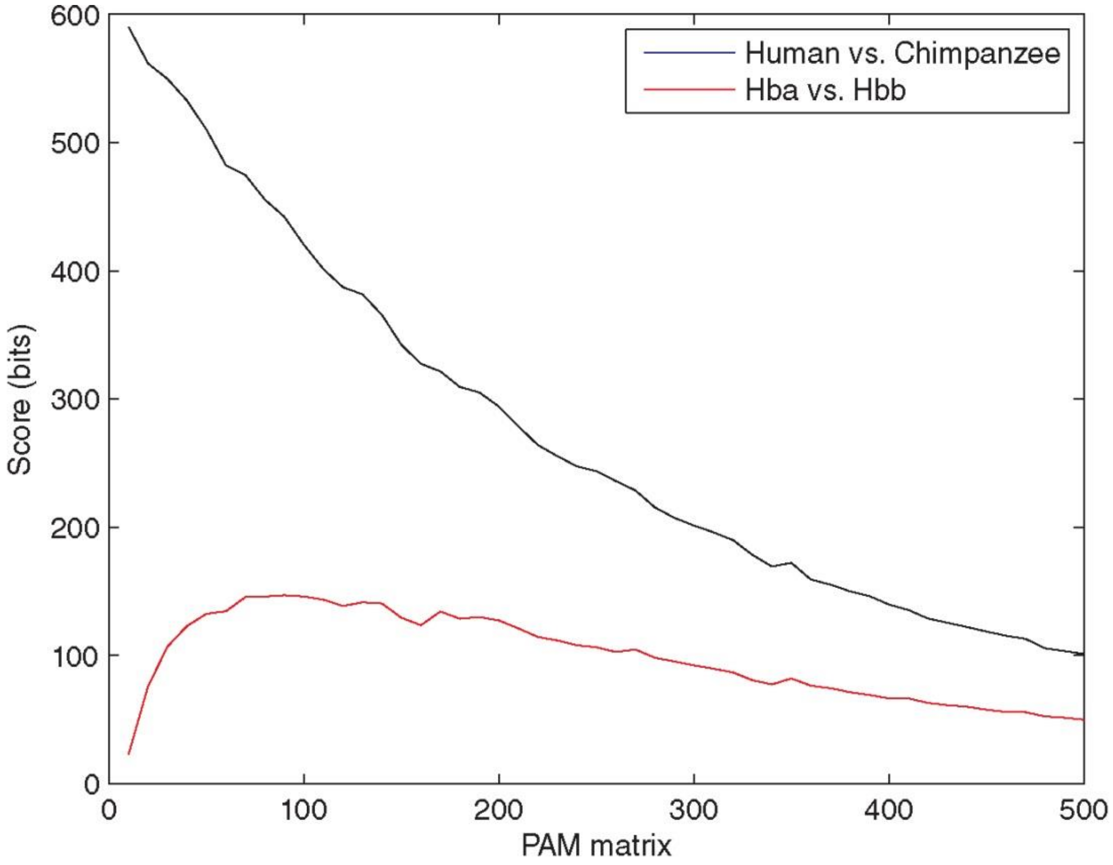
PAM250

- Only 20% identity
- 20% identity is close to what you might get aligning random sequences



		Original amino acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Replacement amino acid	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
	Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
	V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

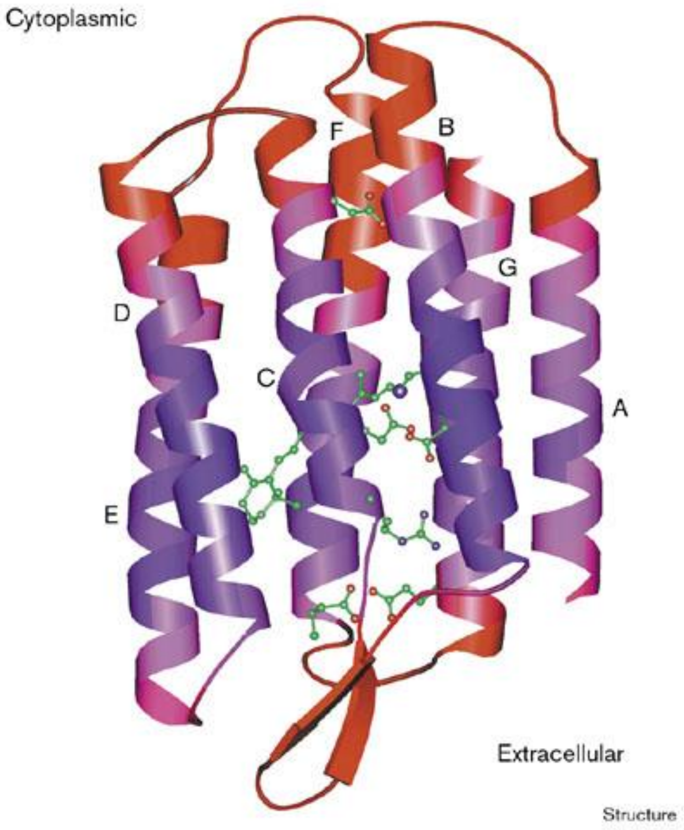
Choice of scale reflects the result



- Human and chimp beta globin—close orthologs
- Human beta and alpha globin – paralogs –further apart

PAM model

- Assumptions
  - Replacement at any site depends only on the a.a. on that site, give the **mutability** of the a.a.
  - Sequences in the training set (and those compared) have average a.a. composition.
- Sources of error
  - Many proteins depart from the average a.a. composition.
  - The a.a. composition can vary even within a protein (e.g. transmembrane proteins).
  - A.a. positions are not “mutated” equally probably; especially in lor evolutionary distances.
  - Rare replacements are observed too infrequently and…
  - …errors in PAM1 are magnified in PAM250.

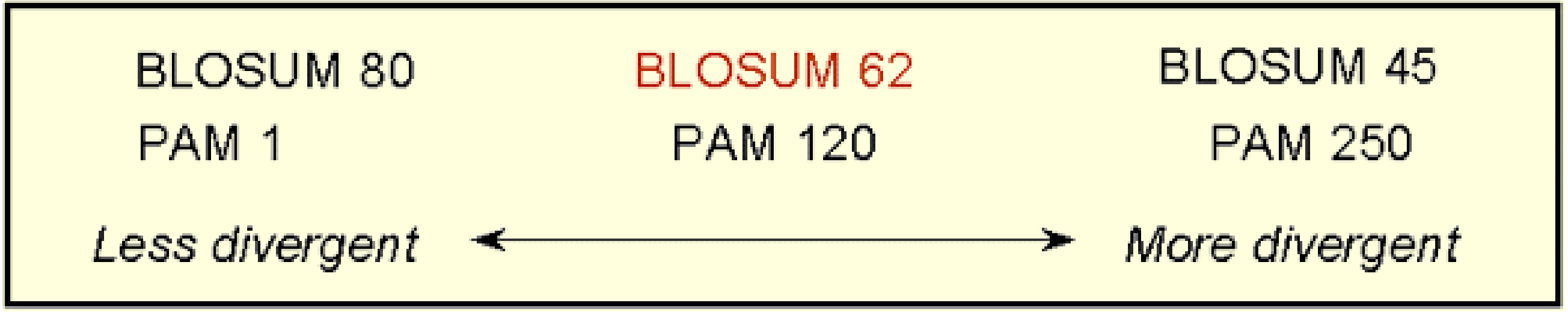


Blocks Substitution Matrices (BLOSUM):

- Log-likelihood matrix (Henikoff & Henikoff, 1992)
- BLOCKS database of aligned sequences used as primary source set.
- Different BOLSUM*n* matrices are calculated independently from BLOCKS (ungapped local alignments)
- BLOSUM*n* is based on a cluster of BLOCKS of sequences that share at least *n* percent identity
- BLOSUM62 represents closer sequences than BLOSUM45
- BLOCKS database contains large number of ungapped multiple local alignments of conserved regions of proteins
- Alignments include distantly related sequences in which multiple base substitutions at the same position could be observed

PAM vs BLOSUM

- PAM is based on closely related sequences, thus is biased for short evolutionary distances where number of mutations are scalable
- PAM is based on globally aligned sequences, thus includes conserved and non-conserved positions; BLOSUM is based on conserved positions only
- Lower PAM/higher BLOSUM matrices identify shorter local alignments of highly similar sequences
- Higher PAM/lower BLOSUM matrices identify longer local alignments of more distant sequences



BLOSUM 80 — BLOSUM 62 — BLOSUM 45

PAM 1 — PAM 120 — PAM 250

Less divergent ⟷ *More divergent*

- Matrices of choice:
  - BLOSUM62: the all-weather matrix
  - PAM250: for distant relatives