

L4

Linear space
Scoring matrices

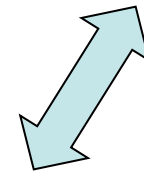
Alignment (Linear Space)

- Score computation

$$S[i, j] = \max \begin{cases} S[i-1, j-1] + C(s_i, t_j) \\ S[i-1, j] + C(s_i, -) \\ S[i, j-1] + C(-, t_j) \end{cases}$$

For $i = 1$ to n

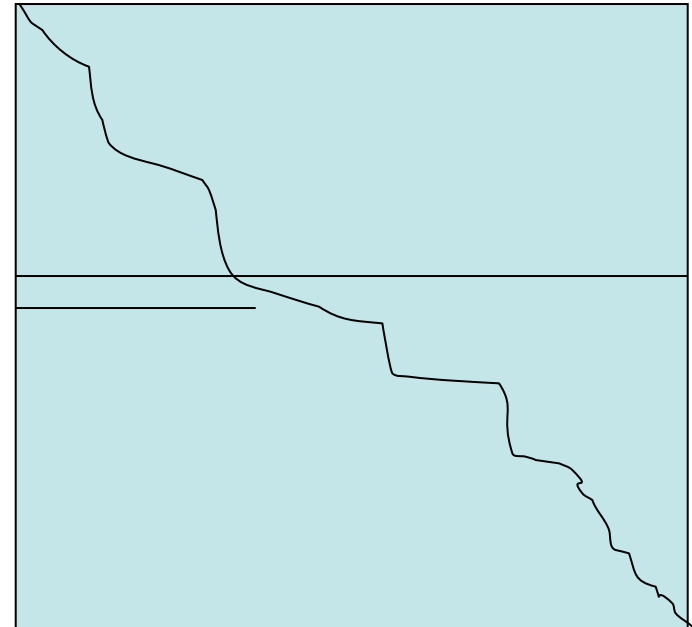
For $j = 1$ to m



$$\begin{cases} i_2 = i \% 2; & i_1 = (i - 1) \% 2; \\ S[i_2, j] = \max \begin{cases} S[i_1, j-1] + C(s_i, t_j) \\ S[i_1, j] + C(s_i, -) \\ S[i_2, j-1] + C(-, t_j) \end{cases} \end{cases}$$

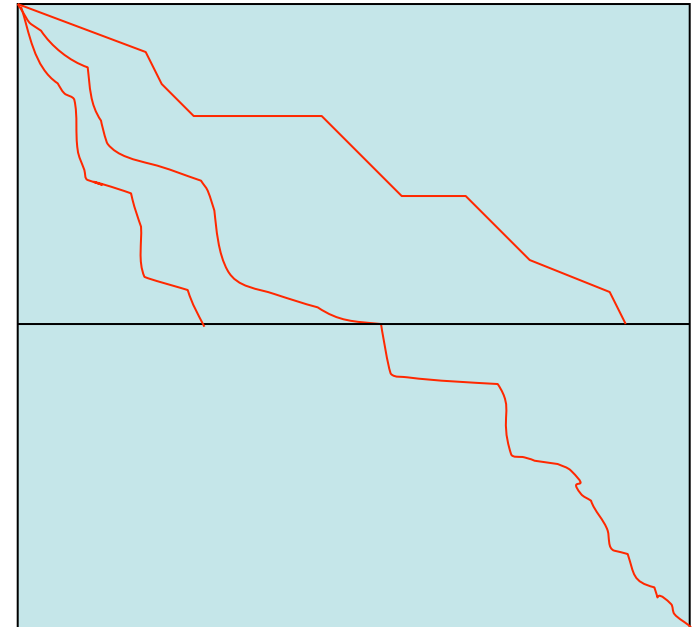
Linear Space Alignment

- In Linear Space, we can do each row of the D.P.
- We need to compute the optimum path from the origin $(0,0)$ to (m,n)



Linear Space (cont'd)

- At $i=n/2$, we know scores of all the optimal paths ending at that row.
- Define $F[j] = S[n/2, j]$
- One of these j is on the true path. Which one?



Backward alignment

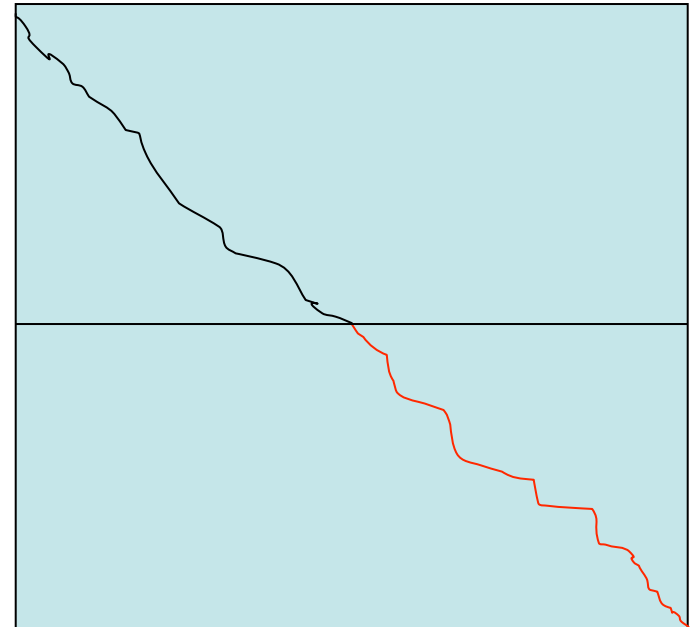
- Let $S_b[i,j]$ be the optimal score of aligning $s[i+1..n]$ with $t[j+1..m]$

$$S_b[i,j] = \max \begin{cases} S_b[i+1,j+1] + C(s_i, t_j) \\ S_b[i+1,j] + C(s_i, -) \\ S_b[i,j+1] + C(-, t_j) \end{cases}$$

- Boundary cases?
- $S_b[n,j]$? $S_b[m,j]$?

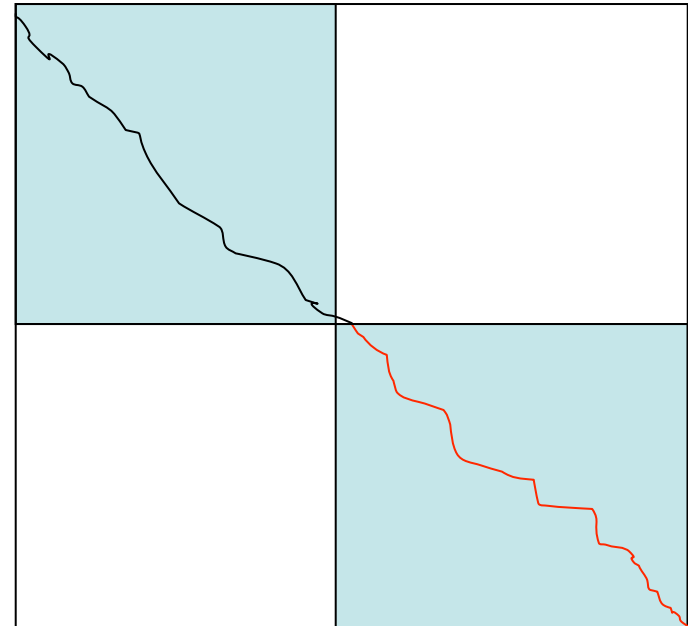
Forward, Backward computation

- At the optimal coordinate, j
 - $F[j]+B[j]=S[n,m]$
- In $O(nm)$ time, and $O(m)$ space, we can compute one of the coordinates on the optimum path.



Linear Space Alignment

- $\text{Align}(1..n, 1..m)$
 - For all $1 \leq j \leq m$
 - Compute $F[j] = S(n/2, j)$
 - For all $1 \leq j \leq m$
 - Compute $B[j] = S_b(n/2, j)$
 - $j^* = \max_j \{F[j] + B[j]\}$
 - $X = \text{Align}(1..n/2, 1..j^*)$
 - $Y = \text{Align}(n/2+1..n, j^*+1..m)$
 - Return X, j^*, Y



Linear Space complexity

- $T(nm) = c.nm + T(nm/2) = O(nm)$
- $\text{Space} = O(m)$

Silly Quiz

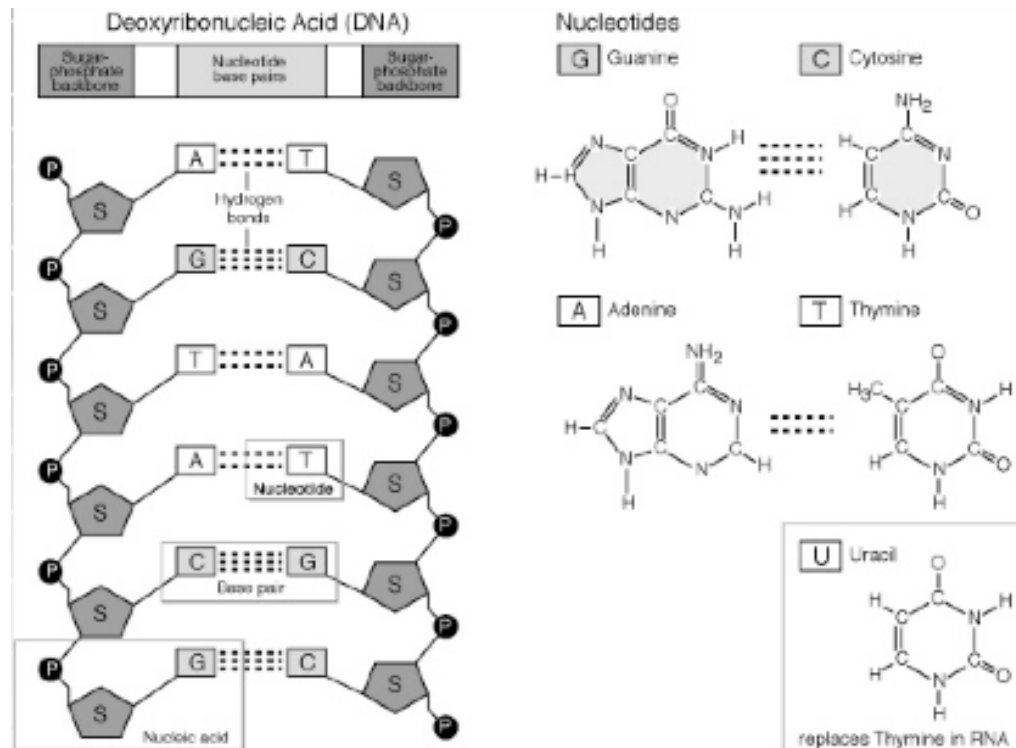
- Name a famous Bioinformatics Researcher

- Name a famous Bioinformatics Researcher who is a woman

Scoring Matrices

- We have seen that affine gap penalties help concentrate the gaps in small regions.
- What about substitution errors. Are all substitutions alike?

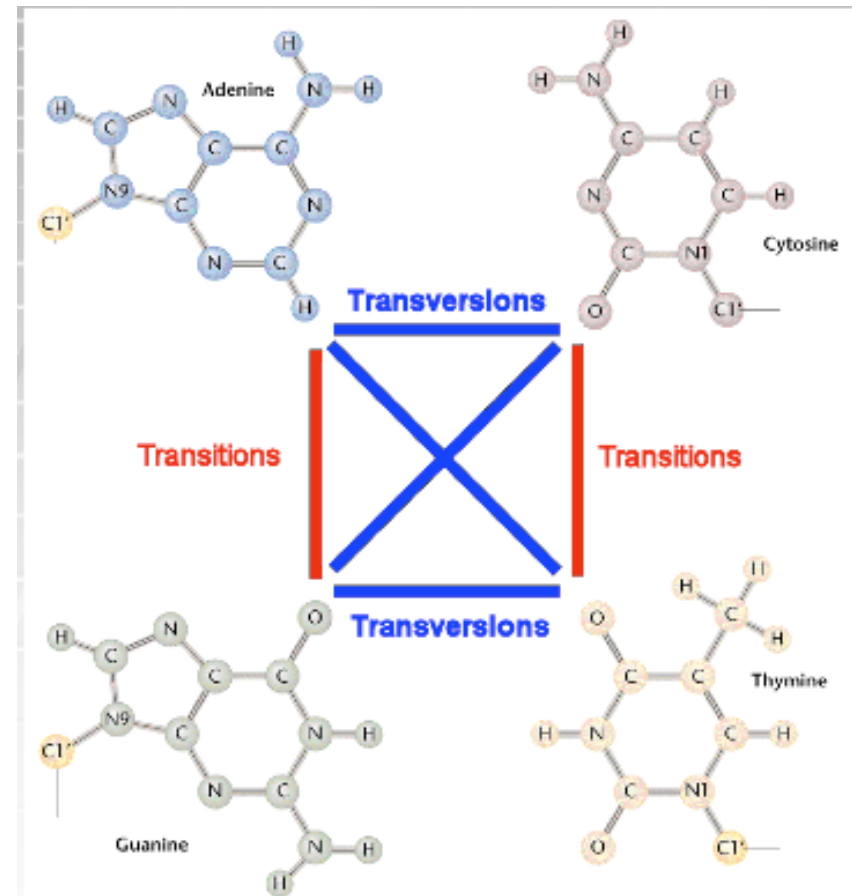
Scoring DNA



- DNA has structure.

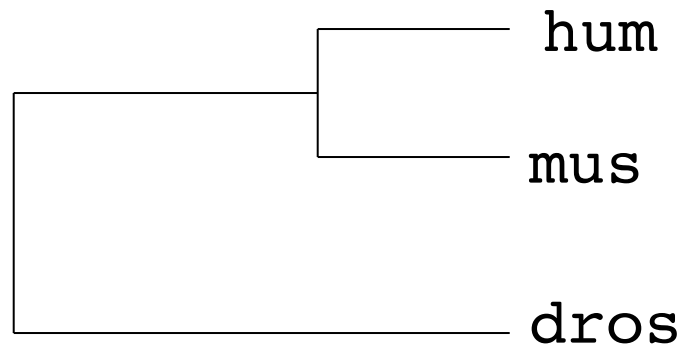
DNA scoring matrices

- So far, we considered a simple match/mismatch criterion.
- The nucleotides can be grouped into Purines (A,G) and Pyrimidines.
- Nucleotide substitutions within a group (transitions) are more likely than those across a group (transversions)



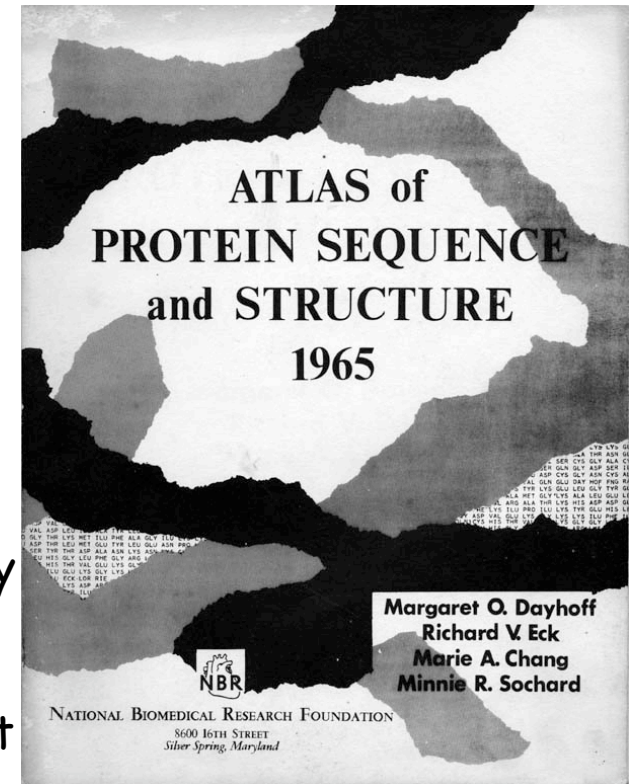
The scoring problem

- When comparing hum and dros., more mismatches are likely just by chance.. If we expect to see 70% of the residues to be mutated, then a 50% identity is great.
- A different scoring function is needed for hum and drosphila



Scoring proteins

- Scoring protein sequence alignments is a much more complex task than scoring DNA
 - Not all substitutions are equal
- Problem was first worked on by Pauling and collaborators
- In the 1970s, Margaret Dayhoff created the first similarity matrices.
 - "One size does not fit all"
 - Homologous proteins which are evolutionarily close should be scored differently than proteins that are evolutionarily distant
 - Different proteins might evolve at different rates and we need to normalize for that



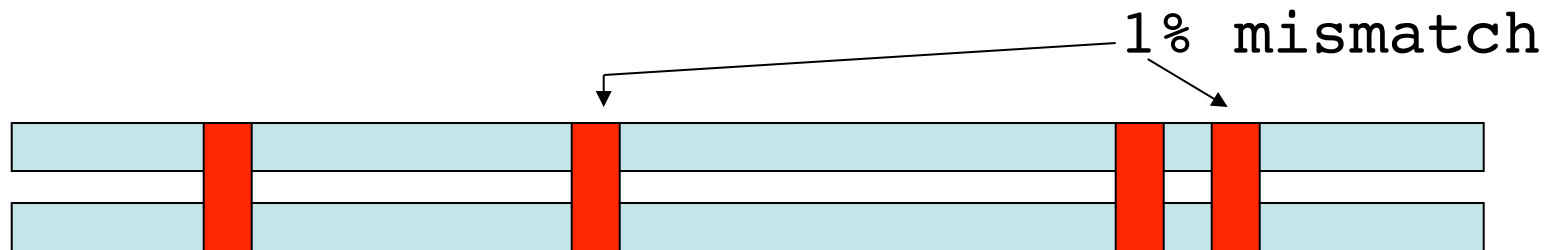
Frequency based scoring

	A	
	B	

- Our goal is to score each column in the alignment
- Comparing against expectation:
 - Think about alignments of pairs of random sequences, and compute the probability that A and B appear together just by chance $P^R(A,B)$
 - Compute the probability of A and B appearing together in the alignment of related sequences (orthologs) $P^O(A,B)$
- A good score function? $\log\left(\frac{P^O(A,B)}{P^R(A,B)}\right)$

PAM 1 distance

- Two sequences are 1 PAM apart if they differ in 1 % of the residues.



- $PAM_1(a,b) = \Pr[\text{residue } b \text{ substitutes residue } a, \text{ when the sequences are 1 PAM apart}]$

PAM1 matrix

- Align many proteins that are very similar
 - Is this a problem?
- 1 PAM evolutionary distance represents the time in which 1% of the residues have changed
- Estimate the frequency $P_{b|a}$ of residue a being substituted by residue b.
- $PAM1(a,b) = P_{a|b} = \Pr(b \text{ will mutate to an } a \text{ after } 1 \text{ PAM evolutionary distance})$
- Scoring matrix
 - $S(a,b) = \log_{10}(P_{ab}/P_a P_b) = \log_{10}(P_{b|a}/P_b)$

PAM 1

- Top column shows original, and left column shows replacement residue = $PAM1(a,b) = Pr(a|b)$

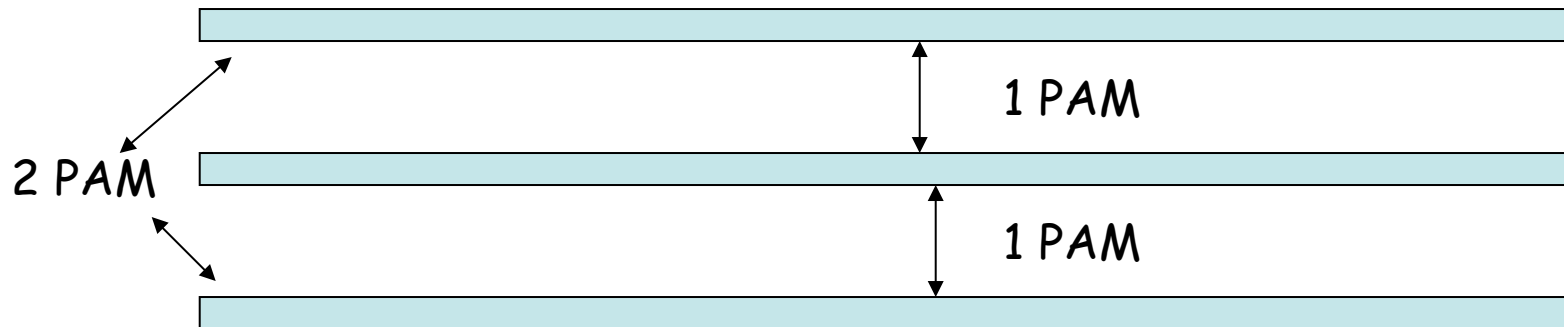
PAM1 Mutation Matrix

1 PAM evolutionary distance

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

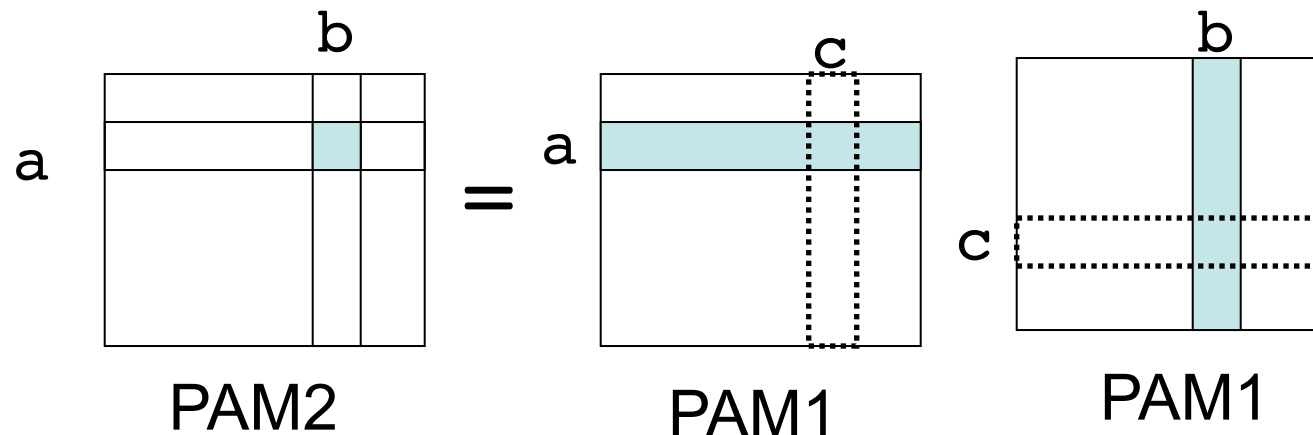
PAM distance

- Two sequences are 1 PAM apart when they differ in 1% of the residues.
- When are 2 sequences 2 PAMs apart?



Generating Higher PAMs

- $PAM_2(a,b) = \sum_c PAM_1(a,c) \cdot PAM_1(c,b)$
- $PAM_2 = PAM_1 * PAM_1$ (Matrix multiplication)
- PAM_{250}
 - $= PAM_1 * PAM_{249}$
 - $= PAM_1^{250}$



PAM250 Mutation Matrix

250 PAM evolutionary distance

	SEQUENCE LOGS LOGS																										
	Ala	Arg	Asn	Asp	Gly	Ile	Leu	Met	Phe	Pro	Thr	Tyr	Val	Ala	Arg	Asn	Asp	Gly	Ile	Leu	Met	Phe	Pro	Thr	Tyr	Val	
Ala	6	12	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	
Arg	6	17	4	4	2	3	4	2	1	4	2	2	2	4	4	4	4	2	3	2	3	4	4	4	4	2	2
Asn	6	4	6	7	2	5	6	4	2	2	2	2	2	2	4	4	4	2	2	2	2	2	2	2	2	2	2
Asp	6	4	6	11	1	7	10	4	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Gly	6	2	1	1	6	1	1	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Ile	6	3	3	3	1	9	12	5	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Leu	6	2	2	2	2	2	2	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Met	6	2	2	2	2	2	2	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Phe	6	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Pro	6	4	4	3	2	5	4	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Thr	6	4	4	4	2	2	2	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Tyr	6	2	2	2	2	2	2	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2
Val	6	2	2	2	2	2	2	2	2	2	2	2	2	4	4	4	4	2	2	2	2	2	2	2	2	2	2

Note: This is not the score matrix:

What happens as you keep increasing the power?

Scoring residues

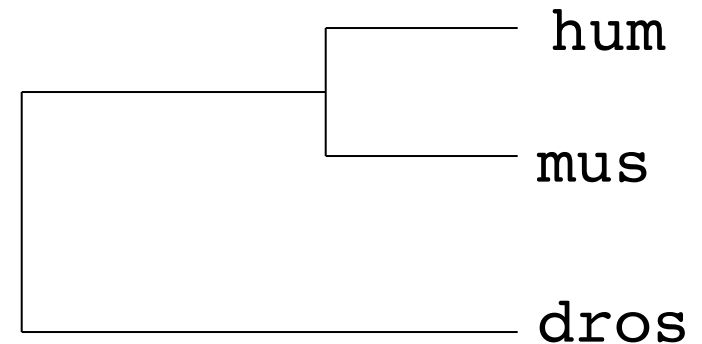
- A reasonable score function $C(a,b)$ is given as follows:
 - Look at 'high quality' alignments
 - $C(a,b)$ should be high when a,b are seen together more often than is expected by chance
 - $C(a,b)$ should be low, otherwise.
- How often would you expect to see a,b together just by chance?
 - $P_a P_b$
- Let P_{ab} be the probability that a and b are aligned in a high-quality alignment
- A good scoring function is the log-odds score
 - $C(a,b) = \log_{10} (P_{ab}/P_a P_b)$

Scoring alignments

- To compute P_{ab} , we need 'high-quality' alignments
- How can you get quality alignments?
 - Use SW (But that needs the scoring function)
 - Build alignments manually
 - Use Dayhoff's theory to extrapolate from high identity alignments

Scoring using PAM matrices

- Suppose we know that two sequences are 250 PAMs apart.
- $S(a,b) = \log_{10}(P_{ab}/P_a P_b) = \log_{10}(P_{a|b}/P_a) = \log_{10}(PAM_{250}(a,b)/P_a)$
- How does it help?
 - $S_{250}(A,V) \gg S_1(A,V)$
 - Scoring of hum vs. Dros should be using a higher PAM matrix than scoring hum vs. mus.
 - An alignment with a smaller % identity could still have a higher score and be more significant



PAM250 based scoring matrix

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

$$\bullet S_{250}(a,b) = \log_{10}(P_{ab}/P_aP_b) = \log_{10}(PAM_{250}(a,b)/P_a)$$

BLOSUM series of Matrices

- Henikoff & Henikoff: Sequence substitutions in evolutionarily distant proteins do not seem to follow the PAM distributions
- A more direct method based on hand-curated multiple alignments of distantly related proteins from the BLOCKS database.
- BLOSUM60 Merge all proteins that have greater than 60%. Then, compute the substitution probability.
 - In practice BLOSUM62 seems to work very well.

PAM vs. BLOSUM

- What is the correspondence?

- PAM1

Blosum1

- PAM2

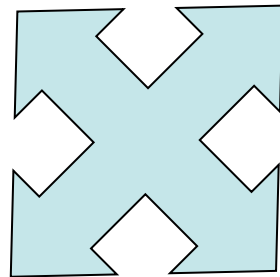
Blosum2

-

Blosum62

- PAM250

Blosum100



END of L4