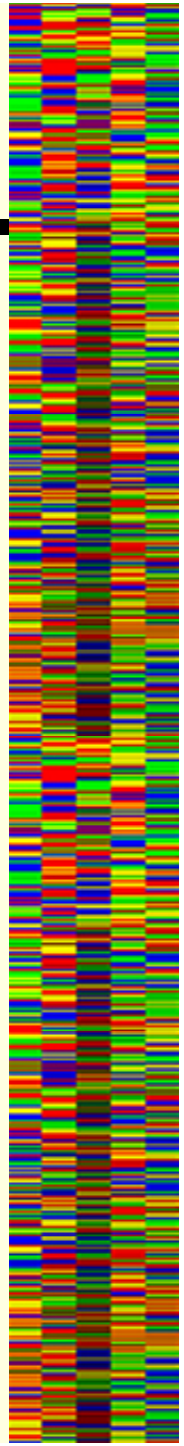
Genomics

8-12 September

6	M 8	MG	Scoring Matrices		Ch 3 and Ch 4
7	W 10	MG	Pairwise Alignment		
8	F 12	MG	Pairwise Alignment	Hw2	

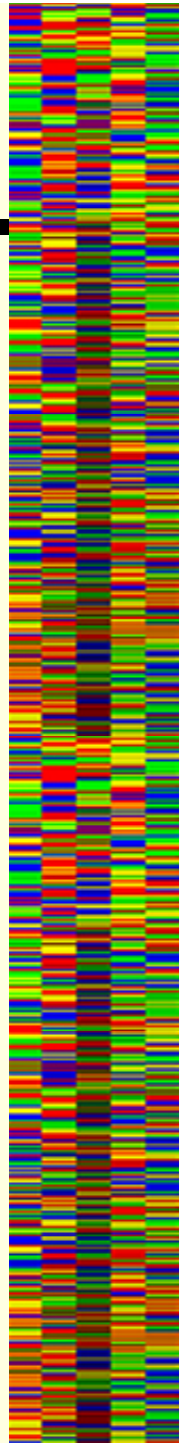
Reading: Mount - ch 3 and 4



Scoring Systems

BLOSUM (BLOcks SUbstitution Matrix)

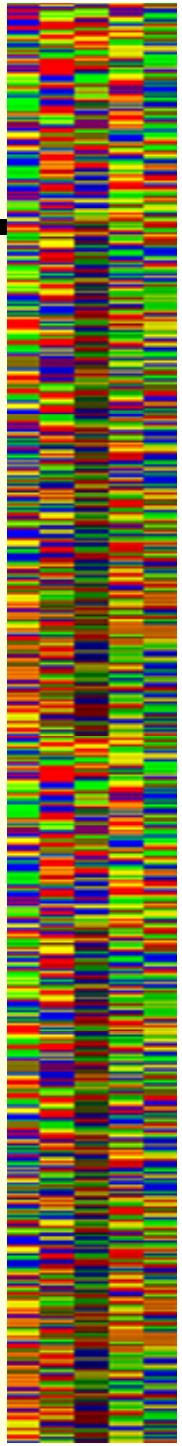
- ***Based on PROSITE signatures***
  - Signatures are short expressions like C-X-X-C-X-X-X-C
- ***Locally align each feature to get "blocks“***
- ***Blocks are locally conserved regions, i.e., more constrained regions likely to be related to structure/function***
- ***Blocks contain sequences at all different evolutionary distances and may be highly biased (e.g. many identical sequences)***



Scoring Systems

BLOSUM Matrices

- ***Dealing with bias and distance***
  - Cluster all sequences with less than X% identities
    - Clustered sequences count as 1 sequence
    - if X is 100% it simply removes identical sequences
    - if X < 100% it reduces the weight on closely related sequences
- ***Calculate substitution frequencies and log-odds matrix***
- ***This gives a BLOSUM X table***
  - BLOSUM 62 - sequences greater than 62% identical are clustered
  - BLOSUM 80 - sequences greater than 80% identical are clustered



Scoring Systems

BLOSUM Matrices

- *METHYLTRANSFERASE BI*

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47
CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

Unweighted (BLOSUM 100) count of transitions for column 1, total $(n^2 - n)/2$ transitions

$c_{FF} = 0$ $c_{FI} = 1$ $c_{FL} = 4$ $c_{FV} = 2$

$c_{II} = 0$ $c_{IL} = 4$ $c_{IV} = 2$

$c_{LL} = 6$ $c_{LV} = 8$

$c_{VV} = 1$

Scoring Systems

BLOSUM Matrices

Unweighted (BLOSUM 100) count (c_{ij}) of transitions for column 1

N=28	F	I	L	V
F	0	1	4	2
I	1	0	4	2
L	4	4	6	8
V	2	2	8	1

N = 28 transitions, $f_{ij} = c_{ij} / N$

	F	I	L	V
F	0.00	0.04	0.14	0.07
I	0.04	0.00	0.14	0.07
L	0.14	0.14	0.21	0.29
V	0.07	0.07	0.29	0.04

Log-Odds $s_{ij} = \log_2(f_{ij} / p_i p_j)$ -
Background frequencies, p_i, from database

- ***$p_F = 0.0397$***
- ***$pI = 0.0529$***
- ***$p_L = 0.0917$***
- ***$p_V = 0.0649$***

	F	I	L	V
F	0.00	4.09	5.29	4.79
I	4.09	0.00	4.88	4.38
L	5.29	4.88	4.67	5.59
V	4.79	4.38	5.59	3.08

Scoring Systems

BLOSUM Matrices

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68

COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47

CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

Count as 1 sequence in BLOSUM80

- ***Sequences are not independent because they are closely related, in this case COMT_EUCGU, CHMT_POPTM, and COMT_MEDSA are all >80 identical, and the other sare more different***
- ***BLOSUM approach accounts for this by treating the entire group as a count of 1***
  - weighted counts in column 1: F=1 I=1 L=2.67 v=1.33
- ***Weighted (BLOSUM 80) count of transitions for column 1***

n=15	F	I	L	V
F	0.00	1.00	2.67	1.33
I	1.00	0.00	2.67	1.33
L	2.67	2.67	2.23	3.55
V	1.33	1.33	3.55	0.22

Scoring Systems

BLOSUM Matrices

- ***BLOSUM80***

	F	**I**	**L**	**V**
F	0.00	0.04	0.14	0.07
I	0.04	0.00	0.14	0.07
L	0.14	0.14	0.21	0.29
V	0.07	0.07	0.29	0.04

transition frequencies

	F	I	L	V
F	0.00	4.99	5.61	5.10
I	4.99	0.00	5.20	4.69
L	5.61	5.20	4.14	5.31
V	5.10	4.69	5.31	1.80

log-odds

higher	
lower	

- ***BLOSUM100 (same as BLOSUM0)***

	F	I	L	**V**
F	0.00	0.07	0.18	0.09
I	0.07	0.00	0.18	0.09
L	0.18	0.18	0.15	0.24
V	0.09	0.09	0.24	0.01

transition frequencies

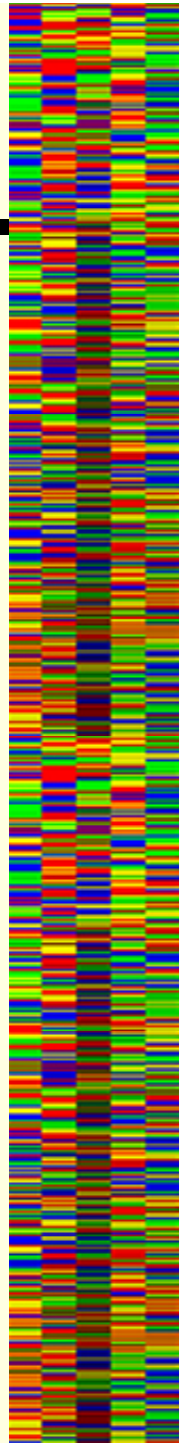
	F	I	L	V
F	0.00	4.09	5.29	4.79
I	4.09	0.00	4.88	4.38
L	5.29	4.88	4.67	5.59
V	4.79	4.38	5.59	3.08

log-odds

Scoring Systems

PAM vs BLOSUM

- ***PAM***
  - Based on explicit evolutionary model
  - Represents a specific evolutionary distance
  - Ranges from identical to completely random
- ***BLOSUM***
  - Based on empirical frequencies
  - Always a blend of distances as seen in the database/PROSITE
  - Narrower range than PAM matrix



Scoring Systems

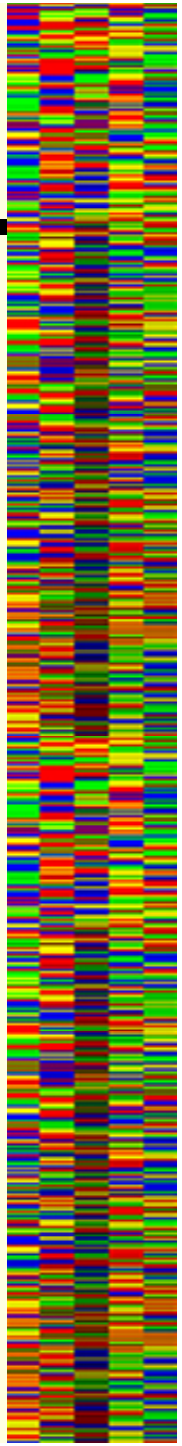
PAM vs BLOSUM

- ***Scoring matrices are difficult to compare because they reflect different target frequencies and evolutionary distances. How do you put the scores on a single scale?***
- ***Relative Entropy***
  - Average information per residue pair

    $H = \Sigma s_{ij} \log_2 (s_{ij} / p_i p_j)$

  - Equivalent PAM and BLOSUM matrices based on relative entropy

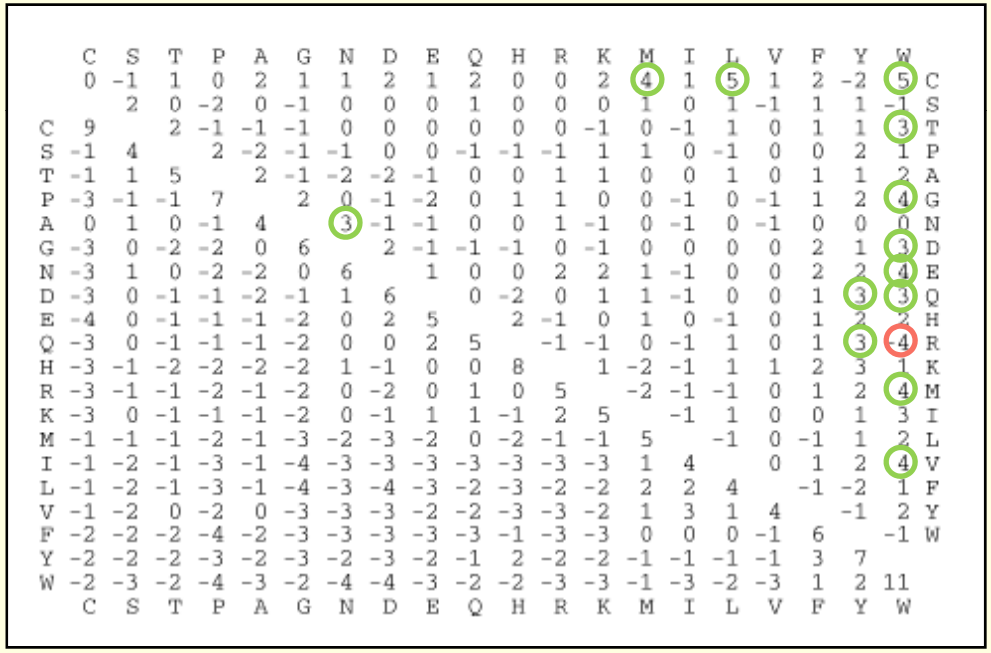
    PAM80 → Blosum100
    PAM100 → Blosum90
    PAM120 → Blosum80
    PAM160 → Blosum62
    PAM200 → Blosum50
    PAM250 → Blosum45
    PAM340 → Blosum35



Scoring Systems

PAM vs BLOSUM

- ***Differences mostly affect rare amino acid residues which were under-counted in 1978***



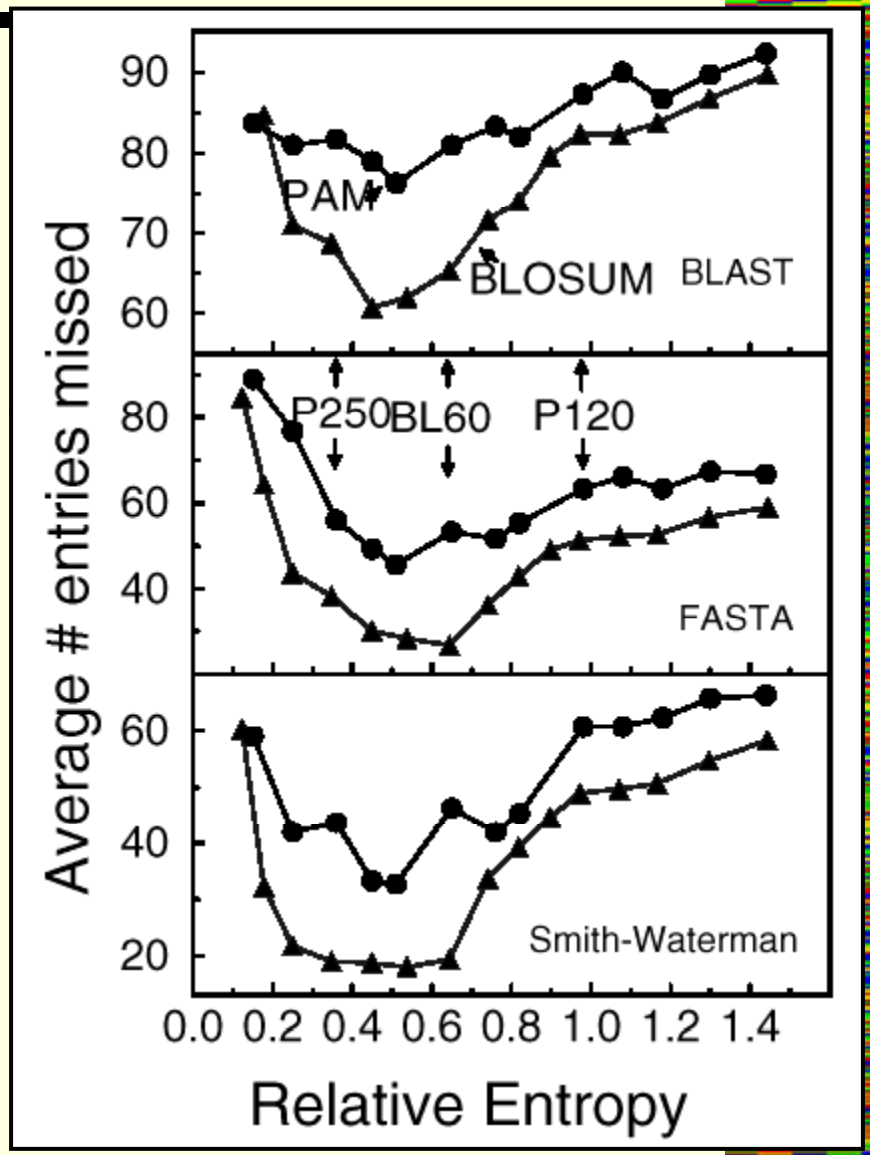
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
		2	0	-2	0	-1	0	0	0	1	0	0	0	1	0	1	-1	1	1	-1	S
C	9		2	-1	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
S	-1	4		2	-2	-1	-1	0	0	-1	-1	-1	1	1	0	-1	0	0	2	1	P
T	-1	1	5		2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	A
P	-3	-1	-1	7		2	0	-1	-2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

  - Upper matrix difference from PAM 160
  - Lower matrix BLOSUM 62

Scoring Systems

PAM vs BLOSUM

- *Use known families. Use one member of family as query.*
- *Count number of family members missed at E=10 (BLAST) or P=.005 (FASTA and SW)*



Scoring Systems

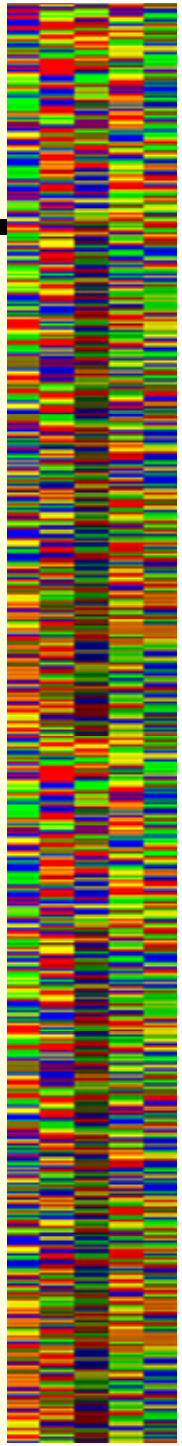
Target frequencies

- *Karlin and Altschul showed that for MSPs, amino acids a_i and a_j will be aligned with frequency approaching*

$$q_{ij} = p_i\, p_j\, e^{-\lambda s}$$

where p_i and p_j are the expected probabilities of observing the amino acid residues and s is the match score

- *A given scoring matrix will "try" to align the residues according to the above equation, so q_{ij} are a characteristic set of target frequencies for the scoring matrix*
- *The best scoring system is the one in which the target frequencies are the same as the frequencies of the actual aligned residues*



Scoring Systems

Target frequencies

- *By rearranging the target frequency equation we get:*

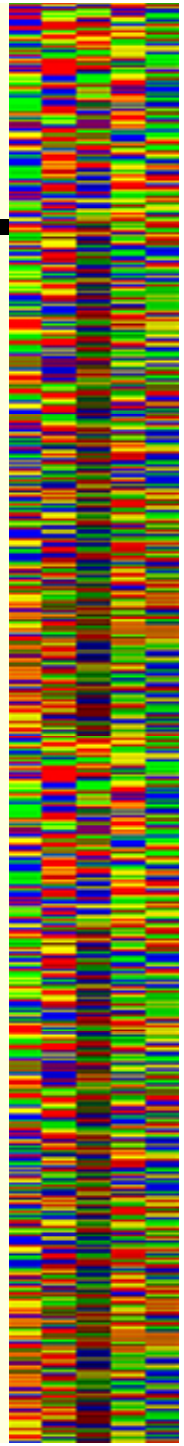
$$s_{ij} = [\ \ln (q_{ij} / p_i\ p_j)\ ] / \lambda$$

- *All scoring systems can therefore be looked on as log-odds matrices with an implied set of target frequencies!*
- *Since multiplying a log-odds scoring system by a constant won't change the relative score for local alignments, λ can be looked on as a scaling factor that we can choose to suit our convenience.*
- *One convenient choice for λ is ln2 , so that the scores can be thought of as representing bits of information*

Scoring Systems

Information

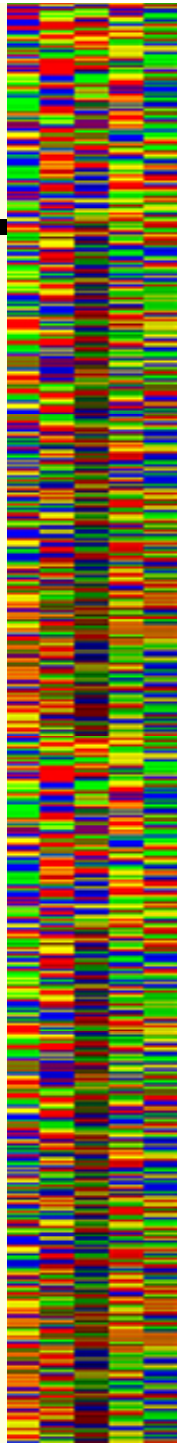
- *A bit of information is the amount of information needed to distinguish between 2 possibilities, i.e., one yes-no question.*
- *Taking λ is ln2 , the Karlin-Altschul equation becomes* $p = KNe^{-\lambda S}$ $p = KN\,2^{-S}$
- *Rearranging gives the score required to find a given number of MSPs with score* $S = \log_2 (K/p) + \log_2 N$
- *K is generally about 0.1 so the first term is negligible*
- *The amount of information required to distinguish an MSP from chance therefore depends entirely on N , the size of the comparison*
- *N is the product of the lengths for two sequences, or the size of the database times the sequence length for a search*



Scoring Systems

Information

- *How much information do you get per base or residue in an alignment?*
- *Relative entropy:* $H = \Sigma q_{ij} s_{ij} = \Sigma q_{ij} \ln (q_{ij} / p_i p_j)$

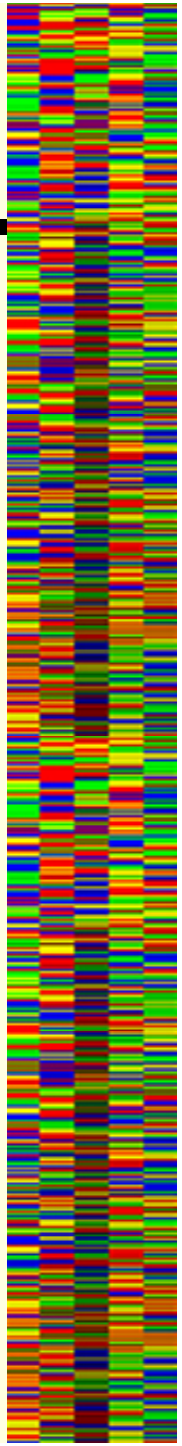


Scoring Systems

Information

- ***Proteins ~3.7 bits/residue (theoretical 4.3)***
  - An MSP of about 16 bits is required for significance in a pairwise comparison of two 250 long sequences
    $log_2 (250^2) = 15.93$ bits
  - for a 250 residue protein sequence and the NCBI nr database,
    $log_2 (1.4 \times 10^8 \times 250) = 35.0$ bits ~ 9 residues

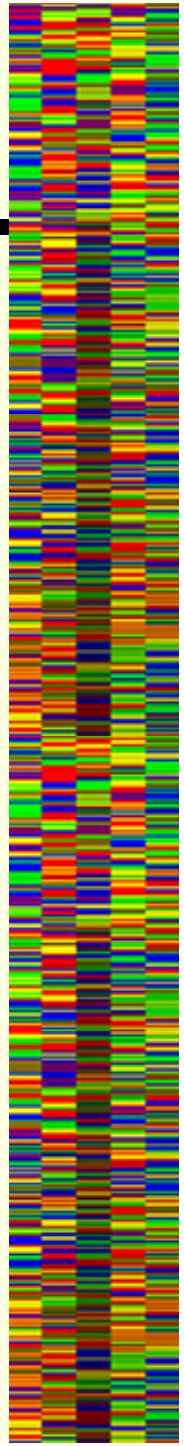
- ***DNA ~ 1.5 bits/base (theoretical 2)***
  - for a 1000 base long DNA sequence and the NCBI nr database,
    $log_2 (9 \times 10^{10} \times 10000) = 49.7$ bits ~ 33 bases



Review

Scoring Matrices - summary

- *All scoring matrices imply "target frequencies", the residue pairs that the matrix will "try" to get in an alignment*
- *Best alignments (both database sensitivity/specificity and correctness) are found when the target frequencies match the observed frequencies for the correct alignment*
- *Our ability to find similar sequences depends almost entirely on the size of the data. This can be roughly measured in bits.*

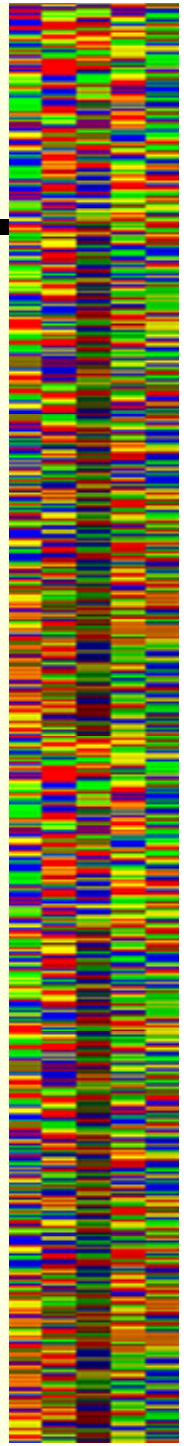


Scoring Systems

PAM vs BLOSUM

Comparison with BLAST and Smith-Waterman (SW)

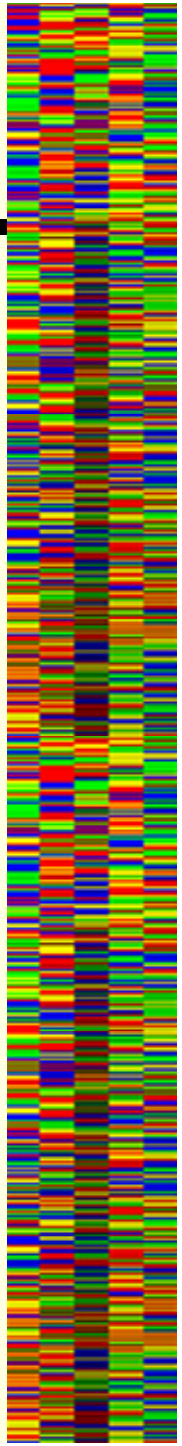
Probe	BLAST ROC50	Best Matrix	SW ROC50	Best PAM Matrix	Open	Ext
FER_ENTHI	0.812	PAM120	0.951	PAM150	24	2
FER_THETH	0.805	PAM120	0.889	PAM150	22	10
ASRA_SALTY	0.521	BLOSUM62	0.455	PAM100	22	22
DCMA_METSO	0.426	PAM250	0.503	PAM200	22	2
FRXB_WHEAT	0.823	BLOSUM62	0.811	PAM200	24	2
PSAC_CHLRE	0.803	BLOSUM62	0.932	PAM200	26	2
AMPR_RHOCA	0.935	PAM250	0.972	PAM200	16	4
BLAA_STRCI	0.690	BLOSUM62	0.743	PAM150	16	8
GLTC_BACSU	0.891	BLOSUM62	0.891	PAM200	18	18
METR_SALTY	0.951	BLOSUM62	0.999	PAM200	26	26
FLIA_SALTY	0.861	BLOSUM62	0.939	PAM100	20	2
RP32_ECOLI	0.851	PAM250	0.959	PAM150	14	2
RP70_ECOLI	0.852	PAM250	0.956	PAM250	26	2
RPSH_BACSU	0.740	BLOSUM62	0.857	PAM250	24	24
RPSK_BACSU	0.873	BLOSUM62	0.944	PAM100	14	2



Sequence Comparison

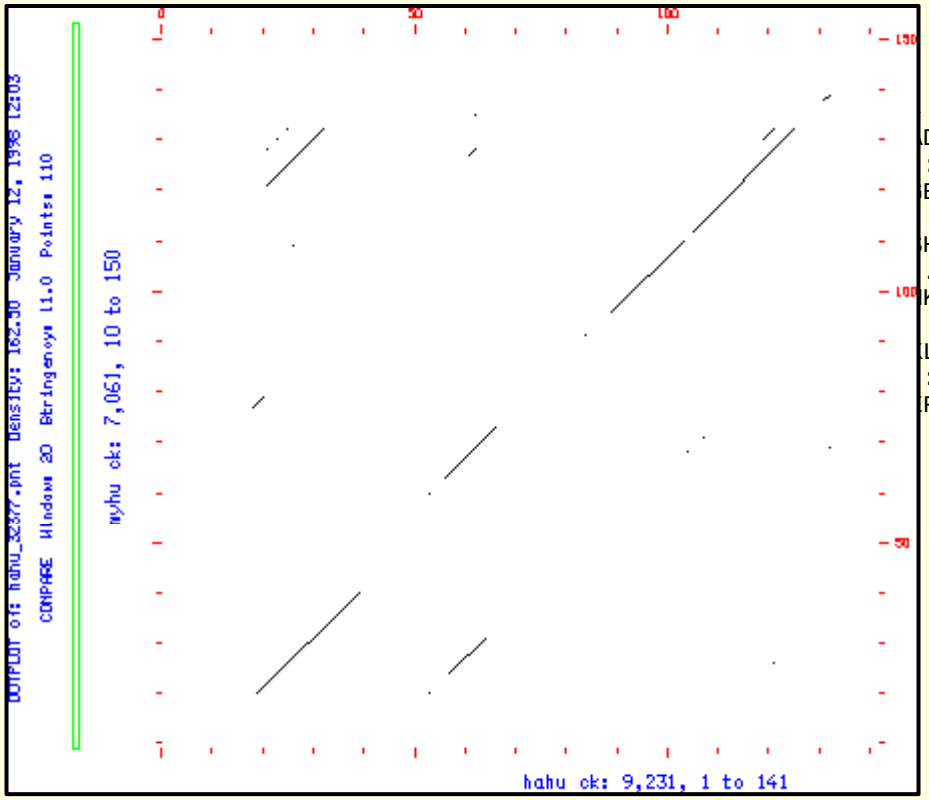
Alignments and DotPlots

- ***We want to match two sequences so that we can see the evolutionary similarity between them***
  - Which functional domains correspond
  - Which functional residues correspond
- ***The matching acts as a map for applying information known about one molecule to the other***
  - What activities can one infer
- ***Two important depictions of sequence matchings***
  - Dot matrix plots
  - Alignments



Sequence Comparison

Alignments and DotPlots

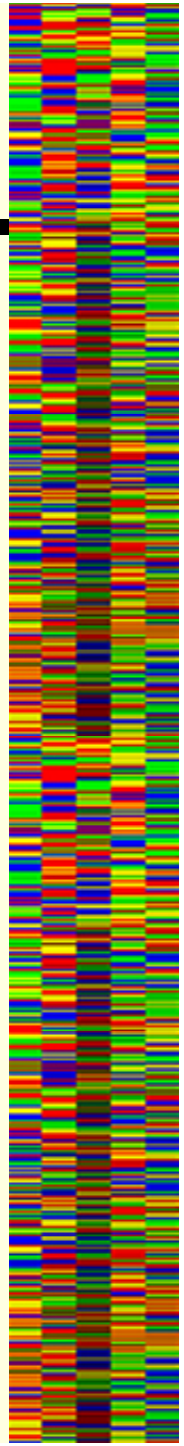



```
LDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.... 46
:.  | ..||||:|. .:.|.|.| |:| : |.| . |..|
EWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLK 50
     .          .          .          .
HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD 94
.:|.::| || .| .||.. : . :. ...:.:|.: || | ::.
KASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIP 100
     .          .          .
LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 141
::|.|:: .|... |::|.:..::.::| |. ... :.|.|:
FISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK 147
```

Sequence Comparison

Alignments and Dot Matrix Plots

- ***Alignments and dot matrix plots produce mappings***
- ***Very powerful way to apply information known about one molecule to another***
- ***Only works if the mapping is correct***
  - Biologically significant - this is what we want to know but we can't ask the question directly
  - Statistically significant - this is the question we can actually ask using a program based on a mathematical model



Statistics: Part II

Is there homology?

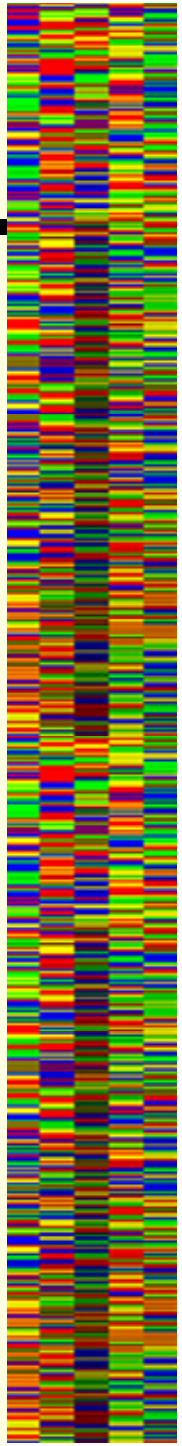
- ***We can answer, "yes," if the result is surprising when compared to unrelated sequences***

What do we mean by surprising?

- ***We are surprised when an event is very unlikely to happen by chance. In this case, we are surprised when the observed level of similarity is very unlikely between unrelated sequences***

This requires a model for unrelated sequences

- ***Most common choice is a random sequence model***



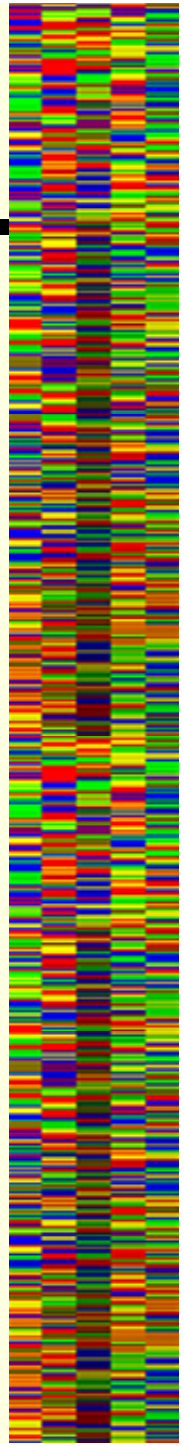
Sequence Comparison

Alignment

- *Provides a one-to-one picture of the residues or bases in the sequences that correspond*

```
 1 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.... 46
   .||.::.  | ..||||:|. .:.|.|.| |:| : |.| . |..|
 1 GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLK 50
             .          .          .          .
47 ..DLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD 94
     |  .:|.::| || .| .||.. : . :. ...:.:|.: || | ::.
51 SEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIP 100
```

- *Computational problem is putting in gaps*

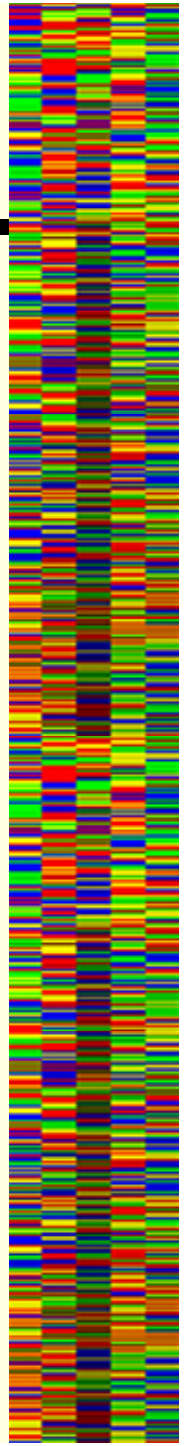


Sequence Comparison

Dynamic Programming Alignment

- ***Much faster than brute force but too slow for searching***

- ***A "rigorous" method***
  - Rigorous = no approximations, all numbers and positions of gaps
  - Every possible alignment is considered
  - Requires time proportional to product of sequence lengths, O(nm)
- ***Optimal Alignment given:***
  - Scores for matches/mismatches
  - Penalties for gaps
  - Score = matches + mismatches + gaps
- ***Affine gap model:***
  - Gap score is Gap_Open + Gap_extend x Gap_Length
  - Generally this is a negative score, i.e., a penalty



Sequence Comparison

Dynamic Programming Alignment

- ***Calculation of alignment score using affine gap model***

```
BOVGH  G G T G C C A C -  T C C C A - - - - - C T G
         : : : : : : : :    : :   :             : : :
A02321 A G T G C C A C C  C C C A A T G C C G C T G
      -3+4+4+4+4+4+4+4-12-3+4+4-3+4  -12-4x4 +4+4+4
```

	A	C	G	T
A	4			
C	-3	4		
G	-3	-3	4	
T	-3	-3	-3	4

matches	52
mismatches	-9
gaps	-40
Overall Score	3

Gap opening = -12
Gap extension = -4