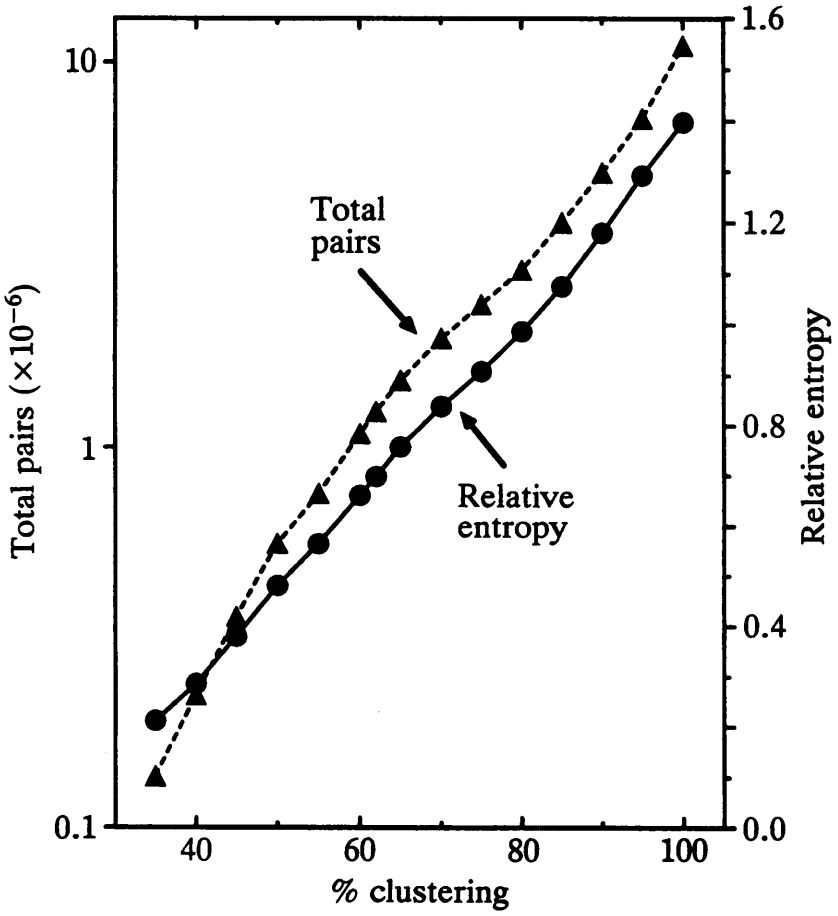
Cluster percentage and AA pair count



BLOSUM62

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

$$\text{round}\{2\log_2 \frac{\widehat{p_{ij}}}{\widehat{q}_i\widehat{q}_j}\}$$

The origin of the BLOCKS

- The aligned (gapless) blocks come from
  - . . . alignning sequences
  - for which we need a scoring matrix . . .
- $2\times$ Henikoff used an iterative approach to circumvent this circular reasonning.

Generating blocks using PROTOMAT

- Input is a group of related proteins
- For *each* group the program MOTIF (Smith, Annau, Chandrasegaran 87) linearly scans for motifs of the form $A_1 - d_1 - A_2 - d_2 - A_3$
  - Overrepresented motifs are determined by a Poisson approximation ($\lambda = nlP_{A_1}P_{A_2}P_{A_3}$) and a user selected significance level
  - The ungapped alignments (blocks) containing the significant motifs are pruned (combining shorter motifs to longer ones)
  - Each surviving block is scored by sum of pairs in each (positively scored) column using a user defined similarity matrix
  - Each block is used to (re)align the group to itself
- The top 50 blocks are extended and merged if possible
- Statistical significance is determined by shuffling the sequences

Group's block assembly in PROTOMAT

- We now have a set of blocks "overlapping in different ways in various subsets of sequences"
- Want to find a best path of nonoverlapping blocks which would serve as a signature for this group
  - Construct a directed graph whose vertices are the blocks
  - Draw a directed edge from block a to b if a fully precedes b in at least x of the sequences ($x \geq \max(n/2, m)$ where m is the MOTIF significance level?)
  - Each vertex has a score: block score $\times$ number of merged motifs
  - Path score is the sum of vertex score times the proportion of sequences in the path.
  - Using DFS (acyclic - why?) score each path and choose best path
- The blocks from the best scoring path are recorded

Using PROTOMAT to construct the BLOCKS

- Raw data included 504 nonredundant groups of proteins from Prosite 8.0
- Using a 0-1 scoring matrix PROTOMAT generates 2205 blocks
- These are used to create a scoring matrix a-la BLOSUM60
- Rerun PROTOMAT with the new scoring matrix to generate 1961 blocks
- Create a new "BLOSUM60" matrix from these
- Use this matrix in PROTOMAT on 559 groups of Prosite 9.0 to generate 2106 blocks (3-60 wide and 2-200+ deep)
- Generate the full range of BLOSUMX matrices.

Markov chains

- A stochastic process X_n, $n = 1, 2, \ldots$ (each X_n is a random variable) is a Markov chain if

$$P(X_n = j | X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) = P(X_n = j | X_{n-1} = i_{n-1})$$

- The state space or simply the states of the chain are all js for which the above is positive for *some* choice of i_ks
- The chain is homogenuous if the transition matrix $P = (p_{ij})$ is independent of n

$$P(X_n = j | X_{n-1} = i) = p_{ij}$$

- Let $P_n(i,j) = P(X_n = j | X_1 = i)$ then,

$$
\begin{aligned}
P_n(i,j) &= \sum_k P(X_n = j, X_{n-1} = k | X_1 = i) \\
&= \sum_k P(X_n = j | X_{n-1} = k, X_1 = i) P(X_{n-1} = k | X_1 = i) \\
&= \sum_k P(X_n = j | X_{n-1} = k) P_{n-1}(i,k) \\
&= \sum_k p_{kj} P_{n-1}(i,k)
\end{aligned}
$$

  - i.e. $P_n = P_{n-1} P$ and by induction $P_n = P^n$
  - Chapman-Kolmogorov equation

Stationary distribution

- A chain is irreducible if for states i, j there exists n s.t. $P_n(i,j) > 0$
- If $X_1 \sim \boldsymbol{q}$ where $\boldsymbol{q}$ is a probability vector then $X_2 \sim \boldsymbol{\mu} = \boldsymbol{q}P$:

$$\mu_j := P(X_2 = j) = \sum_i P(X_2 = j | X_1 = i) P(X_1 = i) = \sum_i p_{ij} q_i$$

  - more generally, $X_n \sim \boldsymbol{q}P^{n-1}$
- A probability vector $\boldsymbol{\pi}$ is a stationary or invariant probability vector of the chain (P) if $\boldsymbol{\pi}P = \boldsymbol{\pi}$
- Steady state: if $X_1 \sim \boldsymbol{\pi}$ then so is $X_n \; \forall n$
- An irreducible homogenuous Markov chain has a unique stationary distribution $\boldsymbol{\pi}$
- Moreover, for any probability vector $\boldsymbol{q}$, $\lim_n \boldsymbol{q}P^n = \boldsymbol{\pi}$

Accepted Point Mutation (Dayhoff et al. 68,72,78)

- "An APM in a protein is a replacement of one AA by another accepted by evolution"
- We want to estimate the
  - probability that given a site with AA A has udergone an APM, the new AA is B
  - the rates each AA undergoes an APM
- Dayhoff et al. estimated those from hypothetically constructed phylogenetic trees
  - originally phylogenetic trees were used to represent evolutionary relationship between species
  - they can be used to represent relationship between sequences
  - trees relating the sequences in 71 families were constructed using the *parsimony* method