Scoring Matrices

The Arrays Used to Find and Evaluate Protein Homologies

Overview of Presentation

- What are scoring matrices?
- Some rudimentary scoring matrices
- More advanced scoring matrices
  - PAM
  - BLOSUM
- Advantages of BLOSUM over PAM

What are Scoring Matrices?

- 20x20 matrix.
  - Each row and column corresponds to an amino acid
- Chart of “similarity” between amino acids

How to Use a Scoring Matrix

- For every pair of amino acids in the two sequences, there is a corresponding score in the matrix.

- Add up the scores for all the aligned pairs in the two sequences you're comparing. The higher the total score, the more "similar" the two sequences are.

	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

Uses of Scoring Matrices

- Used in all processes that involve comparing amino acid sequences.
  - Finding homologies between sequences
  - Finding optimal alignment between sequences
  - Finding repeated regions within a sequence

Unitary Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	1																			
S	0	1																		
T	0	0	1																	
P	0	0	0	1																
A	0	0	0	0	1															
G	0	0	0	0	0	1														
N	0	0	0	0	0	0	1													
D	0	0	0	0	0	0	0	1												
E	0	0	0	0	0	0	0	0	1											
Q	0	0	0	0	0	0	0	0	0	1										
H	0	0	0	0	0	0	0	0	0	0	1									
R	0	0	0	0	0	0	0	0	0	0	0	1								
K	0	0	0	0	0	0	0	0	0	0	0	0	1							
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1					
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	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	1

- A pair of identical amino acids gets a +1 score
- A pair of nonidentical amino acids gets a 0 score

Genetic Code Matrix

- Looks at the number of shared nucleotides between the codons for the two amino acids.
- Scoring for an amino acid pair:
  - +3 if the amino acids are identical
  - +2 if the amino acids have codons that share 2 nucleotides
  - +1 if the amino acids have codons that share 1 nucleotide
  - +0 if no nucleotides are shared

PAM Matrices

- PAM = Percent Accepted Mutation
- Based on observed frequencies of evolutionary change in sequences of amino acids

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

PAM250

Making PAM Matrices

- To Create a PAM Matrix:
  - Arrange a family of closely related proteins (greater than 85% identical) into an phylogenic tree
  - Observe frequencies of amino acid changes across individual evolutionary steps
  - Scores are generated from these frequencies

Drawbacks of PAM Model

- Assumes that all types of mutations are distributed uniformly across proteins.
- Uses data from closely related proteins to infer relationships between very different proteins

BLOSUM Matrices

- Developed by Steven Henikoff and Jorja Henikoff in 1992
- BLOSUM = Block Substitution Matrix
  - Block = ungapped, highly conserved region in a protein family

	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

Making BLOSUM Matrices

- Initial data comes from BLOCKS database
- To Create Matrix:
  - For all blocks in all protein families, calculate the frequency that two related proteins have a specific pair of amino acids aligned within a block region.
  - Divide this by the frequency expected by chance.
  - This will produce a score for each pair of amino acids, which is represented as an entry on the matrix

BLOSUM Matrices – Clustering

- What if a protein family contains a group of very similar proteins?
  - The scores will be skewed away from change. This hides information about variation, which is what we're most interested in.
- Solution: Clustering.
  - Cluster = a group of sequences that agree on a minimum percentage of their amino acids.
  - Members of a cluster are treated as one sequence. Data obtained from them are averaged together

Advantages of BLOSUM over PAM

- BLOSUM matrices are created only using blocks
  - These conserved regions are the most useful areas to give us information about meaningful homologies
- BLOSUM incorporates differences between distantly related proteins
  - Scoring matrices are often used to compare very different proteins, so this is an appropriate source of data.
- BLOSUM gathers data using larger protein families.
  - Generally, more data → more accurate

Summary

- The two major sets of scoring matrices:
  - PAM is based on evolutionary mutation rates across entire proteins
  - BLOSUM is based on differences between conserved regions of related proteins
  - BLOSUM generally works better

Looking ahead

- As we discover new blocks, new proteins, and new scoring methods, these scoring matrices will be updated and inevitably replaced by better models
- Every model is an approximation