Vanderbilt Medical Center

BMIF 310: Foundations of Bioinformatics

Sequence Analysis: Lecture B

Substitution matrices: PAM and BLOSUM

Overview

- Prediction of matrices by theoretical means
- Generation of matrices by empirical means
  - PAM matrices
    - MO Dayhoff, RM Schwartz, BC Orcutt
    - Atlas of Protein Seq. and Struct. (1978) 5:345-352
  - BLOSUM matrices
    - S Henikoff and JG Henikoff
    - PNAS (1992) 89: 10915-10919

Substitution matrix role

- In comparing sequences, one should account for the influence of molecular evolution.
- The probability of *acceptably* replacing an amino acid with a similar amino acid is greater than replacement by a very different one.
- Substitution matrices evaluate potential replacements for protein and nucleic acid sequences.

Jukes-Cantor (1969)

- Assumes that if mutation happens, change to any other letter occurs with equal probability.
- Probability of mutation increases as linear function of time.

	A	C	G	T
A	$1-3\alpha$	α	α	α
C	α	$1-3\alpha$	α	α
G	α	α	$1-3\alpha$	α
T	α	α	α	$1-3\alpha$

Kimura (1980)

- Differentiates transition and transversion rate.
- α: Transition= C ↔ T, G ↔ A
- β: Transversion= T, C ↔ G, A

	A	C	G	T
A	1-α-2β	β	α	β
C	β	1-α-2β	β	α
G	α	β	1-α-2β	β
T	β	α	β	1-α-2β

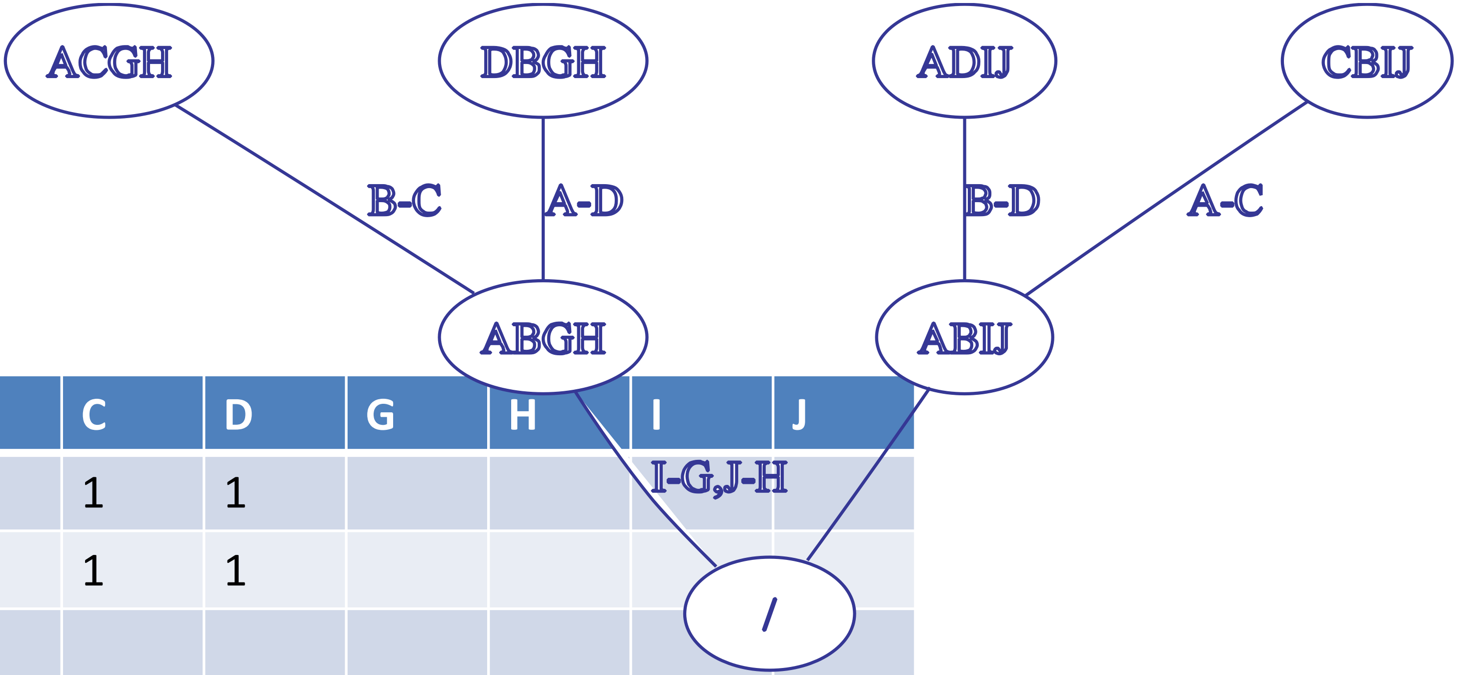
Theoretical vs. empirical substitutions

- Theoretical substitution matrices embody theoretical models of evolution.
- Empirical substitution matrices are constructed from the results of evolution.
- Empirical matrices are more common for proteins because codon position of mutations modulates effect at protein level.

Point Accepted Mutation (PAM)

- “An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection.”
- Aim is to characterize accepted mutations at one PAM (1% of residues changed) distance.
- Dataset included only 1572 mutations in 71 conserved proteins. Sequences in each tree were no more than 15% different. Updates have been computed with more sequences.

Protein sequence trees imply transition matrix



	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

After Dayhoff figs 78 and 79

Relative mutability

- F_i= frequency for AA_i among sequences (Leu is 9.7% of SwissProt, while Trp is only 1.1%)
- $M_{i,j}$= frequency of mutation for $AA_i \rightarrow AA_j$
- Relatedness odds are a ratio of the two: $M_{i,j} / F_i$
- Log odds ratio is $\log_{10}(M_{i,j} / F_i)$

- Asn and Ser showed high mutability, while Cys and Trp showed low mutability.

What PAM-1 means

- Given residues *x* and *y*, we can look up the probability that a given *x* will be replaced by *y* in the time that 1% of residues have mutated.

- Multiplying PAM-1 by itself *n* times yields a matrix showing mutation probabilities over longer intervals of evolution (e.g. PAM-250).

PAM-250 ($\log_{10}$ scale)

- 0 means probability of replacement is what one would expect from random chance.
- 1 means probability of replacement is 10-fold greater than random chance would suggest.
- -1 means probability of replacement is 10-fold lower than random chance would suggest.

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	8								
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	-8	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	8				
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	-4	-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

BLOSUM (blocks substitution matrix)

- PAM models mutations using very similar sequences. Many uses of substitution matrix focus on matching distant sequences.
- Rather than build a distant matrix by raising PAM-1 to a power, why not build a matrix from more distantly related sequences?
- Henikoff and Henikoff used BLOCKS database of short, ungapped sequences to create BLOSUM matrices at several similarity levels.

Substitution without phylogeny

```
LDGADCIMLSGETAKGDYPL
LDGADCIMLSGETAKGDYPL
LDGADCIMLSGETAKGDYPL
LDGADCVMLSGETAKGEYPL
LDGTDCVMLSGETAAGAYPE
FDGTDAIMLSGETAAGIYPV
LDGTDAVMLSGESAKGKYPL
REGADAVMLSGETAHGKYPL
YDGTDCLMLSNETTIGKYPI
```

- If each letter were paired with every other, we would observe **VV**, **VI**, **VL**, **II**, and **IL** pairs.
- These pair counts are accumulated across the positions of multiple blocks into a single table.
- At least 2369 occurrences of any substitution.

Two probabilities make an odds ratio

- Observed probability of occurrence:

q_{ij} = the fraction of table sum found in this cell.

- Expected probability of occurrence:

e_{ij} = the product of the background probabilities of either residue.

- BLOSUM rounds values of $\log_2(q_{ij}/e_{ij})$.
- 0 is expected rate, positive is more than usual.

Different BLOSUM, different data

- The BLOSUM number describes how identical sequences must be to be counted as one sequence. BLOSUM62 = 62% identical.
- When number is low, closely related sequences are condensed, emphasizing most diverse sequences in matrix.
- When number is high, only nearly identical sequences are condensed, emphasizing sequences of lesser diversity.

BLOSUM62 ($\log_2$)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

R E G A D A V M L S G E T A H G K Y P L

-2+2+6+0+6+0+1+5+4+4+0+5+5+0-3+6+5+7+7+2

Y D G T D C L M L S N E T T I G K Y P I

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Summary

- Both theoretical and empirical techniques can yield substitution matrices.
- PAM matrices are designed to model small evolutionary changes, but they can be extrapolated to handle larger scales.
- BLOSUM matrices effectively represent more distant sequence relationships, and BLOSUM62 has become a standard matrix.
- Protein-level comparisons are more common for distant evolutionary relationships.