

Bioinformatics: Life Sciences on Your Computer

Substitution Matrices, Gap Penalties

This is supplemental reading to the video on PAM Matrices.

Substitution scores/gap penalties

There are three major types of mutations but as far as pairwise alignment goes, there are only two. **Substitution** is single amino acid (or nucleotide) change. An **insertion** or **deletion** creates a gap in the alignment. So there are only two things to score--**mismatches** and **gaps**.

Substitution matrices are used to score mismatches in protein sequences. These matrices take advantage of conservative substitutions and that some mismatches better than others. Statistically, some matches better than others in protein sequences, because some amino acids are more common than others.

Gap penalties are broken into two parts. There is a penalty for opening or the existence of a gap and a penalty for extending the gap. There are two ways to define that and that can be confusing.

A gap of length 5 in an alignment can be defined in two ways:

Open/extension

Gap open penalty assessed for opening gap.

Gap extension penalty assessed for extending the gap four more spaces (to make it five spaces).

Existence/length

Gap existence penalty assessed for the existence of a gap.

Gap length penalty assessed for a gap of five spaces in length.

Different algorithms choose one or the other interpretation, but note that the concept is the same!

In the example below, note that match scores are different and mismatch scores are also different. Some mismatches have a positive score. The gap has opening + extend penalties. The raw score calculated but raw scores depend on the scoring system.

(b)

Score = 18.1 bits (35), Expect = 0.015, Method: Composition-based stats
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query	12	V	T	A	L	M	G	K	V	N	D	--	E	V	G	E	A	I	C	R	L	L	33		
		V	+	G	K	V	D												C	E	I	R	L		
Sbjct	11	V	I	N	V	G	K	V	E	A	D	I	P	G	H	C	Q	E	V	L	I	R	L	F	34
match		4	11	5	6	6	5	4	5																sum of matches: +60
mismatch		-1	1	0	-2	-2	-4	0																	sum of mismatches: -13
gap open												-11													sum of gap penalties: -12
gap extend												-1													total raw score: 60 - 13 - 12 = 35

Discussion topic: Any confusions about gap penalties?

Amino acid substitution matrix

The substitution matrix is used to score all amino acid matches and mismatches. The PAM matrices are based on accepted point mutations (but PAM sounds better than APM). The matrices are based on mutation rates. Specifically what point mutations are accepted by nature. Common substitutions are assigned higher scores.

BLOSUM stands for Blocks Amino Acid Substitution Matrices, and are based on conserved amino acid patterns (blocks) from families of proteins. The blocks help determine what amino acid changes tend to be functional in proteins. Changes that maintain function are assigned higher scores.

Discussion topic: From the video lecture, how would you score that alignment using [PAM10](#) and [PAM250](#)? Try some other matrices.

```
Query 1 MDLSELERD
Sbjct 1 MDISQYEMD
```

Shallow matrices

Low number PAM matrices are used to align very related sequences that may not have diverged that long ago. For instance, human, chimp orthologs diverged more recently than human, bacteria orthologs.

Low number PAM matrices are also better for short sequences, for statistical reasons. Shallow matrices have higher match/mismatch penalties and higher scores may be needed to achieve significance. It's not necessarily ideal but necessary for statistical purposes.

Deep matrices have high PAM numbers and are best for finding distant relationships.

BLOSUM matrix

In the BLOSUM matrix, rather than extrapolate based on evolution rate, the numbers are based on ~2000 conserved amino acid patterns (blocks) from proteins in the same family. The actual amino acid changes in the conserved regions are examined by direct observation and then a similar $\log(\text{observed/expected})$ formula is used to generate the matrix. The sequences were derived from protein database from families of related proteins, sometimes distantly related on evolutionary scale.

The numbers are more difficult to interpret. In the BLOSUM 62 matrix sequences used to construct the matrix are less than 62% identical. If two sequences have higher than 62% identity, one of those sequences is discarded and not used for statistical measures. For instance, the human RNase H2 A subunit is 96% identical to the ortholog of the same protein in monkeys. Because 287 of the 300 amino acids are identical, most sites are noninformative. Not only are they noninformative, they may cause an overrepresentation of the primate amino acids in calculating the observed/expected ratio. So only one sequence is counted.

The human RNase H2 A protein is 64% identical and 78% similar to the ortholog in the lizard *Anolis carolinensis*. That's still too close for both proteins to be represented in BLOSUM 62. However, human RNase H2 A is 51% identical and 66% similar to the sea squirt ortholog. That's distant enough that both sequences are accepted.

Now in BLOSUM 80, the human - lizard alignment IS below the required 80% so it can be included in constructing the BLOSUM 80 matrix, as would human and sea squirt. But in the BLOSUM 50 matrix, these particular human-sea squirt proteins are too close to be considered.

If you can wrap your head around that, you would conclude that BLOSUM 50 is based on more distantly related proteins and that the high number BLOSUM includes closer relatives. That's the reverse of PAM. So shallow matrices have low PAM numbers or high BLOSUM numbers. Deep matrices have high PAM numbers or low BLOSUM numbers.

BLOSUM 62 matrix

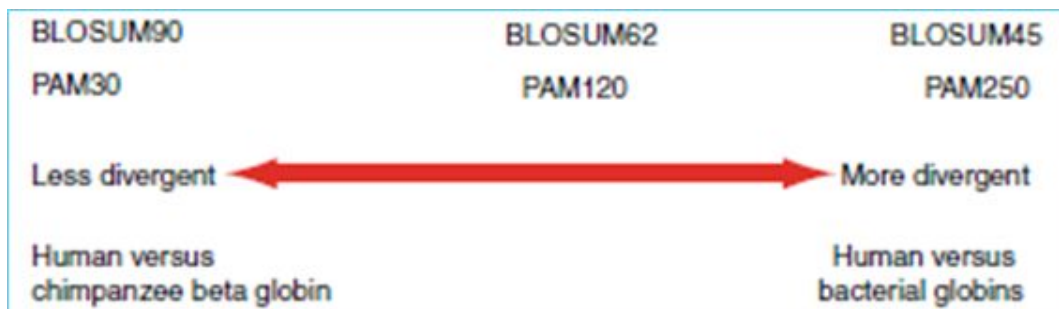
Below is the BLOSUM 62 matrix, the most commonly used substitution matrix.

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

<http://www.acm.org/crossroads/wikifiles/I3-I-1-1/1/1/Image7-BLOSUM62.jpg>

PAM vs. BLOSUM

This is a VERY broad generalization but PAM can be better for phylogenetic analysis of species and BLOSUM can be better for finding related proteins. If proteins are very similar (or short), use low PAM or high BLOSUM--shallow matrices. If proteins are divergent, use high PAM or low BLOSUM--deep matrices. Matrices are for protein only, not DNA, and PAM250, BLOSUM62 are the most common.



--image from Jonathan Pevsner

Discussion topic: PAM vs. BLOSUM, shallow vs. deep....what does it all mean?