More realistic similarity measures

Not all substitutions are equally likely.

- A transition between two purines (A, G) or between two pyrimidines (C, T/U) is more common than a purine-pyrimidine transversion.

- Replacement of one amino acid residue by another with similar size or physiochemical properties is more common than replacement by a dissimilar amino acid residue.

Insertion/deletion of N contiguous amino acid residues or nucleotides is more likely than N independent insertion/deletion events.

Thus, we should have different penalties for opening gap and for extending a gap.

Possible substitution matrices for DNA

	A	C	G	T
A	2	-1	-1	-1
C	-1	2	-1	-1
G	-1	-1	2	-1
T	-1	-1	-1	2

	A	C	G	T
A	2	-2	-1	-2
C	-2	2	-2	-1
G	-1	-2	2	-2
T	-2	-1	-2	2

Relative likelihood and alignment score

Match model (M):

Sequences assumed to be dependent. Residues x_i and y_i at position i in the alignment occur together with probability $p_{x_i y_i}$.

Random model (R):

Sequences assumed to be independent. Residues x_i and y_i at position i in the alignment occur together with probability $q_{x_i} q_{y_i}$.

We can score an alignment using the log of the relative likelihood:

$$S = \log\left(\frac{Pr(x, y|M)}{Pr(x, y|R)}\right) = \log \frac{p_{x_1 y_1} p_{x_2 y_2} \cdots p_{x_n y_n}}{q_{x_1} q_{y_1} q_{x_2} q_{y_2} \cdots q_{x_n} q_{y_n}}$$

$$= \sum_{i=1}^{n} \log\left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}\right) = \sum_{i=1}^{n} s(x_i, y_i)$$

Percent accepted mutations

Expresses scores as log-odds values.

Score of mutation a-b is

$$\log \frac{\textit{observed a-b mutation rate}}{\textit{mutation rate expected from amino acid frequencies}}$$

Frequencies of substitutions of each pair of amino acid residues, extracted from alignments of closely related proteins.

PAM1 reflects the amount of evolutionary change that yields an average of one mutation per 100 amino acids.

Can assume that no position has changed more than once.

Correct for different amino acid abundances.

PAM substitution matrices

Extrapolate to a family of PAMk matrices by multiplying the PAM1 matrix by itself k times.

Different PAM matrices are more suitable when comparing sequences that have diverged by different amounts.

The PAM250 matrix is commonly used.

250 mutations per 100 amino acids.

Sequences still 20% identical:

- some positions change many times, while others don't change at all.
- some positions change one or more times, then revert back to the original amino acid residue.

PAM250

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

BLOSUM substitution matrices

Henikoff, S. and Henikoff, J.G. (1992) “Amino acid substitution matrices from protein blocks”, *Proc. Natl. Acad. Sci. USA*, 89:10915-10919

Based on large collection of multiple alignments of similar ungapped segments.

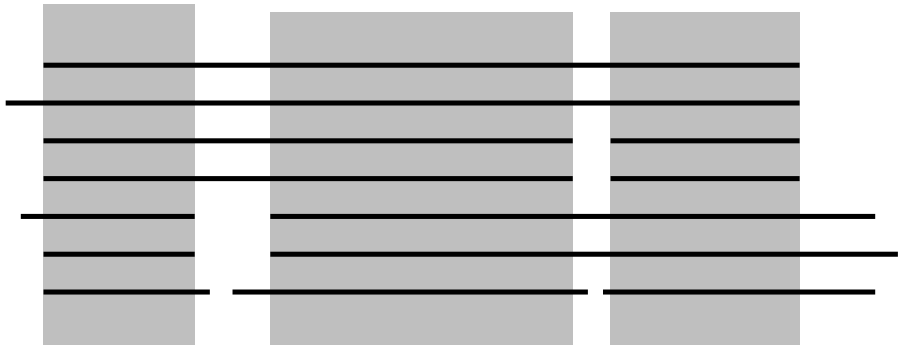
$$score_{ab} = \log \frac{observed\ relative\ frequency\ of\ aligned\ pairs\ ab}{expected\ probability\ of\ pair\ ab}$$

Pairs are only counted between segments that are more than x% identical.

Different values of x give different BLOSUM matrices.

The BLOSUM62 matrix is commonly used.

Deriving a frequency table from blocks



For each column of the block, count the number of matches and mismatches between pairs of sequences.

To illustrate the calculation, suppose we have a single block with one column, containing 9 A residues and 1 S residue.

In this case, there are 36 AA pairs and 9 AS or SA pairs.
That is, $f_{AA} = 36$ and $f_{AS} = 9$

Each column of each block in the blocks database will contribute to the observed frequency counts.

p_i is based on the proportion of residue type i in the whole blocks database.

Computing a logarithm of odds (Lod) matrix

Observed probability for each pair i,j is:

$$q_{ij} = \frac{f_{ij}}{\sum_{k=1}^{20} \sum_{l=1}^{k} f_{kl}}$$

Expected probability for each i,j pair is:

$$e_{ij} = \begin{cases} p_i p_j & \text{if } i = j \\ 2 p_i p_j & \text{if } i \neq j \end{cases}$$

Logarithm of odds is:

$$s_{ij} = \log_2(q_{ij}/e_{ij})$$

s_{ij} is multiplied by 2, then rounded to the nearest integer to give the BLOSUM score.

BLOSUM62

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

Which substitution matrix should I use?

Use a matrix that corresponds to the evolutionary distance between the proteins being compared (usually not known!).

Low PAM matrices are good for finding short, strong similarities.

High PAM matrices are good for finding long, weak similarities.

BLOSUM matrices have been found to perform better for detecting weak homologies than the extrapolated PAM matrices.

BLAST

Basic Local Alignment Search Tool

Less accurate than Smith-Waterman, but over 50 time faster.

1. Find ungapped matches of a small fixed length, w, that score at least T.

2. Extend matches in both directions in an attempt to find an alignment with a score exceeding S.

Segment pairs whose scores cannot be improved by extending or trimming are called high scoring pairs (HSPs).

Typical values for w are 3 when aligning proteins and 11 when aligning nucleic acids.

e-values and p-values

The expected number of HSPs with a score of at least S is given by the formula:

$$E = Kmne^{-\lambda S}$$

Doubling the length of the query sequence (m) or the size of the database (n) should double the number of HSPs.

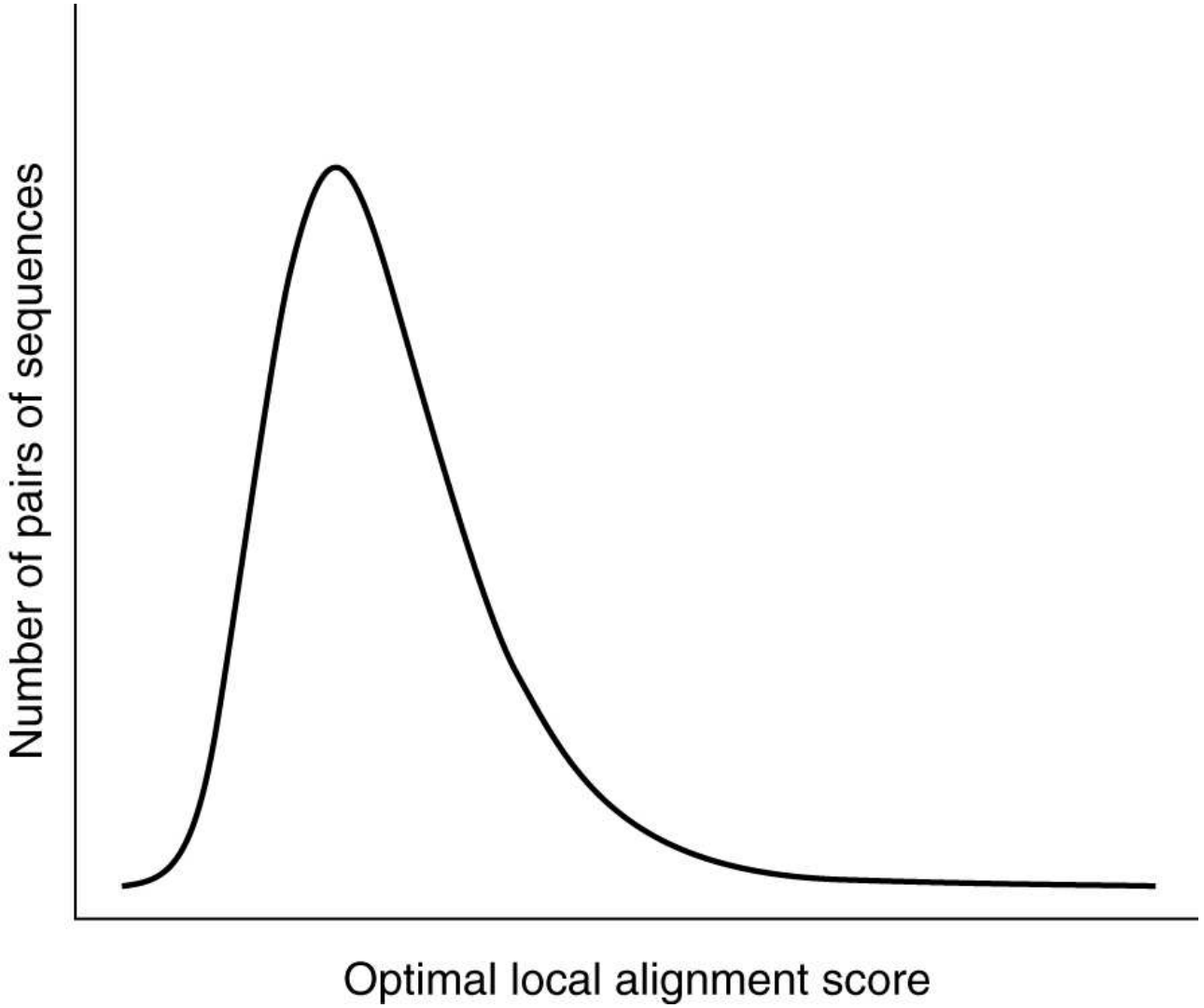
To obtain score $2x$, score x must be obtained twice in a row.
So one expects E to decrease exponentially with score.

The probability of observing a score $\geq S$ is:

$$1 - \exp(-Kmne^{-\lambda S})$$

This is the p-value.

Extreme value distribution



FASTA

k-tuples, strings of length k.

k = 1 - 2 for proteins and 4-6 for nucleic acids.

Construct a look-up table with all k-tuples in the database.

Look up all k-tuples from the query string and mark matching database k-tuples. Sort matches by the difference in their indices (i-j).

Nearby matches on the same diagonal are joined to form an ungapped local alignment region.

Join nearby high scoring regions on different diagonals.

For the best regions, perform dynamic programming in a window around the region.

Applications of pattern matching: DNA

Identifying whether a DNA molecule has a subsequence that will be recognised by a protein.

- Restriction enzymes that cut DNA.
  HindII (the first identified restriction enzyme) cuts “GT[TC][AG]AC”
  EcoRI cuts “GAATTC” between G and A
  http://rebase.neb.com/rebase/rebase.html

- DNA methylation
  e.g. E. coli DNA adenine methyltransferase (DAM) recognises GATC

- Transcription factor binding sites.
  their presence can promote or block transcription

Applications of pattern matching: PROSITE

Members of some protein families can be recognised by the presence of a specific pattern in a protein's sequence, e.g.

ATP/GTP-binding site motif A (PS00017)

```
[AG]-x(4)-G-K-[ST]
```

Zinc finger C2H2 type domain signature (PS00028)

```
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
```

EGF-like domain (PS00022)

```
C-x-C-x(2)-{V}-x(2)-G-{C}-x-C
```

http://prosite.expasy.org/

Patterns vs Profiles

Patterns are qualitative

— they either match or they don’t!

Profiles are quantitative

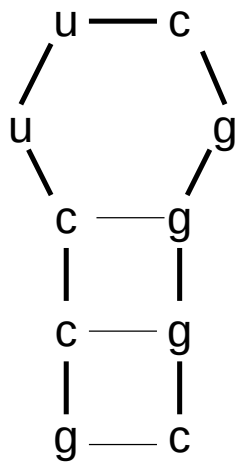
— numerical weights are associated with matches and mismatches at various positions

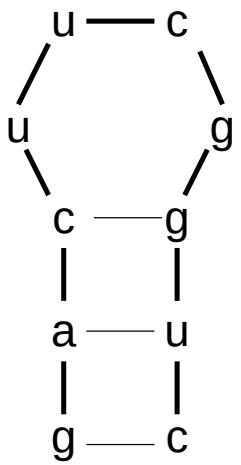
— can give greater sensitivity, allowing family membership to be detected even if the family has only a few highly conserved sequence positions

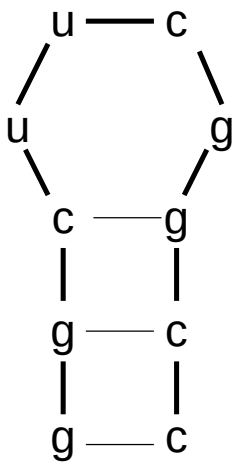
— Hidden Markov Model are commonly used to derive profiles

Covariance

Can give clues about base pairing and RNA secondary structure.








```
  *       *
gccuucgggc
gacuucgguc
ggcuucggcc

(((....)))
```