Sequence comparison 1/25

Sequence comparison amounts to comparing the content of 2 sequences.

In simple term, a sequence is a string of letters.

DNA sequence : A T G G C (4 letters)

Protein sequence : W C Y L V I (20 letters)

Approach 1 : look at sequence composition.

DNA: single nucleotide content
→ weak relationship to specie

Protein: content as tripeptide (3 AA)
– distinguish domains on the tree of life (bacteria
eukaryots
⋮)

All this is qualitative at best, and a more quantitative approach is needed.

Approach 2 quantification.

We need to start with quantifying the difference between two letters in the sequence alphabet.

1 Identity.

Build a table:

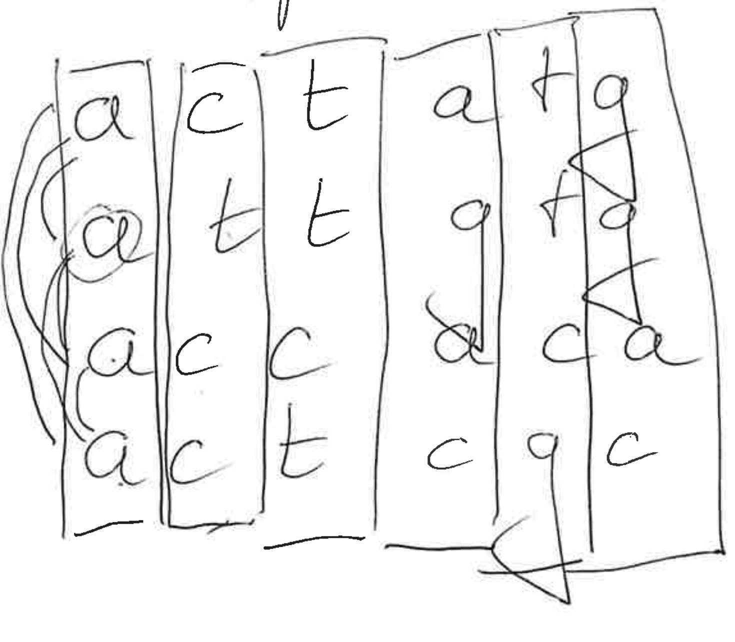
	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

2 Account for chemistry

	A	T	G	C
A	4	0	0.5	0
T	0	1	0	0.5
G	0.5	0	1	0
C	0	0.5	0	1

3 Probe nature

Example:



	A	T	G	C
A	||||| |		|	|||
T		||||	||	||||||
G	|||		|	|||
C		||	|	|||

	A	T	G	C
A	7	0	1	3
T	0	4	2	6
G	3	0	1	3
C	0	2	1	3

	A	T	G	C
A	$\frac{7}{11}$	0	$\frac{1}{11}$	$\frac{3}{11}$
T	0	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$
G	$\frac{3}{7}$	0	$\frac{1}{7}$	$\frac{3}{7}$
C	0	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

In the table above, I was able to define quantities of the form P_{ij} that represents the probability to replace a letter i with a letter j

I really want:

$$S_{ij} = \frac{P_{ij}}{P_i \, P_j}$$

For convenience,

$$\boxed{B_{ij} = \log S_{ij} = \log \frac{P_{ij}}{P_i \, P_j}}$$

Which sequences should we use to build those preferences?

The most common sequences and corresponding tables are refered ~~to~~ as BLOSUM:

a) Use sequences that are at least 90% identical

→ BLOSUM 90

b) Use sequences that are at least 50% identical

→ BLOSUM 50

The best "universal" matrix is

BLOSUM 62