Sequence Alignment and Database Search

Many Bioinformatics problems require as the starting point an alignment of two or more sequences. An alignment refers to a display of a collection of sequences, all of the same type, with one sequence written over another showing the similarities among the different members of the collection. Although the rule is not absolute, it is true in many cases that similar nucleotide sequences or protein sequences share similar structure and functions. A good sequence alignment can therefore communicate important information not only about evolutionary relationship, but also functional commonalities.

Finding a good alignment between even a pair of sequences is a formidable task for the human eyes. However, with the computing power available to us at present, good sequence alignments between a pair of sequences can be obtained so quickly that one can align a query sequence against each sequence in a database holding thousands of them within a very reasonable amount of time (say, a few minutes). Such a process of database search is getting very popular among geneticists and molecular biologists as they can derive useful information for a newly sequenced stretch of DNA from other similar sequences that had been previously studied.

We shall devote the first section of this chapter to familiarize the reader with the basic dynamic algorithm used in many popular sequence alignment programs. Section 2 will describe the statistics involved in evaluating the significance of the amount of similarity between sequences described by an alignment. Section 3 turns the attention to database search programs which is perhaps the most used application of sequence alignment. Finally, we discuss some multiple alignment techniques in Section 4.

6.1 Sequence Alignment and Similarity

Consider the pair of DNA fragments AGTAGTCAAGA and AGAAGCTCAAGA of length 10 and 11 nucleotide bases respectively. One cannot help noticing that these sequences kind of "look alike", and hence one would describe them as "similar" to each other. The similarity between the fragments are much more obvious to our eyes if we display them as follows:

(Alignment A)

```
A G T A G _ T C A A G A
|   |   :   |   |       |   |   |   |   |   |
A G A A G C T C A A G A
```

A display of this kind is called an alignment. In an alignment, one sequence is stacked on top of the other. Since the two sequences may have different lengths, gaps are inserted at various places as necessary. Sandwiched between the two sequences is a line of symbols indicating whether the letters on the two sequences at corresponding positions are matches (|) or mismatches (:).

The above display is only one of the numerous possible alignment of the given pair of DNA sequences. For example, these two DNA fragments can also be displayed as

(Alignment B)

```
A G T A G _ _ _ _ T C A A G A
      | |         | | | | | |
_ _ _ A G A A G C T C A A G A
```

Indeed, if we allow ourselves to slide the first sequence on top of the second and introduce gaps at any arbitrarily places as necessary, we can generate an enormous number of different alignments. However, some of the alignments can better reveal the similarity between the pair of sequences than others. For our example, alignment A obviously reveals the similarity between the sequence pair better than alignment B. The goal of sequence alignment is to find the best alignment that reveal the highest amount of similarities between the two sequences. Sometimes there are actually more than one such best alignments. This brings up the question of how do we measure the similarity between two sequences when we are given an alignment of them.

Measure of Sequence Similarity

A simple way to measure the similarity expressed by an alignment is to assign a score to each individual position of the alignment according to whether there is a match, a mismatch, or a gap. The total of the scores at all the individual positions will give an overall score of the entire alignment. If we are interested only in a particular portion of the alignment, we can simply sum the scores in that portion. For example, if we assign a score of 1 to a match, -1 to a mismatch, and a -2 to a position with the gap letter in one of the sequences, we will get a score of 1+1-1+1+1-2+1+1+1+1+1+1 = 7 for Alignment A and a score of -2-2-2+1+1-2-2-2-2+1+1+1+1+1+1 = -6 for Alignment B. Clearly Alignment A expresses more similarity of the sequence pair than Alignment B.

Scoring functions of this kind, which depends only on the count of matches, mismatches, and gap letters, do not take into account the various degrees of similarity in biochemical properties among the different pairs of bases. This is particularly important when we are aligning amino acid sequences because some of the 20 different amino acids are much more similar to each other than others in their biochemical properties. Commonly, the amino acids are grouped into four families:

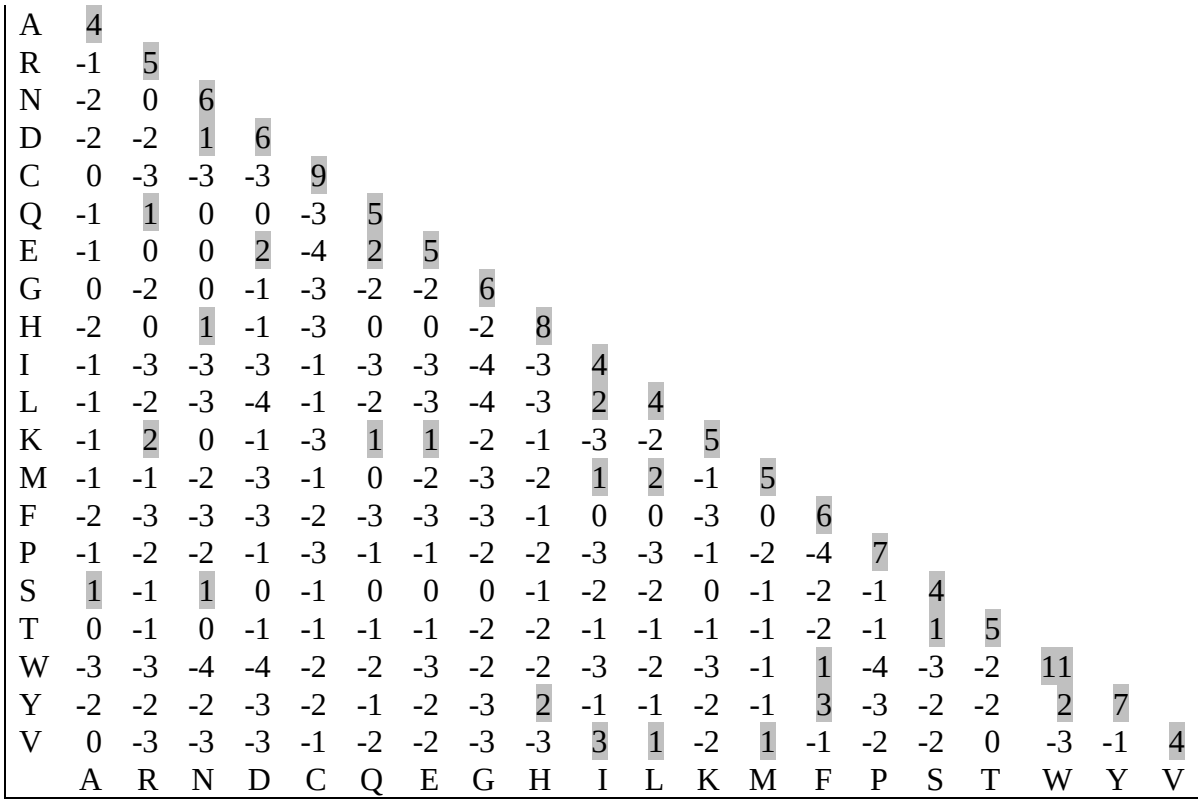
Family	**Members**
Acidic	Aspartic acid, Glutamic acid
Basic	Lysine, Arginine, Histindine
Uncharged Polar	Asparagine, Glutamine, Serine, Threonine, Tyrosine
Nonpolar	Alanine, Valine, Leucine, Isoleucine, Proline, Phenylalanine, methionine, tryptophan, cysteine

Members of the same family share similar characteristics. For example, glutamic acid would be much more similar to aspartic acid (both being acids) than to say, cysteine. Leucine and Isoleucine are almost identical in structure and they can easily substitute each other without altering too much of the chemical properties of the proteins. To take into account the various degrees of similarity and dissimilarity among amino acids, we

make use of the scoring matrices. These are 20 by 20 matrices in which each entry indicates the similarity between the amino acid on the row and that on the column. Because of symmetry, it is sufficient to give the entries above and including the diagonal, or those below and including the diagonal. The other entries can be inferred by symmetry.

The two big classes of scoring matrices are the PAM (Dayhoff 1972) and BLOSUM (Heinikoff and Heinikoff 1992) families of matrices. These families of matrices are constructed based on statistical analysis of a carefully collected database and the biologists knowledge of the evolutionary relationship of the sequences in the collection. We shall explain in detail the construction processes of these matrices in the section 6.3. Figure 6.1 shows the BLOSUM 62 matrix, a popularly used member of the BLOSUM family.

When we are allowed to introduce gaps in an alignment, we need to assess how much is the similarity affected by the insertion of a gap, and the length of the gap. It is believed that extending the length of an already opened gap does not cause as devastating an effect as opening a new gap. While there are no general rules dictating what gap opening penalty and gap length extension penalty to use, most sequence alignment programs use a gap penalty function given by $w(k) = a + bk$ for a gap of length k. Here a and b are respectively the gap opening and gap extension penalties which are free parameters for the users to choose values for. In practice, when we try different values for these parameters and examine the alignments obtained, we generally get a feeling of which values will produce the alignments that exhibit the similarities of the sequence under comparison.



	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

Figure 6.1 The BLOSUM 62 amino acid substitution matrix.