CHAPTER II

CONCEPTS AND THEORIES OF SEQUENCE ALIGNMENT

This chapter provides basic concepts and theories that help us understand the next chapters. Discussion starts from sequence alignment concepts and biological background that will be explained in section 2.1. Since sequence alignment is related to a sequence, in section 2.2 will be discussed sequence theories first. When this *skripsi* only discusses alignment of two sequences, it means that we need to explain about pairwise sequence alignment rules. Further explication about this is available in section 2.3. In section 2.4 we will exhibit some scoring schemes to find the best pairwise sequence alignment. Discussion of this chapter ends in local alignment which is a method used in this *skripsi* to find optimal alignment.

2.1 SEQUENCE ALIGNMENT

The heredity of all living organisms, except some viruses, is carried by DNA molecules. DNA usually consists of two complementary chains twisted around each other to form a helix. Each chain is a polynucleotide consisting of four nucleotides. They are divided into two purines: adenine (A), and guanine (G), and two pyrimidines: thymine (T), and cytosine (C). The two chains are linked to each other by hydrogen bonds between pairs of nucleotides.

During the process of reproduction, DNA strands are normally copied. Rarely, errors occur that give new sequences to exist. These errors are called mutations. Mutations may be classified by the types of change such as: (1) substitutions, when nucleotides are replaced by some other nucleotides, (2) deletions, when some nucleotides are removed from the DNA, and (3) insertions, when new nucleotides are added to the existing ones in a sequence [LG 91].

A basic process in the evolution of DNA strands is the modification in nucleotides as time passed by. This process deserves a detailed consideration since changes in nucleotide are used in molecular evolutionary studies both for estimating the rate of evolution and for reconstructing the evolutionary history of organisms [LG 91]. But the process of nucleotide modification is usually very slow, so it is difficult to be observed within a researchers's life. Therefore, to uncover changes in a DNA sequence, comparative methods is applied where a given sequence is compared with another which they might shared a common ancestry in the evolutionary past. Such comparisons require mathematical computation.

Comparisons of two homologous sequences involves the identification of the locations of deletions and insertions that might have occurred in either of the two lineages since their divergence from a common ancestor. This process is referred to as sequence alignment. Based on the formal definition, sequence alignment is a way of arranging the primary sequence of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationship between the sequences.

Sequence alignment is used as a first and vital step in protein structure prediction, phylogenetic reconstruction, analysis of protein domains and identification of functional sites in genomic sequences, to declare just a few important applications [Mor 04]. Therefore, it is deeply essential in all aspects of DNA and protein sequence analysis.

The main purpose of sequence alignment is finding a homology (similar structures due to shared ancestry) among sequences. Homology among proteins and DNA is often concluded on the basis of sequence similarity, especially in bioinformatics. For example, in general, if two or more genes have highly similar DNA sequences, it is likely that they are homologous. But sequence similarity may arise from different ancestors. Such sequences are similar but not homologous.

An alignment involves a series of paired bases, one base from each sequence. There are three types of aligned pairs: (1) pairs of matched bases, where a matched pair implies a site that has not changed, (2) pairs of mismatched bases, where a mismatched pair denotes a substitution, and (3) pairs consisting of a base from one sequence and a null base from the other, a null pair indicates that a deletion has occurred at this position in one of the two sequences or an insertion in the other which null bases are denoted by gap (-) [LG 91].

There are two types of alignment based on the number of sequences being compared, namely pairwise and multiple sequence alignment. The pairwise alignment compares only two DNA or protein sequences whereas the multiple alignment compares a collection of at least three DNA or protein sequences. To find a homology among these set of sequences we need to determine an optimal multiple alignment for the whole collection.

Various methods have been developed to obtain sequence alignment, such as global and local alignment. Those methods are based on dynamic programming procedures but have different approach. The global alignment compares all possible individual residues between two sequences whereas local alignment compares all possible segments of two sequences.

This *skripsi* will give details about how DIALIGN produce pairwise local alignment. Here are some basics on pairwise alignment.

2.2 SEQUENCE

Before explaining pairwise sequence alignment, let us reconsider the theory of sequence itself. Suppose a sequence Z,

$$Z = Z_1 Z_2 \ldots Z_k \in$$

where Z_i is the *i*-th element of sequence *Z* and is the alphabet of the sequence. Since we discuss about nucleic acid sequence, = {A, G, C, T} whereas on protein sequence has member of 20 alphabet. Sequence *Z* above can be written in the form (z_i), $i \in K$ where $K = \{1, 2, ..., k\}$.

Subsequence is a new sequence which is formed from the original sequence by deleting some of the elements without disturbing the positions of the remaining elements. Formally in mathematics, suppose that (z_i), $i \in K$ is a sequence where $K = \{1, 2, 3, ..., k\}$. Then, a subsequence of (z_i) is a sequence of the form $\left(z_{k_r}\right)$ where (k_r) is a strictly increasing sequence in the index set *K*.

Segment is a subsequence which preserves the consecutive number of index of the original sequence. Example 2.1 will clarify the difference of subsequence and segment.

Example 2.1

Let Z = A T G C T

Sequence *Z* has length of five. Character A is the first element of sequence *Z*, character T is the second element of sequence *Z*, and so on until character T is the fifth element of sequence Z. By deleting alphabet A of sequence Z we have T G C T as a new sequence (see Figure 2.1.a), by deleting alphabet T and C of sequence Z we have A G T as a new sequence (see Figure 2.1.b), and by deleting both alphabet T of sequence Z we have C G A as a new sequence (see Figure 2.1.c).

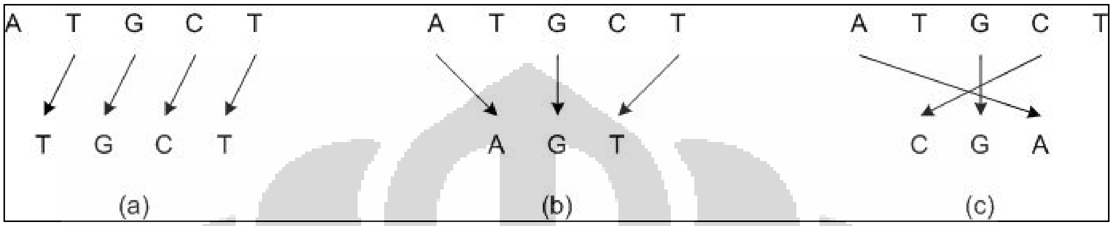

Figure 2.1 The new sequences which are formed from sequence Z

Subsequence in Figure 2.1.a is called segment since it preserves the consecutive number of index of the original sequence. Subsequence in Figure 2.1.b is an example of subsequence but not segment because it does not preserve the consecutive number of index. And subsequence in Figure 2.1.c cannot be considered as subsequence nor segment for the reason that it disturbs the positions of the remaining elements.

Some terminologies here will be utilized in sequence alignment especially when using segment-to-segment approach. Further description is available in next section.

2.3 PAIRWISE SEQUENCE ALIGNMENT

Several terminologies are contained in this field, such as segment pair, diagonal, consistency, etc. This section will explicate those one by one.

Given two nucleic acid sequences:

$X = X_1 X_2 \ldots X_i \ldots X_m$, and

$Y = Y_1 Y_2 \ldots Y_j \ldots Y_n$.

Whenever two sequences are to be compared, there is a pair of segment taking from a pair of sequence that usually called segment pair which might exhibit regions of similarity.

Diagonals are segment pairs which have the same length that are to be compared. Such pairs of segments are referred to diagonals since they would appear as diagonal runs in the respective pairwise comparison matrices.

Alignment is a collection of consistent diagonals. If two sequences are to be aligned, consistency is given if diagonals are ordered in the sense that the end positions of one of the diagonals are both smaller than the respective starting positions of the other one [Mor 98]. Example 2.2 will describe diagonal and its consistency.

Example 2.2

Let,

X = A C T C A G G G C T T A

Y = A C C C G G C T T A A G

By taking a pair of equal length segment from sequence X and sequence Y we have some diagonals (see Figure 2.2).

C T C	G C T T A	G G G C
A C C	C T T A A	T A A G
(a)	(b)	(c)

Figure 2.2 Some possible equal length segment pairs (diagonals)

Diagonal (a) in Figure 2.2 is consistent with diagonal (b) in Figure 2.2 since they are ordered in the sense that the end positions of diagonal (a) are both smaller than the respective starting positions of diagonal (b), and diagonal (a) in Figure 2.2 is also consistent with diagonal (c) in Figure 2.2. But diagonal (b) in Figure 2.2 is not consistent with diagonal (c) in Figure 2.2 because the end positions of diagonal (b) are both bigger than the starting positions of diagonal (c).

Comparisons of two DNA sequences usually cannot tell us whether a deletion had occurred in one sequence or an insertion had occurred in the other. Therefore, the outcomes of both types of events are collectively referred to as gaps.

A sequence alignment satisfies the following two properties: (1) the length of each of the aligned sequence must be the same after an alignment, (2) each of the aligned sequence before and after an alignment is identical if gap(s) is (are) ignored [ACL$^+$ 02]. According to these properties, sequence alignment procedures can be applied under these rules:

1. Residues from one sequence are assembled with residues from another and residues must be allowed to be aligned not only to other residues but also to gaps.

   example 2.3:

   We use sequence *X* and sequence *Y* from example 2.2.

   During the alignment process, it is likely to have a result like

   ```
   A   C   T   C   A   G   G   G   C   T   T   A   --  --
   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   A   C   C   C   --  G   G   --  C   T   T   A   A   G
   ```

   where each residues of the first sequence are paired with each residues of the second and some are paired to gaps.

2. There is a possibility of the presence of gaps in front of, behind, and between residues that represents either an insertion or deletion event.

   example 2.4:

   From example 2.3, alignment obtained

   ```
   A   C   T   C   A   G   G   G   C   T   T   A   --  --
   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   A   C   C   C   --  G   G   --  C   T   T   A   A   G
   ```

   where gaps in the first sequence could represent that sequence X was undergoing deletions or sequence Y was undergoing insertions and gaps in the second sequence could represent that sequence X was undergoing insertions or sequence Y was undergoing deletions.

3. The order of residues on aligned sequence is exactly the same as before.

   example 2.5:

   Checking the alignment obtained from example 2.4,

   the second residue A in the first sequence is located after residue C and before residue G, also the last residue C in the second sequence is located after residue G and before residue T. The order of residues on aligned sequence and original sequence are of the same position.

Having those ways of aligning sequences, there would be many possible alignments. How can we choose among them? Which one is better alignment?

The best possible alignment between two sequences is the one in which the number of mismatches and gaps is minimized according to certain criteria. Unfortunately, reducing the number of mismatches usually results in an increase in the number of gaps, and vice versa. Therefore to find the best alignment we need to be able to score all possible alignments. Then the alignment that have the highest score is by definition the best or optimal one (there may be more than one such alignment).

In the scoring scheme, each paired bases has its own value (score). Next section will describe a complete scoring scheme.

2.4 SCORING SCHEME

In this section we will show some scoring schemes to find the optimal sequence alignments. Different scoring schemes may result different optimal alignments. Consider, for example, the following two sequences from example 2.2

X = A C T C A G G G C T T A

Y = A C C C G G C T T A A G

We can reduce the number of mismatches to zero as follows,

A C T -- C A G G G C T T A -- --

A C -- C C -- G G -- C T T A A G

The number of gaps in this case is six. The number of gaps can be reduced with a consequent increase in the number of mismatches such as

A C T C A G G G C T T A -- --

A C C C -- G G -- C T T A A G

In this example, we have unavoidable gaps but the number of mismatches is two.

Alternatively, we can choose an alignment that minimizes neither the number of gaps nor the number of substitutions. For example,

A C T C A G G G C T T A -- --

A C C C -- -- G G C T T A A G

In this case, the number of mismatches is one and the number of gaps is four.

So, which of the three alignments is preferable? It is obvious that they may not be compared directly. As a consequence, we must find a common measurement to compare gaps and substitutions. This common measurement is called the gap penalty.

However, an optimal alignment may change when the scoring scheme is changed. As an example, given a scoring scheme of M = s(a,a) = 2 (the score of a match), MM = s(a,b) = -1 if a ≠ b (the score of a mismatch), and G = s(-,a) = s(a,-) = -2 (the gap penalty). The following examples seem to be more informative.

From example 2.2 there are some possible alignments when applying scoring scheme above, three of them are

Alignment 1:

A C T C A G G G C T T A -- --

A C C C -- G G -- C T T A A G

Alignment 2:

A C T C A G G G C T T A -- --

A C C C -- -- G G C T T A A G

Alignment 3:

A C T C A G G G C T T -- A --

A C C C -- G -- G C T T A A G

These three alignments have an equal maximum alignment score of 9, so they could be considered as the optimal ones.

Sequence alignment is a method of data exploration instead of an analytical method that will lead to a single best solution, therefore there is no objective way of choosing the right scoring scheme. Through a local alignment method, this *skripsi* discuss an algorithm using segment-to-segment approach that ignores the gap penalties in its scoring scheme. The picture of this approach is available in next section.

2.5 LOCAL ALIGNMENT

There are various methods using local alignment procedures, one of them is known by DIALIGN. This alignment algorithm is based on segment-to-segment comparison instead of the commonly used residue-to-residue comparison. It avoids the wellknown difficulties concerning the choice of appropriate gap penalties. Here, gaps are not treated explicitly but remain as those parts of the sequences that do not belong to any of the aligned segments [MDW 96]. In DIALIGN, the diagonals will be evaluated to obtain the optimal local alignment. Further explication about DIALIGN algorithm is provided in chapter three.