CHAPTER I

PRELIMINARIES

1.1 BACKGROUND

Biology scientists have known millions of organisms but as time changes the number of organisms has been growing more and more. We can identify them, to what species they belong, from their gene structures and functions. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species.

Biologists often discover the sequence of a new protein with unknown function. If that sequence can be associated with a known protein sequence we will have informations about structure and/or function. Similar structure indicates homology which provides critical information about the functions of these sequences. Since 1977, DNA sequences of hundreds of organisms have been decoded and stored in databases. With the growing amount of data, it became impractical to analyze DNA sequences manually. Here, we use bioinformatics.

In the last two decades, bioinformatics is a most active area of recent studies in science. It is the combination of mathematics, statistics, computer science, artificial intelligence, and chemistry to solve biological problems on molecular level. Major research in this field include sequence alignment, protein structure prediction, gene expression, etc.

The first step towards inferring homology is to look for sequence similarity. To see if they are similar, we have to properly align them. In bioinformatics, this process is known as sequence alignment.

A 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. It is a powerful tool for identify mutations that cause genetic diseases. Therefore, bioinformatics is much needed in medical field.

When sequences starts to evolve, their residues can undergo substitutions (when residues are replaced by some other residues), insertions (when new residues appear in a sequence in addition to the existing ones), and deletions (when some residues disappear). When we are trying to produce the best possible alignment between two sequences, residues must be allowed to be aligned not only to other residues but also to gaps. The presence of a gap represents either an insertion or deletion event. To choose best alignment we need to be able to score any possible alignments and the alignments that have the highest score are by definition the optimal ones (there may be more than one such alignments). And a scoring scheme must be biologically relevant in order to produce a sensible alignment [Isa 06].

There are two approaches in aligning sequences, namely global alignment (introduced by Needleman and Wunsch) and local alignment (introduced by Smith and Waterman). Both are based on dynamic programming. Alignment by dynamic programming guarantees that the resulting alignment is the optimal alignment or one of the equally optimal alignment. Although dynamic programming is extendable to more than two sequences, it performs slow for large numbers or extremely long sequences.

Many algorithms have been developed to solve alignment problem such as CLUSTAL W, MUSCLE, T-COFFEE, etc. In this *skripsi* we use DIALIGN that inspired by diagonals in the so-called dot matrix. A dot matrix plot is a method of aligning two sequences to provide a picture of the homology between them. This algorithm is based on segment-to-segment comparison instead of the commonly used residue-to-residue comparison and avoids the well-known difficulties concerning the choice of appropriate gap penalties [MDW 96].

This *skripsi* discusses how DIALIGN algorithm produces optimal alignment for a pair of DNA or protein sequence.

1.2 PROBLEM STATEMENT

Does DIALIGN algorithm produce an optimal local alignment for a pair of sequence?

1.3 OBJECTIVE OF WRITING

The objective of this *skripsi* is to examine how DIALIGN algorithm aligns two DNA sequences. From here, we want to know whether DIALIGN algorithm produces optimal sequence alignment accurately and how long it takes.

1.4 SCOPE OF DISCUSSION

This *skripsi* will discuss alignment algorithm for two sequences only. The main focus is to align a pair of nucleic acid sequences using DIALIGN algorithm.

1.5 STRUCTURE

This *skripsi* is splited into five chapters. Chapter two discusses the basic concept and theory of sequence alignment. Chapter three explains the detail of DIALIGN algorithm. Chapter four shows the implementation of the algorithm. Chapter five covers the conclusion from previous chapters and suggestions.