Title: Sequence Alignment using *Align*

Author: Richard M. Salter
Computer Science Department
Oberlin College
rms@cs.oberlin.edu

Year: 2008

Funding Source: National Science Foundation CCLI DUE 0618252

Math/Sci Level: College First-year

Abstract: According to Krane & Raymer, *Fundamental Concepts of Bioinformatics*:

> *In a very real sense, any alignment between two or more nucleotide or amino acid sequences represents an explicit hypothesis regarding the evolutionary history of those sequences. As a direct result, comparisons of related protein and nucleotide sequences have facilitated many recent advances in understanding the information content and function of genetic sequences. For this reason, techniques for aligning and comparing sequences, and for searching sequence databases for similar sequences, have become cornerstones of bioinformatics.*

Align is a Java application that presents 3 essential alignment algorithms. In this module students will learn about the algorithms used in pairwise sequence alignment. Students will use the *Align* tool to help them understand how these algorithms work, and to experiment with sequence matching under various constraints.

This module consists of 4 laboratory exercises distributed as PDF files.

Part I introduces sequence alignment.
Part II describes the basic dynamic programming algorithm.
Part III discusses the scoring schemes used in sequence alignment.
Part IV considers semi-global and local variations of the basic algorithm.

Each part contains worked-out examples and several exercises with solutions. Please see the Instructors Guide for more information.

Keywords:	bioinformatics, sequence alignment, algorithm, dynamic programming
Problem Statement:	An *alignment* of two DNA or protein sequences is a way of placing the sequences next to each other so that corresponding elements match. Given two such sequences, find an optimal alignment of the pair under various constraints.
Background Information:	See Align1, pp. 4 – 9.
Model:	See Align1, pp. 10 – 12.
Solution Methodology:	The solution is based on an implementation of three well-known sequence alignment algorithms: Needleman-Wunsch global alignment, Needleman-Wunsch semi-global alignment, and Smith-Waterman local alignment. These algorithms all use a common solution technique based on *dynamic programming.*
Conceptual Questions:	1. What are some situations in which a molecular biologist might wish to perform a pairwise sequence alignment? 2. Why might matrix scoring be more important to protein alignment than to DNA alignment? 3. How are PAM and BLOSUM matrices determined? 4. What sort of evolutionary questions are best handled by local searches?
Problems:	Each of the four parts contains several exercises with solutions.
Project:	Learn about the BLAST algorithm and visit the BLAST Website http://www.ncbi.nlm.nih.gov/blast/. Submit a sequence to the BLAST database.
Suggestions to Instructors:	See enclosed Instructors Guide.
Glossary:	**sequence**: a string of symbols representing nucleotides or amino acids. **sequence alignment**: a pairwise match between the characters of each sequence. **mutation**: a replacing of one character with another between two sequences. **insertion**: addition of one or more positions in one of the sequences. **deletion**: elimination of one or more positions in one of the sequences. **gap**: an anonymous entry added to a sequence to represent an

insertion or deletion.
scoring: assignment of a numerical value to reflect the degree of similarity between the two sequences.
optimization problem: a problem with many potential solutions, where the goal is to find the one that achieves the highest score.
combinatorial explosion: a problem in which the number of cases that must be checked is so large that it cannot be solved in a reasonable amount of time even with the fastest computer.
divide and conquer: a solution strategy in which a problem is broken into smaller sub-problems, the solutions to which are used to construct the solution of the main problem.
dynamic programming: a divide-and-conquer-based problem solving technique in which each sub-problem solution is recorded so that it does not have to be computed more than once.
origination penalty: the score assessed for starting a new series of gaps.
length penalty: the score assessed for a series of gaps.
scoring matrix: a table specifying the scores to be assessed for matches and mismatches between specific sequence entries.
global algorithm: an alignment algorithm in which the two sequences are aligned over their entire lengths.
semi-global alignment algorithm: a modification of the global alignment algorithm that ignores misalignment at the beginnings and ends of the sequence pair.
local alignment: an alignment algorithm that focuses on matching subsequences of the sequence pair.

References: See the Instructors Guide.