

Sequence Alignment Term Project
for
CHEM 434 – BIOINFORMATICS
Prof. Jamil Momand and Prof. Nancy Warter-Perez

Due date: Thursday, June 11th 12 noon

If you successfully complete the base project, you will receive a B on your project. If you successfully complete both the base project and one extension, you will receive an A on your project. If you successfully complete *one* further extension you will receive a hundred percent on your project. Note: Presentation grades are separate (Overall grade: 90% project and 10% presentation).

Base project (Global alignment using dynamic programming):

Implement global alignment using dynamic programming (similar to Smith-Waterman's algorithm). Use PAM and BLOSSUM scoring matrices and assume a fixed gap penalty (given by the matrix).

The program should prompt the user to enter the scoring matrix file name and two sequences. The program should display the aligned sequences, showing gaps (-) in each sequence and the matches (|) between the sequences, and the alignment score.

Extensions (Select 1 of the following):

1. Extend your program to support local alignment algorithm (Smith-Waterman).
2. Extend local alignment algorithm to print out alignments with maximum score.
3. Extend your program to support affine gap penalties. Use the gap penalty from the scoring matrix for your gap open penalty.
4. Modify your program to work with a query sequence and a database. In this case, instead of prompting the user for two sequences, the program should prompt the user for the query sequence and the database flat file. The format *of each entry in the database flat file* is:

> sequence identifying information
{sequence without spaces or newline characters}

The program should also prompt the user for a score threshold. All sequences that exceed that score will be displayed (ideally in order from highest score to lowest score).

Project Presentation

Each group will give a 10 minute presentation followed by a 5-minute question period. Each group member should participate in the presentation. The presentation should include a demonstration of your software and a discussion of the algorithms used. Suggestions for organization of presentations:

- Brief description of the particular aspects of sequence alignment you are presenting with a biological motivation.

- Presentation of the algorithm used and any unique issues concerning implementation.
- Demonstration and/or experimental results from database search. The database and query string will be provided (however, you should test your program on a simple query and database that you create).
- Contributions of group members (**required**).
- Conclusions

No formal project report is required. You should submit your power point presentation and a disk containing your project source code.

References

1. Needleman, S.B. and Wunsch, C.D. A General Method Applicable to the Search for Similarities in Amino Acid Sequence of Two Proteins. *J. Mol. Biol.*, 48, pp. 443-453, 1970.
2. Smith, T.F. and Waterman, M.S. Identification of Common Molecular Subsequences. *J. Mol. Biol.*, 147, pp. 195-197, 1981.