

**Problem set V:** Algorithms for nucleotide and peptide sequence analysis in Perl.  
Team project. Do any one.  
Note that as many as four students may sign up for each problem.

Given below are some landmark articles in the (brief) history of the bioinformatics discipline. These articles have presented strategies (algorithms) for sequence analysis, alignment and multiple alignment. Pick any one article. You will probably want to consult the original article (you may need to use the UCSD Library if the journal is unavailable at SDSU or from the Internet), but you may also want to use web resources such as course sites that describe the algorithm. Please provide a brief (1 paragraph) synopsis (a word .doc) of the paper cited to describe the strategy used by the author(s). Please cite all references, including websites (e.g. <http://www.ncbi.nlm.nih.gov/BLAST/>), used to prepare your report. Each team need only turn in one written report and one floppy disc. Every team member should contribute in some fashion to the final report.

Write the simplest Perl program you can create to demonstrate the algorithm presented by the authors. As necessary, create your own sample (short) data files to demonstrate operation of your coded solution; or, use a segment of published sequence data retrieved from public databases. Your program should first print out the original test data sequence, then perhaps intermediate results or choices/options to be selected by the user, and then any calculated results. It is essential that you comment liberally in your program regarding your solution strategy! This is expected to be a team effort. Note too that you are expected only to demonstrate the fundamental algorithm in the simplest way possible, not to duplicate the software developed by the original authors.

1. Altschul, S.F., W. Gish, W. Miller, E.W. Meyers and J.D. Lipman, A basic local alignment search tool. *J. Mol. Biol.* 215: pp. 403-410 (1990)
2. Corpet, F., Multiple sequence alignment with hierarchical clustering. *Nucl. Acids Res.* 16: pp. 10881-10890 (1988)
3. Henikoff, S. and J.G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci. USA* 89: pp. 10915-10919 (1992)
4. Higgins, D.G. and P.M. Sharp, CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73: pp. 237-244 (1988)
5. Karlin, s. and S.F. Altschul; methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci. USA* 87, pp. 2264-8 (1990)
6. Lipman, D.J. and W.R. Pearson, Rapid and sensitive protein similarity search, *Science* 227: pp. 1435-1441 (1985)
7. Lipman, D.J., S.F. Altschul and J.D. Kececioglu: A tool for multiple sequence alignment. *Proc. Nat. Acad. Sci. USA* 86, pp. 4412-5 (1989)
8. Needleman, S.B. and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* 48: pp. 443-453 (1970)
9. Parida, L., A. Floratos and I. Rigoutsos, An approximation algorithm for alignment of multiple sequences using motif discovery. *J. Combinatorial Optimization* 3: pp. 247-275 (1999).
10. Smith, F.F. and M.S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* 147: pp. 195-197 (1981)
11. Staden, R.: Methods to define and locate patterns of motifs in sequences. *Comput. Appli. Biosci.* 4, pp. 53-60 (1988)
12. Thompson, J.D., D.G. Higgins and T.J. Gibson: CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22: pp. 4673-4680 (1994)