1. **Semi-global alignment [25 pts].**
   In this question we consider a variant of the global alignment problem, where we impose no penalty on gaps at the ends of one of the sequences. Consider the following two alternative alignments:
   
   sequence1: CAGCA-CTTGGATTCTCGG  
   sequence2: ---CAGCGTGG--------  
   sequence1: CAGCACTTGGATTCTCGG  
   sequence2: CAGC-----G-T----GG  
   
   When doing global alignment under the scoring scheme of +1 for match, -1 for mismatch, and -1 for a gap the second alignment is preferred, despite our intuition that the first alignment is more biologically relevant. If the gaps on the ends of sequence2 are not penalized, then the first alignment scores higher. This approach is called *semi-global* alignment. Note that in this approach the end gaps in one of the sequences in the alignment
   
   sequence1: ACGTCAT---  
   sequence2: ---TCATGCA  
   
   will be penalized. Show how to modify the Needleman-Wunsch algorithm to compute a semi-global alignment (including the initialization of the matrix, and the traceback operations). Illustrate your algorithm on the sequences **ACAGATAC** and **AGTT** using the above simple scoring scheme.

2. **Homology [10 pts].**
   Is homology transitive? (i.e. if A and B are homologous, and B and C are homologous, are A and C homologous?)

3. **Global alignment with a limited number of gaps [20 pts].**
   For a parameter $k$, suggest a dynamic programming algorithm that finds an optimal global alignment between two sequences subject to the constraint that the alignment contains at most $k$ blocks of consecutive indels.

4. **Local alignment [50 pts].**
   Implement a local alignment algorithm with affine gap penalties.
   Parameters for your program should be:
Homework 2

(a) A file in Fasta format that contains the input sequences. The Fasta format is a common format for sequence data. The format is described in http://en.wikipedia.org/wiki/FASTA_format; for our purposes you can assume the file won’t contain comment lines, which begin with a semicolon. There is a simple example on the homework page.


(c) Gap opening and gap extension penalties.

Use your program to align the Aniridia protein from H. sapiens to the eyeless protein from D. melanogaster. The sequence of these proteins can be obtained from the uniprot protein database using the following links: http://www.uniprot.org/uniprot/P26367 (aniridia) and http://www.uniprot.org/uniprot/018381 (eyeless). Note that Uniprot has an option for saving an entry in Fasta format. Use a BLOSUM 62 substitution matrix with a gap opening penalty of 7 and gap extension penalty of 1.

• What is the alignment and alignment score you obtained? While in this case it’s quite clear that the sequences are related, that is not always so easy to determine.

• Run your program on 10 randomly generated protein sequences with the same length and amino acid composition as the Aniridia protein, and compare the results with those of the two PAX6 proteins.

Email your program to the instructor.