

---

# CS/ST 480

## HOMWORK 2 (DUE FEBRUARY 10TH)

---

### 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
```

Under the simplest 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 `ACAGATA` and `AGT` 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 an algorithm that finds a 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 the Smith-Waterman local alignment algorithm, using fixed gap costs, and a fixed mismatch penalty (the same for all amino acids/nucleotides). Use your program to align the following DNA sequences:

## HOMEWORK 1

---

```
TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC
AATTGCCGCCGTCGTTTTTCAGCAGTTATGTCAGATC
```

Explore the parameter space of the algorithm, and comment on the alignments you obtain (e.g. what kind of parameter values give a global alignment, vs parameter values that give a local alignment). In addition, use your program to align *random* DNA sequences of length 100, and repeat the same exploration of parameter space. Comment on the differences as compared to the above two sequences.

Attach a printout of your program, as well as email the code to our TA. For those in search of a greater challenge, implement the affine gap version.