Problem Set #1

Due Tuesday, January 12, 2010, at the **beginning** of class. Assignments turned in more than 10 minutes after the beginning of class will be penalized.

1. (10 points) You can find a copy of the BLOSUM62 protein substitution matrix at ftp://ftp.ncbi.nih.gov/blast/matrices/BLOSUM62. Which amino acid has the most negative scores associated with it? Why? (give an evolutionary answer, not a biochemical one)

2. (15 points)

   ```
   RLINLMP----WVLATEYKNY
   QFFPLMPPAPYWILATDFENY
   ```

   Score the above protein alignment using

   - BLOSUM62 and a linear gap penalty of -4
   - BLOSUM80 with affine gap penalties: gap open of -9 and gap extension of -1.

   You can find the BLOSUM80 protein substitution matrix at ftp://ftp.ncbi.nih.gov/blast/matrices/BLOSUM80. Be sure to show your work.

3. (20 points) Draw and fill in the dynamic programming matrix to align these two sequences: `CGTTC` and `CAATC`. Use this substitution matrix:

	A	C	G	T
A	2	-7	-3	-7
C	-7	2	-7	-3
G	-3	-7	2	-7
T	-7	-3	-7	2

   and use a fixed gap penalty of -5. What is the score of the optimal global alignment?

4. (10 points) Write a program that takes as input the first three command line arguments (after the program name) and prints them in uppercase letters with each argument on a separate line. For example:

   ```
   > python get-three-args1.py con stan tinople
   CON
   STAN
   TINOPLE
   ```

5. (15 points) Write a program similar to the previous one, but print the three arguments on one line **without** spaces between.

   ```
   > python get-three-args2.py con stan tinople
   CONSTANTINOPLE
   ```

6. (15 points) Write a program that takes as input three command line arguments: the first argument is a DNA or protein sequence, and the second and third represent a range of positions in the sequence. Print the range of characters in the given sequence.

```
> python get-subsequence.py cantankerous 5 7
ank
```

7. (15 points) Write a program that takes as input two command line arguments, counts how many times the second one appears inside the first one, and then tells the user how many there are.

```
> python count-substrings.py acgtacgtttgacgtacc acg
The sequence acg appears in the sequence acgtacgtttgacgtacc 3 times.
```

8. **Challenge questions (no points)** Write a program that takes as input a single string and makes as output the reverse of the string. Write a program that takes as input a DNA sequence and makes as output its reverse complement. Do the same thing except retain the case of the input sequence in the output (case is often used to indicate things like exons vs introns). Do the same thing but impose a requirement that the input characters be valid nucleotides ('A', 'C', 'G', or 'T' or their lowercase equivalents) - provide a useful error message if not.

```
> python reverse.py acgtac
catgca

> python reverse-complement1.py acgtac
gtacgt

> python reverse-complement2.py acGTac
gtACgt

> python reverse-complement3.py acGTac
gtACgt

> python reverse-complement3.py acJTac
Input error: character J not a valid nucleotide
```