

COMP 5970/6970 Project 1: 100 points 20% Credit

Submission due before 11:59 PM Friday February 7

Instructions:

1. This is an individual project. You should do your own work. Any evidence of copying either from a public source or from the works of other without due credits will result in a zero grade and additional penalties/actions.
2. **Submissions by email or late submissions (even by minutes) will receive a zero grade.** No makeup will be offered unless prior permission has been granted, or there is a valid and verifiable excuse.
3. **No show for your project presentation will receive a zero grade. There is also a penalty for missing a presentation day in which you are not presenting.**

Submission:

For 5970, you are required to upload the following to canvas before **11:59 PM Friday February 7:**

1. **Source Code:** Python source files (upload .zip file in case of multiple files) containing your code only (no test data needed) and ReadMe.txt file (template provided) describing how to run your code. Note that we will NOT debug your code. If your code does not execute as described in ReadMe.txt, you will receive a zero grade.
2. **Presentation Slide:** One slide only in PPT/PPTX/PDF format to be used during the oral presentations (see below). If you submitted file spans more than a page, we will extract the first page for the oral presentation.

For 6970, you are required to upload the following to canvas before **11:59 PM Friday February 7:**

1. **Source Code:** Python source files (upload .zip file in case of multiple files) containing your code only (no test data needed) and ReadMe.txt file (template provided) describing how to run your code. Note that we will NOT debug your code. If your code does not execute as described in ReadMe.txt, you will receive a zero grade.
2. **Presentation Slide:** One slide only in PPT/PPTX/PDF format to be used during the oral presentations (see below). If you submitted file spans more than a page, we will extract the first page for the oral presentation.
3. **Project Report:** Completed report document in PDF format using template provided. Make sure to have all necessary sections of scientific writing: abstract, introduction, methods, results, discussion, references.

Presentations:

Presentations will be during the class on **Monday February 10** and **Wednesday February 12.**

Attendance is mandatory during all the presentation days. Missed presentation days without university-approved excuse will result in a penalty of 25 points for each missed class. Note that this penalty will be applied when you miss a presentation day in which you are not presenting. **No show for your project presentation will receive a zero grade.**

Everyone is required to deliver 3 minutes flash presentation accompanied by the submitted slide following the Three Minute Thesis (3MT) format, with additional 2 minutes for Q&A:

1. Your presentation should at least contain methods (i.e., implementation), results (e.g., output), and conclusion.
2. Having appropriate graphics and visuals (e.g., figures, plots) in the presentation slides to help illustrate key concepts or results will be positively graded.
3. Any additional scientific insights and/or challenges faced and/or limitations of your implementation and/or efficiency analyses and/or comparisons with alternative approaches will be positively graded.
4. Practice your talk not to exceed the time limit or finish too early.
5. No need to bring your slides. We will set things up and decide the presentation sequence.

Biological Sequence Alignment using Dynamic Programming

Implement Needleman-Wunsch and Smith-Waterman dynamic programming algorithms for global and local sequence alignment respectively for protein sequences.

Note: You must use standard Python programming language. You are NOT allowed to use non-standard packages or libraries (e.g. Biopython, scikit-learn, SciPy, NumPy, etc.).

A: Input format:

Protein sequences are provided in the FASTA format. In the FASTA format, the first row is the tag of the sequence with a leading character '>'. The following rows are the actual sequence. For example, it may look like:

```
>test protein X
EEEE
KKKK
AAAA
FFF
```

It represents a test protein sequence "EEEEKKKKAAAAFFF". Notice the length of each row is variable. Your program should accept two FASTA files for the alignment. Please use the BLOSUM62 (<http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt>) scoring matrix as the objective function.

B: Output format:

The output should be the resulting alignment score and an alignment between the two input sequences. For amino acids (characters) whose matches have positive scores in BLOSUM62, use '|' to indicate a positive match; otherwise use '*'. For alignment of amino acid and gap, no symbol should be placed. For example:

```
Score: 12345
EEEEKKKK
||||
EEEE-----

AAAAAFFF
*****|||
BBBBBFFF
```

Each row should contain 80 alignment columns such that the print out does not get messed up.

C: Test case:

Three sets of test protein sequences containing various deadly viruses are supplied:

Pair 1: Dengue virus and Zika virus

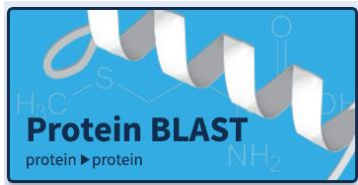
Pair 2: Human papillomavirus (HPV) and Human immunodeficiency virus (HIV)

Pair 3: Poliovirus and Rhinovirus

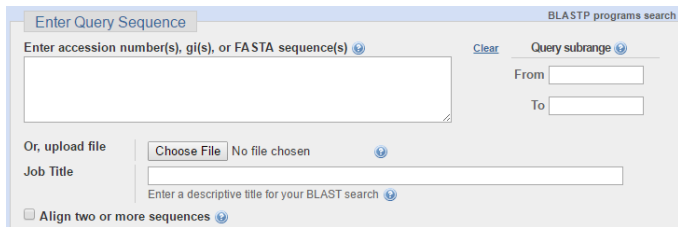
For each pair, perform global and local alignments using Needleman-Wunsch and Smith-Waterman dynamic programming, respectively, and compare the resulting alignments with the alignment generated by Basic Local Alignment Search Tool (BLAST) (see below).

To use BLAST, go to the online server at <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

Click on “Protein BLAST”



Check the box that says “Align two or more sequences” (at the bottom of the figure).



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

From

To

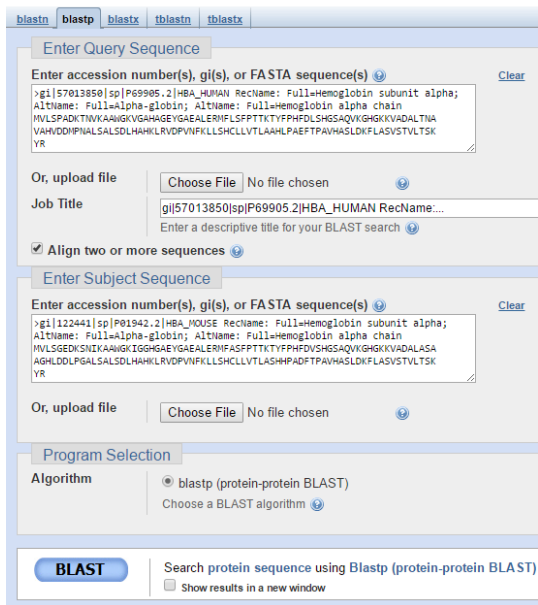
Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Copy the sequences into the boxes.



blastn | **blastp** | blastx | tblastn | tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

```
>gi|57013850|sp|P69905.2|HBA_HUMAN RecName: Full=Hemoglobin subunit alpha;
AltName: Full=alpha-globin; AltName: Full=Hemoglobin alpha chain
MVLSPADKTVKAAIGKVGAGHAGEYGAEALERIFLSPPTTKTYFPHFDLSHSGSAQVKGHGKVVADALTNIA
VAHVDOPNIALSALSDLHAHKLKRVDPVNFKLLSHCLLVTLAAHLPAETTPAVHASLQKFLASVSTVLTSLK
YR
```

Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☒ Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

```
>gi|122441|sp|P01942.2|HBA_MOUSE RecName: Full=Hemoglobin subunit alpha;
AltName: Full=alpha-globin; AltName: Full=Hemoglobin alpha chain
MVLSEEDKSLKAAIGKVGAGHAGEYGAEALERIFLSPPTTKTYFPHFDLSHSGSAQVKGHGKVVADALASA
AGHLDLPGALSALSDLHAHKLKRVDPVNFKLLSHCLLVTLASHHPADFTPAVHASLQKFLASVSTVLTSLK
YR
```

Or, upload file No file chosen [?](#)

Program Selection

Algorithm ☒ blastp (protein-protein BLAST) [?](#)

☐ Choose a BLAST algorithm [?](#)

BLAST Search protein sequence using Blastp (protein-protein BLAST)

☐ Show results in a new window

Scroll down and view the alignment generated by BLAST.

gi|122441|sp|P01942.2|HBA_MOUSE RecName: Full=Hemoglobin subunit alpha; AltName: F
Sequence ID: Query_160837 Length: 142 Number of Matches: 1

Range 1: 1 to 142 [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
253 bits(645)	4e-93	Compositional matrix adjust.	122/142(86%)	131/142(92%)	0/142(0%)
Query 1	MVLSPADKTNVKAANGKVGAGHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG				60
Sbjct 1	MVLSGDEKSNIKAAWGKIGGHGAEYGAELERMFASFPTTKTYFPHFDVSHGSAQVKGHG				60
Query 61	KKVADALTNAAVHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP				120
Sbjct 61	KKVADAL +A H+DD+P ALSALSDLHAHKLRVDPVNFKLLSHCLLVTLA+H PA+FTP				120
Query 121	AVHASLQKFLASVSTVLTSKYR				142
Sbjct 121	AVHASLQKFLASVSTVLTSKYR				142

D: Analysis:

For each of the three sets of test protein sequences, compare alignments generated by your implementations of local and global alignments with that generated by BLAST to identify and discuss similarities and differences.