

## Homework 1: Implementing the sequence alignment algorithms

**Objective: Implement the global sequence alignment algorithm (with banding) and local sequence alignment algorithm. For both programs please implement the affine gap penalty strategy.**

You are NOT allowed to use existing packages or libraries. You can use any programming language; a 2-point bonus will be given the 3 correct submissions that run the fastest.

### A: Input format:

The sequence is 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:

```
>dog protein X
EEEEEE
KKKKK
AAAAA
FFF
```

It represents a dog protein sequence "EEEEEEKKKKAAAAAFFF". Notice the length of each row is variable. The 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 for match and mismatch; use -11 as the gap opening penalty and -1 as gap extension penalty.

### B: Output format:

The output is 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
EEEEEEKKKKK
| | | |
EEEE-----

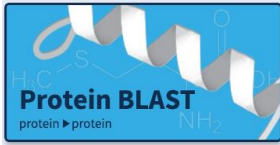
AAAAAFFF
***** | | |
BBBBBFFF
```

Each row should contain 50-80 alignment columns such that the print out will not be messed up.

### C: Test case

You can download two protein sequences (human hemoglobin alpha unit and mouse hemoglobin alpha unit) from our course website. Go to BLAST 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 BLASTP programs search

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  
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSQAQVKGHGKVDALTA  
VAHVDLDPGALSALSDLHAHLKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLT  
SKYR

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  
MVLSGEDKSNIKAAWGKIGGHGAEYGAELERMFASFPPTTKTYFPHFDVSHGSAQVKGHGKVDALASA  
AGHLLDLPGALSALSDLHAHLKLRVDPVNFKLLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLT  
SKYR

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	MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSQAQVKGHG				60
MVLS	DK+N+KAAWGK+G H EYGAEALERMF SFPTTKTYFPHFD+SHGSAQVKGHG				
Sbjct 1	MVLSGEDKSNIKAAWGKIGGHGAEYGAELERMFASFPPTTKTYFPHFDVSHGSAQVKGHG				60
Query 61	KKVADALTNAAVAHVDDMPNALSALSDLHAHLKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP				120
KKVADAL	+A H+DD+P ALSALSDLHAHLKLRVDPVNFKLLSHCLLVTLA+H PA+FTP				
Sbjct 61	KKVADALASAAGHLLDLPGALSALSDLHAHLKLRVDPVNFKLLSHCLLVTLASHHPADFTP				120
Query 121	AVHASLDKFLASVSTVLTSKYR 142				
AVHASLDK	FLASVSTVLTSKYR				
Sbjct 121	AVHASLDKFLASVSTVLTSKYR 142				

**D: Submission**

- Send your source code to my email (cczhong at KU dot edu) with title “EECS 730 HW1 submission” by **11:59 PM, Sep 30<sup>th</sup>**.
- Include a README file describing how to run your program (platform to compile, command line to run, *etc*).