



This lab is meant to acquaint you with the NCBI Basic Local Alignment Search Tool (BLAST) suite. In lecture we've discussed metrics, and we'll soon discuss the Smith-Waterman dynamic programming alignment algorithm. Although that algorithm has been superseded by BLAST, its core ideas (gap cost, insertion cost, deletion cost, etc.) are all part of BLAST, too.

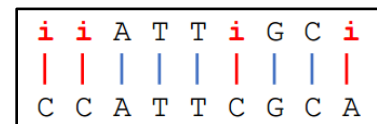


As was done for previous labs, use a word editor to compose answers to the questions throughout this lab, and submit your writeup via Canvas.

## I. Alignment

Assume the following scenario, where SOURCE specifies a sequence that you'd like to align with another one, Target 1.

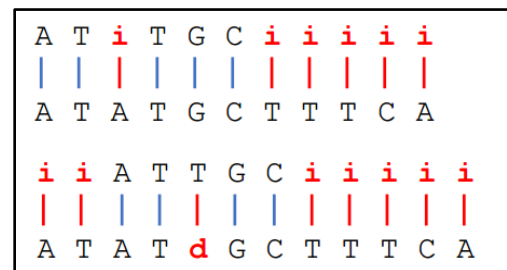
SOURCE 1 :        A T T G C  
Target 1 :        C C A T T C G C A



You could argue that there is a “best” way to align SOURCE 1 with Target 1, shown above right. Two insertions, labelled **i**, before the SOURCE's first A, an insertion between T and G, and a last insertion after the C, are the modifications needed so that the SOURCE perfectly matches Target 1. The term **edit distance** measures the number of changes that must be made to alter a sequence so that it is like another. In this case the edit distance is 4 if we assume that each insertion incurs a cost of 1.

Other types of modifications to the SOURCE, or modifications to the Target, might be needed, to make a quality alignment, as for example in the following SOURCE 2 and Target 2:

SOURCE 2 :        A T T G C  
Target 2 :        A T A T G C T T T C A



In this case, multiple alignments are possible; two are shown right. The top alignment involves 6 insertions into the SOURCE strand, for an edit distance of 6. The bottom alignment involves 7 insertions into the SOURCE strand, and a deletion from the target strand, specified **d**, for an edit distance of 8, assuming deletions and insertions incur the same cost of 1. Although the top alignment has a lower edit distance than the lower alignment, if a researcher is interested in finding alignments without introducing gaps into the SOURCE, then the lower alignment might be considered better.

**Q1 : What changes to the cost of the operations (insertion, deletion, replacement, etc.) would result in an edit distance for the lower alignment for SOURCE 2 / Target 2 that is smaller than the edit distance for the upper alignment in SOURCE 2 / Target 2?**

## II. Smith-Waterman

The Smith-Waterman (to be discussed in lecture) dynamic programming solution to the SOURCE 2 / Target 2 alignment problem is shown right. The algorithm finds the optimal (smallest edit distance) solution, but unfortunately the computational resources needed (creating matrix, populating it, and then “reading” the solution from the matrix) make the approach impractical when aligning a SOURCE sequence against a database of millions of sequences.

	-	A	T	A	T	G	C	T	T	T	C	A
-	0	0	0	0	0	0	0	0	0	0	0	0
A	0	5	4	5	4	3	2	1	0	0	0	5
T	0	4	10	9	10	9	8	7	6	5	4	4
T	0	3	9	8	14	13	12	13	12	11	10	9
G	0	2	8	7	13	19	18	17	16	15	14	13
C	0	1	7	6	12	18	24	23	22	21	20	19

## III. BLAST

Since the advent of high throughput sequencing in the 1990s, the use of Smith-Waterman and related dynamic programming approaches became obsolete due to the time needed to perform a dynamic program run for each SEARCH – database entry pair. Smith-Waterman has been superseded by the Basic Local Alignment Search Tool approach to quickly find good alignments of a source sequence with millions of sequences in a database.

The basic premise of BLAST is the following: for a SOURCE sequence  $S$  of length  $n$ , generate all of the possible words (subsequences) with length  $m$ , where  $m$  is an integer such that  $m < n$ . For example

SOURCE = TAGCTCGGT ( $n = 9$ )

words = {TAG, AGC, GCT, CTC, TCG, CGG, GGT} (all words have length  $m = 3$ )

Using a scoring scheme in which letter matches are +5 and letter mismatches are -4, a similarity measure is computed for each entry in words when it is compared against ALL possible  $m$ -length matching words (in this case, there are  $4^3$  of them, AAA, AAC, AAT, AAG, ACA, ATA, AGA, etc.). For example, for the word TAG, the following similarity scores are calculated

TAG versus matching word AAA for a similarity score of  $-4 + 5 - 4 = -3$

TAG versus matching word AAG for a similarity score of  $-4 + 5 + 5 = 6$

TAG versus matching word AGA for a similarity score of  $-4 - 4 - 4 = -12$

TAG versus matching word GAA for a similarity score of  $-4 + 5 - 4 = -3$

...

The matching words with the highest similarity score are retained, and an exact match for them is searched for in a database. When a match is found, the algorithm continues to calculate the similarity score between the database entry and the matching word, until the similarity score decreases. For example, consider the SOURCE, one of the words, its matching word, and a database entry:

SOURCE : TAGCTCGGT  
word : TAG (one of several)  
matching word : AAG (with highest similarity score)  
database entry : AAGCAAGCTAGGGGCCCC (one from among millions)

The SOURCE, the word, the matching word, and the database entry align the following way:

```

TAGCTCGGT
TAG
AAG
AAGCAAGCTAGGGGCCCC
```

For which the following match is found between the SOURCE and database entry

```

iTAGCTCGGT
TAG
AAG
AAGCAAGCTAGGGGCCCC
```

At this point, BLAST would specify entry AAGCAAGCTAGGGGCCCC in the database as having a certain degree of a match with SOURCE. A user might then specify to view the top 100 matches.

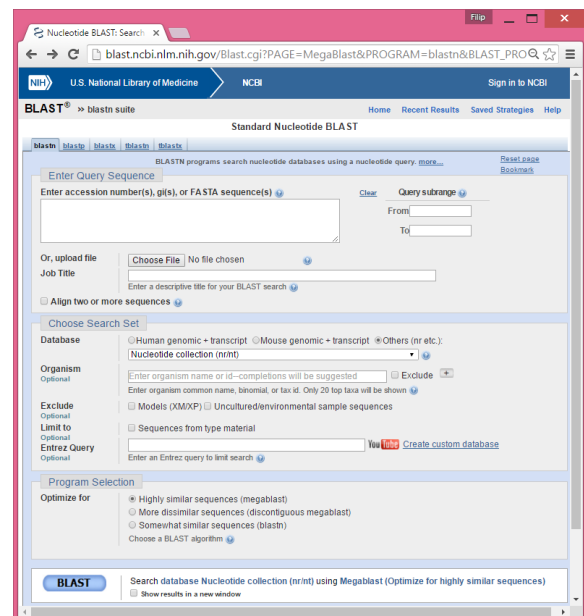
The premise of BLAST is the following : **minimize the time spent looking at sequences in a database whose similarity with the SOURCE sequence has little chance of generating a good alignment.**

#### IV. NCBI BLAST

NCBI's suite of BLAST tools are at the following URL:  
<http://blast.ncbi.nlm.nih.gov/>

Because BLAST is a general technique (although the name has been trademarked) rather than a tool used just for aligning nucleotide sequences, you have the option to select from among multiple BLAST implementations.

Select **nucleotide blast**. The basic Graphical User Interface (GUI, shown right) is easy-to-use (as all GUIs should be!)



In the textbox entry labeled **Enter accession number(s), gi(s), or FASTA sequence(s)**, input the following sequence (the query); specify to show results in a new window if you want the results in a new tab:

GTCTGACCACTCTGATCCTGTTATGGGCAACCGTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT

**Future screen shots are not necessarily identical to what your BLAST query might return.**

When you submit a BLAST query to the server, your job is assigned an ID, and a status message panel will be displayed (shown right).

Request ID	GAGBGHSH014
Status	Searching
Submitted at	Wed Apr 6 13:03:45 2016
Current time	Wed Apr 06 13:03:55 2016
Time since submission	00:00:09

This page will be automatically updated in 2 seconds

An alignment/match search is being performed against the many entries in the gene bank, so depending on the query sequence, jobs might take tens of seconds, or even minutes to complete.

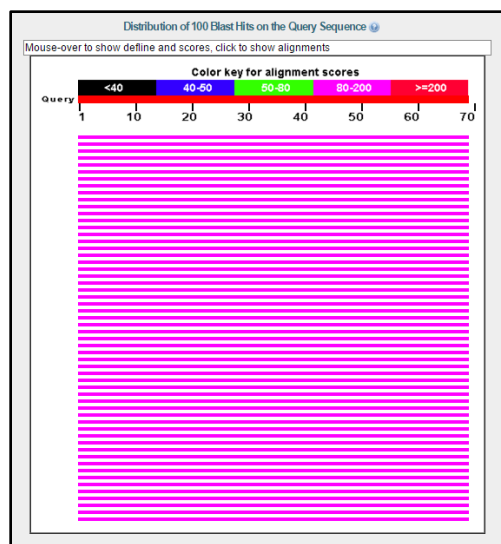
<b>Nucleotide Sequence (70 letters)</b>	
RID	<a href="#">GAGH8AXW014</a> (Expires on 04-08 01:06 am)
Query ID	<a href="#"> cl_Query_163751</a>
Description	None
Molecule type	nucleic acid
Query Length	70
Database Name	nr
Description	Nucleotide collection (nt)
Program	BLASTN 2.3.1+ <a href="#">► Citation</a>
Other reports: <a href="#">► Search Summary</a> <a href="#">Taxonomy reports</a> <a href="#">Distance tree of results</a>	
<a href="#">+ Graphic Summary</a>	
<a href="#">+ Descriptions</a>	
<a href="#">+ Alignments</a>	

The results panel (shown above right) includes batch job details (ID, a URL link that is valid for a certain amount of time, the query length, etc.)

**Q2 : What is the BLAST search ID for the alignment/job you submitted?**

### Q3 : How long did the BLAST search take?

You can delve into the details of the alignment(s) found via a graphical summary, descriptions of the results (matches), or by inspecting the alignments in detail. The graphic summary (shown right), displays color-coded bars where the color of each bar represents the alignment score for a sequence that was found to have a high match.




Mouse hover over the graphical view, and select and click on the top-most bar. That will take you to highest-scoring alignment results portion of the page for that alignment (show right).


Download - GenBank Graphics


Parvovirus isolate S beta globin (bb) gene, exons 1, 2 and partial sequence to [gb217382.1](#); Length: 1012 Number of Matches: 1


Range: 818 to 887 GenBank Statistics View History | Previous Match

Sequence	Coverage	Gap	Pha/Sha
115 (82)	7e-33	87.75 (96%)	0.0 (0%)

Query 1  60

Subject 818  877

Query 62  70

Subject 878  887

**Related Information**

NCBI Virus - aligned genome context

**Q4 : What is the organism and gene that is the best match for the query sequence?**

**Q5 : The 67-sequence query aligns best with what range of bases in the found (Subject) sequence?**

**Q6 : The alignment contains how many exact nucleotide matches?**

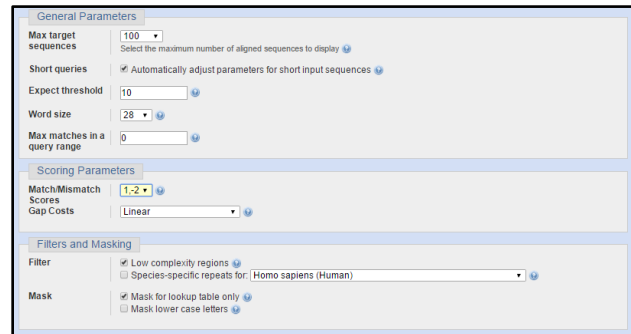
**Q7 : The alignment contains how many nucleotide mismatches?**

**Q8 : The alignment contains how many gaps in the Query sequence (be careful here ... notice the starting base number of the query sequence)?**

## V. NCBI BLAST parameters

Referring back to the introduction of this lab, and to the lectures about scoring alignments, you are now aware that BLAST (and most alignment tools) assign cost values to gaps, insertions, deletions, matches, etc. to help score and identify a sequence that is the best match from among millions.

Return to the BLAST input page, and expand the **Algorithm parameters** section, shown right.



**Q9 : What is the default word size?**

**Q10 : What “penalty” value for use in the distance score is imposed for a base mismatch?**

**Q11 : What “reward” value for use in the distance score is imposed for a base match?**

**Q12 : What are the possible gap cost choices? Explain what “existence” and “extension” refer to.**

Change the BLAST parameters to the following:

Word size : 32

Match/Mismatch Scores : 2, -3

And perform a new BLAST search (use the BLAST button in the Algorithms section) using the same sequence provided to you earlier in this lab.

**Q13 : Inspect the top alignment result. How does it differ from the result that you received when the default BLAST parameters were used? EXPLAIN the cause of the difference. If there is no difference, why might that be the case?**

## VI. Submission and rubric

Upload to Canvas your answers to the 13 questions. Only .pdf, .docx, or .doc files are accepted

Component of Lab	Points
Q1-12, each 2 points	24 points
Q13	4 points
Total	28 points