

## CS196-1 HW7

*Due: Thursday, April 24<sup>th</sup> 11:59pm (Midterm given out in class on Tuesday)*

### 0 Reading

- BLAST handout: Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990)
- **Optional** Textbook section 9.8: BLAST: Comparing a Sequence Against a Database

### 1 BLAST

In this problem you will be implementing the BLAST algorithm, and then testing your code against a series of databases.

**The BLAST Algorithm** The BLAST algorithm is a heuristic search method that seeks words of length  $k$  that score at least  $T$  when aligned with the query and scored with a substitution matrix. Words in the database that score  $T$  or greater are extended in both directions in an attempt to find a locally optimal ungapped alignment or MSP (maximum segment pair) with a score of at least  $S$ . MSPs that meet these criteria will be reported.

**The Substitution Matrix** As discussed in class, we will use the following DNA substitution matrix:

$$\delta(x_1, x_2) = \begin{cases} +5, & x_1 = x_2 \\ -4, & x_1 \neq x_2 \end{cases}$$

Roughly, the algorithm is as follows:

1. Break down the query string into all possible *words* of length  $k$ . Now consider each word,  $w$ , individually.
2. Create a list of *search terms*, comprised of  $w$  and all other words that, when compared against  $w$ , have a score of at least  $T$ .
3. Identify all the places in the database where one of search terms appears, and extend in each direction to find the maximum segment pair (MSP).
4. If the MSP has a score of at least  $S$ , report it as a match.

You will need to complete the following:

1. Implement the BLAST algorithm, which will accept query strings and search for matches in the database, using the index you constructed. The parameters should be as follows:

$$K = 4$$

$$T = 10$$

$$E = 0$$

$$S = 0.80 * 5 * |query\_string|$$

2. Using the sample database of 20 genes (500 bases each) found on the website:  
<http://www.cs.brown.edu/courses/cs196-1/hw7.html>  
and the following query strings:

#	Query String
1	GAGAACCA
2	AAATAAAATGTGGGAGTGTG
3	AAAGGGGGCTTCCCTGTGAGTCGCTG
4	CCAGCATAGGGCTGTTATCTAAGTACCTGTGCATCAGCACGCCCCAGGCCTGC
5	GGATTCATTTGTATCATGCATGGAA

Report the search results. Specifically, each search result should include:

- Score
  - Database entry number
  - Offset
3. Now let  $K = 3$ ,  $S = 0.60 * 5 * |query\_string|$  and search for the string: TAGCAATCTGGC-CATTGGCCCTATCCGGG. What score/database/offset do you get?
  4. **Extra Credit (5 pts):** Compared to the frequency of searches, the underlying database changes relatively infrequently. So it can be extremely useful to create an index of the database beforehand, which the BLAST algorithm can reference to look up word locations. Write a program that constructs an index of the database.