Lab 5: Imperfect Motif Discovery

Due date: Tuesday, November 25.

About the Lab

Time	CSC 448	CHEM 441
November 6 lab	Discuss assignment/map out requirements	
November 6 – 13	Design, development	Data preparation
November 13 lab	Implementation of core algorithms	
November 13 – 18	Implementation of core algorithms	
November 18 lab	Testing	
November 18 - 20	Testing	Software use
November 20 lab	Software delivery	Software use
November 21 – 25	Software refinement	Software use

Lab Assignment

In Lab 4 you helped your BIO 441 partners by providing software that searched for *exact string matches* in DNA fragments. The short string matches you were looking for are called *motifs*. The motifs that show up significantly more often than expected in DNA sequences are usually theoretized to have specialized function, and therefore are important for DNA analysis and comparative DNA studies.

In many cases, however, the true *motifs* present in the DNA sequences are not *exact*. Due to a variety of reasons (genetic code degeneracy, evolutionary mutations), a specific specialized function may be ascribed to an *imprecise motif*.

Where, a precise (exact) motif is a short sequence of nucleotides, e.g., $ACGTCTCTCTA$, an *imprecise* motif may be expressed by mutliple strings that are similar in structure, but also differ on one or more codon.

Example. Consider the following English-language definition of a possible imprecise motif:

> *A specialized function is ascribed to any DNA sequence that starts with codons $ACTC$, followed by either an A or a T, followed by GTA, followed by another A or T, followed by TT.*

The motif description above matches four different strings:

$ACTC\mathbf{A}GTA\mathbf{A}TT$
$ACTC\mathbf{A}GTA\mathbf{T}TT$
$ACTC\mathbf{T}GTA\mathbf{A}TT$
$ACTC\mathbf{T}GTA\mathbf{T}TT$

Biologists often use an extended nucleotide alphabet to describe imprecise motifs. The alphabet is outlined in the table below:

Symbol	Meaning	Nucleic Acid
A	A	**A**denine
T	T	**T**hymine
C	C	**C**ytosine
G	G	**G**uanine
M	A or C	a**M**ino
R	A or G	pu**R**ine
W	A or T	**W**eak
S	C or G	**S**trong
Y	C or T	p**Y**rimidine
K	G or T	**K**eto
V	A or C or G	not T (V)
H	A or C or T	not G (H)
D	A or G or T	not C (D)
B	C or G or T	not A (B)
N	A or C or G or T	a**N**y
-		gap in sequence

Using the alphabet above, the motif described in our example can be represented as a string

$$ACTC\mathbf{W}GTA\mathbf{W}TT.$$

Lab Assignment

Your partners want to extend the ability of your software to identify motifs to include imprecise motifs expressed using the extended nucleotide alphabet above.

While the specific expectations for the software will, as always, be provided by your BIO 441 partners, the core algorithmic problem you will need to solve is as follows:

Given a (possibly long) DNA fragment D in nucleotide alphabet, *and a (short) imprecise motif sequence P in extended nucleotide alphabet, find all occurrences in D of P **or sequences similar to** P.*

While it is possible to solve this problem using suffix trees[1], because imprecise motif search on suffix trees will no longer follow a single path in the tree, the procedure may be computationally expensive.

Instead, there is a relatively simple solution to the imprecise motif finding problem by coopting the sequence alignment problem that we are studying in class.

Note: The assignment is handed out to you a week before we will actually be covering the sequence alignment problems and the appropriate algorithms. Out November 6 class concentrates on preliminaries of dynamic programming techniques for problem solving and on the definitions of global and local alignment problems. To compensate for this, the length of the lab has been extended.

Your task for this lab is to do the following:

1. Map the imprecise motif finding problem to one of the alignment problems to be covered in class (global alignment or local alignment).
2. Determine the appropriate algorithm to implement (Nudelman-Wensch for global alignment, Smith-Waterman for local alignment).
3. Determine necessary inputs to the algorithm and work with your BIO 441 partners to obtain them[2].
4. Implement the appropriate alignment algorithm.
5. From the output produced by the alignment algorithm of your choice, produce the output that actually answers the imprecise motif finding question in a way that is convenient for your BIO 441 partners.

Submission Instructions

As usual, two sets of deliverables are needed: one for `handin` and one - for `Piazza`. The core deliverables are the code and two requirements documents: the original document provided to you by your BIO 448 partners and the final version.

handin deliverables. The `handin` deliverables are: *source code*, *compilation/running instructions (README)*, *user documentation* and *both requirements documents*. Submit them using the following command:

```
$handin dekhtyar 448-lab5 <files>
```

[1] A question like this may find its way onto the final exam, for example.

[2] In addition to the two input strings there is at least one more input that your partners will need to provide for you - once you identify what it is, you may work jointly on developing it.

Deadlines. There is only one deadline: submit everything by the end of the calendar day on Tuesday, November 25. Please note, that your **BIO 441** partners may want to work with preliminary versions of the software before the submission deadline. **Also, note that Lab 6 assignment may be distributed to you either during November 20, or during November 25 classes, so you may need to use the November 20–26 period to work on two assignments concurrently.**