Artificial Intelligence Term Project Spring 2017

Background

You are working for a state-of-the-art pharmaceutical company which is working on a new lung cancer treatment based on gene therapy. You are given a set of the control sequences in a gene family that the company believes will have impact on the growth of lung cancer cells. Your job is to identify from these control sequences the important patterns that may carry significant biological meanings.

Issues you must understand and resolve

A candidate pattern in control sequences is considered an important pattern only if it does not appear in the entire genome (just think of this as a very large set of biological sequences for now) at random. Therefore, you first need to define what you mean by "not at random" in your own sense. There is no easy or clear answer to that even for a REAL molecular biologist.

To simplify the real-world problem to our search problem, we assume:

(a) the length of each control sequence is the same,
(b) the length of each pattern in a control sequence is fixed and is provided by the company in advance,
(c) the exact number of occurrences of the patterns in a control sequence may or may not be known beforehand, depending on the patterns we are looking for, and
(d) the content/configuration of a pattern may vary between different control sequences.

A toy example:

You are given a set of 3 control sequences in a gene family as follows. Each sequence has only 4 possible characters (i.e. 4 nucleotides in DNA sequences)
.

```
POS:1--------10 -------20-------30--------40--------
S1: AAATTTTGGGAAATTTAAACCCGCATGAAAATTTTTAAAAATTTTTAAA
S2: CAAATTTAATAAATTTTAAAAATTTTTAAAACGCATGAGAAATTTAACC
S3: TTTAAATTTACGCATGAAATTTAAACCCGGAAATTTAAATTTCAAATTT
```

Your job is to search the 3 control sequences for important patterns, each with 6 nucleotides (i.e. 6 characters in length). In addition, you are told each sequence must contain at least one copy of the pattern.

Obviously, there are numerous candidate patterns in the 3 control sequences,

e.g. AAATTT. However, recall that a "significant" pattern does not exist at random, and in a sense, this also implies that a significant pattern will not appear all over the places on the control sequences.

It is easy to tell that the control sequences are AT-rich, i.e. A and T are all over the places, compared with G and C, which gives us an idea that AAATTT is less likely to be a significant pattern. If you check the sequences, you can identify 12 copies in total. Similarly, you can identify 8 copies of TTTAAA.

Now consider CGCATG, and we find each control sequence has only a copy of CGCATG. Based on the qualitative criteria, "not random and contained in each sequence," compared with AAATTT or TTTAAA, CGCATG is more likely to be a significant pattern.

Note that a significant pattern may vary on different control sequences.

In the example above, the "significant" pattern is CGCATG, and it exists on each control sequence.

However, in the real biological world, the significant pattern on each control sequence may be degenerate, i.e. it has minor variance.

Take the following control sequences for example.

```
POS:1--------10 -------20-------30--------40--------
S1: AAATTTTGGGAAATTTAAACCCGCATGAAAATTTTTAAAAATTTTTAAA
S2: CAAATTTAATAAATTTTAAAAATTTTTAAAACGCTTGAGAAATTTAACC
S3: TTTAAATTTACGCATCAAATTTAAACCCGGAAATTTAAATTTCAAATTT
```

The real pattern is CGCATG, but it varies a bit on S2 and S3.

To solve the pattern search problem, you must take degeneracy into account.

The objective of this project is to search/identify the "significant" patterns from the given sequences.

Define your own "significant" patterns, and implement your search strategy to find them.

Input Data/Information

For each search problem, you'll be given the following:

(a) A set of control sequences, each of the same length.

(b) A pseudo-genome, i.e. a set of biological sequences, as your reference basis to measure "randomness" of patterns. Hint: "random" typically means "everywhere."

(c) The fixed length of patterns you are looking for

(d) Whether the patterns have degeneracy or not

There are 3 problems to solve for this project, each with different levels of difficulty.

Your Output and What to Turn in

(a) A copy of your source code written in C/C++ only. Make sure it is correct and can be compiled in Dev++ or CodeBlock.

(b) A report that describes your design for this project, e.g. (i) what are your states and how do you represent them, (ii) what are the operators you apply, (iii) your goal test, (iv) your starting state, (v) any heuristics, (vi) any novel design idea that was not discussed in class, etc.

The report should be clear and precise about your design to promote your contributions.

(c) A list of "significant" patterns you identify from the given control sequences, e.g. S1:{(CGCATG,22)}, S2:{(CGCTTG,32)},…, where the number indicates the position of the pattern in the sequence.

Search Problems

For the following search problems, you are given a pseudo genome represented by a file named "genome.data" posted on e3.

Q1. The control sequences are listed in the file named "Q1.data" on e3.

The length of the patterns in fixed to 15 characters (i.e. 15 nucleotides).

The pattern on each control HAS NO degeneracy, i.e. it is the same on each control sequence.

Q2. The control sequences are listed in the file named "Q2.data" on e3.

The length of the patterns in fixed to 15 characters (i.e. 15 nucleotides).

The pattern on each control MAY HAVE degeneracy, i.e. it may not be the same on each control sequence.

The degeneracy may occur on at most FIVE random positions within the pattern.

For example, the real pattern is AAATTTGGTTACGGG. However, its appearance on the 1st control sequence may be ATATATCCTTACGGC (5 mutations), GAATTCGGCTACAGG (4 mutation), AAACTTGATTACTGG (3 mutation), AAATCTGGTAACGGG (2 mutations), AAATTTGATTACGGG (1 mutation) or AAATTTGGTTACGGG (0 mutation).

Similarly, its appearance on the 2nd control sequence may be AAAAAAGGTTTTGGG (5 mutations), CGCTTTGGTTACTGG (4 mutations),

AAATGGGGTTATGGG (3 mutations), AACTTGGTTACGGA (2 mutations), AAATTTGGGTACGGG (1 mutation) or AAATTTGGTTACGGG (0 mutation).

Q3. The control sequences are listed in the file named “Q3.data” on e3.
The length of the patterns in fixed to 15 characters (i.e. 15 nucleotides).
The pattern on each control MAY HAVE degeneracy, i.e. it may not be the same on each control sequence.
The degeneracy may occur on at most SEVEN random positions within the pattern.