Algorithms for Molecular Biology
© Ron Shamir
School of Computer Science
Tel Aviv University
December 24, 2001

# Algorithms for Molecular Biology: Assignment No. 3
## due on January 3rd

**General Guidelines:** This assignment is part of your test in the course. It should be done **INDEPENDENTLY**, without any help from others.

**Credit:** Solve 6 items out of 7 for 12/75.

**Practical exercise guidelines:**

- Useful links to this exercise:
  - EBI (European Bioinformatics Institute, UK) home: `http://www.ebi.ac.uk/`.
  - NCBI (National Center for Biotechnology Information, USA) home: `http://www.ncbi.nlm.nih.gov/`.
  - TAU Bioinformatics course by Dr. Racheli Kreisberg-Zakarin, with similar assignments, with solutions: `http://www.tau.ac.il/lifesci/bioinfo/teaching/bioinfo/course_winter_2001.html`.
- Please submit printouts of the relevant browser screens. When needed to answer specific questions, or to mark some findings, you can handwrite on the printout. If you prefer handing in a pen-free assignment, you may edit the html or a separate document.
- Occasionally, the biological knowledge presented in the course and the results of your database queries so far do not lead to a definitive *Yes* or *No* answer beyond reasonable doubt. In that case, *Possible* is a legitimate (and in fact the best) answer.

**Background:** You are a biologist who studies a rare human disease, called *Algmbitis*. The symptoms are excessive problem solving and lack of free time, with the mean age being 20-30 years. After years of work, you understand that is it associated with specific malfunctioning cells. You harvest mRNA from such a cell, and get a cDNA, which you sequence. You reverse and complement the result to obtain the coding strand, and get:

```
tagtttacagtatatggattctatgaaaaaatgtccacaacacagtaattggtcacttgttaagaaaaattt
atccttgtaagtatcttcaaagttgatatttggaactttattccaaaatagtgcatgtggagaaagaatcta
gactttcttgtatacatttttctcttctccagtaataaacaattacctttcatttaaaaaaaaaaaaaaaaa
```

**Questions:**

1. Search this sequence for open reading frames (NCBI ↦ tools ↦ ORFinder) that make sense, and translate them into a protein. Are the following statements correct/incorrect/possible for the sequence/a subsequence of it:
   - It is a coding region.
   - It is an exon.
   - It is an intron.
   - It is a 5' untranslated region.
   - It is a 3' untranslated region.

   Explain your answer.

2. Search the databases for your full length molecule. Which database should you use? Which of the following is correct/incorrect/possibly correct:
   (a) You have found a new gene.
   (b) You have found a mutated version of a known gene. (What are the mutations?)
   (c) Your sequencing machine made sequencing errors.(What are they?)
   (d) The gene is not human, but rather an infection of a virus in the test tube. (What virus?)
   (e) The database contains multiple records for your gene. (Why?)
   (f) Your sequence is a part of a much longer clone, which is not dedicated to a specific gene.
   (g) Your gene has a homologous counterpart in mouse.
   (h) Your gene has a homologous counterpart in the fruit fly.
   (i) Your gene has a homologous counterpart in a plant. (What plant?)
   (j) Your gene has a homologous counterpart in the E.Coli bacterium.

**Note:** For identifying species you may need the help of NCBI $\mapsto$ Taxonomy.

3. Search the database with the full-length gene (not the sequence you were given). Try at least two of the four possibilities DNA/Protein vs. DNA/Protein. Compare the searches you performed. Should you revise the answers you have given to the previous question regarding homologous counterparts?

4. Search your protein against SwissProt. Which sure homologous sequences did you find? Which alignments that you found may be random?

5. To resolve ambiguities perform a $\Psi$-Blast search against SwissProt. Repeat iterating until convergence. Which of your hits are significant? Which are you still uncertain of? Try playing with the $\Psi$-Blast threshold, and iterate further till convergence to a set of significant proteins which is more clearly separated from random hits. What is the E-score of your top random hit? What is the probability of the best random hit to have such a score (or better)? Be careful not to use a threshold that is too permissive. Explain this last warning.

6. Look at the multiple alignment of the proteins in your set of significant $\Psi$-Blast hits (EBI $\mapsto$ tools $\mapsto$ ClustalW). Recall, that the active, or function-determining part of a protein is subject to selective, conservative pressure more than other parts. Where in your protein can you find candidates for active sites? Compare to a search versus the Conserved Domains Database (NCBI $\mapsto$ structure $\mapsto$ CDD), and a view of the multiple alignment of the family you find there.

7. Find the relevant records for the gene and the protein. Does your gene have a homologous counterpart in yeast? Find your gene in the Human genome (`http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/euk_g.html`). Where does your gene lie? Find, by a text-based search (EBI $\mapsto$ tools $\mapsto$ SRS, or NCBI $\mapsto$ Entrez), more human proteins with the same function. Do they lie anywhere near the *Algmbitis*-related gene along the genome?