

Algorithms in Bioinformatics I, WS2006/7

Assignment sheet # 4

Tobias Dezulian

November 6, 2006

1 Expected number of matches (1 point)

For BLAST, the E -value of an HSP can be expressed in terms of the so-called bit-score S' as follows:

$$E = mn2^{-S'}$$

Please show how to derive this from the definitions given in the lecture.

2 Preparation: download a genome and a proteome (2 points)

Please go to the NCBI homepage at <http://www.ncbi.nlm.nih.gov/> and sequentially to sections “Genomic Biology”, “Whole Genomes”, “Genomes”, “Complete Genome Sequences”, “Bacteria”, “Chromosome”, and download the complete genome sequence (nucleotides) of *Escherichia coli K12* in FASTA format.

Next, go to the homepage of NCBI’s European counterpart, the European Bioinformatics Institute (EBI) at <http://www.ebi.ac.uk/> and then to “Services”, “Databases”, “Nucleotide Databases”, “Genomes Server”, “Bacteria”. From there, download all amino acid sequences which are encoded in the genome of *Yersinia pestis Antiqua* (Tax ID: 360102)—the complete proteome—in (multi-) FASTA format (file 25054.Y_pestis_Antiqua.fasta).

How many nucleotides does the *E. coli* genome contain? How many amino acid sequences are contained in the given *Y. pestis* proteome?

3 BLASTing amino acids against a genome (4 points)

Use your *Y. pestis* sequences as BLAST queries against your database of the *E. coli* genome. Be careful to use a BLAST program that is suitable for amino acid queries and a nucleotide database. Use an E-value cutoff of e^{-30} . Assume that the best hit for each query sequence in the *E. coli* genome using this E-value cutoff is an ortholog. What is an ortholog? How

many orthologs of *Y. pestis* proteins are contained in the *E. coli* genome? Why could we expect to find orthologs?

4 Java wrapper / Automation (3 points)

Using Java, it is not easy to interact with external programs such as BLAST. Why? What are common problems? Write your own Java wrapper around the appropriate BLAST program so that you can execute the above task (# 3) simply by starting your Java program. Your Java program should take as arguments two file names (query file and BLAST database) and deliver as its only output the number of orthologs. Hint: use shell redirection ">" to write the BLAST output into a file and then parse its contents from Java.

Have a look at the Java class `ExternalWrapper` that tries to make external program calls easier. Execute it on a linux system. Why is this class so complicated?

Due by 10am, Monday, 13 Nov 2006