Databases and Data Mining

Assignment 4

23-11 2008

Due: Friday 21-12 2009
Grading: This assignment will be graded from 0 to 10.
Notes: Please read carefully:

- Groups of 1-3 students are allowed.
- Use C, C++ (MS Visual C++, or gcc), Python, or JAVA together with an HMM Toolkit of your choice (GHMM [4] is recommended). Also MatLab code is allowed, if your MatLab code installs and runs under the available LIACS license on the Linux student machines.
- Put the complete code with **short** and **clear** instructions on how to compile and execute it in a single directory called ***"<your student number(s)><your last name(s)>_4"***. If you used any derived data files, please add them.
- Write down your report for this assignment in a *.pdf* file with the following name ***"<your student number(s)><your last name(s)>_4.pdf"*****,** e.g., ***"012345jansen_4.pdf"*** and put it in the same directory as the code**.** (A *.doc* file is also allowed.)
- Compress the complete directory into a single zip file called ***"<your student number(s)><your last name(s)>_4.zip".***
- Send this *.zip* file as an attachment of an e-mail with subject **"DBDM_4"** to erwin@liacs.nl.
- Grading will be based on the originality, and quality, of your code, the quality of your analysis and results, and the argumentation, validity, and clarity of your **final report**. Do not forget to clearly state the references you used for your work!

Introduction

Hidden Markov Models (HMMs) have been used in many application areas to model very different types of data. Especially, in the field of continuous speech recognition remarkable successes have been obtained by applying HMMs. A very nice tutorial on HMMs applied to the field of speech recognition is by L.R. Rabiner [1].
In the field of Bioinformatics, the first applications of HMM for analyses of DNA sequences, finding protein-binding sites, finding genes, etc., stem from the late 80's - early 90's. One of the first studies on finding genes in DNA is by A. Krogh et al. [2]. In this work the HMM consists of two parts, a kind of strict Markov model for the gene regions, combined with a more generic HMM for the intergenic regions. It was shown that the approach was very effective in finding the exact locations of genes and intergenic regions in the E. coli genome.

At the time of the study of A. Krogh only around a third of the E. coli genome was sequenced, less than 25% of the genes were known, and several sequencing errors had to be expected. In this assignment we will use the complete genome sequence of the *Escherichia coli* O157:H7 strain EDL933, as described in the January 25, 2001 issue of Nature. This genome was sequenced during the E. coli Genome Project of the University of Wisconsin – Madison [3].
The goal of this assignment is to automatically find genes in the E. coli genome using an HMM that is specially designed and trained for this task by you.

Datasets

The *data_assignement_4.zip* file contains four data files (from [3]). The two files *AEOO5174v2-1.fas*, and *AEOO5174v2-1.fas* are FASTA formatted files containing the two contigs 1 and 2, respectively, that together constitute the complete E. coli (O157:H7 strain EDL933) genome. This basically gives a particular linearized 'TCGA'-sequence of the complete E. coli genome (which is circular).
Please note, that there is a gap of about 4 kbp between contig 1 and contig 2, and the end of contig 2 and the beginning of contig 1 overlap by 527 bp (base pairs) to complete the circular chromosome.
The two other files each contain the same table with all the annotated genes (ORFs and RNAs) in the current EDL933 sequence with information derived from the gene annotations (e.g., location, name, product, function, etc.). *edl_genes.zip* contains the original tab delimited file with all the annotated genes, whereas *edl_genes.xls* is a *MS Excel* file containing the same information. This file can be used for an easy extraction of single columns.

Statistical Report

Please, give statistics on the frequency of individual nucleotides (C, G, T, A, respectively) in contig 1, contig 2, and the complete genome, respectively. Also give the statistics for the frequency of the codons (i.e., nucleotide triplets) in each of the **6 reading frames*.** Report these statistics in tables of the form:

	Contig 1	Contig 2	All
C			
G			
T			
A			

and

	Contig 1	Contig 2	All
AAA			
AAC			
AAT			
…			

You may also add the same kind of statistics for the genic and intergenic regions (see also [2]].

*__Note:__ you can scan the CGTA-sequence for codons from left-to-right and right-to-left. Furthermore, in each direction you can start the scanning process for codons at position 1, 2, or 3, respectively. This gives a total of 6 (2 x 3) so called **reading frames**.

HMM Modeling

It is encouraged that for designing and modeling your HMM for finding genes in the E. coli genome, you use the General Hidden Markov Model library (GHMM) [4]. GHMM is a freely available LGPL-ed C library implementing efficient data structures and algorithms for basic and extended HMMs with a Python interface. Of course you are also allowed to use another HMM Toolkit of your choice (e.g. HTK, HMM under Matlab, etc.).
Please give very clear instructions on how you modeled, trained and tested your HMM! In principle your fellow students should be able to check your work and results very easily using these instructions.

HMM Training & Testing

Use the first file *AEOO5174v2-1.fas* together with the gene table for training of the HMM. Use the second file *AEOO5174v2-2.fas*, together with the gene table for testing. Report your results in a table.

References

[1] L.R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77, No. 2, pp 257-286, February 1989.
[2] A. Krogh, I. Saira Mian, D. Haussler, A Hidden Markov Model that finds genes in E. coli DNA, Nucleic Acids Research, Vol. 22, ,pp. 4768-4778, 1994.
[3] http://www.genome.wisc.edu/sequencing/o157.htm
[4] http://ghmm.sourceforge.net/ and http://ghmm.org/