Final exam Computational Biology – BIA – winter 2012

Unequal usage of codons in the coding regions appears to be a universal feature of the genomes across the phylogenetic spectra. This bias results mainly in the uneven usage of the amino acids in the existing proteins and the uneven usage of synonymous codons. The bias in the usage of the synonymous codons correlates with the abundance of the corresponding tRNAs.
By comparing the frequency of codons in a region of a species' genome, read in a given frame with the typical frequency of codons in the species genes, it is possible to estimate a likelihood of the region to be coding for a protein in such a frame. Regions in which codons are used with frequencies similar to the *typical species codon frequencies* are likely to code for proteins (or part of proteins) while regions which codons distribution is *uniform* -- **1/64** (0.015625) -- could be considered as non-coding regions (introns, intergenic or non protein coding exons like exons corresponding to untranslated regions of the transcript).

Let S be a sequence of DNA, the higher the number of codons in S matching the more frequent codons in the table, the higher the coding potential of the sequence S (the probability to be encoding a protein). This value can be computed as:

$$cod\_pot = \frac{\mathrm{p(C1)p(C2)p(C3)\ldots p(Cm)}}{\left(\frac{1}{64}\right)\left(\frac{1}{64}\right)\left(\frac{1}{64}\right)\cdots\left(\frac{1}{64}\right)}$$

where C1, C2, C3, … , Cm are the m codons. This expression is known as likelihood ratio and it is usually computed in logarithmics terms to reduce the magnitude of the numbers (**log-likelihood ratio**).

Write a Perl script able to test the coding potential of a **single** sequence stored into a FASTA formatted file. **You will be given 3 files**:

Fasta1.txt, Fasta2.txt and HCUT.txt . The HCUT.txt file contains the human codon usage table. Fasta1.txt and Fasta2.txt are files containing ‘unknown’ human sequences.

- The **objective** of this exam is to decide *if a protein coding region is contained in Fasta1.txt or Fasta2.txt* using a Perl script **and** to find evidences supporting your conclusions. (I suggest the use of the *blat* aligner available in the UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgBlat?command=start).
- The script should be able to compute the coding potential for each of all the possible reading frames. Why?
- Once identified the (eventually present) protein coding region use the fasta for a BLAT or BLAST search against the human genome and find the IDENTIFIER and the EXON(S) spanned by the alignment.
- Write a **very short** report containing the pseudocode (or the equivalent flowchart) of the proposed solution, the output produced by the script using the fasta files as input, your interpretation of the produced results. In the case a putative protein coding region is found try to blat the fasta sequence in order to verify if it can be aligned to a protein coding gene (and put in the report the identifier of this gene).

The files are available at:
http://homes.di.unimi.it/~re/Corsi/CB13mat/FinalExam_Erasmus/