

# Algorithms in Bioinformatics 2, SoSe2007

## Assignment sheet # 2

Tobias Kloeppe

April 23, 2007

This weeks assignment has its focus on finding/predicting genes in a genome.

### 1 ORF finder (5 points)

Please download [www-ab.informatik.uni-tuebingen.de/teaching/ss07/albi2/java/assign02.zip](http://www-ab.informatik.uni-tuebingen.de/teaching/ss07/albi2/java/assign02.zip). Please write a Java program `albi2.genes.getORFs` that takes as input the name of a file containing a DNA sequence in FastA format, and a minimal length  $n$ , and produces as output a list of all ORFs (open reading frames) of length  $\geq n$  that can be found on the forward or reverse strand of the input sequence.

The output should be one ORF per line, specifying the start and finish position of the nucleotides involved in the ORF (i.e., the position of the first nucleotide included in the start codon and the last included in the stop codon), the length (i.e., number of codons) and the corresponding amino-acid sequence: *start end length amino-acid-sequence*

Let the first nucleotide of the input sequence have position 0. For an ORF in forward direction, your program should report  $start < end$ , whereas for an ORF on the opposite strand, please report  $start > end$ .

Use the standard genetic code to translate the DNA into amino-acids:

“gc[agtc]”=A, “tg[ct]”=C, “ga[tc]”=D, “ga[ag]”=E, “tt[tc]”=F, “gg[agct]”=G, “ca[tc]”=H, “at[atc]”=I, “aa[ag]”=K, “ct[agtc]|tt[ag]”=L, “atg”=M, “aa[tc]”=N, “cc[atgc]”=P, “ca[ag]”=Q, “cg[agct]|ag[ag]”=R, “tc[agct]|ag[ct]”=S, “ac[agct]”=T, “gt[agct]”=V, “tgg”=W and “ta[tc]”=Y. Additionally, use “ta[ga]|tga”=*stop* and “atg”=*start*.

Note that “atg” codes both for M and *start*. Please report only *maximal* ORFs, that is, ORFs that cannot be extended to the right without introducing an in-frame stop codon. In such an ORF, interpret internal “atg” codons as coding for M.

First run your program on the file `test.fna` and compare your output with the sample output contained in file `test.txt`. Then download the complete genome DNA sequence of *Escherichia coli 536* (accession NC 008253) and apply your program using a minimum ORF length of 100.

### 2 Comparison to NCBI (1 point)

BLAST 3 of your predicted ORFs against known protein sequences for *Escherichia coli* using the NCBI website, to see whether the ORFs and/or their function are known. Are any of the matches exact?

### 3 Run Genscan on annotated sequence (2 points)

Please view the NCBI entry of the “nucleotide sequence of the human ornithine decarboxylase gene” with accession number X16277. Refer to the gene described there as the *actual gene*. Run the complete DNA sequence through Genscan using <http://genes.mit.edu/GENSCAN.html>. Refer to Genscan’s result as the *predicted gene*. Please compare the exons of the actual and predicted gene, reporting the number of *true positive*, *true negative*, *false positive* and *false negative* classifications of individual bases, and also of whole exons.

Do the same for the NCBI accession J03733, which contains the mouse ornithine decarboxylase gene.

### 4 Homology-assisted gene finding (2 points)

To obtain a better prediction of the mouse ornithine decarboxylase gene present in the nucleotide sequence of accession J03733, use the homology-assisted gene finder *Genomscan* running at <http://genes.mit.edu/genomscan.html>.

To do so, paste in the DNA sequence of J03733 as target sequence and the *protein* sequence reported in X16277 as the protein homolog sequence.

Compare the output of this program with the actual gene structure reported in J03733. Can you explain the improvement?

Assignments due: **Monday, April 30, 10h**