

# Algorithms in Bioinformatics II, SS2004

## Assignment sheet # 3

Daniel Huson

May 3, 2004

### 1 ORF finder (6 points)

Please download [www-ab.informatik.uni-tuebingen.de/teaching/ss04/abi2/java/assign03.zip](http://www-ab.informatik.uni-tuebingen.de/teaching/ss04/abi2/java/assign03.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 left 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 *Nanoarchaeum equitans Kin4-M* (accession NC 005213) 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 *Nanoarchaeum equitans* using the NCBI website, to see whether the ORFs and/or their function are known. Are any of the matches exact?

### 3 E-values for ORF predictions (3 points)

Assume we are hunting for ORFs in a given DNA sequence of length  $N$ . Please derive a formula for an *E-value* associated with ORF predictions: For an ORF of a given length  $n$  (number of codons, including start and stop codons), this should be the *number of* ORFs of length  $\geq n$  that we would *expect* to find in a *random* DNA sequence of length  $N$ . (Your formula should depend only on  $n$  and  $N$  but not, e.g., on the actual amino acids of the ORF.)

What values do you get for  $N = 400kb$  and  $n = 20, 50, 100, 300$ ?

Assignments due: **Monday, May 10, 10am**