TEL AVIV UNIVERSITY

אוניברסיטת תל-אביב

RFM$_{app}$ User Guide

1 Introduction 3
2 RFM Brief Description.. 4
3 Supported Organisms 5
3.1 Model Eukaryotes.. 5
3.2 Model Prokaryotes 5
4 Input 6
4.1 Initiation Rate 7
4.2 Organism Coding Sequence.. 8
4.3 Organism Codon Translation Times 9
4.4 Chunk Size.. 10
4.5 Output Folder 10
4.6 Graphical Output 10
5 Submission.. 10
6 Output 12
6.1 Text File.. 12
6.2 Graphical predicted RD Profile 14
7 Command Line Version 15
7.1 Command Line Execution Instructions 15
8 References... 16

1 Introduction

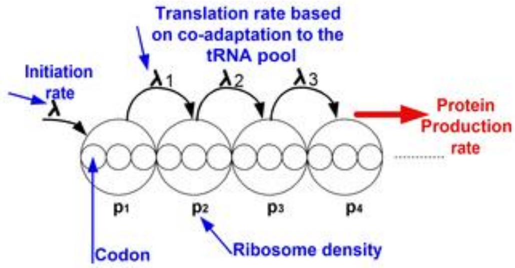
Ribosome Flow Model Application (RFMapp), is a graphical user interface application, enabling the prediction of fundamental features of the gene translation process, including translation rates, ribosomal densities, protein abundance levels, and the relation between all these variables, based on the approach presented in [1]. RFM affords the first large scale analysis of gene translation based on a model taking into account the physical and dynamical nature of this process. Considering the stochastic nature of the translation process, the RFM model is aimed at capturing the effect of codon order and composition on translation rates.
The RFM is a simple, physically plausible computational model that is *solely based on the coding sequence* (*i.e.* a vector of codons in each gene). The model allows for a computationally efficient analysis of the translation process on a genome-wide scale and across all species. Focusing on the coding sequence, we by no means wish to imply that it is the only factor taking place in the determination of translation rates. Nevertheless, since it has been widely recognized as a prime factor in the translation elongation process, we chose to concentrate on it. The RFM is aimed at capturing the effect of codon order on translation rates, the stochastic nature of the translation process and the interactions between ribosomes.
One advantage of the *RFM* is its amenability to *both* analytical and numerical analysis. In particular one can study ribosome density profiles and protein production rates from the equilibrium dynamics of the translation process (see the methods section of [1]).

RFMapp is available as a distributable cross platform Java application, requiring only a JVM http://www.java.com/en/download/index.jsp. Installations are available for Windows, Linux and generic Unix. In addition, we provide a Windows installation including a JVM. We acknowledge using the Java ODE library of Dr Michael Thomas Flanagan at http://www.ee.ucl.ac.uk/~mflanaga. A command line version of RFMapp is also available, and its execution is described in section 7.

The usage of RFMapp is straightforward, starting with the submission of genomic data for a certain organism: the input is genes' coding sequences, codon translation times, initiation rate (that can either be global/identical for all the organism's genes, or specific per gene) and number of approximated translation sites (chunk size), for a certain organism. The output is a text file of the predicted translation features, and an optional graphical predicted ribosomal density profile. RFMapp provides the genomic data (coding sequences and codon translation times) for six model organisms, and supports both organisms across all the domains of life, and synthetic simulations.

2 RFM Brief Description

The RFM has two free parameters: the initiation rate λ and the number of codons C at each 'site' (proportional to the size of the ribosome). Each site has a corresponding transition rate λ_i that is estimated based on the co-adaptation between the codons of the site and the tRNA pool of the organism (see the methods section of [1]). The output of the model consists of the steady state occupancy probabilities of ribosomes at each site and the steady state translation rates, or ribosome flow through the system.



The input to the RFMapp includes a coding sequence, estimated translation time for each codon, initiation rate (λ), and chunk size (*C*). The mRNA molecules (coding sequences) are coarse-grained into sites of codons according to the chunk size and the coding sequence (in the figure, C = 3). Ribosomes arrive at the first site with a global initiation rate λ but are only able to bind if this site is not occupied by another ribosome. The rate of each chunk is computed by the RFMapp based on the translation times of the organism's codons and the codon composition of the chunk. In this application, the transition time of a codon can be determined by the the tRNA pool of the organism. Briefly, taking into account the affinity between tRNA species and codons, we assume that the translation rate of a codon is proportional to the abundance of the tRNA species that recognize it [1] (see also the next section). According to the RFM, a ribosome that occupies the *i-th* site moves, with rate λ_i, to the consecutive site provided the latter is not occupied by another ribosome. Denoting the probability that the *i-th* site is occupied at time *t* by $P_i(t)$, it follows that the rate of ribosome flow into/out of the system is given by: $\lambda(1-P_1(t))$ and $\lambda_n P_n(t)$ respectively. The rate of ribosome 'flow' from site *i* to site *i+1* is given by: $\lambda_i P_i(t)(1-P_{i+1}(t))$. RFM finds the steady state solution of the equations below and specifically the rate of protein production at steady state.

$$\begin{cases} \dfrac{dp_1(t)}{dt} = \lambda[1-p_1(t)] - \lambda_1 p_1(t)[1-p_2(t)] \\ \dfrac{dp_i(t)}{dt} = \lambda_{i-1} p_{i-1}(t)[1-p_i(t)] - \lambda_i p_i(t)[1-p_{i+1}(t)] \; 1<i<n \\ \dfrac{dp_n(t)}{dt} = \lambda_{n-1} p_{n-1}[1-p_n(t)] - \lambda_n p_n(t) \end{cases}$$

3 Supported Organisms

The RFMapp supports both organisms across all the three domains of life, and synthetic simulations. For the users' convenience we have included six model organisms' codon translation times to RFMapp.
The codon translation times were based on the tAI [2], which were adjusted o the RFM model in [1] as described below:
Let n_i be the number of tRNA isoacceptors recognizing codon *i*. Let *tCGNij* be the copy number of the *j*th tRNA that recognizes the *i*th codon, and let S_{ij} be the selective constraint on the efficiency of the codon-anticodon coupling. We define the *absolute adaptiveness*, W_i, for each codon *i* as:

$$W_i = \sum_{j=1}^{n_i} (1-S_{ij}) tCGN_{ij}$$

The S_{ij}-values can be organized in a vector (S-vector) as described in [2]; ; each component in this vector is related to one wobble nucleoside-nucleoside paring: I:U, G:U, G:C, I:C, U:A, I:A, etc.
From W_i we obtain p_i, which is the probability that a tRNA will be coupled to the codon:

$$p_i = \frac{W_i}{\sum_{j=1}^{61} tCGN_j}$$

The expected time on codon *i* $t_i = 1/p_i$.

3.1 Model Eukaryotes

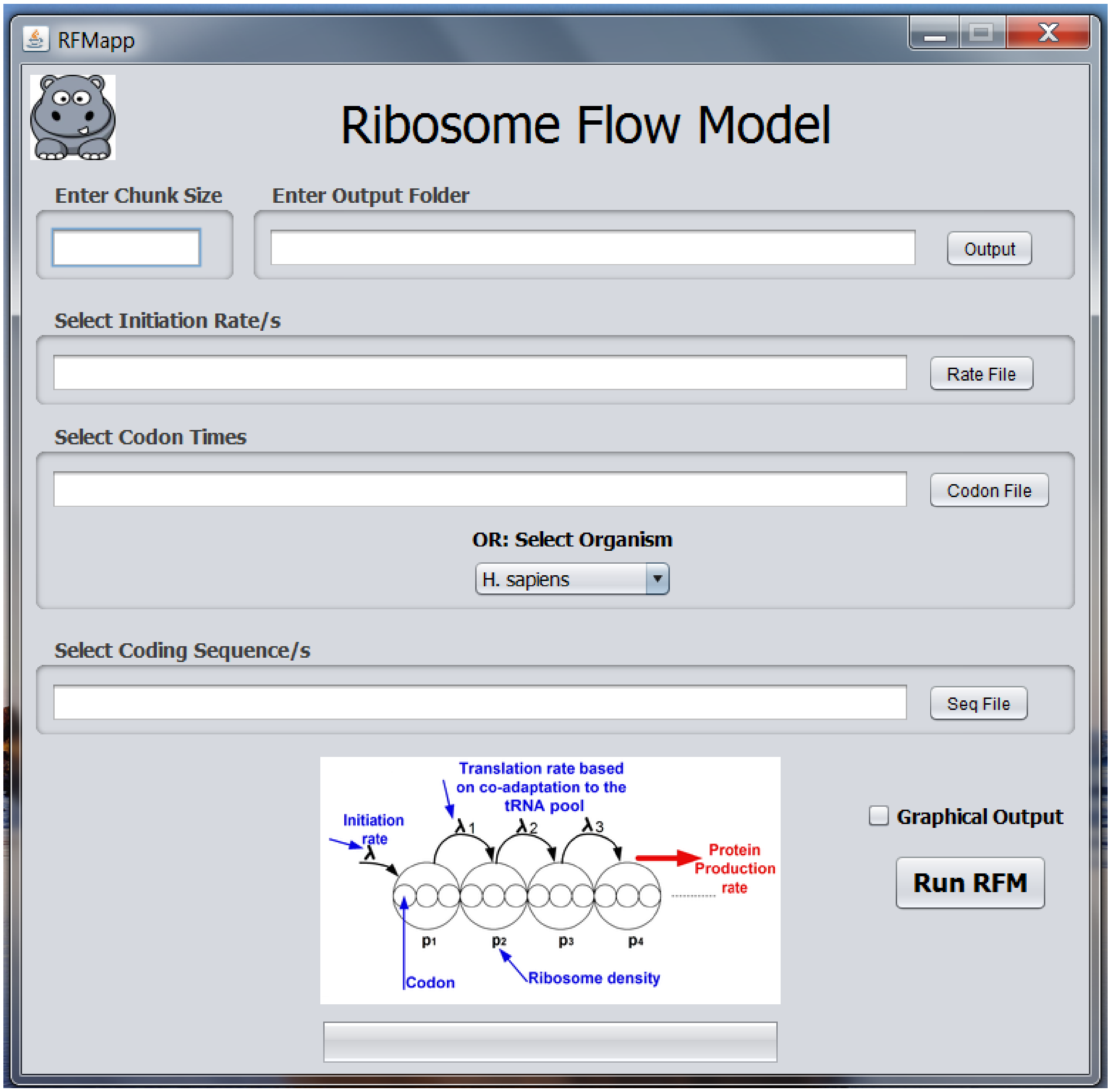
1. H. sapiens
2. S. cerevisiae
3. S. pombe
4. C. elegans
5. A. thaliana

3.2 Model Prokaryotes

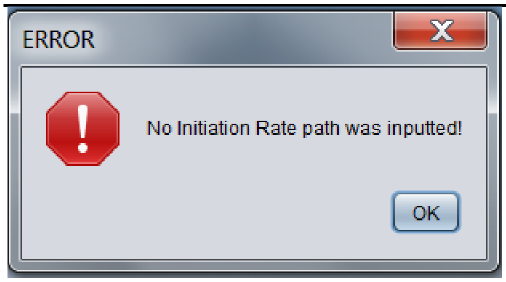
1. E. coli K12

4 Input

The RFM$_{app}$ GUI contains five input fields, a checkbox indicating if graphical output is required, one submission button, and a progress bar:



If any of the fields are not filled the user will receive the relevant error message, for example if the Initiation Rate is not supplied, the following error message will be displayed:



No Initiation Rate path was inputted!

4.1 Initiation Rate

A global/identical Initiation Rate can be assumed for all inputted genes, or specific per gene rates can be provided.
For the specific option, a two column tab delimited file containing the gene name in the first column, and the initiation rate in the second, for example:

YAL001C	0.06
GENE_NAME2	0.05
GENE_NAME3	0.07

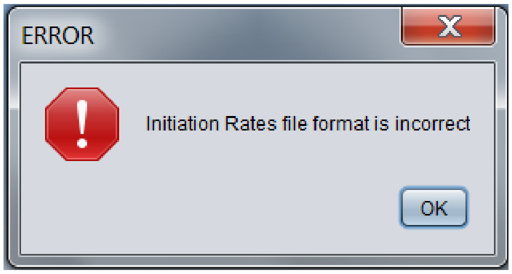
.
.
.

In this mode, every gene in the coding sequence FASTA file must have a corresponding initiation rate. If this is not maintained, genes with missing initiation rates will not be processed (details in the Output section).

For the global option a single tab delimited row containing two columns, in the first any single string (such as Global), and in the second the global initiate rate, for example:

```
GLOBAL_RATE	0.06
```

If the file is in an incorrect format, the analysis will not be performed, and the following Error Message will be displayed:



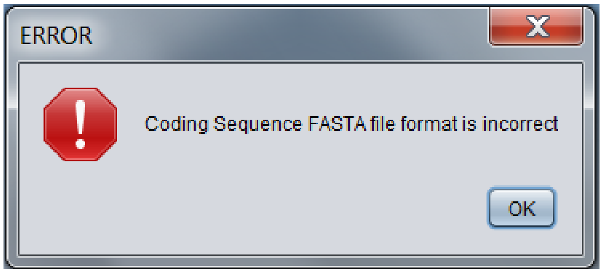
Initiation Rates file format is incorrect

4.2 Organism Coding Sequence

A FASTA format file, for example:

```
> YAL001C
ATGGTACTGACGATTTATCCTGACGAACTCGTACAAATAGTGTCTGATAAAATTGCTTCAAATAAGGGAAAAATCACTTTGAATCAGCTGTGGGATATATCTG
GTAAATATTTTGATTTGTCTGATAAAAAAGTTAAACAGTTCGTGCTTTCATGCGTGATATTGAAAAAGGACATTGAGGTGTATTGTGATGGTGCTATAACAAC
TAAAAATGTGACTGATATTATAGGCGACGCTAATCATTCATACTCGGTTGGGATTACTGAGGACAGCCTATGGACATTATTAACGGGATACACAAAAAAGGA
GTCAACTATTGGAAATTCTGCATTTGAACTACTTCTCGAAGTTGCCAAATCAGGAGAAAAAGGGATCAATACTATGGATTTGGCGCAGGTAACTGGGCAAGA
TCCTAGAAGTGTGACTGGACGTATCAAGAAAATAAACCACCTGTTAACAAGTTCACAACTGATTTATAAGGGACACGTCGTGAAGCAATTGAAGCTAAAAAA
ATTCAGCCATGACGGGGTGGATAGTAATCCCTATATTAATATTAGGGATCATTTAGCAACAATAGTTGAGGTGGTAAAACGATCAAAAAATGGTATTCGCCA
GATAATTGATTTAAAGCGTGAATTGAAATTTGACAAAGAGAAAAGACTTTCTAAAGCTTTTATTGCAGCTATTGCATGGTTAGATGAAAAGGAGTACTTAAAG
AAAGTGCTTGTAGTATCACCCAAGAATCCTGCCATTAAAATCAGATGTGTAAAATACGTGAAAGATATTCCAGACTCTAAAGGCTCGCCTTCATTTGAGTATG
ATAGCAATAGCGCGGATGAAGATTCTGTATCAGATAGCAAGGCAGCTTTCGAAGATGAAGACTTAGTCGAAGGTTTAGATAATTTCAATGCGACTGATTTAT
TACAAAATCAAGGCCTTGTTATGGAAGAGAAAGAGGATGCTGTAAAGAATGAAGTTCTTCTTAATCGATTTTATCCACTTCAAAATCAGACTTATGACATTGC
AGATAAGTCTGGCCTTAAAGGAATTTCAACTATGGATGTTGTAAATCGAATTACCGGAAAAGAATTTCAGCGAGCTTTTACCAAATCAAGCGAATATTATTTA
GAAAGTGTGGATAAGCAAAAAGAAAATACAGGGGGGTATAGGCTTTTTCGCATATACGATTTTGAGGGAAAGAAGAAGTTTTTTAGGCTGTTCACAGCTCAG
AACTTTCAAAAGTTAACAAATGCGGAAGACGAAATATCCGTTCCAAAAGGGTTTGATGAGCTAGGCAAATCTCGTACCGATTTGAAAACTCTCAACGAGGAT
AATTTCGTCGCACTCAACAACACTGTTAGATTTACAACGGACAGCGATGGACAGGATATATTCTTCTGGCACGGTGAATTAAAAATTCCCCCAAACTCAAAAA
AAACTCCGAATAAAAACAAACGGAAGAGGCAGGTTAAAAACAGTACTAATGCTTCTGTTGCAGGAAACATTTCGAATCCCAAAAGGATTAAGCTAGAGCAGC
ATGTCAGCACTGCACAGGAGCCGAAATCTGCTGAAGATAGTCCAAGTTCAAACGGAGGCACTGTTGTCAAAGGCAAGGTGGTTAACTTCGGCGGCTTTTCT
GCCCGCTCTTTGCGTTCACTACAGAGACAGAGAGCCATTTTGAAAGTTATGAATACGATTGGTGGGGTAGCATACCTGAGAGAACAATTTTACGAAAGCGTT
TCTAAATATATGGGCTCCACAACGACATTAGATAAAAAGACTGTCCGTGGTGATGTTGATTTGATGGTAGAAAGCGAAAAATTAGGAGCCAGAACAGAGCCT
GTATCAGGAAGAAAAATTATTTTTTTGCCCACTGTTGGAGAGGACGCTATCCAAAGGTACATCCTGAAAGAAAAAGATAGTAAAAAAGCAACCTTTACTGAT
GTTATACATGATACGGAAATATACTTCTTTGACCAAACGGAAAAAAATAGGTTTCACAGAGGAAAGAAATCAGTTGAAAGAATTCGTAAGTTTCAGAACCGCC
AAAAGAATGCTAAGATCAAAGCTTCAGATGACGCTATCTCTAAGAAGAGTACGTCGGTCAACGTATCAGATGGAAAGATCAAAAGGAGAGACAAAAAAGTG
TCTGCTGGTAGGACAACGGTGGTCGTGGAAAATACTAAAGAAGACAAAACTGTCTATCATGCAGGCACTAAAGATGGTGTTCAGGCTTTAATCAGAGCTGTT
GTAGTTACTAAAAGTATTAAAAATGAAATAATGTGGGACAAAATAACAAAATTATTTCCTAATAATTCTTTAGATAACCTAAAAAAGAAATGGACGGCACGGC
GAGTAAGAATGGGTCATAGTGGTTGGAGGGCATATGTCGATAAGTGGAAAAAAATGCTCGTTCTAGCCATTAAAAGTGAAAAGATTTCACTGAGGGATGTT
GAAGAACTAGATCTTATCAAATTGCTTGATATTTGGACCTCTTTTGATGAAAAGGAAATAAAAAGGCCGCTCTTTCTTTATAAGAACTACGAAGAGAATAGAA
AAAAATTTACTCTGGTACGTGATGACACACTTACACATTCTGGCAACGATCTGGCCATGTCTTCTATGATTCAAAGAGAGATCTCTTCTTTAAAAAAAACTTAC
ACTAGAAAGATTTCCGCTTCTACTAAGGACTTATCGAAGAGTCAAAGCGACGATTATATTCGCACAGTGATCCGGTCCATATTAATAGAAAGTCCTTCGACCA
CTAGAAATGAAATAGAGGCGTTGAAGAACGTTGGAAACGAATCAATAGATAACGTCATCATGGATATGGCTAAGGAAAAGCAAATTTATCTCCATGGCTCAA
AACTTGAATGTACTGATACTTTACCAGACATTTTGGAAAATAGAGGAAATTATAAAGATTTTGGTGTAGCTTTTCAGTATAGATGTAAGGTTAATGAATTATTG
GAGGCCGGAAACGCTATTGTTATCAATCAAGAGCCGTCCGATATATCCTCTTGGGTTTTAATTGATTTGATTTCGGGAGAGCTATTGAATATGGATGTAATTC
CAATGGTGAGAAATGTTCGACCTTTAACGTATACTTCAAGGAGATTTGAAATACGAACATTAACTCCCCCTCTGATTATATATGCCAATTCTCAGACAAAATTG
AATACAGCAAGGAAGTCTGCTGTCAAAGTTCCACTGGGCAAACCATTTTCTCGTTTATGGGTGAATGGATCTGGTTCCATTAGGCCAAACATATGGAAGCAG
GTAGTTACTATGGTCGTTAACGAAATAATATTTCATCCAGGGATAACATTGAGTAGATTGCAATCTAGGTGTCGTGAAGTACTTTCGCTTCATGAAATATCAG
AAATATGCAAATGGCTCCTAGAAAGACAAGTATTAATAACTACTGATTTTGATGGCTATTGGGTCAATCATAATTGGTATTCTATATATGAATCTACATAA
.
.
.
```

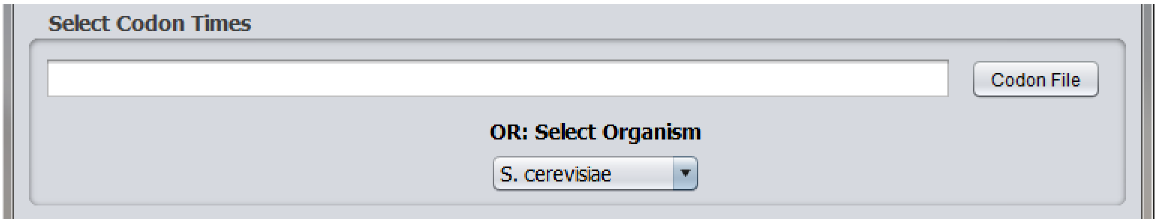
The coding sequence must terminate with a stop codon, and contain no stop codons along the sequence. Irregular genes which do not conform to those requirements will not be processed (details in the Output section). If the FASTA text file is not in the correct format, the analysis will not be performed, and the following Error Message will be displayed:



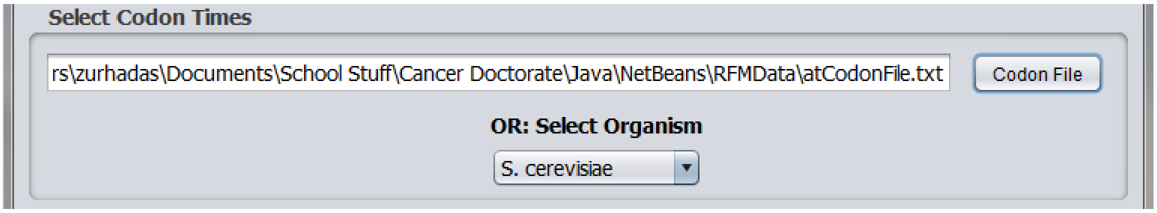
Coding Sequence FASTA file format is incorrect

4.3 Organism Codon Translation Times

Either submit your own codon translation times file, in the format described below; or select one of the six model organisms. If no codon translation times file is submitted, whichever organism is selected in the organism list box will be used:



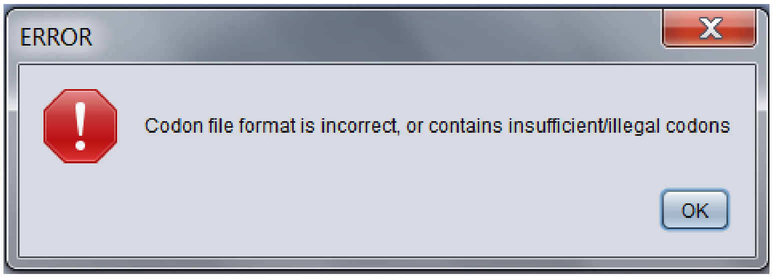
If a file path is provided, then the organism selected in the list box is ignored:


rs\zurhadas\Documents\School Stuff\Cancer Doctorate\Java\NetBeans\RFMData\atCodonFile.txt

The user submitted codon translation times file is a two column tab delimited file containing the codon in the first column, and the translation time in the second, for example:

```
AAA 39
AAC 27.3
AAG 16.810345
.
.
.
```

The codon file must contain the 61 coding codons (excluding stop codons). If the 61 submitted codons are non-standard, or the file is in incorrect format, the analysis will not be performed, and the following Error Message will be displayed:


Codon file format is incorrect, or contains insufficient/illegal codons

4.4 Chunk Size

As aforementioned mRNA molecules (coding sequence) are coarse-grained into sites of *C* codons (chunk size), the number of codons (*C*) are supplied by the user. The chunk size should be a whole number. If a whole number is not supplied it will be rounded to the nearest integer. The number of codons per chunk (*C*) should be smaller than the number of codons in the coding sequence of a gene, if this is not maintained, that gene will not be processed (details in the Output section). According to the RFM model the chunk size must be larger than 1.

4.5 Output Folder

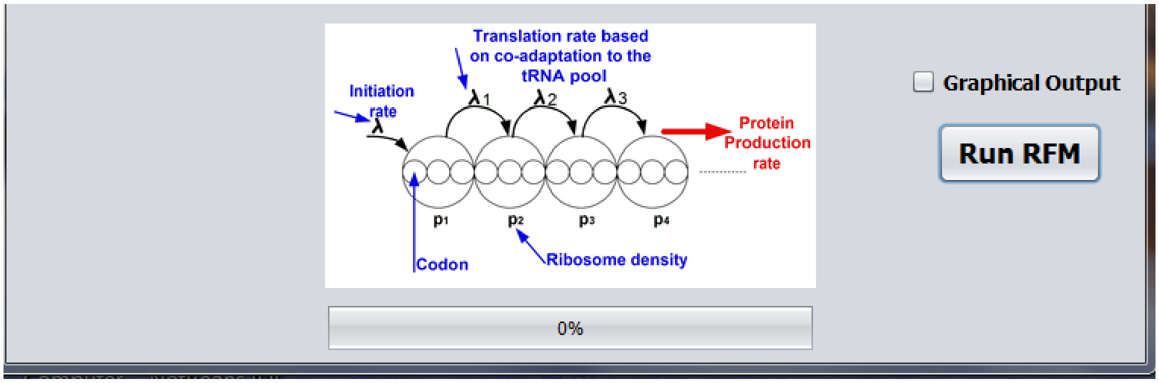
The user must specify an output folder with appropriate write permissions.

4.6 Graphical Output

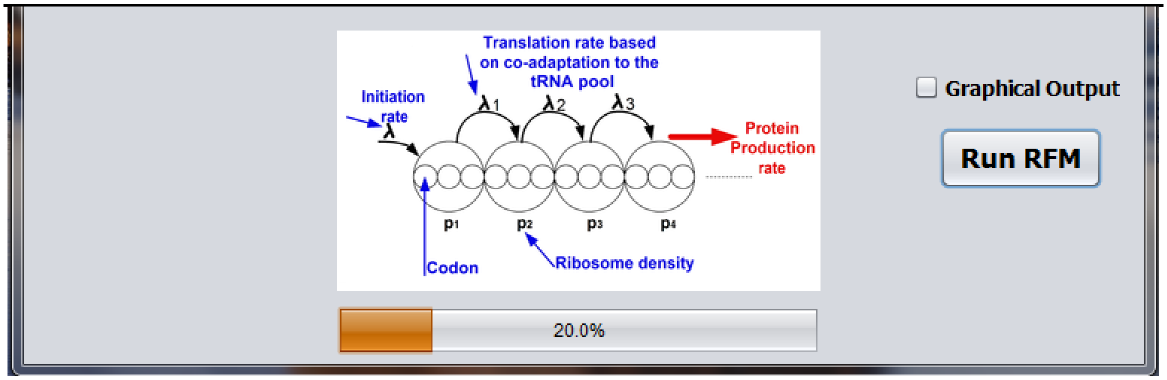
If the Graphical Output checkbox is checked, a graphical predicted ribosomal density profile of the translation chunks will be created per gene. This increases the computational time, and when analyzing large-scale quantities of genes may be superfluous, as an image is created per gene.

5 Submission

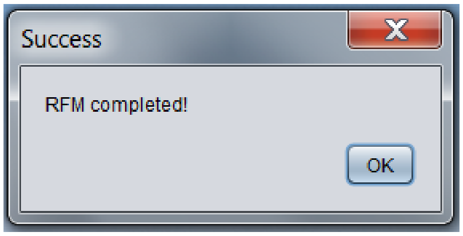
Following the pressing of the 'Run RFM' button, the progress bar will indicate the beginning of the analysis by the appearance of 0%:



It may take several seconds for the progress bar to calculate the analysis time, following which it will begin indicating the percentage of the analysis completed:



When running whole genomes analysis times may span several hours. The progress calculation is per gene, thus it is not evenly distributed time-wise; longer genes will take longer to process. On completion the following message will appear:



RFM completed!

The output described in the following section has then been created.

6 Output

6.1 Text File

RFMapp enables the prediction of fundamental features of the gene translation process, namely translation rates and ribosomal densities (occupation probabilities). Protein abundance levels are proportional to the multiplication of the predicted translation rates and mRNA levels, and can be extrapolated if mRNA levels are available.

Following data submission, upon successful completion of the requested analysis, a text output file described below, is created in the specified output folder. The initiation rate inputted for gene YAL001C was 0.06, and the chunk size 25.

```
    > YAL001C
    Translation Rate = 9.0E-5
    Occupation probabilities = 0.998492   0.819088     0.777816     0.823769
        0.828705     0.780669     0.795613     0.771546     0.612972
        0.759309     0.75751      0.817203     0.854445     0.81854
        0.479676     0.239879     0.155567     0.160358     0.151535
        0.170784     0.156234     0.156535     0.154398     0.121385
        0.168602     0.163115     0.135808     0.132196     0.151252
        0.108624     0.246533     0.396007     0.291048     0.283705
        0.215123     0.09482      0.238309     0.131557     0.212368
        0.142124     0.227242     0.48151      0.150279     0.163151
        0.336536     0.163557     0.05661
    > GENE_NAME2
    Translation Rate = 1.14E-4
    Occupation probabilities = 0.997722   0.81348      0.640365     0.744531
        0.706568     0.732177     0.716249     0.756527     0.804936
        0.683814     0.810202     0.701883     0.693049     0.763732
        0.790828     0.681179     0.741116     0.678163     0.723075
        0.689518     0.672252     0.805379     0.778826     0.725289
        0.640299     0.773168     0.716706     0.658815     0.587155
        0.459353     0.649751     0.712617     0.650337     0.746821
        0.592343     0.340243     0.461852     0.317132     0.294183
        0.372651     0.204811     0.266831     0.220449     0.294788
        0.330145     0.31496      0.377669     0.439881     0.406272
        0.380543     0.164954
.
.
.
```

The following are mandatory conditions, which when not maintained will prohibit processing and the respective Error Message will be displayed:

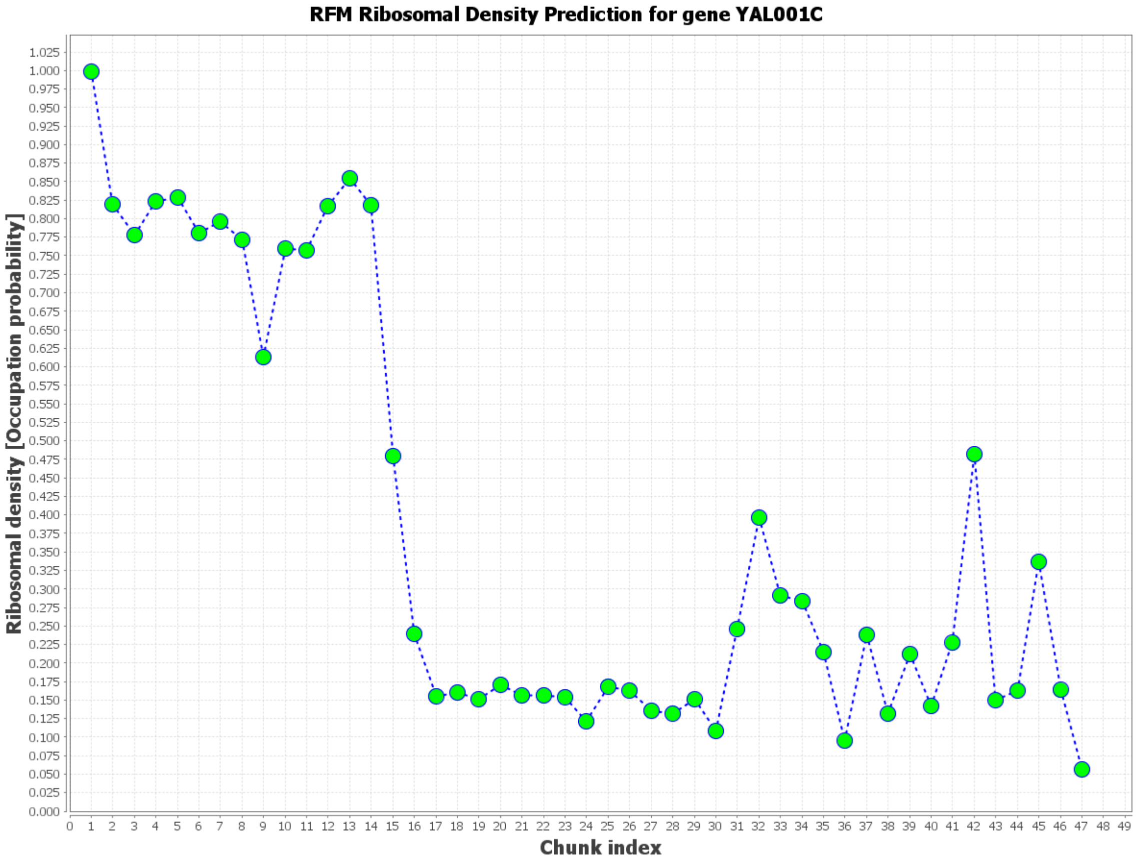
1. The coding sequence must terminate with a stop codon:

```
  ERROR: Last codon of coding sequence is not a Stop codon
```

2. The coding sequence must contain no stop codons along the sequence:
   ```
   ERROR: The coding sequence contains a Stop codon
   ```
3. The coding sequence must not contain illegal codons:
   ```
   ERROR: The coding sequence contains an illegal codon
   ```
4. The coding sequence (in codons) must be longer than the Chunk Size:
   ```
   ERROR: The Chunk Size is longer than the coding sequence
   ```
5. The coding sequence length is not divisible by three, indicating is not composed of codon nucleotide triplets:
   ```
   ERROR: The coding sequence is not composed of codon nucleotide
   Triplets
   ```
6. When the gene specific Initiation Rate option is exercised, and a gene specified in the coding sequence FASTA file does not have a corresponding initiation rate in the Initiation Rates file:
   ```
   ERROR: Initiation Rate was not provided
   ```
7. There is a problem with the coding sequence, possibly the length excluding the stop codon is for example 25 codons, if one uses a chunk size of 25, this creates an illegal chunk size 1:
   ```
   ERROR: The chunk size is possibly 1, or there is another
   problem with the coding sequence
   ```
8. An unknown error has occurred:
   ```
   ERROR: An unknown error occurred
   ```

6.2 Graphical predicted RD Profile

An optional graphical predicted ribosomal density profile of the translation chunks can be created per gene. An example predicted RD profile for gene YAL001C, based on initiation rate 0.06, and chunk size 25:



7 Command Line Version

7.1 Command Line Execution Instructions

To run the command line version of the RFMapp, supply the following 7 arguments in the following order:

1. Codon times file: either a path to your own codon translation times file, or the string 'empty' and then specify the required organism number in parameter 7.

2. Coding sequence/s path

3. Chunk size

4. Initiation rate/s path

5. Output path

6. 1 if graphical output is required, and 0 if it is not

7. Specify an organism number if wanting to use internal codon times, specify 0 if you supplied you own codon translation file path in parameter 1.
Organism numbers:
1: H. sapiens
2: S. cerevisiae
3: S. pombe
4: C. elegans
5: A. thaliana
6: E. coli

8 References

1. Reuveni, S., et al., *Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model.* PLoS computational biology, 2011. **7**(9): p. e1002127.
2. dos Reis, M., R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection.* Nucleic acids research, 2004. **32**(17): p. 5036-5044.