DETECTION OF FRAMESHIFTS AND IMPROVING GENOME ANNOTATION

A Thesis
Presented to
The Academic Faculty

by

Ivan Valentinovich Antonov

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computational Science and Engineering

Georgia Institute of Technology
December 2012

DETECTION OF FRAMESHIFTS AND IMPROVING GENOME ANNOTATION

Approved by:

Professor Mark Borodovsky, Advisor
School of Computational Science and Engineering
Georgia Institute of Technology

Professor Kostas T. Konstantinidis
School of Civil and Environment Engineering
Georgia Institute of Technology

Professor Brian Hammer
School of Biology
Georgia Institute of Technology

Professor Le Song
School of Computational Science and Engineering
Georgia Institute of Technology

Professor King I. Jordan
School of Biology
Georgia Institute of Technology

Professor Pavel Baranov
Biochemistry Department
University College Cork, Ireland

Date Approved: December 2012

To my family back in Russia

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Borodovsky for providing me the opportunity to study Bioinformatics at the Georgia Institute of Technology and for his guidance and motivation. I would like to thank Dr. Pavel Baranov and Arthur Coakley for conducting the experiments for the selected programmed frameshift candidates. I would like to thank Alex Lomsadze for many useful discussions.

I would like to thank all my soccer friends with whom I played for the last two years. These games gave me a diversion from the challenging journey of the PhD.

I am grateful to the members of my committee, Dr. Pavel Baranov, Dr. Brian Hammer, Dr. King Jordan, Dr. Kostas Konstantinidis and Dr. Le Song, for their time and effort reviewing this thesis.

Contents

DEDICATION . . . . . **iii**

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **xii**

GLOSSARY . . . . . **xix**

SUMMARY . . . . . **xx**

I INTRODUCTION **1**

1.1 Sequencing errors 2

1.2 Indel mutations 2

1.3 Programmed frameshifting – an example of recoding 3

1.3.1 Transposable elements: Insertion Sequences (IS) and retrotransposons 5

1.3.2 Bacterial *prfB* gene encoding Release Factor 2 6

1.3.3 Bacterial *dnaX* gene encoding DNA polymerase III subunits τ and γ 8

1.3.4 Eukaryotic ornithine decarboxylase antizyme 10

1.3.5 Other examples of programmed frameshifting in prokaryotes 11

1.3.6 Other examples of programmed frameshifting in eukaryotes . 11

1.3.7 Mechanisms of programmed ribosomal frameshifting (PRF) . 13

1.3.8 Mechanism of programmed transcriptional realignment (PTR) 18

1.3.9 The biological purpose of programmed frameshifting 20

1.4 Translational coupling in prokaryotic mRNAs 22

1.5 Phase variation in prokaryotic genomes 22

1.6 Frame shifting alternative splicing in eukaryotes 23

1.7 Existing approaches for frameshift identification 24

1.7.1 Similarity search based programs 25

1.7.2 *Ab initio* frameshift prediction programs 25
1.7.3 Programs for finding programmed frameshifts 26

II GENETACK: FRAMESHIFT IDENTIFICATION IN PROTEIN CODING SEQUENCES BY THE VITERBI ALGORITHM . . . 29

2.1 GeneTack algorithm 29
2.2 GeneTack-GM algorithm 32
2.2.1 Parameter estimation 35
2.2.2 High-GC genomes 37
2.3 Results 40
2.3.1 Datasets 40
2.3.2 GeneTack-GM performance: comparison with other programs 40
2.4 Discussion 44
2.4.1 Can GeneTack predict programmed frameshifts? 44
2.4.2 Insensitivity zones 44
2.4.3 Filter effectiveness 46

III IDENTIFYING THE NATURE OF READING FRAME TRANSITIONS OBSERVED IN PROKARYOTIC GENOMES 49

3.1 Introduction 49
3.2 Results 49
3.2.1 The set of frameshifts predicted in 1,106 genomes 49
3.2.2 About 50% of fs-genes could be clustered 51
3.2.3 Clusters identified as programmed frameshift clusters 53
3.2.4 Genes with known programmed frameshifts 53
3.2.5 New genes that may utilize programmed frameshifting 63
3.2.6 Other large clusters of fs-genes 66
3.2.7 Pseudogene clusters 70
3.2.8 Singletons: authentic indel mutations or sequencing errors? . 72
3.2.9 Distribution of relative frameshift coordinates in fs-genes . . 73
3.3 Materials and Methods 76

3.3.1 Translation of predicted fs-genes; BLASTp and Pfam confirmations . . . 76
3.3.2 Ribosome binding site (RBS) of the downstream ORF . . . 77
3.3.3 Clustering . . . 78
3.3.4 Functional characterization of the GeneTack clusters . . . 79
3.3.5 Identification of clusters of fs-genes with non-standard mechanisms of transcription and translation . . . 79
3.3.6 A new measure of motif periodicity . . . 83
3.3.7 Inferring a type of frameshifting mechanism . . . 85
3.3.8 Clusters of sequences with frame transitions determined by phase variation and translational coupling . . . 86
3.3.9 Experimental verification of predicted programmed frameshifting . . . 88
3.4 Discussion . . . 89

IV COMPARATIVE GENOMICS ANALYSIS OF EUKARYOTIC MRNAS WITH FRAMESHIFTS . . . 91

4.1 Introduction . . . 91
4.2 Materials and methods . . . 93
4.2.1 Sequence data . . . 93
4.2.2 HMM structure and parameters . . . 94
4.2.3 Preparation of the test set with artificial frameshifts . . . 96
4.2.4 Filters . . . 96
4.2.5 Clustering . . . 96
4.2.6 Exon mapping . . . 97
4.3 Results . . . 97
4.3.1 Testing quality of frameshift prediction in eukaryotic mRNAs 97
4.3.2 Predicting frameshifts in the eukaryotic mRNAs . . . 98
4.3.3 Rediscovery of known programmed frameshifting events . . . 98
4.3.4 Frame shifting alternative splicing isoforms and indel mutations 99
4.3.5 Dual-coding mRNA sequences . . . 102

4.4 Discussion 106

V GENETACK DATABASE: GENES WITH FRAMESHIFTS IN PROKARYOTIC GENOMES AND EUKARYOTIC MRNA SEQUENCES 108

5.1 Introduction 108

5.2 Database statistics and usage 111

5.3 Tools for frameshift prediction 114

5.4 Application of the tools and database 116

5.5 Availability 116

REFERENCES 117

VITA 133

List of Tables

1 Types of states used in the GeneTack HMM and the properties of the emission probabilities. 31

2 An example of emission probabilities calculation for overlap of genes carrying the genetic code in frames 1 and 2, (for the *1-2* hidden state). The pattern of frequencies (F_1, F_{12}, F_2) repeated for the whole sequence carrying overlapping genes is shown in bold font. 34

3 Frameshift prediction accuracy estimation for 17 prokaryotic genomes (sorted by GC content) each containing 400 genes longer than 1000 nt with simulated frameshifts (dataset_1000). The Sn and Sp values were calculated for GeneTack-GM, FrameD, FSFind and FSFind-BLAST programs. The programs were compared based on average sensitivity and specificity $(Sn + Sp)/2$. Bold numbers indicate the best performance. *FSFind-BLAST results for *R. solanacearum* were not available because of a runtime error, thus the average values were computed for 16 genomes. 42

4 Frameshift prediction accuracy estimation for 18 prokaryotic genomes (sorted by GC content) each containing 400 genes of length between 600 and 1000 nt with simulated frameshifts (dataset_600_1000). The programs were compared based on average sensitivity and specificity (Sn+Sp)/2. Bold numbers indicate the best performance. *FSFind-BLAST results for *R. solanacearum* were not available because of a runtime error, thus the average values were computed for 17 genomes. 43

5 Examples of known frameshift prone patterns: PRF - programmed ribosomal frameshifting, PTR - programmed transcriptional realignment. *DNA alphabet with standard symbols for ambiguous bases is used for convenience. Note that the table gives just a few examples of frameshift sites and not all genes are listed. 52

6 Correspondance between GeneTack clusters and clusters from Sharma et al. established based on BLASTn search (e-value threshold 10^{-20}). 54

7 The largest GeneTack programmed frameshift clusters that correspond to known cases of programmed frameshifting. **Cluster ID** – unique identifier of a cluster; **Function** – expected gene function derived for the corresponding fs-proteins from Pfam domains and BLASTp hits against the NCBI nr database; **Size** – number of fs-genes in the cluster (FS), number of different genera (G); **D** – frameshift direction (+1 or -1); **BR** – possible biological role (PTR – programmed transcriptional realignment, PRF – programmed ribosomal frameshifting, TC – translational coupling); **FS coord** – median value of the relative frameshift coordinate for all frameshifts in the cluster; **SD** – standard deviation of the relative frameshift coordinate; **Heptamer** – overrepresented heptamer (the fraction of the cluster's fs-genes that contain the heptamer is shown in parentheses), contrary to Table 5 we specify consensus sequence rather than regular expression pattern; **Frameshift site Logo** – Logo of the frameshift site (see text for details); **Sharma et al clusters** – ID(s) of the corresponding Sharma et al clusters. 58

8 Programmed frameshift clusters predicted by GeneTack that were selected for experimental verification. **Experimental results** – summary of the results shown on Fig. 18 and Fig. 19 (FS – number of tested fs-genes that showed ribosomal frameshifting; TC – number of tested fs-genes that showed translational coupling) 59

9 Inserts cloned in between GST and MBP genes that showed highest frameshifting efficiency. **FS %** – frameshifting efficiency detected in experiments. 62

10 The largest clusters containing 100 or more fs-genes. **Size** – number of fs-genes in the cluster, **#G** – number of different genera in the cluster; **D** – frameshift direction; **%AT** – fraction of fs-genes with 7+ nt poly-AT stretch located near predicted frameshift; **%R** – fraction of fs-genes with tandem repeats located near predicted frameshifts; **%S** – fraction of fs-genes with ORF2 start codon ATG ([1]GTG) located within 10nt ([2]20nt) from the ORF1 stop codon; **%B** – fraction of fs-proteins validated by BLASTp against NCBI nr database; **BR** - biological role (PF – programmed frameshifting, PV – phase variation, TC – translational coupling); **?** – putative prediction of biological role; *experimentally verified. 67

11 GeneTack clusters with members representing known cases of phase variation. **Query gene name (Organism)** – the name of a gene with known phase variation; **# hits** – number of fs-proteins found by BLASTp search (how many of them belong to clusters is specified in brackets); **Main cluster** – name of the fs-cluster with the largest number of hits; **Size** – size of the cluster (number of fs-proteins found by the BLASTp search is specified in brackets). 68

12 Features (the first column) used to classify predicted frameshifts into Types (the Type names are given in the top two rows). **H-pseudo** – hypothetical pseudogene; **n/r** – the feature is not required; **n/a** – the feature is not applicable; *a cluster must contain at least one annotated pseudogene; **>50% of cluster fs-genes must be validated by BLASTp; ***manual verification includes functional analysis of the fs-proteins and literature survey. 72

13 Heptamers with maximum score that were used to select the 7 A-rich motifs (plus the *prfB* motif) that could cause programmed frameshifting 82

14 Examples of known eukaryotic dual coding genes. **Genes** – pair of genes in two different frames that share the dual-coding region; **DC (nt)** – length of the dual-coding region (in nucleotides); **Ori** – origin of the dual coding region: AS (alternative splicing isoform), IG (internal gene); **Refs** – references. 92

15 GeneTack clusters that correspond to known cases of programmed frameshifting. **Cluster ID** – unique identifier of a cluster; **Name** – cluster name; **Size** – number of fs-genes in the cluster; **Species** – number of different genera; **D** – frameshift direction (+1 or -1); **FS Site** – frameshift site; **Ref** – references. 99

16 Largest clusters containing 10 or more fs-genes from at least 5 different species. **Cluster ID** – unique identifier of a cluster; **Name** – cluster name; **Size** – number of fs-genes in the cluster; **#S** – number of different species; **D** – frameshift direction; **%E** – fraction of cluster's fs-genes that have exon annotation; **EJC** – exon junction distance (average distance in nucleotides from the frameshift to the nearest exon-exon junction); **Type** – possible biological nature of the cluster (DC – cases of possible dual coding, AS – alternative splicing, Indel – indel mutation inside exon, AS & Indel – mixture of AS and Indel frameshifts, FP – GeneTack false positive) 101

17 Statistics on eukaryotic and prokaryotic sections of the GeneTack database 111

List of Figures

1 Diagram of RF2 frameshift site conservation, the height of symbols indicates conservation of nucleotides, while their weight shows the relative frequency of nucleotides at corresponding positions. The diagram was build using MEME from 428 sequences of RF2. 7

2 Regulatory feedback provided for RF2 biosynthesis by the frameshifting mechanism. The first ORF has a UGA stop codon. The regulation is autonomous and the level of RF2 biosynthesis depends on its own concentration. 8

3 Regulation of cellular polyamine levels using antizyme +1 frameshifting as a sensor. High polyamine levels stimulate +1 frameshifting required for the synthesis of functional antizyme 1 (AZ1). AZ1 binds ornithine decarboxylase (ODC) and triggers its degradation by the 26S proteasome, being itself recycled. As ODC catalyzes the first step of the polyamine biosynthesis pathway, its degradation leads to a decrease in polyamine levels, which in turn reduces frameshifting efficiency. . . . 9

4 Examples of PRF "frameshift signals". Each signal consists of frameshift site and two stimulatory sequences (stimulators). (A) "-1" programmed frameshift is utilized in *dnaX* gene to express two subunits of DNA polymerase III. The Logo for (A) was derived from aligned sequences from 9 genera (*Escherichia*, *Salmonella*, *Neisseria*, *Vibrio*, *Shigella*, *Citrobacter*, *Enterobacter*, *Yersinia*, *Serratia*). The frameshift signal consists of conserved frameshift pattern AAA_AAA_G ("slippery sequence") and two stimulators. The upstream stimulator is a Shine-Dalgarno like sequence that interacts with ribosome and the downstream stimulator makes a hairpin secondary structure. (B) "+1" programmed frameshift is utilized in *prfB* gene to auto regulate expression of Release Factor 2. The Logo for (B) was derived from aligned 413 sequences from 138 genera. The frameshift signal consists of conserved frameshift pattern CTT_TGA and two stimulators. The upstream stimulator is also a Shine-Dalgarno sequence while the downstream stimulator is represented by a single cytosine that forms the "weakest" termination context. 14

5 Model of Ty1 frameshift mechanism in *S. cerevisiae*. The three ribosomal tRNA-binding sites (E, P, and A) are diagrammed as dotted rectangles binding tRNAs cartooned as T's; the anticodon of each tRNA appears above the tRNA running 5'-3' from right to left. On the left, the P site is shown occupied by peptidyl-$tRNA^{Leu}_{UAG}$; the identity of the tRNA in the E site does not influence frameshifting and it is shown with XXX as anticodon. Two alternatives exist for the next step of elongation. Above, $tRNA^{Arg}_{CCU}$ is shown occupying the A site, leading to in-frame decoding; this reaction is shown as reversible because wobble mispairing in the P site appears to block cognate acceptance. Below, $tRNA^{Gly}_{GCC}$ is shown occupying the A site, also reversibly; its binding can lead to +1 frameshifting. Binding of this tRNA to the A site is shown requiring the middle A of the shifty heptamer to be excluded from the A site to allow the GGC anticodon to bind there, facilitating out-of-frame binding of $tRNA^{Gly}_{GCC}$. The number of tRNA shown corresponds to their relative concentration in the cell; $tRNA^{Gly}_{GCC}$ is present at approximately 16-times the concentration of $tRNA^{Arg}_{CCU}$. 17

6 Direct control of transcriptional realignment by competing NTP substrates. In this example, which uses the DNA sequence of the *E. coli pyrBI* initially transcribed region, the competing substrates are GTP and UTP. After synthesis of the AAUUU transcript, it can reversibly slip one base upstream due to a weak RNA-DNA hybrid. Addition of the template encoded G residue at position +6 of the completely aligned AAUUU transcript results in an RNA-DNA hybrid that is stable enough to prevent further transcript slippage, allowing the AAUUUG transcript to be extended into full-length transcripts. Conversely, addition of a U residue at position +6 of the slipped transcript prevents addition of a G residue and entry into the productive mode of transcription through a mechanism that remains obscure. Subsequently, the AAUUUU transcript is either released from the transcription initiation complex or it slips upstream, allowing addition of another extra U residue. This process can be repeated many times, with each AAUUUUn transcript eventually released from the transcription complex. 19

7 Schematic representations of co-translational control and genetic organization of IS911. Cartoon illustrating cotranslational binding. IRL (terminal inverted repeat left) and IRR (terminal inverted repeat right) are indicated as is the indigenous promoter pIRL, located partially in IRL. RNA polymerase (RNAP), ribosome and mRNA are also indicated. The nascent peptide is shown in black. The cartoon is not to scale. 21

8 Development of the GeneTack HMM structure. The complexity of the model increased from a simple 3-state structure to the final version of the model consisting of 28 states – Fig. 9. 30

9 Hidden Markov model used in GeneTack consists of 28 states. States 1, 2 and 3 correspond to the three global frames of reading the genetic code. The type of shading of a state reflects its frame. There is no single frame related to n/c and overlap states, thus they have no shading. 31

10 (A) GeneTack-GM algorithm overview; (B) splitting genome into fragments; (C) description of filters used to reduce number of false positive predictions (filters listed in the order they applied). 33

11 Calculation of transition probabilities for GeneTack HMM. * Transition probability to stop codon (upon approaching TAA, TAG, TGA) is 1. ** Transition probabilities to start codon in frames 2 and 3 (upon approaching ATG, GTG, TTG) are 0.0001 (0.001 for high GC genomes). 36

12 An example of the GeneMarkS gene prediction for a gene with simulated frameshift in a high GC genome. The figure shows the coding potentials in all six frames as determined by the GeneMark program. Black bars on the horizontal axis indicate predicted genes. A frameshift was introduced at position 1,848 in the gene on direct strand. There is a clear jump of coding potential from frame 3 to frame 1 at the location where the frameshift was introduced. However, there is a gene predicted in frame 2 on the opposite strand. Such artifacts are corrected by the modification in GeneTack-GM for high GC genomes as described in the text. 38

13 Dependence of the number of correctly predicted frameshifts on the distance from the artificially made frameshift to the gene border (either start or end). 45

14 Performance of the filters for 18 prokaryotic genomes (genomes are shown along the X axis, sorted by GC content). A domain of life is indicated in parenthesis ("A" stands for Archaea and "B" for Bacteria) (A) Filtering false positive predictions. The fraction of false positives with respect to the total number of genes in a genome, before (gray bars) and after (black bars) filtering are shown for each species. (B) Relative impact of filtering on true positives and false positives. For each genome percentages of removed false positives (with respect to false positives before filtering) and kept true positives (with respect to the number of true positives before filtering) are shown. The filters are supposed to remove as many false positives and as few true positives as possible. Thus, the sum of heights of two bars reflects the filters performance for a given genome. The best performance was observed for *Caulobacter crescentus*, the worst performance was for *Pyrobaculum aerophilum*. 47

15 Number of frameshifts predicted in a prokaryotic genome correlates with genome length (data from analysis of 1,106 genomes). Total number of predicted frameshifts was 206,991. Genomes shorter than 1 Megabase were not considered as not possessing sufficient amount of sequence for the GeneMarkS self-training. 50

16 Possible outcomes of the BLASTp and Pfam searches for a conceptual translation of fs-gene. If a frameshift position is covered by BLASTp hit (or Pfam domain) the predicted frameshift is considered to be validated by BLASTp (or Pfam) and is likely to be a true positive prediction. 51

17 Alignment of poly-A motifs from the 9 fs-genes of DNA polymerase III cluster with frameshift motif from *Thermus thermophilus* for which transcriptional realignment was previously shown. 56

18 Western blot analysis and quantification of frameshift products. . . . 60

19 Anti-his western blot analysis of frameshift products and internal initiation (translational coupling) products. 61

20 Distribution of the number of annotated and predicted pseudogenes among prokaryotic genomes. Black bars were obtained for pseudogenes annotated in GenBank. The white bars show the updated distribution with 4,806 pseudogenes identified in this work added to the annotated pseudogenes. The largest change in the distribution has been observed in the genomes with less than 10 pseudogenes annotated in GenBank. 69

21 Classification of predicted frameshifts was done by using features specified in Table 12. One of the most important properties of a predicted fs-gene was its membership in a cluster. Singletons are likely to be result of indel mutation or sequencing error, while clustered fs-genes could represent programmed frameshifts, phase variation and translational coupling, as well as pseudogene clusters or clusters of genes with indel mutations. 71

22 (A) Empirical distributions of frameshift coordinates relative to fs-gene lengths for 1/ all the predicted frameshifts (206,991 fs-genes), 2/ singletons (104,260 fs-genes) and 3/ frameshifts in clusters containing 10 or more members (47,278 fs-genes in total). (B) Distribution of relative coordinates of all the predicted frameshifts (A1) is shown along with the theoretical distribution combining a uniform distribution of coordinates of random frameshifts and distribution of false positive predictions (the $(x-t)/(x+y+z)$ distribution – see text). The random frameshifts correspond to indel mutations and sequencing errors while the false positives are predicted for adjacent genes (with overlapping ORFs). The theoretical curve has good fit to the observed distribution, with the value of parameter $\alpha = 0.005$. See text for more details. . . . 74

23 Distribution of frequencies of GeneTack magnitudes of errors in predicting frameshift positions. We have applied GeneTack to 400 *E. coli* genes (longer than 1,000 nt) with a single frameshift created at random in a position separated by at least 150 nt from the gene border. The program successfully predicted 351 frameshifts (the remaining 49 frameshifts were predicted as adjacent genes). This distribution shows that in 83% of cases a prediction is located within 5 nt from the true frameshift position. 75

24 An example of a frameshift box. Predicted frameshift is flanked by two stop codons in different frames (frame 2 TAG stop codon upstream of the frameshift and the frame 1 TGA stop codon downstream of the frameshift). The true frameshift position is always located inside the region between the two stop codons. We call this region "frameshift box". 80

25 Masks used to calculate motif periodicity of a motif 84

26 GeneTack HMM that was used to predict frameshift in eukaryotic mRNAs. The HMM is based on the assumption that eukaryotic mRNAs contain one gene only. 94

27 Genera-specific models for GeneTack were generated from 5% GC% ranges, each containing at least 1,000 mRNAs. In the example below, for the mRNA with GC% lower than 35 the 35_39 model would be used and for sequences with GC% higher than 64 the 60_64 model would be used. 95

28 Pipeline of the work (see text for more details). 98

29 Discrimination between a frameshift caused by alternative splicing (AS_FS) and indel mutation inside exon (Indel_FS). Three-frame translation of exons where frameshifts were predicted is shown (the images were obtained using http://arbl.cvmbs.colostate.edu/molkit/translate/). White lines indicate the translation path predicted by GeneTack with arrows indicating frameshift locations. Purple and green dashes indicate stop and start codons respectively. (A) The translation of a terminal exon started in frame 3 while the correct frame is 2. -1 frameshift (from frame 3 to frame 2) is predicted near the beginning of the exon suggesting that this is likely to be frame shifting AS isoform (note that there is no stop codons in frame 2 upstream of the frameshift position); (B) An example of an exon with indel mutation. The translation of the exon started in the correct frame 1, because there are stop codons in frame 2 and 3 upstream of the predicted frameshift position. +1 frameshift predicted in the middle part of the exon corresponds to indel mutation. 100

30 Distribution of the distance between predicted frameshift and the closest exon-exon junction for all mRNAs with known exon locations. The frameshifts predicted near exon-exon junctions are more likely to be caused by alternative splicing. 102

31 An alignment of the alternative frame translations derived from the SRCAP cluster (cluster ID 275483014) – one of the dual coding candidates. 103

32 Distribution of the K_a/K_s values calculated for the main frame and for the two alternative frames for 15,576 pairwise alignments of human-mouse homologous genes from HomoloGene database. The distribution demonstrates that K_a/K_s values can efficiently discriminate main (coding) frame from alternative frames. 104

33 Comparison of K_a/K_s values and ORF lengths between the main frame and alternative frames obtained from pairwise alignments of homologous human-mouse genes from HomoloGene database. A single pairwise alignment produces two dots on each plot – one dot for the main frame (blue) and one – for an alternative frame (red). For a pair of human-mouse genes the longest ORF was found in a given frame (main, +1 or +2) and the average length between human and mouse longest ORFs was used for this frame as a Y value. Black dot on the top plot corresponds to the alternative frame of a confirmed case of dual-coding – the ALEX protein. This dot can be used as a reference point to find other coding candidates. The green dots correspond to the alternative frames of the dual coding candidates found among fs-gene clusters in normal, single coding genes. 105

34 The GeneTack database entries: fs-genes predicted in genome of *Escherichia coli str. K-12 substr. DH10B*. **FS_ID** – unique fs-gene identificator, **Coord** – frameshift coordinate in the input sequence, **D** – frameshift direction (+1 or -1), **GeneL** – coordinate of left border of fs-gene (gene start for '+' strand, gene end for '-' strand), **GeneR** – coordinate of right border of fs-gene (gene end for '+' strand, gene start for '-' strand), **S** – the fs-gene strand, **F** – frameshift coordinate in fragment (the sequence used as input to GeneTack), **G** – frameshift coordinate in fs-gene, **P** – frameshift coordinate in fs-protein, **BLASTp** – information on the BLASTp hit covering frameshift position in the fs-protein, **Pfam** – information on the Pfam domain covering frameshift position in the fs-protein, **COF** – cluster ID (if available), **RBS** – RBS score of the downstream gene defined by GeneMarkS. 113

35 (A) Logo of the conserved motif and (B) distribution of coordinates of frameshifts in 428 fs-genes of Release Factor 2 collected in a cluster (ID 474411093) [1]. Red bars in (B) correspond to frameshift positions and green bars show the total length of fs-proteins. The small green bars indicate existence of subgroups of longer fs-proteins. 115

GLOSSARY

bp Base pairs, p. 5.

DNA Deoxyribonucleic Acid, p. 3.

FN False Negative, p. 40.

FP False Positive, p. 40.

GC% Percentage of G and C nucleotides, p. 34.

HMM Hidden Markov Model, p. 25.

kB Kilobases, p. 115.

mRNA Messenger Ribonucleic Acid, p. 2.

NCBI National Center for Biotechnology Information, p. 49.

nt Nucleotide, p. 27.

ORF Open Reading Frame, p. 2.

PRF Programmed Ribosomal Frameshifting, p. 3.

PTR Programmed Transcriptional Realignment, p. 3.

RBS Ribosomal Binding Site, p. 34.

RNA Ribonucleic Acid, p. 3.

rRNA Ribosomal Ribonucleic Acid, p. 4.

Sn Sensitivity ($Sn = TP/(TP + FN)$), p. 40.

Sp Specificity ($Sp = TP/(TP + FP)$), p. 40.

TP True Positive, p. 40.

tRNA Transfer Ribonucleic Acid, p. 7.

URL Uniform Resource Locator, p. 63.

UTR Untranslated region, p. 91.

SUMMARY

We developed a new program called GeneTack for *ab initio* frameshift detection in intronless protein-coding nucleotide sequences. The GeneTack program uses a hidden Markov model (HMM) of a genomic sequence with possibly frameshifted protein-coding regions. The Viterbi algorithm finds the maximum likelihood path that discriminates between true adjacent genes and a single gene with a frameshift. We tested GeneTack as well as two other earlier developed programs FrameD and FSFind on 17 prokaryotic genomes with frameshifts introduced randomly into known genes. We observed that the average frameshift prediction accuracy of GeneTack, in terms of (Sn+Sp)/2 values, was higher by a significant margin than the accuracy of the other two programs.

GeneTack was used to screen 1,106 complete prokaryotic genomes and 206,991 genes with frameshifts (fs-genes) were identifed. Our goal was to determine if a frameshift transition was due to (i) a sequencing error, (ii) an indel mutation or (iii) a recoding event. We grouped 102,731 genes with frameshifts (fs-genes) into 19,430 clusters based on sequence similarity between their protein products (fs-proteins), conservation of predicted frameshift position, and its direction. While fs-genes in 2,810 clusters were classified as conserved pseudogenes and fs-genes in 1,200 clusters were classified as hypothetical pseudogenes, 4,730 fs-genes from 146 clusters possessing conserved motifs near frameshifts were predicted to be recoding candidates. Experiments were performed for sequences derived from 20 out of the 146 clusters; programmed ribosomal frameshifting with efficiency higher than 10% was observed for four clusters.

GeneTack was also applied to 1,165,799 mRNAs from 100 eukaryotic species and

45,295 frameshifts were identified. A clustering approach similar to the one used for prokaryotic fs-genes allowed us to group 12,103 fs-genes into 4,087 clusters. Known programmed frameshift genes were among the obtained clusters. Several clusters may correspond to new examples of dual coding genes.

We developed a web interface to browse a database containing all the fs-genes predicted by GeneTack in prokaryotic genomes and eukaryotic mRNA sequences. The fs-genes can be retrieved by similarity search to a given query sequence, by fs-gene cluster browsing, etc. Clusters of fs-genes are characterized with respect to their likely origin, such as pseudogenization, phase variation, programmed frameshifts etc.

All the tools and the database of fs-genes are available at the GeneTack web site http://topaz.gatech.edu/GeneTack/

Chapter I

INTRODUCTION

Analysis of intronless gene sequences available in the public databases (such as RefSeq [2]) revealed that some protein coding regions contain frameshifts, i.e. sudden frame transition from one reading frame to another.

There are several potential reasons for the existence of frame transitions in a gene. Sequencing and assembly errors resulting in discrepancies between reported and real nucleotide sequence is an obvious non-biological reason for the existence of frameshifts. Frame transitions may occur when a gene contain a recent indel mutation. Even when such a mutation inactivates the encoded protein product, framing constrains may still be detected if there was insufficient time for neutral mutations to accumulate and deteriorate the protein coding signal. The frameshifts may also be detected within the genes that use non-standard mechanisms of transcription or translation, such as in case of Recoding where these mechanisms are used for gene expression purposes and often play a regulatory role [3, 4, 5, 6, 7].

The three frameshift types (sequencing errors, indel mutation and recoding events) can be observed in both prokaryotes and eukaryotes. Prokaryote specific types of frameshifts include cases of phase variation and translational coupling. Frame transitions may also be evident for genes that utilize phase variation, e.g. when members of the same population of bacteria have genomes that differ at a specific hypermutable location [8, 9]. This mechanism provides bacterial population with a possibility to diversify their population proteome beyond the limits of a single cell proteome.

Frame transitions are also expected to be detected when open reading frames of two genes overlap in prokaryotic genome. Although in these cases, a sequence

within an overlap or between frames may not exhibit a clear framing signal. Genes with overlapping ORFs tend to have conserved colocation producing polycistronic mRNAs that keep equal amount of protein product from each gene. This regulatory mechanism is observed in prokaryotic genomes and is known as translational coupling.

In eukaryotic mRNA sequences frame transitions are observed in some isoforms of alternatively spliced genes. In this case, frameshift is predicted not because of indel mutation but due to presence or absence of an exon.

All the frameshift types are discussed in more details below.

1.1 Sequencing errors

Sequencing error induced frameshifts are of significant interest. A volume of genomic data is increasing dramatically with advent of the next generation sequencing technologies (454 [10], Illumina [11], SOLiD [12]). Still, the assembly of a huge mass of short sequence reads may result in less homogeneous sequence coverage and higher rate of sequence errors than at a time of "slow sequencing". Errors of insertion or deletion type that occur inside protein-coding regions lead to frameshifts (unless the indel size is a multiple of three) and to erroneous gene prediction. It is highly desirable to detect frameshifts early and to correct predicted errors before genome sequence release.

1.2 Indel mutations

Frameshifts due to indel mutations inside protein coding regions have a special interest. They may significantly change the corresponding protein sequence resulting in truncated nonfunctional products.

The indel mutations during replication occur more frequent at homonucleotide runs, particularly at the poly-A sites. Frequently they lead to production of truncated proteins and result in gene pseudogenization. Such mutations in human somatic cells may lead to cancer. For example in 23 cases of colorectal cancer, 6 were caused by

insertion of additional A in a stretch of 8 A's in the human APC gene [13]. Apparently these mutations appeared during replication of somatic cells DNA.

The only reliable method to distinguish sequencing errors from authentic indel mutations is resequencing of the region where sequencing error is suspected. In a study of *Mycobacterium smegmatis* genome, where resequencing was done, 28 out of 73 detected frameshifts appeared to be caused by sequencing errors [14]. In the analysis of *Bacillus subtilis* genome resequencing was done for 522 fragments where frameshift sequencing errors were suspected [15]. 131 fragments appeared to contain 284 indel sequencing errors (68% of deletions and 32% of insertions).

1.3 Programmed frameshifting – an example of recoding

Recoding events may take place at different levels of gene expression. Alterations of transcript sequences can be accomplished through a range of RNA editing mechanisms (slippage [16, 17], guided RNA editing [18, 19], Adenosine and Cytosine deamination [20, 21, 22]) while the readout of RNA transcripts can be a subject to a variety of translational recoding mechanisms (ribosomal frameshifting, codon redefinition, translational bypass, stop-go). Here we concentrate only on the mechanisms that effect transitions between reading frames – transcriptional realignment and ribosomal frameshifting. In those cases where utilization of these mechanisms is deemed to have functional role, they described as programmed, e.g. Programmed Ribosomal Frameshifting (PRF) and Programmed Transcriptional Realignment (PTR) [23].

During PRF, ribosomes change the reading frame at specific locations in mRNA. While the frame can be changed only in two directions and +1 and -1 frameshifting are known as predominant mechanisms, -2, +2 and even +50 (commonly known as hopping) have been well documented and studied for over two decades [24, 25, 26, 27]. PRF has been reported in organisms from all kingdoms of life and is likely to occur in virtually all organisms but is especially frequent in viruses [5, 28]. High efficiency

of ribosomal frameshifting is modulated by a range of stimulatory signals, most frequently RNA structures [29, 30], but also signals that affect readout of mRNA either through interactions with ribosome components [31, 32], through direct complementary mRNA:rRNA interactions or through encoded peptide inside the ribosome [33, 34].

PTR (also termed transcriptional slippage [17], stuttering [35], molecular misreading [36] and reiterative transcription [37]) occurs when growing RNA chain realigns to the DNA template within the RNA polymerase ternary complex and this realignment results in insertion or deletion (indel) of a single or multiple nucleotides relative to the DNA template [38, 39]. The indels usually occur in characteristic motifs such as homopolymeric runs of adenines or thymines.

Programmed frameshifting of both types is utilized in Insertion Sequences (IS) and Transposable elements (prokaryotic and eukaryotic).

The best known examples of recoded genes in prokaryotes are those encoding Release Factor 2 (*prfB* gene) and DNA polymerase III (*dnaX* gene). Ornithine decarboxylase antizyme is a well-known example of eukaryotic gene utilizing recoding for its translation. Notably, many other prokaryotic and eukaryotic genes utilizing programmed frameshifting have also been reported (see Table 5 and Table 15).

The recoded genes are particularly interesting as they often cannot be predicted by conceptual triplet translation of corresponding nucleotide sequences and often requires specialized tools [40, 41, 42]. In addition to identification of novel protein coding genes, a search for Recoding may reveal novel stimulatory sequences that often are required for non-standard mechanism to achieve efficiency comparable to that of standard decoding. Identification of such sequences is important at least for two reasons. First, understanding the mode of their action could shed light on the functions of components of transcription and translation machinery. Second, such sequences provide means for manipulation of gene expression for synthetic biology

purposes. Therefore, we put particular emphasis on identification of novel recoding candidates.

1.3.1 Transposable elements: Insertion Sequences (IS) and retrotransposons

Insertions sequences (ISs) are small (800 – 2700 bp long) ubiquitous bacterial transposable elements and to date represent the largest group of chromosome encoded genes utilizing recoding. Sequence comparison and functional analysis lead to the sorting of known insertion sequences into 19 different families [43]. Members from four of these families appear to use programmed ribosomal -1 frameshifting to express their transposase, the enzyme required for their mobility.

Frameshifting was first identified in IS1 where it occurs at a rather low frequency (below 1%) because of a rather inefficient slippery motif and because of a lack of a proper stimulator [44]. In a few members of the IS1, IS5, and IS630 families, use of programmed -1 frameshifting is suspected whereas in the large and widespread IS3 family (27% of the known insertion sequences) such frameshifting appears to be the general rule. Occurrence of frameshifting was recognized, and demonstrated, at about the same time in several insertion sequences related to IS3 of *Escherichia coli* [45, 46, 47, 48]. Since then, members of that group were found in many bacterial species from all branches of the bacterial evolutionary tree.

The vast majority of the members of the IS3 family possess two consecutive and overlapping genes, with the second, orfB, being in frame -1 relative to the first, orfA. Strikingly, nearly all of these appear to contain a potential frameshifting signal in the orfA-orfB overlap region [45]. Direct evidence that frameshifting does indeed generate a OrfA::OrfB hybrid protein, the OrfAB transposase, has been obtained for a few insertion sequences. The OrfAB protein catalyzes the excision of the IS, generating an IS circle that is subsequently re-inserted at a new location through the combined action of the OrfA and OrfAB proteins [49, 50]; note that no function was

found for the OrfB protein, though it is indeed synthesized in the case of IS911.

Some eukaryotic transposable elements also employ recoding for their expression, particularly LTR retrotransposons. They frequently use programmed frameshifting at the boundary between *gag* and *pol*, two genes found in all retrotransposons. *gag* is the 5'-most gene and encodes structural proteins that form the virus-like particle. *pol* is located 3' of *gag*, and encodes enzymes such as reverse transcriptase, which are required for replication. In most retrotransposons, there is no independent initiation of *pol* translation; rather, Pol is expressed as part of a Gag-Pol polyprotein. The level of Pol relative to Gag is critical for retrotransposon functionality because particle assembly requires many more copies of Gag than Pol [51]. The *Saccharomyces cerevisiae* Ty1 and Ty3 retrotransposons utilize +1 frameshifting to synthesize Gag-Pol [52, 53]. Retrotransposons with *pol* in the -1 frame relative to *gag* are limited to Ty3/gypsy and DIRS-type elements that are widespread in the animal kingdom [54]. It should be noted that most of the retrotransposons with *pol* in -1 frame relative to *gag* originate from *Drosophila* and *B. mori*. No -1 retrotransposons were found in *C. elegans*. This differential distribution might be due to differences in cellular translational machinery in different animal hosts or differences in the types of retrotransposons that colonize certain hosts. Finally, an equal percentage of elements with *gag* and *pol* in a single frame or in -1 or +1 overlapping frames were found in fungi [55].

1.3.2 Bacterial *prfB* gene encoding Release Factor 2

The *Escherichia coli* gene *prfB* encodes Release Factor 2 (RF2) and was among the first discovered chromosomal genes requiring programmed ribosomal frameshifting for its expression [1]. In bacteria, two class-I release factors are responsible for recognition of codons specifying termination of translation, Release Factor 1 and Release Factor 2. These factors are semi-specific, they both recognize TAA stop codons. TAG

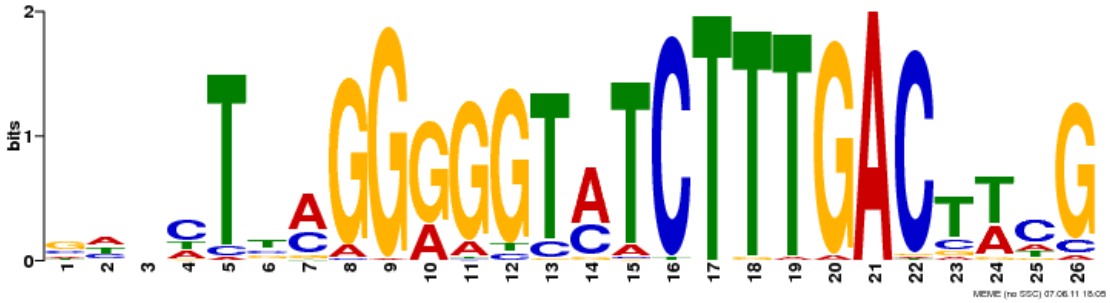


Figure 1: Diagram of RF2 frameshift site conservation, the height of symbols indicates conservation of nucleotides, while their weight shows the relative frequency of nucleotides at corresponding positions. The diagram was build using MEME from 428 sequences of RF2.

is recognized exclusively by Release Factor 1, while TGA recognition is specific to Release Factor 2 [56, 57]. In *E. coli* and most (≈87%) other bacteria, RF2 is encoded in two overlapping ORFs [40]. While the main portion of RF2 protein is encoded in the second long ORF, this ORF does not have its own translation initiation site. Initiation of translation takes place at the start of the first short ORF whereas the second ORF can be translated only if elongating ribosomes shift reading frame in the +1 direction at the end of the first ORF.

The shift sequence where frameshifting takes place is CTT_TGA_C (the underscores separate codons) – see Fig. 1. The key element responsible for sensitivity of frameshifting efficiency to the cellular concentration of RF2 is the TGA stop codon. When ribosomes approach the end of the first ORF and the stop codon occupies the ribosomal A-site, either of two major events occur: termination of translation or +1 slippage of P-site tRNA which directs translation to the longer ORF. These two events are in competition, so that increasing termination efficiency results in decreasing frameshifting efficiency and vice versa. As termination efficiency is directly influenced by the concentration of release factors, frameshifting efficiency also depends on the concentration of release factors. Since TGA is not recognized by Release Factor 1, frameshifting efficiency is solely dependent on the concentration of

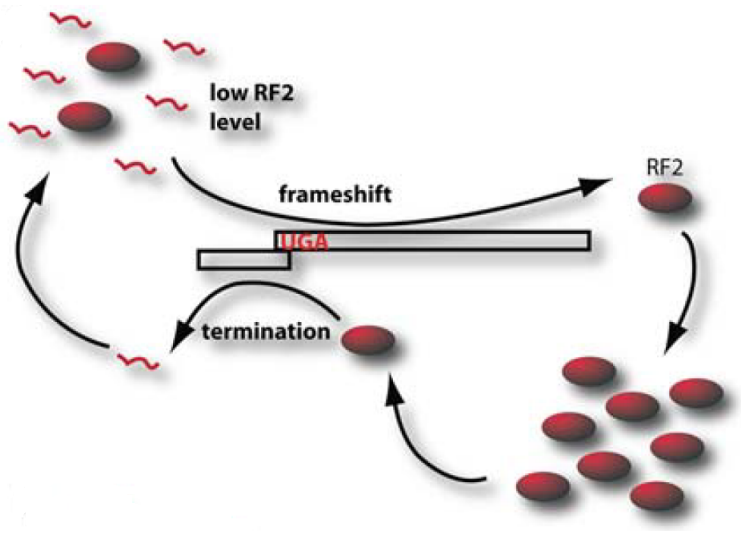


Figure 2: Regulatory feedback provided for RF2 biosynthesis by the frameshifting mechanism. The first ORF has a UGA stop codon. The regulation is autonomous and the level of RF2 biosynthesis depends on its own concentration.

RF2. This mechanism creates an elegant regulatory feedback loop, as illustrated on Fig. 2 (taken from [3]), where the level of RF2 biosynthesis depends on the cellular concentration of RF2.

1.3.3 Bacterial *dnaX* gene encoding DNA polymerase III subunits τ and γ

dnaX gene that encodes τ and γ subunits of DNA polymerase III is another well-known case of programmed frameshifting in bacteria. The τ subunit is the full-length product of the gene. The γ subunit is produced from the *dnaX* gene as well; the N terminal of γ subunit is identical to the τ subunit. However, the C terminal of the γ subunit is shorter and generated via -1 programmed frameshift within the τ reading frame [58, 59, 60]. The γ subunit appears to be associated with distributive synthesis on the lagging strand while the full-length τ subunit provides the extreme processivity required for the leading strand. In *E. coli* frameshifting occurs during translation (PRF mechanism) on the frameshift motif A_AAA_AAG [61]. On another hand, in *T. thermophilus* a single base is removed during transcription (PTR mechanism) on

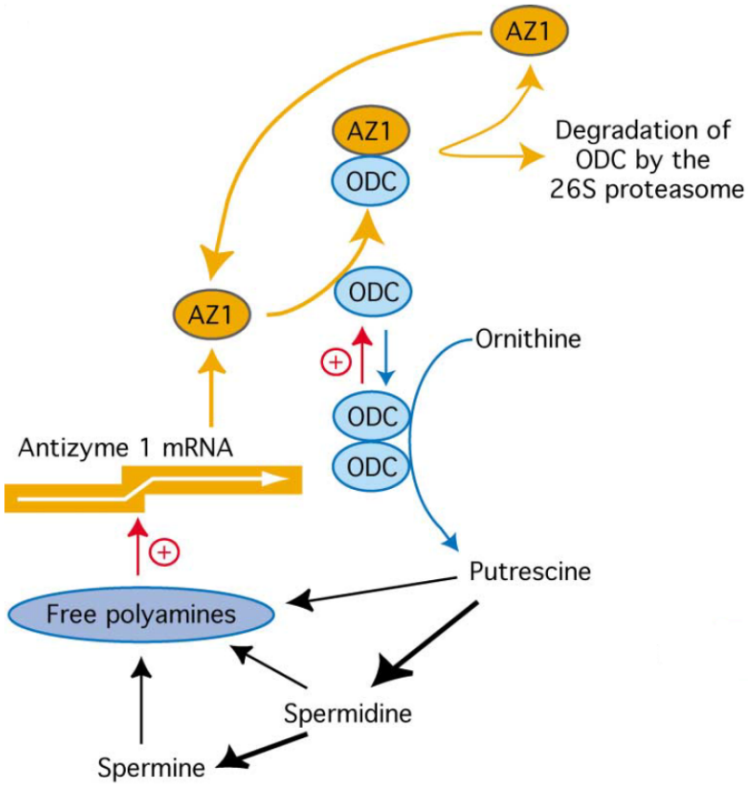


Figure 3: Regulation of cellular polyamine levels using antizyme +1 frameshifting as a sensor. High polyamine levels stimulate +1 frameshifting required for the synthesis of functional antizyme 1 (AZ1). AZ1 binds ornithine decarboxylase (ODC) and triggers its degradation by the 26S proteasome, being itself recycled. As ODC catalyzes the first step of the polyamine biosynthesis pathway, its degradation leads to a decrease in polyamine levels, which in turn reduces frameshifting efficiency.

a stretch of 9 A's [62]. The frameshifting efficiency in *dnaX* was reported up to 50% keeping ratio between τ and γ subunits about 1:1 [63].

Fewer examples of recoding were reported for *dnaX*, as compared with the number of annotated programmed frameshifting in *prfB*. It is unclear whether this is because the recoding mechanism is less spread or less studied in *dnaX*. However, interestingly, both ribosomal frameshifting (PRF) and transcriptional realignment (PTR) are known to be used in *dnaX* from different bacteria suggesting that these mechanisms are interchangeable.

1.3.4 Eukaryotic ornithine decarboxylase antizyme

Perhaps the best known example of eukaryotic gene utilizing programmed frameshifting is ornithine decarboxylase antizyme 1 (AZ1) [64]. The antizyme gene consists of two ORFs joined by a +1 shifty stop frameshift site. The frameshifting occurs when stop codon TGA is in ribosome A-site. The stop codon is conserved among eukaryotic species while the codon located in ribosome P-site (the one that is immediately upstream of the TGA stop codon) varies for different prokaryotic clades. The examples of antizyme frameshift site are GCG_TGA_C (*Saccharomyces cerevisiae*) and TCC_TGA_T (*Homo sapeins*) [65]. Ribosomes synthesizing antizyme start in one ORF and at the frameshift site move one nucleotide forward to a second and partially overlapping ORF which encodes most of the protein.

Like RF2, AZ1 frameshifting is linked to a feedback mechanism, although in this case as a sensor to regulate polyamine levels in eukaryotic cells. The antizyme protein binds to and directs the proteasomal degradation of ornithine decarboxylase (ODC) in the presence of excess spermidine, an eventual downstream product of ODC (Fig. 3, taken from [6]). At high concentrations, +1 frameshifting is increased, promoting the synthesis of AZ1, which in turn degrades ODC, leading to a reduction in cellular polyamine concentrations. The exact mechanism of frameshifting as well as how polyamines stimulate the frameshift event remains to be determined.

Frameshifting is employed in the expression of all known antizymes from mammals [64] to *Drosophila melanogaster* [66] to *Caenorhabditis elegans* and yeasts [67, 68, 69, 70] and in all cases to date, the process is used as a sensor of free polyamines. The conservation of this mechanism throughout evolution highlights a crucial role for frameshifting in the regulation of polyamine levels.

1.3.5 Other examples of programmed frameshifting in prokaryotes

B. subtilis cdd gene encodes cytidine deaminase – zinc-containing enzyme involved in the pyrimidine salvage pathway and catalyzes the formation of uridine and deoxyuridine from cytidine and deoxycytidine, respectively. The translation of the gene involves -1 ribosomal frameshifting at the CGA_AAG site resulting in the synthesis of a product extended by 13 amino acids [71]. The physiological relevance of the *cdd* frameshifting event is uncertain, since the C-terminal extensions has no apparent effect on cytidine deaminase activity. It has been speculated that the *cdd* frameshift allows translational regulation of the following gene, *bex*, since the 3' end of *cdd* overlaps the 5' end of *bex*.

Fu and Parker have found efficient +1 frameshifting at the beginning of *E. coli argI* mRNA [72]. Ribosomes that shift in the +1 direction produce a short truncated polypeptide. No functional role was suggested and this frameshifting has been considered as a highly efficient translational error [73].

Another bacterial gene utilizing programmed frameshifting include *pheL* [33] from *E. coli* and *mxiA* [74] and *mxiE* [75] genes from *S. flexneri*.

To our knowledge, there is only one example of programmed frameshifting in Archaea. The *fucA1* gene of *S. solfataricus* contains a typical slippery sequence A_AAA_AAT followed by a putative stem-loop that acts as a frameshifting stimulator [76]. The frameshifting signal in the *fucA1* gene differs slightly from those reported in bacteria (particularly in the *dnaX* gene) and probably functions in a similar way.

1.3.6 Other examples of programmed frameshifting in eukaryotes

A requirement for +1 frameshifting in telomerase activity has been shown in the last few years. The synthesis of telomeres in *S. cerevisiae*, as in many organisms, depends upon telomerase, a reverse transcriptase that uses an internal RNA as a template. The yeast enzyme is a ribonucleoprotein composed of at least four proteins (Est1p,

Est2p, Est3p, Cdc13p) and the template RNA (TLC1). Est3p is a stable component of the telomerase holoenzyme and essential for the maintenance of telomeres *in vivo*. The +1 frameshifting mechanism employed in the expression of the protein involves an essential, short slippery sequence, CTT_AGT_T [77]. As the AGT codon in this stretch is decoded by a low abundance tRNA, a ribosomal pause is thought to occur, promoting +1 slippage of peptidyl tRNALeu from the CTT to the overlapping TTA codon. The organization of EST3 genes and the utilization of +1 frameshifting are conserved among different *Saccharomyces* species suggesting a key role for frameshifting in telomere maintenance [6].

The *ABP140* gene of *S. cerevisiae* [78] encodes an actin binding protein whose expression requires +1 frameshifting at the slippery sequence CTT_AGG_C. This sequence is one of the most underrepresented heptameric stretches in the *S. cerevisiae* genome [79] and promotes highly efficient frameshifting, with about one in three ribosomes changing frame at this site.

Another interesting example of +1 programmed frameshifting comes from the human *IL-10* gene, encoding an immunosuppressive, anti-inflammatory cytokine [80]. It has been shown that a cytotoxic T cell epitope generated from the *IL-10* gene by +1 frameshifting could activate autoagressive T cells leading to the elimination of the subset of cells producing this cytokine.

There are known examples of -1 programmed frameshifting as well. It is used in expression of many animal, plant and bacterial viruses and a number of mobile elements genes. In *S. cerevisiae* -1 PRF is used to express the endogenous L-A doublestranded RNA virus [81]; as with the retrotransposons, frameshifting occurs between the structural and enzymatic genes of this virus.

One example of a cellular gene utilizing programmed -1 frameshifting is mouse *Edr* ("Embryonal carcinoma Differentiation Regulated") and its human ortholog *PEG10* ("Paternally Expressed Gene" 10) [82, 83]. The frameshift signal is present between

two long overlapping ORFs and resembles a typical retrovirus frameshift signal, with a slippery sequence G_GGA_AAC and a potential pseudoknot-forming region five bases downstream. Pausing at the pseudoknot is believed to promote simultaneous -1 slippage of both ribosome-bound tRNAs in manner similar to that described for the *dnaX* frameshift signal. The function of the gene is unknown but the conservation of the frameshifting site in mouse and human and the expression pattern during development and in adult tissues argues for an important role for frameshifting.

A functional -1 ribosomal frameshifting signal was found in the human paraneoplastic gene *Ma3* [84]. *Ma3* is a member of a family of six genes in humans whose protein products contain homology to retroviral Gag proteins. The -1 frameshift site and pseudoknot structure are conserved in other mammals. Although the functions of the *Ma* genes are unknown, the serious neurological effects of ectopic expression in tumor cells indicate their importance in the brain.

1.3.7 Mechanisms of programmed ribosomal frameshifting (PRF)

Programmed ribosomal frameshifting (PRF) occurs in certain mRNAs from diverse organisms when the ribosome dynamically diverted into an alternative reading frame at specific sites. Where utilized for regulatory purposes or to produce an additional protein, the ribosomal frameshifting involved is often 'programmed' to occur at high efficiency by signals embedded in the same mRNA. Programmed frameshifting is induced by a specific sequence that sometimes called "frameshift signal". The sequence consists of a "frameshift site" (also known as "slippery sequence") and optional "stimulatory sequences" (or just "stimulators") that may be present around the frameshift site (Fig. 4). Frame transition occurs within the frameshift site, the most conserved and relatively short (6-10 nucleotides) part of the motif. Examples of stimulatory sequences are Shine-Dalgarno like sequences, sequences forming pseudo-knots at RNA level that could pause a ribosome at the frameshift site, etc.

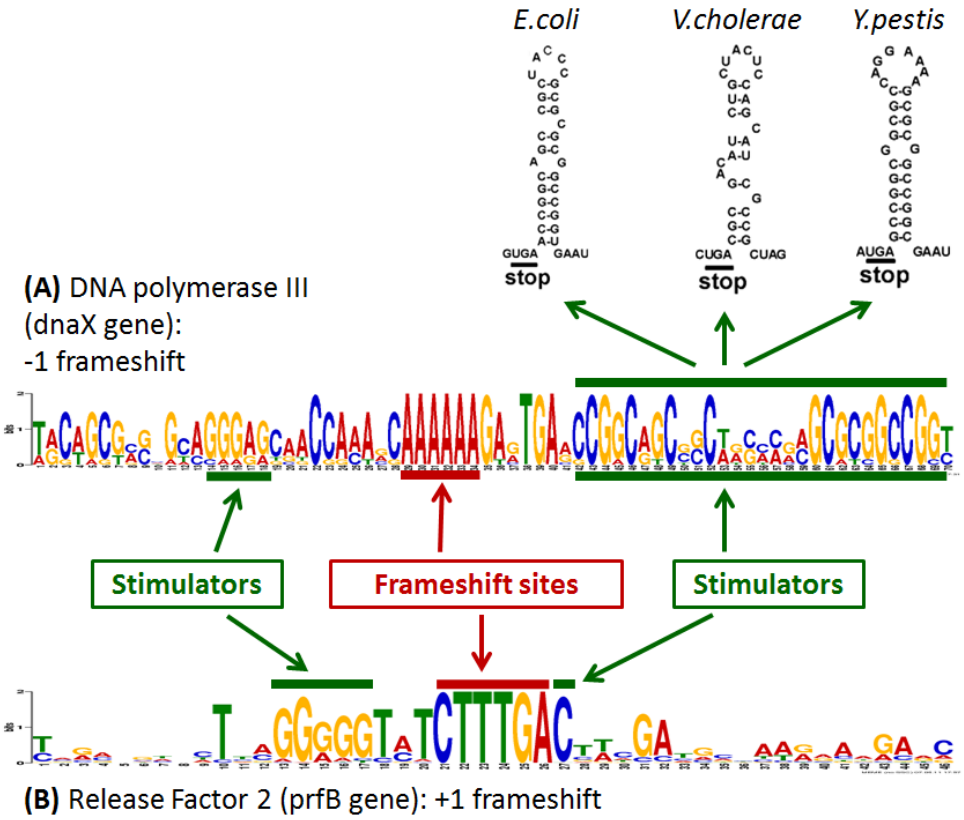


Figure 4: Examples of PRF "frameshift signals". Each signal consists of frameshift site and two stimulatory sequences (stimulators). (A) "-1" programmed frameshift is utilized in *dnaX* gene to express two subunits of DNA polymerase III. The Logo for (A) was derived from aligned sequences from 9 genera (*Escherichia*, *Salmonella*, *Neisseria*, *Vibrio*, *Shigella*, *Citrobacter*, *Enterobacter*, *Yersinia*, *Serratia*). The frameshift signal consists of conserved frameshift pattern AAA_AAA_G ("slippery sequence") and two stimulators. The upstream stimulator is a Shine-Dalgarno like sequence that interacts with ribosome and the downstream stimulator makes a hairpin secondary structure. (B) "+1" programmed frameshift is utilized in *prfB* gene to auto regulate expression of Release Factor 2. The Logo for (B) was derived from aligned 413 sequences from 138 genera. The frameshift signal consists of conserved frameshift pattern CTT_TGA and two stimulators. The upstream stimulator is also a Shine-Dalgarno sequence while the downstream stimulator is represented by a single cytosine that forms the "weakest" termination context.

In programmed ribosomal frameshifts (PRFs) not only frameshift site sequence but also the phase with respect to the reading frame of genetic code is important. For example, in the gene of Release Factor 2 the consensus of the frameshift site motif is CTT_TGA_C where triplet TGA is a stop codon with important regulatory role.

In case of PRF the length of the frameshift site is frequently 7nt (e.g. CTT_TGA_C for +1 frameshift in RF2 or A_AAA_AAG for -1 frameshift in *dnaX* gene) that includes two codons located inside ribosome P- and A-sites before frameshifting (the original frame) and one more nucleotide defines the codon that will be in P- or A-site after frameshifting occurs (the shifted frame).

In case of Release Factor 2 there are stimulatory elements that are responsible for elevation of the absolute level of frameshifting efficiency, which in their absence would be insignificant even at low concentrations of release factors. The element whose role in the frameshifting mechanism is relatively easy to understand is the TGA stop-codon. The ribosome pauses at the TGA stop codon when concentration of release factor 2 proteins is low. Unlike all sense codons that are recognized by RNA molecules via complementary interactions, stop-codons are recognized by protein molecules. Notably the nucleotide 3'-adjacent to the TGA stop codon affects termination efficiency. Since frameshifting efficiency negatively correlates with termination efficiency, it is not surprising that the weakest termination context, the cytosine, has been selected in the RF2 frameshift site during its evolution [85]. It can be seen in Fig. 1 that the 3' nucleotide adjacent to the stop codon is nearly always C, which has been shown to be the most inefficient context codon for termination in eubacterial organisms [86].

Another important stimulatory element in the RF2 frameshifting cassette is the internal Shine-Dalgarno (SD) sequence located upstream of the shift site. Normally SD sequences are used for the initiation of translation in bacteria and are located upstream of initiator codons [87]. The main role of the internal SD is clearly to target elongating ribosomes. One particular important aspect of the SD stimulatory

effect on frameshifting efficiency is the location of the SD relative to the frameshift site. The length of the spacer between the SD sequence and the P-site tRNA during the frameshift is shorter than the distance between the SD and initiator codons [88]. It is reasonable to assume that the distance between an SD and an initiator codon is optimal for the relaxed conformation of the ribosomal RNA during the initiation. If so, the shorter distance between the internal SD and the shift site should create tension in the ribosomal RNA between the anti-SD and the decoding center of the ribosome. Such tension likely acts in a manner of a compressed spring, whose relaxation is achieved by a progressive movement of tRNA with the decoding center of the ribosome toward the 3'-end of mRNA. This movement would explain the stimulatory effect of SD on +1 frameshifting.

In case of *dnaX* gene the mechanism is somewhat similar to the RF2 case, but in this case the ribosome shifts into -1 frame. The ribosome pauses at the frameshift site, A_AAA_AAG because it encounters mRNA secondary structure located downstream of the frameshift site. The SD-like sequence is located at a larger than the optimal distance between SD and ribosome, but close enough for their interaction.

+1 frameshifting in *S. cerevisiae* Ty1 retrotransosons occurs at the heptameric frameshift CTT_AGG_C site without help of stimulatory sequences. The frameshifting occurs because of "wobbling mispairing" between tRNA and mRNA template inside the ribosome P-site (see Fig. 5, taken from [3]). The frameshift signal consists of two codons in the normal or zero reading frame and a third overlapping codon in the shifted frame. The frameshift occurs when the CTT codon occupies the ribosomal P site with the ribosome selecting a tRNA recognizing the +1 frame GGC codon rather than the zero frame AGG codon (Fig. 5). Peptidyl-tRNA reading CTT would slip +1 onto the overlapping TTA codon allowing reading of the then in-frame GGC. The major tRNA for CTT codon decoding in yeast is $tRNA^{Leu}_{TAG}$ [89]. This tRNA forms the weak U:U pair on CTT. It also has the ability to recognize the overlapping

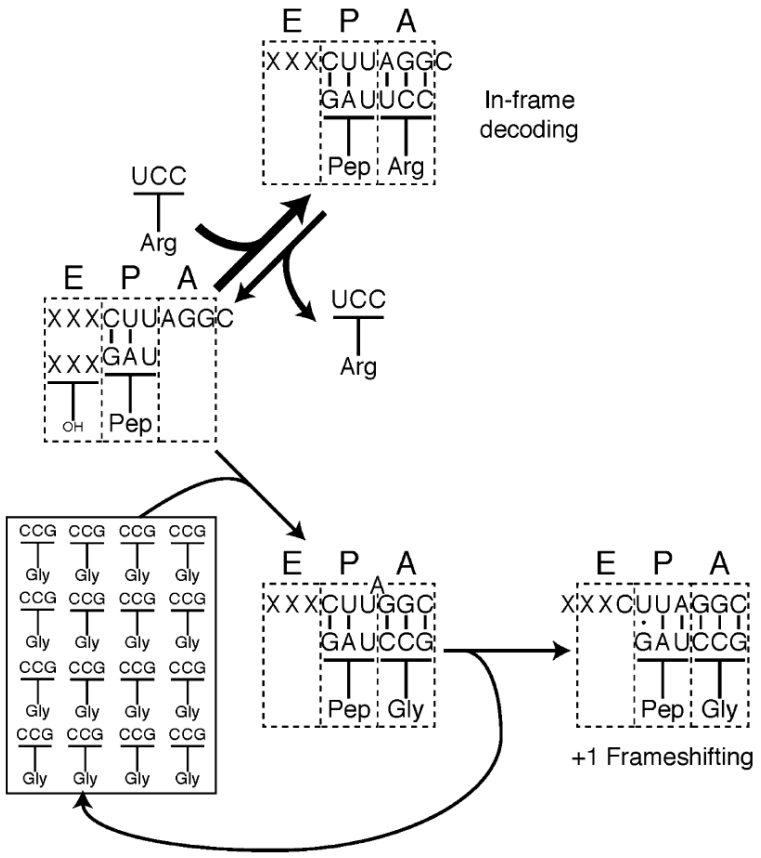


Figure 5: Model of Ty1 frameshift mechanism in *S. cerevisiae*. The three ribosomal tRNA-binding sites (E, P, and A) are diagrammed as dotted rectangles binding tRNAs cartooned as T's; the anticodon of each tRNA appears above the tRNA running 5'-3' from right to left. On the left, the P site is shown occupied by peptidyl-$tRNA^{Leu}_{UAG}$; the identity of the tRNA in the E site does not influence frameshifting and it is shown with XXX as anticodon. Two alternatives exist for the next step of elongation. Above, $tRNA^{Arg}_{CCU}$ is shown occupying the A site, leading to in-frame decoding; this reaction is shown as reversible because wobble mispairing in the P site appears to block cognate acceptance. Below, $tRNA^{Gly}_{GCC}$ is shown occupying the A site, also reversibly; its binding can lead to +1 frameshifting. Binding of this tRNA to the A site is shown requiring the middle A of the shifty heptamer to be excluded from the A site to allow the GGC anticodon to bind there, facilitating out-of-frame binding of $tRNA^{Gly}_{GCC}$. The number of tRNA shown corresponds to their relative concentration in the cell; $tRNA^{Gly}_{GCC}$ is present at approximately 16-times the concentration of $tRNA^{Arg}_{CCU}$.

TTA codon. Both of these factors could explain the tendency toward frameshifting. Another important factor that is that concentration of tRNAs decoding GGC codon located in +1 frame is 16 times higher than the concentration of tRNA for in-frame AGG codon. Combination of these factors results in 40% efficient frameshifting that is caused by a 7nt frameshift site only. Frameshifting mechanism of this type is called "near cognate decoding model".

The precise frameshifting mechanism in antizyme 1 is not known, but it may follow the near cognate decoding model as well. All +1 programmed frameshift events in *S. cerevisiae* occur when the ribosomal P site contains a near-cognate peptidyl-tRNA, one that fails to form a legal wobble base pairing interaction. The *S. cerevisiae* OAZ1 frameshift signal GCG_TGA_C that involves a putative P site codon, GCG and a poorly recognized TGA_C tetranucleotide known to be competent to stimulate 37% frameshifting [90]. In antizyme 1 frameshifting requires slow recognition of the termination sequence TGA_X by yeast peptide release factor 1 (eRF1) and eRF3 [91].

1.3.8 Mechanism of programmed transcriptional realignment (PTR)

PTR occurs during transcription when nucleotides are repetitively added to the 3' end of a nascent transcript due to upstream transcript slippage. It is typically modulated by interactions between RNA polymerase and its nucleoside triphosphate substrates without the involvement of regulatory proteins (Fig. 6, taken from [37]). Usually slippage occurs between a homopolymeric sequence in the transcript and at least three complementary bases in the template. Although PTR can involve the addition of any nucleotide (at least under certain conditions), addition of U or A residues appears to occur most frequently. This preference presumably reflects a requirement in the reaction for disruption of the RNA-DNA hybrid within the transcription complex which would be facilitated by relatively weak U:A or A:T base pairing.

The location where transcriptional realignment occurs can be compared with PRF

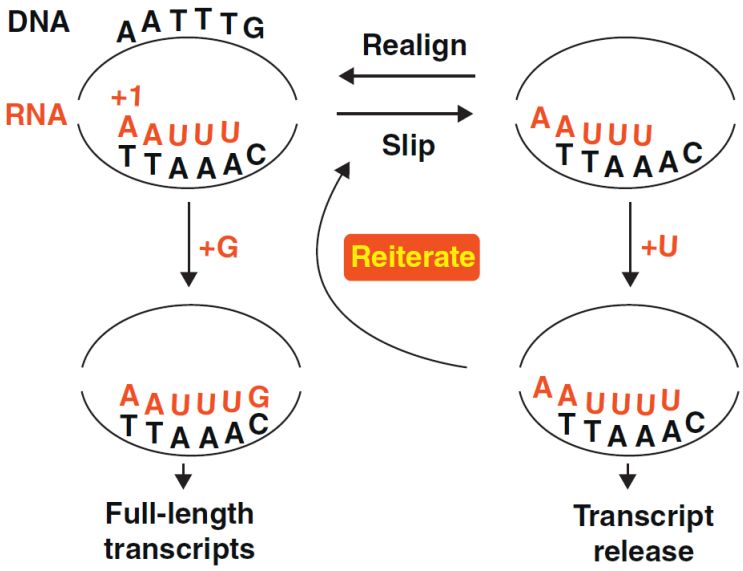


Figure 6: Direct control of transcriptional realignment by competing NTP substrates. In this example, which uses the DNA sequence of the *E. coli pyrBI* initially transcribed region, the competing substrates are GTP and UTP. After synthesis of the AAUUU transcript, it can reversibly slip one base upstream due to a weak RNA-DNA hybrid. Addition of the template encoded G residue at position +6 of the completely aligned AAUUU transcript results in an RNA-DNA hybrid that is stable enough to prevent further transcript slippage, allowing the AAUUUG transcript to be extended into full-length transcripts. Conversely, addition of a U residue at position +6 of the slipped transcript prevents addition of a G residue and entry into the productive mode of transcription through a mechanism that remains obscure. Subsequently, the AAUUUU transcript is either released from the transcription initiation complex or it slips upstream, allowing addition of another extra U residue. This process can be repeated many times, with each AAUUUUn transcript eventually released from the transcription complex.

frameshift signal and using the PRF terminology the PTR frameshift signal can be viewed as a signal consisting of a frameshift site only (homopolymeric sequence) without surrounding stimulators. Since transcriptional realignment occurs during transcription the sequence of PTR frameshift site is specified without indicating the reading frame (Table 5).

1.3.9 The biological purpose of programmed frameshifting

Genes use programmed frameshifts for a variety of purposes. Frameshifting produces two primary translation products, one form by normal translation and a second less abundant form through frameshifting. In cases when both products are functional (like in *dnaX*) the function of frameshifting is to define the stoichiometric ratio between them. In cases when only one product is functional (the longer one) the frameshifting may be utilized in regulation of gene expression. In case of *E. coli prfB* gene autogenous control insures that sufficient amount of RF2 is continuously present in the cell. Regulation of expression of the antizyme gene is based on polyamine concentration. The exact mechanism of the regulation is not known, but it was shown that in at high concentrations frameshifting efficiency is increased, promoting the synthesis of antizyme.

Use of programmed frameshifting significantly decreases expression of the functional product that is useful for transposases. The purpose of a transposable element is to maintain and propagate itself, whereas its host would rather eliminate it since it does not encode any cellular proteins. Insertion sequences and retrotransposons are not under selective pressure as suggested by the frequent presence of mutated or deleted transposable elements in genome. To stay functional the mobile elements use transposition to keep constant their number of active copies and to have a chance to colonize other hosts, but it should not transpose too frequently. The important issue with any mobile element is control of transposition at a level ensuring survival and

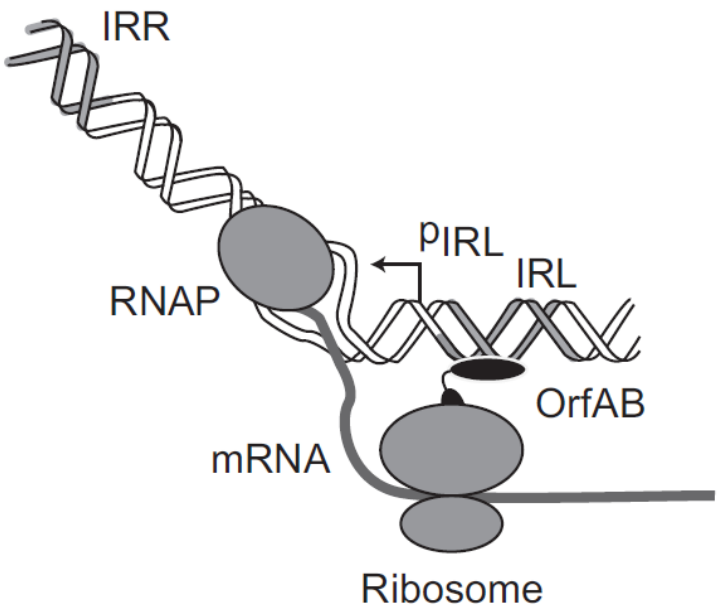


Figure 7: Schematic representations of co-translational control and genetic organization of IS911. Cartoon illustrating cotranslational binding. IRL (terminal inverted repeat left) and IRR (terminal inverted repeat right) are indicated as is the indigenous promoter pIRL, located partially in IRL. RNA polymerase (RNAP), ribosome and mRNA are also indicated. The nascent peptide is shown in black. The cartoon is not to scale.

propagation of itself while not being detrimental, and sometimes being beneficial, to the host. In this sense programmed frameshifts may be useful acting as inhibitor of transposition. Efficiency of programmed frameshifting around 1-3% results in low level of functional transpose expression.

It was also hypothesized that programmed frameshifting provide *cis* action of bacterial insertion sequences. One IS regulatory phenomenon is a preference of their transposases for action on the element from which they are expressed (*cis*) rather than on other copies of the same element (*trans*) [92]. For example in IS911 activity in *cis* is 200 fold higher than in *trans* [93]. Translational frameshifting pause signal influences *cis* preference presumably by facilitating sequential folding and co-translational binding of the transposase (Fig. 7, taken from [93]).

Programmed frameshifting in one gene could affect translation of the downstream

gene in polycistronic mRNAs. For example the 3' end of *B. subtilis cdd* gene overlaps the 5' end of the *bex* gene. The SD-like sequence present upstream of the *cdd* slippery sequence may be used to initiate translation at the *bex* gene. Under these conditions, a ribosomal pause at the SD-like sequence during translation of the *cdd* gene could prevent initiation at the *bex* gene.

For many know cases of programmed frameshifting the exact biological purpose is still needs to be determined. Moreover, in some cases the frameshifting is currently considered as translational error (e.g. *E. coli argI* and *pheL* genes [72, 33]).

1.4 Translational coupling in prokaryotic mRNAs

Translational coupling (also known as translational re-initiation) occurs in polycistronic mRNAs when start of the downstream gene is located close to the end of the upstream gene. After finishing translation of the upstream gene, the ribosome moves several nucleotides backward or forward to the start codon of the next gene and begins translation. This mechanism helps maintain a specific ratio between the concentrations of expressed coupled proteins.

In some cases there is an interesting cooperation of translational coupling and programmed frameshifting. As was mentioned earlier, the bacterial insertion sequences of IS3 family utilize -1 programmed frameshifting to express a fusion product orfAB. It was shown that translational coupling also occurs at the same location producing two separate proteins orfA and orfB [94]. This is the only example of these two processes to occur in the same gene. Our results indicate that translational coupling and programmed frameshifting could appear together more often than it was thought before.

1.5 Phase variation in prokaryotic genomes

Phase variation is a reversible and inheritable change of bacterial phenotype. It is often considered a random process that has evolved to facilitate immune evasion in a

host. Phase variation has been a focus of study mainly in bacteria pathogens, however, occurrence of phase variation in commensal species or species that do not reside in or on a host cannot be ruled out [95]. Majority of known phase variable moieties are exposed to the environment: proteins involved in capsule, fimbriae, pili, flagella and other outer surface proteins such as transporters, receptors, porins. However, some encoded protein variation for which there is no evidence of association with changes in the cell surface could occur as well, such as phase variation of DNA modification and metabolism associated genes [96]. Notably, many of the large clusters included fs-genes for cell surface and secretory proteins.

Among several molecular mechanisms responsible for phase variation (homologous recombination, inversion of DNA elements, insertion/excision of genetic elements from chromosome) slipped strand mispairing (SSM) seems to be the major one. During replication SSM may occurs at repeat units (such as short sequence repeats, microsatellites or variable number tandem repeats). The repeat unit could be as simple as a homopolymer sequence (for example poly-A in the *p78* gene of *M. fermentas* [97] or poly-C/poly-G in the type III methyltransferases genes [98]) or a repeat of more complex subunits (for example AGTC is repeated over 30 times in *H. influenzae mod* gene). Insertion or deletion of a repeat unit upon replication creates frameshift mutation that turns the gene on or off and consequently changes the phenotype of the bacteria. Some changes allow pathogenic bacteria avoid immune system of the host. It should be noted that the phase variation is a reversible process and the wild type will be restored after several generations.

1.6 Frame shifting alternative splicing in eukaryotes

Alternative splicing (AS) produces alternative isoforms by using different combinations of exons. In most cases each exon is designed to be read in one particular frame while other two frames contain many premature termination codons (PTCs). It is

estimated that 75% of mammalian genes are alternatively spliced [99]. Functional important mRNAs show frame preserving preference, i.e. the exons are concatenated in a synchronized way preserving reading frame. However, in up to one-third of alternative spliced mRNA variants there is no reading frame synchronization between concatenated exons [100, 101]. In such mRNAs coding potential shifts at the exon-exon junctions where reading frame synchronization is lost leading to formation of premature termination codons (PTCs). These transcripts are targets of nonsense-mediated decay (NMD), a surveillance mechanism that selectively degrades nonsense mRNAs [102]. During mRNA processing, exon-exon splice junctions are marked with exon junction complexes that serve the dual purpose of facilitating export to the cytoplasm and remembering gene structure [103]. As translation occurs, the ribosome displaces all exon junction complexes in its path. If a complex remains after a pioneering round of translation [104], a series of reactions lead to transcript degradation. Thus, transcripts that contain premature termination codons, that is, termination codons >50 nucleotides 5' of the final exon, are candidates for NMD.

1.7 Existing approaches for frameshift identification

Frameshifts – changes of reading frame in protein-coding genes – can be classified by origin into natural and artificial. Artificial frameshifts are caused by sequencing and assembling errors that may occur even in high X coverage sequencing (errors of length not divisible by 3). Early detection of frameshifts related to sequence errors could improve the quality of the assembly process and subsequence annotation of the sequence.

Several programs have been developed to detect frameshifts of both kinds. These programs can be divided into two groups with respect to the approach they use: (i) comparative genomics (similarity search) based, (ii) single sequence based (*ab initio*). A number of programs have been developed to predict a special kind of natural

frameshifts – genes utilizing programmed frameshifting.

1.7.1 Similarity search based programs

Similarity search based programs use translation of concatenated ORFs located in the same DNA strand as a query for a protein database search. The search may identify a database protein with statistically significant similarity region (a hit) that overlaps the junction in the "chimeric" query. Such an outcome indicates either a frameshift or naturally occurred events of gene fusion or gene fission. To discriminate between these events further analysis of conservation of the protein primary structures in multiple species is required.

Development of the similarity search based approaches included the initial heuristic program DETECT using 3-frame translations of potentially frameshifted sequence in protein database searches [105]; introduction of frameshift dependent scoring matrices for protein sequence alignment algorithm [106]; refinement of translated DNA to protein alignment techniques to detect both frameshifts within codons and between codons [107]; implementation of dynamic programming algorithm for correct alignment of the protein translation of DNA in three frames to a homologous protein [108, 109].

1.7.2 *Ab initio* frameshift prediction programs

The similarity based methods have a clear limitation: it is impossible to detect frameshifts in genes of orphan proteins that do not have known homologs. *Ab initio* (single sequence based) approach does not have this limitation; it was implemented in the programs FSED [110], ProFED [15], FrameD [111] and FSFind [112].

FSED program uses k-tuple frequencies to identify the frame of genetic code along the genomic sequence. ProFED predicts frameshifts using posterior probabilities of the reading frames determined by the GeneMark program. FrameD is an HMM-based gene prediction algorithm that allows to predict genes in the presence of frameshifts.

FSFind predicts frameshifts from scanning the posterior probabilities determined by the GeneMark algorithm.

Hidden Markov models (HMMs) are one of the most powerful tools for analysis of biological sequences. The structure of an HMM corresponds to a specific biological feature of interest (such as gene, donor/acceptor sites of exons, frameshifts etc). Several algorithms are utilized to apply HMMs to sequences in order to obtain biological meaningful predictions. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states that produce a given sequence. In the scope of frameshift prediction problem, the most likely locations of reading frame transitions (frameshifts) can be derived from the output of Viterbi algorithm.

It should be noted that FrameD and FSFind have an option to additionally verify the initial set of predicted frameshifts using similarity search. With this option the programs can be considered as a combination of *ab initio* and similarity search algorithms.

1.7.3 Programs for finding programmed frameshifts

A number of computer programs have been specifically designed to identify new instances of programmed frameshifting. Typically, these programs use a combination of two broad approaches: (1) search for genes that bear homology to known genes that use recoding (e.g. ARFA and OAF programs); (2) search for specific signals within the nucleotide sequence that resembles signals known to stimulate frameshifting.

FSFinder [113] and its descendant FSFinder2 [114] are programs specifically designed for the identifications of cases of programmed ribosomal frameshifting. FSFinder searches regions of two overlapping ORFs for patterns characteristic of particular types of frameshift signals, such as -1 frameshift cassette (comprising a X_XXY_YYZ heptanucleotide where XXX is a run of any nucleotides, YYY is either UUU or AAA and Z is usually not G followed by a stem loop or RNA pseudoknot structure [115]), an

RF2 frameshift cassette (a Shine-Dalgarno-like sequence upstream of a CTT_TGA_C motif), and an antizyme frameshift cassette (TTT_TGA or YCC_TGA followed by an RNA structure).

Several studies have been performed to identify genes utilizing -1 ribosomal frameshifting caused by X_XXY_YYZ pattern. Jacobs et al identified over 1,000 genes in *S. cerevisiae* genome as a candidates to use -1 frameshifting [116]. Later, a specialized program KnotInFrame was developed to identify these type of genes [42].

To identify genes of RF2 and antizyme in different organism two special programs have been developed – the Automated Release Factor Annotation (ARFA) [40] and Ornithine decarboxylase Antizyme Finder (OAF) [41]. These programs use combination of similarity search tool HMMER and search for specific frameshift pattern to identify the location of frameshifting.

Disadvantage of the above studies is that they are limited by the assumption about structure of the frameshift site and consequently are not able to predict genes with different PRF mechanism.

FSscan program is more flexible in this sense. It doesn't search for a specific sequence patterns but rather estimates the possibility of +1 frameshifting based on a thermo dynamical model. FSscan calculates a frameshifting score for every 16 nt fragment from a protein coding that was used to find new recoded genes [117]. The authors analyzed *E. coli* genome and selected 6 candidate genes with highest score to utilize +1 frameshifting.

The work by Shah et al [79] is an attempt to predict genes with PRFs without any prior knowledge about the underlying frameshifting mechanism. Motivation for this work was an assumption that motifs causing programmed frameshifting should be avoided in coding regions of genes that do not use PRF. They compiled a list of most underrepresented heptanucleotides (in comparison with random sequence generated by zero order Marko model) in coding regions of *S. cerevisiae* genes. Notably, several

known PRF motifs (CTT_AGG_C and CTT_AGT_T) were found among the obtained heptamers. The disadvantages of this work are that it was done on *S. cerevisiae* genome only and although new signals that may cause programmed frameshifting were identified, no new genes that may utilize PRF had been proposed. Moreover, it was shown that motifs causing programmed frameshifting are actually abundant in coding regions [118] that contradicts to the original assumption made by Shah et al. But it should be noted that, indeed, PRF motifs are avoided in coding regions of highly expressed genes.

Here we present a new algorithm and the program for *ab initio* frameshift detection in nucleotide sequences containing intronless genes (prokaryotic genomes, metagenomes, phage genomes, EST sequences). The GeneTack program (tack – a zigzag movement) is an HMM-based approach and designed to run on a DNA fragment with all genes located in the same strand. To analyze the whole genome we use a combination program, GeneTack-GM, a wrapper around GeneTack utilizing earlier developed program GeneMarkS [119] (GM) which makes a parse of the whole new genome into fragments with collinear genes.

It should be noted that GeneTack predicts all types of frameshifts (sequencing errors, indel mutations, programmed frameshifts). In some cases additional analysis of the predictions reveals the true nature of the frameshift.

Chapter II

GENETACK: FRAMESHIFT IDENTIFICATION IN PROTEIN CODING SEQUENCES BY THE VITERBI ALGORITHM

A function to encode a protein imposes constrains on a genomic sequence. These constrains are phase dependent, for example, stop codons are avoided in one of the three reading frames, if one considers only one strand of genomic DNA or six for two strands. Because of these constrains it is possible to infer which reading frame is likely to be translated by analyzing sequence of a protein coding gene without prior knowledge regarding the sequence of that protein. Recently, we have designed an HMM-based computational approach for identification of locations in DNA where protein coding constrains transit from one frame to another.

2.1 GeneTack algorithm

The problem of predicting protein-coding regions has been successfully solved by the algorithms employing hidden Markov models (HMMs) [119, 120, 121]. Some of these algorithms include provisions for finding frameshifts (EcoParse [122], EasyGene [123], FrameD [124]). The accuracy of frameshift finding by these programs was not systematically assessed.

The logic of the GeneTack algorithm is as follows. The program takes as an input a fragment of a genomic sequence containing collinear genes in the direct strand. Such fragments could be selected based on gene predictions by GeneMarkS. Assuming that the frameshift may result in prediction of two (overlapping or not) adjacent genes located in the same strand, we designed GeneTack to discriminate between correctly

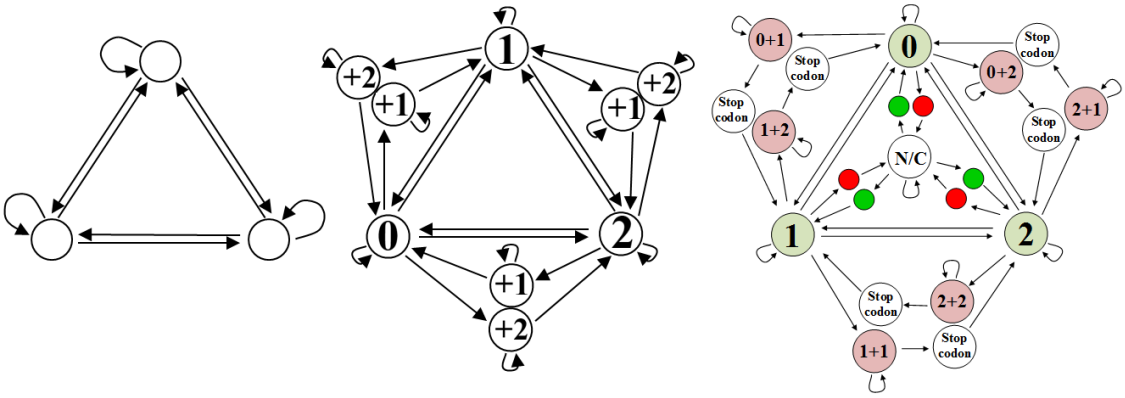


Figure 8: Development of the GeneTack HMM structure. The complexity of the model increased from a simple 3-state structure to the final version of the model consisting of 28 states – Fig. 9.

predicted ingenious adjacent genes and those adjacent genes that are predicted due to a sequence error and a split of a single gene by a frameshift.

To implement this idea we tried several different HMM structures increasing the complexity over time (Fig. 8). We have also tried different order of models (from 0 order to 5th order). In terms of average sensitivity and specificity we found that the best performance is observed for the 4th order model and the 28 state HMM (Fig. 9).

The algorithm uses a probabilistic model (HMM) that allows for three alternative scenarios: presence of true overlapping genes, true non-overlapping adjacent genes, and adjacent genes (overlapping or not) predicted due to the presence of a frameshift (Fig. 9).

The HMM consists of 28 states divided into four groups (Table 1): (i) states 1, 2 and 3 emit protein-coding sequence and correspond to the three possible "global" reading frames; (ii) the state denoted as n/c emits a non-coding sequence; (iii) six states designated as $i - j$, where $i, j = (1, 2, 3)$ and $i \neq j$, emit sequences where two adjacent genes overlap (here numbers i and j indicate the global frames of the upstream and downstream genes respectively); (iv) 18 states emitting nucleotides of start and stop codons (shown as triangles and squares on the diagram).

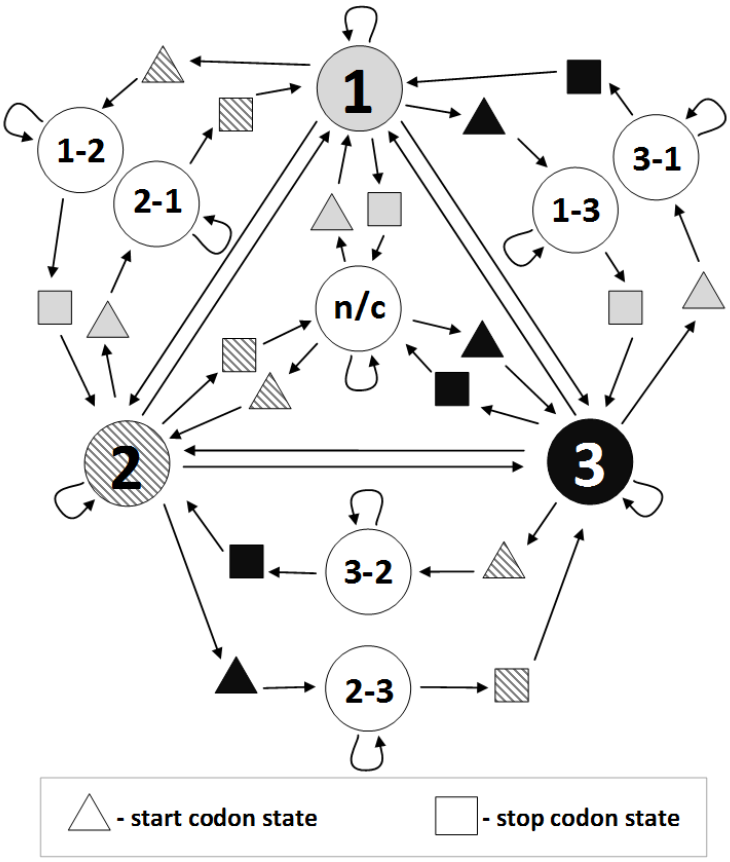


Figure 9: Hidden Markov model used in GeneTack consists of 28 states. States 1, 2 and 3 correspond to the three global frames of reading the genetic code. The type of shading of a state reflects its frame. There is no single frame related to n/c and overlap states, thus they have no shading.

Table 1: Types of states used in the GeneTack HMM and the properties of the emission probabilities.

Type	State(s)	Periodicity	Order
Coding states	1, 2, 3	3	4
Non-coding state	n/c	1	4
Start/stop states	9 start and 9 stop states	3	2
Overlapping states	"i-j", i,j = 1,2,3; (ij)	3	4

Each hidden state emits a single nucleotide. The emission probability of a nucleotide X depends on the type of hidden state as well as the nucleotides emitted earlier. If this probability depends on s previous nucleotides then the probability of emission of a nucleotide X from a hidden state k, $P_k(X)$ is numerically equal to the value of transition probability for the order s Markov chain model: $P_k(X_i) = P_k(X_i|X_{i-s}, X_{i-s+1}, ..., X_{i-1})$ defined for the state K. In the computations described below we used 4th order three-periodic Markov chains [124] as emission probabilities for the states 1, 2 and 3 (see Table 1).

The maximum likelihood path through the HMM with respect to the given DNA sequence is determined by the Viterbi algorithm [125]. This path makes a decoding of the DNA in terms of which nucleotide corresponds to which hidden state. Notably, the transitions between hidden states in the determined maximum likelihood path carry important information. Direct transitions between states 1, 2 and 3 correspond to frameshifts; transition between states 1, 2 and 3 through the n/c state(s) indicates a presence of non-overlapping adjacent genes; transition between states 1, 2 and 3 through $i - j$ states indicates a presence of overlapping adjacent genes (Table 2).

The initial (terminal) hidden state in the analysis of the DNA fragments with collinear genes is supposed to be either n/c or *start* or *stop* state. Input for GeneTack program includes a file with genome-specific parameters of the HMM on Fig. 10.

2.2 GeneTack-GM algorithm

For rather long genomic sequences where genes may reside in both strands we need an initial run of the self-training GeneMarkS program [119] to estimate the GeneTack-GM parameters and to make a parse of the whole sequence into fragments with collinear genes. The logic of operations of GeneTack-GM running on a new genomic sequence is shown on Fig. 10.

The GeneMarkS program runs in several iterations to determine parameters of

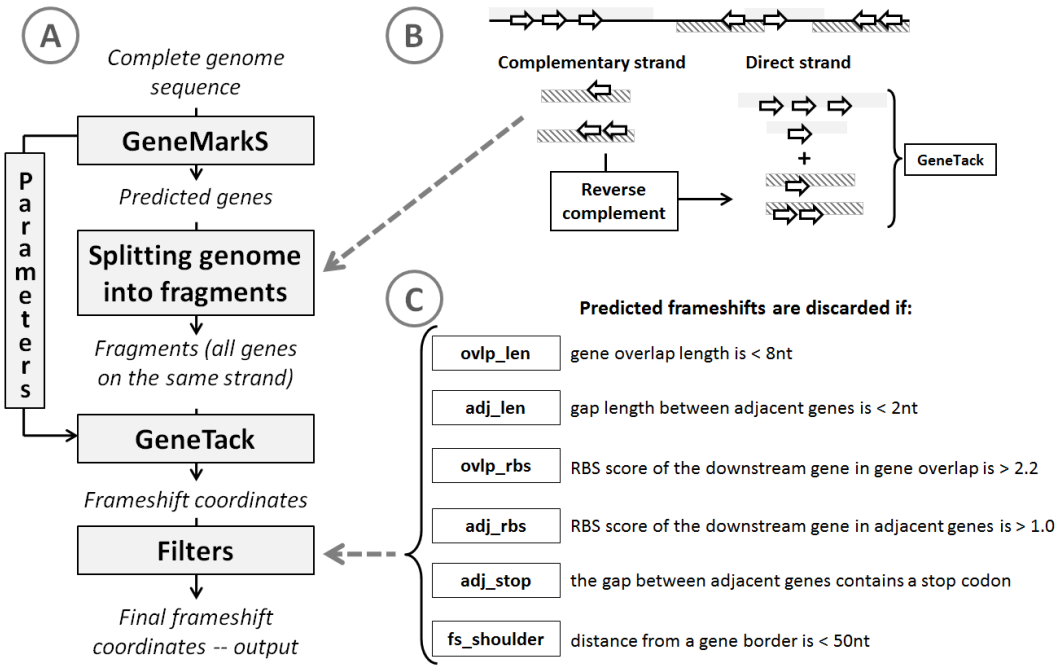


Figure 10: (A) GeneTack-GM algorithm overview; (B) splitting genome into fragments; (C) description of filters used to reduce number of false positive predictions (filters listed in the order they applied).

Markov chain models for protein-coding and non-coding regions [119] that will be used also in the GeneTack run. In the end of the training process GeneMarkS defines the final set of predicted genes. Since GeneMarkS is not designed to recognize frameshifts, so instead of a gene with a frameshift, a pair of adjacent genes in the same strand (overlapping or not) will be predicted. All sequence fragments that contain collinear genes predicted by GeneMarkS may contain a frameshift. The output of GeneMarkS is used to split the genomic sequence into genomic fragments (Fig. 10B) that carry collinear genes augmented by non-coding flanks on both sides (500nt or less). It is convenient to analyze sequences with gene in direct strand, therefore, reverse complements are used if the original fragments contain genes in complementary strand. Further, the GeneTack program is applied to each fragment to identify possible frameshifts. Finally, to reduce the number of false positive predictions, we apply several filters (Fig. 10C). The decision rules and parameters of the filters

Table 2: An example of emission probabilities calculation for overlap of genes carrying the genetic code in frames 1 and 2, (for the *1-2* hidden state). The pattern of frequencies (F_1, F_{12}, F_2) repeated for the whole sequence carrying overlapping genes is shown in bold font.

Position	0	1	2	3	4	5	6	7	8	9
Position % 3	0	1	2	0	1	2	**0**	**1**	**2**	0
Gene in frame 1	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1
Gene in frame 2					F_1	F_2	F_3	F_1	F_2	F_3
State *1-2*					F_{12}	F_2	F_1	F_{12}	F_2	F_1

were determined from the results of the analysis of GeneTack-GM predictions for the *E. coli* genome with artificially introduced frameshifts. The predicted frameshifts possessing the following features were discarded:

1. in the sequences carrying predicted non-overlapping gene pairs:
   (a) if the gap (an intergenic region) between genes is less than 2 nt;
   (b) if there is a stop codon in the upstream region (the gap length + 20 nt) to the start codon of the downstream gene (in the same frame);
   (c) if the score of the RBS for the downstream gene is larger than 1.0;
2. in the sequences carrying predicted overlapping gene pairs:
   (a) if the overlap region length is shorter than 8 nt;
   (b) if the score of the RBS for the downstream gene is larger than 2.2;
3. in the sequences from both A/ and B/ classes – if the predicted frameshift location is closer than 50 nt to a border of the coding region.

Our analysis has shown that the indicated values of parameters produce sufficiently accurate results for genomes with various GC% content (see below).

2.2.1 Parameter estimation

Emission probabilities for the states 1, 2, 3 and n/c, the coding and non-coding states, are defined in the run of GeneMarkS. To compile a standard training set for estimation of emission probabilities for overlapping regions is difficult since overlaps longer than 1 nt and 4 nt are rare in real genomic sequences. To overcome this difficulty we used the following heuristic model that uses emission probabilities of nucleotides defined for non-overlapping coding states. Presence of two overlapping genetic codes reduces probability of accumulating so-called neutral mutations (usually mutations in the third position of codon) because the mutation would also touch either the 1st or the 2nd position of a codon in another gene. In the model for the gene overlapping states the first and the second positions of a codon are considered as the "strong" ones and the third position as the "weak" one. We assume that in an overlapping region strong codon positions dominate weak positions, i.e. if the first (strong) position of the upstream gene overlaps the third (weak) position of the downstream gene, the emission probabilities typical for the first position will be used. Note that two weak positions never overlap. If two strong positions overlap (for example the 2nd position of upstream gene and 1st position of downstream gene) then an emission probability, F_{12}, is calculated as an average between two "strong" probability values:

$$F_{12}(\alpha|\text{prefix}) = \frac{1}{2}(F_1(\alpha|\text{prefix}) + F_2(\alpha|\text{prefix})) \quad (1)$$

where α is a nucleotide, *prefix* is a string of upstream letters with the length equal to the order of the Markov chain model of the coding region, F_1 and F_2 are probabilities of nucleotide emission from a coding state for the first and second codon positions respectively. This heuristic approach is illustrated in Table 2.

Another important group of the HMM parameters are transition probabilities between hidden states. The sum of probabilities of all outgoing transitions for each state must be equal to 1. Since the GeneTack HMM is symmetrical, only two transition

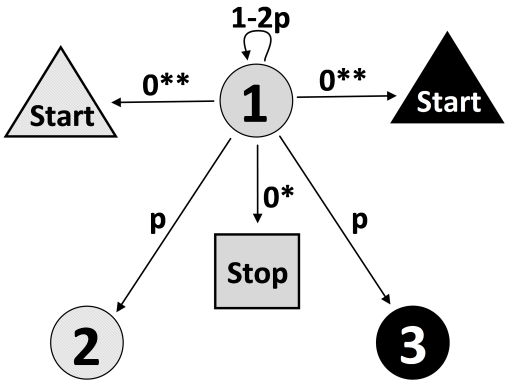


Figure 11: Calculation of transition probabilities for GeneTack HMM. * Transition probability to stop codon (upon approaching TAA, TAG, TGA) is 1. ** Transition probabilities to start codon in frames 2 and 3 (upon approaching ATG, GTG, TTG) are 0.0001 (0.001 for high GC genomes).

probabilities are needed to be defined: probability of transition between coding states (p) and probability of transition to the start codon (Fig. 11). As it was mentioned above a frameshift is predicted if a direct transition between coding states does occur. Therefore, the value of p can be interpreted as a probability of a frameshift. It can be different for different genomes because number of frameshifts can be different in different sequence data (e.g. number of indel sequencing errors depends on sequencing technique). The default value of parameter $p = 0.0006$ was chosen to minimize the frameshift prediction errors in experiments with the *E. coli* genomic sequences. Although we expect this value to be different for other genomes, the difference is apparently very small, given the comparable to the *E. coli* case figures of frameshift prediction accuracy in other genomes where we have used the same value p (see below).

Probability of return (i.e. transition from a state to itself) for a coding state is $1 - 2p$ (Fig. 11). All around, there are 10 circular transition probabilities defined for the 3 coding states, the n/c state and the overlap states. In the current implementation all of them have the same values.

Notably, each coding state has two more probabilities, the ones that control transition to the *stop* state of the same frame and the *start* state of the overlapping downstream gene. These transition probabilities are sequence dependent; the transition probabilities are equal to zero in each position in a sequence which does not complete a start or stop codon triplet. For example, with a sequence NNTGA and T in the first position of a codon, emission of the A is made from a *stop* state upon transition from preceding coding state with probability 1. Upon approaching a possible start codon, transition probability 0.0001 to start codon (0.001 for high GC genomes, see below) is used.

2.2.2 High-GC genomes

To analyze genomes with high GC content we have made two modifications. First, for genomes with GC content higher than 65% instead of 0.0001 we use 0.001 as the sequence dependent transition probability to a start codon state (Fig. 11). This choice reduces the number of false positive predictions in high GC genome where the frequency of AT reach triplets such as start (as well as stop) codons is lower than in low and mid GC genomes. The lower (0.0001) value of transition probability to start codon makes less likely prediction of gene overlap and forces the program to make frameshift predictions more frequently. Second, we have observed that for high GC genomes the parse of a genome into segments with collinear genes, as predicted by GeneMarkS, does not deliver all the candidates for frameshift detection. In some cases a gene split by a frameshift is interpreted by the GeneMarkS program as not as a pair of genes in the same strand, but, surprisingly as a pair of genes in different strands (Fig. 12).

This misinterpretation is explained as follows. First, in a gene in a high GC genome the third position of a codon is occupied by C or G nucleotides in 80 to 90% of cases. Thus, the reading frame in the complementary strand which mirrors

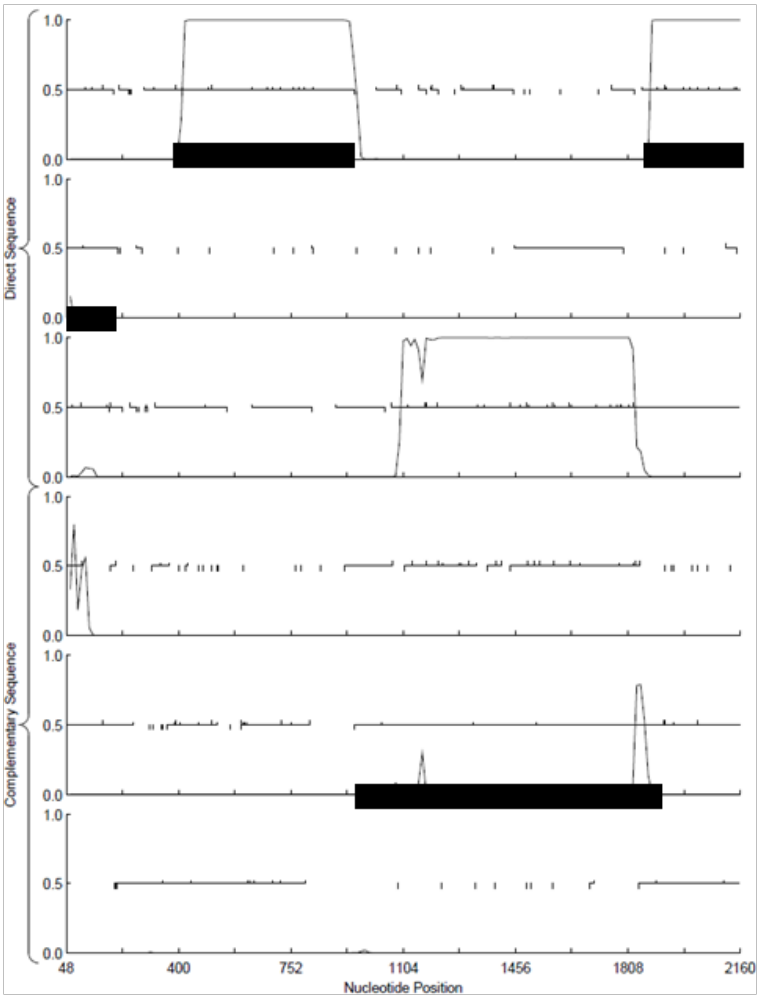


Figure 12: An example of the GeneMarkS gene prediction for a gene with simulated frameshift in a high GC genome. The figure shows the coding potentials in all six frames as determined by the GeneMark program. Black bars on the horizontal axis indicate predicted genes. A frameshift was introduced at position 1,848 in the gene on direct strand. There is a clear jump of coding potential from frame 3 to frame 1 at the location where the frameshift was introduced. However, there is a gene predicted in frame 2 on the opposite strand. Such artifacts are corrected by the modification in GeneTack-GM for high GC genomes as described in the text.

the reading from of a true gene has a strong three-periodicity of C and G as well. Second, with diminished frequency of stop codons, we observe long ORFs that occur by chance; an appearance of such ORF in the mirror frame in the complementary strand makes a candidate for false positive gene prediction. Third, interruption of the true gene by a frameshift does not preclude an initial reading frame from continuation to a significant distance, 100 to 200 nucleotides, until a stop codon type triplet would occur at random. Since a coding potential exists only in a true coding section of this elongated ORF it can be omitted by the Viterbi algorithm in favor of the shorter but actually non-coding ORF with lower coding potential located in the complementary strand. Thus, the pair of genes predicted in the place of a frameshifted gene turns out to be not a collinear gene pair; moreover one of the predicted genes has no coding region at all. This frameshift related prediction of a gene in a wrong strand poses an obvious problem. Now the parse of a sequence (Fig. 10B) splits the pair of coding regions originated from a frameshifted gene by placing them into separate sequence fragments, thus making the frameshift detection impossible. Such outcomes result in a drop in Sensitivity observed in the computational modeling.

To deal with the problem we have modified the parsing procedure for high GC content genomes. We use the output of the GeneMark program [124] to calculate an average coding potential for each gene predicted by GeneMarkS. If an average coding potential for a predicted gene is less than 0.4 while it is larger than 0.6 for an ORF in the opposite strand, the thresholds 0.4 and 0.6 chosen heuristically, we reassign the predicted strand of a gene. This reassignment effectively elongates the upstream part of the fragment with collinear genes and thus includes the earlier missed position of a potentially frameshifted gene into the sequence fragment for the GeneTack analysis.

2.3 Results

2.3.1 Datasets

The accuracy of the GeneTack-GM predictions was assessed on 17 prokaryotic genomes with GC content ranging from 28% to 75% (note that the *E. coli* genome, that was used to estimate program parameters, is not included in the dataset in order to keep training and test datasets separate). From this set we generated datasets to test program performance at different gene length ranges.

Dataset_1000 included 17 genomes with frameshifts simulated in 400 genes longer 1,000 nt. Dataset_600_1000 included 17 genomes with frameshifts simulated in 200 genes with length ranges from 600 to 1,000 nt.

In both datasets frameshifts were simulated by insertion of a single nucleotide into a randomly selected gene at a random position located at a distance of at least 180 nt (dataset_1000) or 100 nt (dataset_600_1000) from either gene end. The accuracy of the frameshift detection for the case when a frameshift was located closer than 100 nt to one of the gene borders was studied separately (see below).

2.3.2 GeneTack-GM performance: comparison with other programs

GeneTack-GM as well as two earlier developed frameshift prediction programs, FrameD [111] and FSFind [112] were applied to both dataset_1000 and dataset_600_1000.Coordinates of predicted frameshifts were compared with precisely known coordinates of simulated frameshifts. A predicted frameshift was considered as a true positive (TP) if it was located not farther away than 50 nt from the real frameshift, otherwise prediction was considered as false positive (FP). A simulated frameshift was classified as "not found", a false negative (FN) prediction case, if no predicted frameshifts were reported in the 50 nt vicinity of the simulated frameshift.

Program performance was characterized by conventional characteristics Sensitivity (Sn) and Specificity (Sp). The value of Sn is defined with respect to the actual number

of simulated frameshifts and the value of Sp is defined with respect to the number of predictions made.

$$Sn = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TP}{TP + FP} \tag{3}$$

To compare the GeneTack performance with the performance of the FrameD program genomic sequences with artificial errors were submitted to the FrameD web server (http://bioinfo.genotoul.fr/apps/FrameD/FDM.pl) for model generation. The models were used in a local copy of the FrameD program. The FSFind program was installed and run on the local server.

For all 17 genomes in dataset_1000 GeneTack-GM has shown better average values of $(Sn + Sp)/2$ than FrameD and FSFind (run in the *ab initio* mode) with margins of 8,7% and 11.2% respectively (Table 3). Notably, every genome in the test set could contain additional inherent frameshifts. For instance, the *E. coli* genome contains 33 annotated programmed frameshifts. Since we could not know the locations of additional frameshifts, we considered the frameshifts predicted in the locations not coinciding with the artificial frameshifts as false positive for all programs. Therefore, the actual performance of each tested program could be even better in terms of Specificity than it appears in Table 3. Additional comparison was done with the FSFind program running in the mode of verification of predicted frameshifts via BLAST analysis by search for similarity to tentative translations of frameshifted genes [112] in the nr database. This step improves the Sp value. Still the overall average $(Sn + Sp)/2$ on 17 genomes is not as high as we have observed for GeneTack run in the purely *ab initio* mode (Table 3).

The data on program performances on dataset_600_1000 are shown in Table 4. It indicates that performance of the same set of program, though reduced, is ranked in

Table 3: Frameshift prediction accuracy estimation for 17 prokaryotic genomes (sorted by GC content) each containing 400 genes longer than 1000 nt with simulated frameshifts (dataset_1000). The Sn and Sp values were calculated for GeneTack-GM, FrameD, FSFind and FSFind-BLAST programs. The programs were compared based on average sensitivity and specificity $(Sn + Sp)/2$. Bold numbers indicate the best performance. *FSFind-BLAST results for *R. solanacearum* were not available because of a runtime error, thus the average values were computed for 16 genomes.

	GC %		GeneTack-GM		FrameD		FSFind		FSFind-BLAST	
Methanosphaera stadtmanae	28	Sn	71.3	77.2	62.5	72.5	65.5	72.4	64.5	**77.5**
		Sp	83.1		82.5		79.2		90.5	
Campylobacter jejuni	31	Sn	81.7	**73.3**	60.2	60.1	64.9	63.2	63.4	71.9
		Sp	64.9		60.0		61.5		80.3	
Staphylococcus aureus Mu50	33	Sn	79.8	**80.1**	49.5	68.4	63.0	69.7	60.5	75.6
		Sp	80.4		87.2		76.4		90.6	
Picrophilus torridus	36	Sn	83.8	75.1	68.0	64.4	84.8	74.8	85.3	**85.5**
		Sp	66.3		60.7		64.7		85.7	
Streptococcus pyogenes	39	Sn	77.3	**75.8**	42.0	61.4	58.8	67.0	56.8	72.2
		Sp	74.3		80.8		75.1		87.6	
Pasteurella multocida	40	Sn	83.8	**81.9**	54.8	71.6	73.5	77.8	70.8	81.5
		Sp	80.0		88.3		82.1		92.2	
Bacillus subtilis	44	Sn	79.5	**71.4**	40.5	52.3	62.0	58.2	60.3	66.1
		Sp	63.2		64.0		54.4		71.9	
Thermotoga maritima	46	Sn	82.8	76.9	77.5	72.8	76.0	67.1	73.3	**78.3**
		Sp	71.0		68.1		58.1		83.2	
Archaeoglobus fulgidus	49	Sn	89.3	68.3	70.0	60.0	82.5	65.4	81.0	**77.5**
		Sp	47.2		50.0		48.3		74.0	
Pyrobaculum aerophilum	51	Sn	85.2	64.7	60.3	46.8	61.4	55.4	54.6	**65.7**
		Sp	44.2		33.2		49.3		76.8	
Thermococcus kodakaraensis	52	Sn	86.0	81.2	77.8	73.8	78.5	75.0	76.8	**83.0**
		Sp	76.3		69.7		71.4		89.2	
Salmonella typhimurium	52	Sn	85.3	71.8	64.5	66.5	75.5	70.3	74.0	**79.3**
		Sp	58.2		68.4		65.1		84.6	
Methanopyrus kandleri	61	Sn	87.0	**73.3**	74.2	66.6	72.9	55.8	70.7	58.1
		Sp	59.5		59.0		38.6		45.4	
Ralstonia solanacearum	67	Sn	93.0	**88.5**	95.0	86.5	79.8	70.1	n/a*	n/a*
		Sp	84.0		78.0		60.4		n/a*	
Caulobacter crescentus	67	Sn	96.0	**87.3**	95.5	82.8	86.3	65.6	83.5	70.1
		Sp	78.5		70.1		44.9		56.7	
Clavibacter michiganensis	73	Sn	98.5	**81.0**	98.3	78.6	66.5	59.2	61.0	61.3
		Sp	63.4		58.9		51.8		61.6	
Anaeromyxobacter dehalogenans	75	Sn	97.3	**82.5**	98.0	77.3	59.8	53.0	52.0	56.3
		Sp	67.7		56.6		46.1		60.5	
AVERAGE:		Sn	85.7	**77.1**	69.9	68.4	71.3	65.9	68.0	72.5*
		Sp	68.4		66.8		60.4		76.9	

Table 4: Frameshift prediction accuracy estimation for 18 prokaryotic genomes (sorted by GC content) each containing 400 genes of length between 600 and 1000 nt with simulated frameshifts (dataset_600_1000). The programs were compared based on average sensitivity and specificity (Sn+Sp)/2. Bold numbers indicate the best performance. *FSFind-BLAST results for *R. solanacearum* were not available because of a runtime error, thus the average values were computed for 17 genomes.

	GC %		GeneTack-GM		FrameD		FSFind		FSFind-BLAST	
Methanosphaera stadtmanae	28	Sn	71.3	77.2	62.5	72.5	65.5	72.4	64.5	**77.5**
		Sp	83.1		82.5		79.2		90.5	
Campylobacter jejuni	31	Sn	81.7	**73.3**	60.2	60.1	64.9	63.2	63.4	71.9
		Sp	64.9		60.0		61.5		80.3	
Staphylococcus aureus Mu50	33	Sn	79.8	**80.1**	49.5	68.4	63.0	69.7	60.5	75.6
		Sp	80.4		87.2		76.4		90.6	
Picrophilus torridus	36	Sn	83.8	75.1	68.0	64.4	84.8	74.8	85.3	**85.5**
		Sp	66.3		60.7		64.7		85.7	
Streptococcus pyogenes	39	Sn	77.3	**75.8**	42.0	61.4	58.8	67.0	56.8	72.2
		Sp	74.3		80.8		75.1		87.6	
Pasteurella multocida	40	Sn	83.8	**81.9**	54.8	71.6	73.5	77.8	70.8	81.5
		Sp	80.0		88.3		82.1		92.2	
Bacillus subtilis	44	Sn	79.5	**71.4**	40.5	52.3	62.0	58.2	60.3	66.1
		Sp	63.2		64.0		54.4		71.9	
Thermotoga maritima	46	Sn	82.8	76.9	77.5	72.8	76.0	67.1	73.3	**78.3**
		Sp	71.0		68.1		58.1		83.2	
Archaeoglobus fulgidus	49	Sn	89.3	68.3	70.0	60.0	82.5	65.4	81.0	**77.5**
		Sp	47.2		50.0		48.3		74.0	
Pyrobaculum aerophilum	51	Sn	85.2	64.7	60.3	46.8	61.4	55.4	54.6	**65.7**
		Sp	44.2		33.2		49.3		76.8	
Thermococcus kodakaraensis	52	Sn	86.0	81.2	77.8	73.8	78.5	75.0	76.8	**83.0**
		Sp	76.3		69.7		71.4		89.2	
Salmonella typhimurium	52	Sn	85.3	71.8	64.5	66.5	75.5	70.3	74.0	**79.3**
		Sp	58.2		68.4		65.1		84.6	
Methanopyrus kandleri	61	Sn	87.0	**73.3**	74.2	66.6	72.9	55.8	70.7	58.1
		Sp	59.5		59.0		38.6		45.4	
Ralstonia solanacearum	67	Sn	93.0	**88.5**	95.0	86.5	79.8	70.1	n/a*	n/a*
		Sp	84.0		78.0		60.4		n/a*	
Caulobacter crescentus	67	Sn	96.0	**87.3**	95.5	82.8	86.3	65.6	83.5	70.1
		Sp	78.5		70.1		44.9		56.7	
Clavibacter michiganensis	73	Sn	98.5	**81.0**	98.3	78.6	66.5	59.2	61.0	61.3
		Sp	63.4		58.9		51.8		61.6	
Anaeromyxobacter dehalogenans	75	Sn	97.3	**82.5**	98.0	77.3	59.8	53.0	52.0	56.3
		Sp	67.7		56.6		46.1		60.5	
AVERAGE:		Sn	85.8	**77.0**	69.5	68.2	71.5	66.1	68.3	72.8*
		Sp	68.2		66.9		60.6		77.3	

the same way. Specifically, the observed $(Sn + Sp)/2$ values are 77.0%, 68.2%, 66.1% and 72.8% for GeneTack, FrameD, FSFind and FsFind-BLAST, respectively.

2.4 Discussion

2.4.1 Can GeneTack predict programmed frameshifts?

We applied GeneTack to the 23 DNA sequences, from 19 different species, retrieved from the RECODE database [126] containing +1 and -1 annotated programmed frameshifts. GeneTack successfully predicted annotated frameshifts in 18 sequences. The five sequences where GeneTack did not predict frameshifts had in fact no frameshifts on DNA level; in all five cases the coding region lengths were multiples of three. Notably, the notion of a frameshift was used by the authors [126] in a general sense: shifting the reading frame in the process of translation. In these five cases the ribosome could either translate a gene from start to end, or, under certain conditions, the ribosome could change the frame at a certain point and quickly get to a stop codon. This type of translation regulation has been experimentally observed and was documented in RECODE. This case study indicates that GeneTack can be used in a pipeline for prediction of programmed frameshifts. The pipeline could also contain filters to decrease the number of false positives by checking for presence of signal sequences in the vicinity of programmed frameshifts.

2.4.2 Insensitivity zones

It is difficult to detect frameshifts located close to the gene start or end; thus we have defined two insensitivity zones for GeneTack at the borders of a gene. To determine the characteristic length of the insensitivity zone (expected to be of about the same size at both ends) we conducted the test on individual genes flanked by 500 nt of non-coding sequence and with frameshifts introduced at a distance from the gene border ranging with step 5 nt from 1 to 200 nt. The analysis was done for 400 genes from the *E. coli* genome longer than 1,000 nt (Fig. 13).

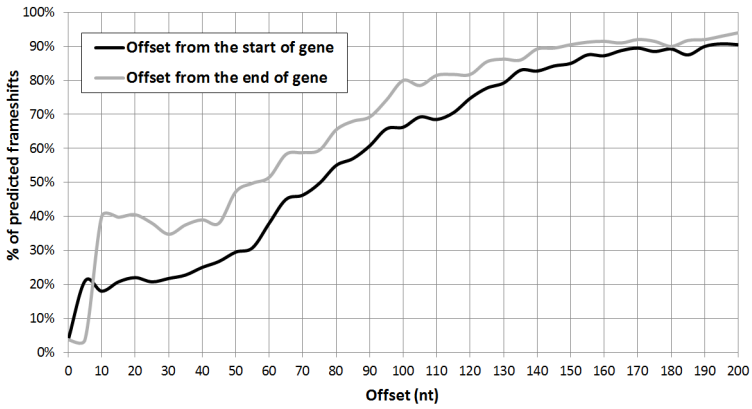


Figure 13: Dependence of the number of correctly predicted frameshifts on the distance from the artificially made frameshift to the gene border (either start or end).

It is seen that GeneTack correctly detects frameshifts with offset 60 nt from the gene end in $\approx 50\%$ of genes and frameshifts with offset 75 nt from the gene start also in $\approx 50\%$ of genes. The accuracy increases steadily as the offset grows and at 180 nt the performance reaches saturation (>90%). The length of 180 nt was chosen as the minimal distance for a simulated frameshift from the gene borders in the accuracy tests described above.

We observed (Fig. 13) that GeneTack is able to detect frameshifts located close to the gene end better than the ones simulated in the beginning of a gene. The observation can be explained as follows. We need to show that it is easier to predict adjacent genes (overlapping or not) if a frameshift is located at a given distance downstream from a true start of a gene than if a frameshift is located at the same distance upstream to the gene end. We have to consider the expected distance L_s from a frameshift down to the random stop codon forming the short upstream gene in the adjacent gene pair in the "start" case. On the other hand, we have to consider the expected distance L_e from a frameshift up to the random start codon forming the downstream gene in the gene pair in the "end" case. Obviously, with three stop codons and one (two) start(s) in the genetic code L_s is smaller than L_e. Therefore, in the "start" case a larger part of a short gene in the gene pair will be occupied by the

true coding region. Thus, the chance of predicting the short gene making the gene pair (hence no frameshift) is larger in the "start" case.

2.4.3 Filter effectiveness

We have assessed a filter performance by the percentage of eliminated false positive predictions and the percentage of true positive predictions it keeps in the list. These values were calculated for each of the 17 genomes from the dataset_1000 plus *E. coli* genome (Fig. 14).

On average the filters remove 72% of false positives and keep 91% of true positives initially predicted by GeneTack (Fig. 14B). At the same time filters have different effectiveness for different genomes.

One of the reasons for the variability in effectiveness is that the same filter parameters, optimized for *Escherichia coli* genome, are used for genomes with different GC content.

Also, the level of conservation of the RBS site is variable between genomes. There are two filters, ovlp_rbs and ajd_rbs (Fig. 10C) that rely on the RBS score determined by GeneMarkS; these two filters do not work efficiently for genomes with weak RBS. For example, for *Pyrobaculum aerophilum*, the species that has a weak RBS for genes inside operons and no RBS at all for the first genes in operons due to the use of leaderless transcripts, GeneTack-GM predicts the largest number of false positive frameshifts, with only 49% of false positives filtered out. Similarly, these two filters work poorly for *Archaeoglobus fulgidus* with adj_rbs filtering out 26 FP and 10 TP. In contrast, for *Thermococcus kodakaraensis* and *Thermotoga maritima* the ovlp_rbs and ajd_rbs filters remove more than 84% of false positive predictions. Together, these two filters eliminate 217 false positives and 16 true positives frameshifts for *Thermococcus kodakaraensis* and 286 false positives and 19 true positives for *Thermotoga maritime*.

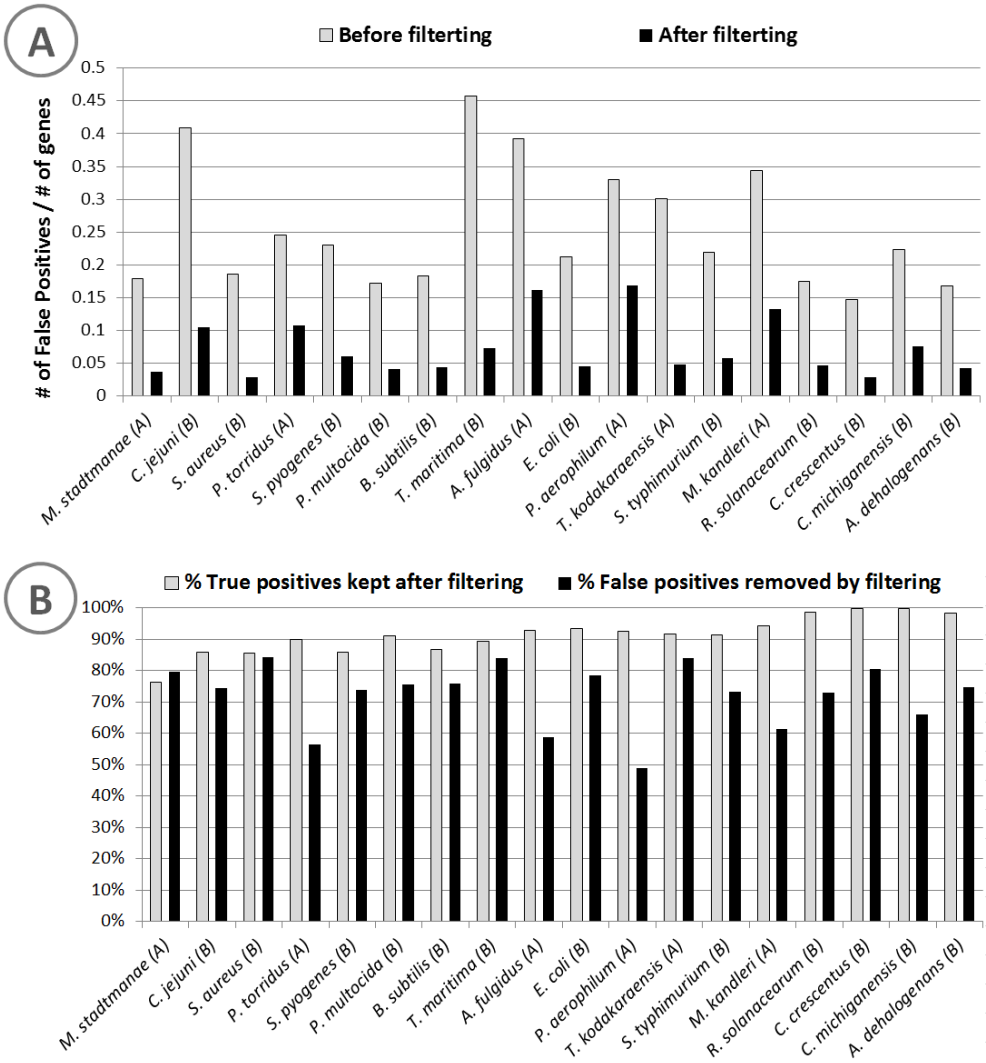


Figure 14: Performance of the filters for 18 prokaryotic genomes (genomes are shown along the X axis, sorted by GC content). A domain of life is indicated in parenthesis ("A" stands for Archaea and "B" for Bacteria) (A) Filtering false positive predictions. The fraction of false positives with respect to the total number of genes in a genome, before (gray bars) and after (black bars) filtering are shown for each species. (B) Relative impact of filtering on true positives and false positives. For each genome percentages of removed false positives (with respect to false positives before filtering) and kept true positives (with respect to the number of true positives before filtering) are shown. The filters are supposed to remove as many false positives and as few true positives as possible. Thus, the sum of heights of two bars reflects the filters performance for a given genome. The best performance was observed for *Caulobacter crescentus*, the worst performance was for *Pyrobaculum aerophilum*.

The GeneTack-GM program can be adapted for analysis of other genomic sequences with intronless genes, such as metagenomic sequences as well as EST sequences. For metagenomic sequences GeneTack-GM can use the heuristic models [120] that allow for quite accurate gene prediction in short sequences, i.e. without a knowledge of full genomic context for estimating parameters of the three-periodic Markov chain model of the coding region. For the EST sequences, that belong to one and the same species, the training procedure of GeneMarkS has been modified to account for the Kozak pattern at the gene start (Ter-Hovhannisyan and Lomsadze, unpublished). The models thus derived in the training on the sequenced transcripts (EST) can be immediately used to run GeneTack to detect the frameshifts. Note that alignment of EST sequences to genomic sequence helps to correct majority of frameshifts. Still, the genome projects that focus on sequencing EST only will benefit from using GeneTack-GM to correct the gene and protein predictions.

Chapter III

IDENTIFYING THE NATURE OF READING FRAME TRANSITIONS OBSERVED IN PROKARYOTIC GENOMES

3.1 Introduction

In the present work we apply a phylogenetic approach to understand and classify frame transitions observed in bacterial genomes. This is conceptually similar to a recent study where such an approach was used to classify bacterial genes annotated in GenBank as having disrupted ORFs [23].

In this study we identified 206,991 genes with frame transitions (fs-genes) in 1,106 complete bacterial and archael genomes screened by the *ab initio* frameshift prediction program GeneTack. Using comparative sequence analysis to detect phylogenetic conservation, we were able to cluster 102,731 fs-genes and to classify many clusters with respect to the likely nature of frame transitions and, particularly, to produce a set of candidate recoded genes. We also experimentally tested several candidate recoding cassettes in *E. coli*. The results suggest that our dataset of recoding candidates is significantly enriched with *bona fide* recoded genes.

3.2 Results

3.2.1 The set of frameshifts predicted in 1,106 genomes

On April 12, 2010, 1,106 genomes (76 Archaeal and 1,030 Bacterial) longer than 1Mb were downloaded from the NCBI web site[1] (draft genomes were excluded). The GeneTack program [127] with default settings was applied to all 1,106 genomic sequences.

[1] ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.gbk.tar.gz

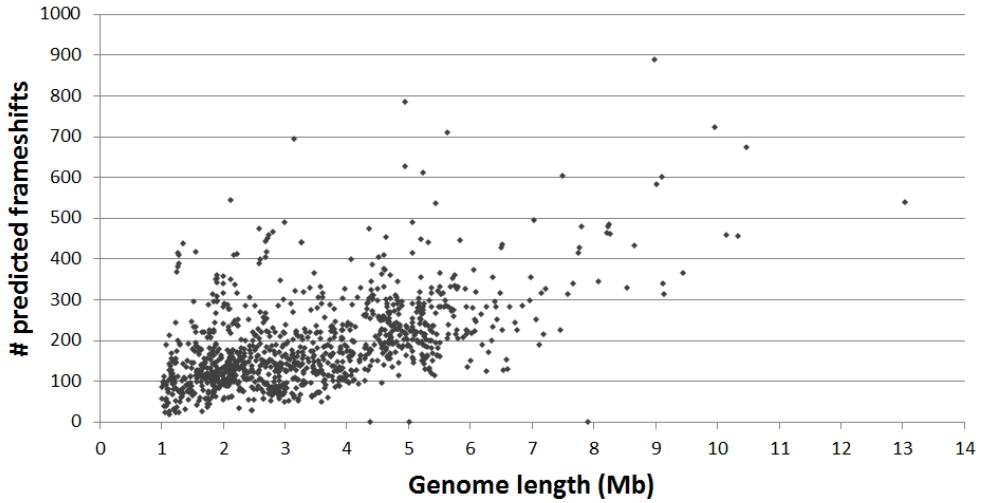


Figure 15: Number of frameshifts predicted in a prokaryotic genome correlates with genome length (data from analysis of 1,106 genomes). Total number of predicted frameshifts was 206,991. Genomes shorter than 1 Megabase were not considered as not possessing sufficient amount of sequence for the GeneMarkS self-training.

In total 206,991 frameshifts were predicted (see Fig. 15). Since the GeneTack accuracy in frameshift detection is characterized by 85.8% Sensitivity and 32.8% False discovery rate, we should have expected about 1/3 of the predictions to be related to i/ frame transition between adjacent overlapping genes (type A) while 2/3 of the frameshift predictions would be related to ii/ sequencing error, indel mutation or functional programming frameshift (type B). BLASTp and Pfam searches were used to delineate predictions of the later type (see Methods and Fig. 16). The total number of frameshifts with downstream ORF RBS score -1 or less was 40,544. Many of them were likely to be caused by type B frame transitions.

For 36,668 fs-genes, their extended fs-proteins had a similar protein in the nr database detected by BLASTp; also in 16,307 fs-genes Pfam domains covering the predicted frameshift were detected; finally both continuous BLASTp hits and Pfam domains existed for 10,434 fs-genes. Thus, only 17.7% of all frameshift predictions made were identified by BLASTp as type B while the expected percentage of such GeneTack predictions was 68%. Both type A and type B fs-genes and fs-proteins

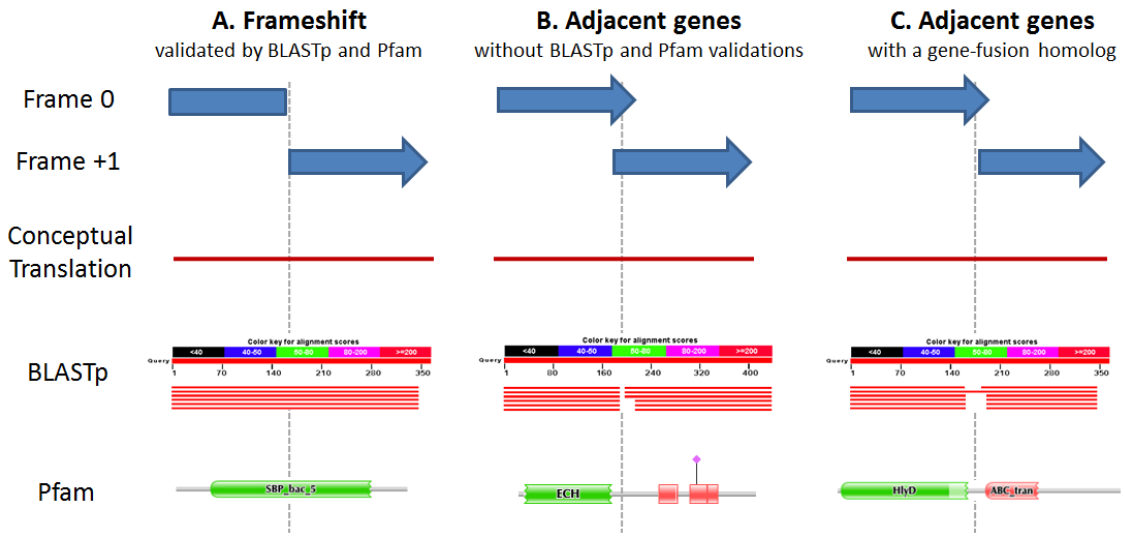


Figure 16: Possible outcomes of the BLASTp and Pfam searches for a conceptual translation of fs-gene. If a frameshift position is covered by BLASTp hit (or Pfam domain) the predicted frameshift is considered to be validated by BLASTp (or Pfam) and is likely to be a true positive prediction.

were used in the analysis described below.

3.2.2 About 50% of fs-genes could be clustered

One of our main goals was to determine the nature of correctly predicted frameshifts: was it a sequencing error, a pseudogenization mutation, a programmed frameshift, or a phase variation?

We assume that the presence of frameshifts in several homologous genes provides evidence against sequencing errors being the explanation for their frameshift categorization. A clustering procedure described in Methods identified 19,430 clusters with 102,731 fs-genes. The remaining fs-genes (about 50% of the total number of predicted frameshifts) did not form clusters; these "singletons" are likely to be sequencing errors and will be discussed below.

The majority of the 19,430 clusters contained a small number of fs-genes, 48% of the clusters (9,282) contained only two fs-genes while close to 75% of the clusters (14,441) contained less than five fs-genes (35,775 fs-genes total). The abundance of

Table 5: Examples of known frameshift prone patterns: PRF - programmed ribosomal frameshifting, PTR - programmed transcriptional realignment. *DNA alphabet with standard symbols for ambiguous bases is used for convenience. Note that the table gives just a few examples of frameshift sites and not all genes are listed.

Shift-prone patterns*	Efficiency	Genes
Y.TT_T.RA_N PRF [1]	up to 50%	*E. coli*: *prfB* gene (Release Factor 2)
TTC_CCC_TGA PRF [33, 118]	up to 60%	*E. coli*: *pheL* gene and artificial constructs
AGG_AGG PRF [129, 130]	up to 50%	*E. coli*: frameshifting efficiency depends of the level of gene expression
CG.A_AA.G PRF [71]	16% (*cdd* gene)	*B. subtilis*: *cdd* gene [71], Insertion sequence IS1222 [131]
A_AA.A_AA.R PRF [61, 132]	up to 40%	*E. coli*: *dnaX* gene; Bacteria: transposable elements of IS3 family
A_n, n >7; +1 or -1 PTR [62]	50%	*T. thermophilus*: *dnaX* gene
A_n, n >7; -1 PTR [16, 62]	?	*C. thermocellum*: IS120 (IS3 family)
A_n, n >7; +1 or -1 PTR [74]	up to 70%	*S. flexneri*: *mxiA*, *spa13* and *spa33* genes
T_n, n >8; -1 PTR [75]	30%	*S. flexneri*: *mxiE* gene

small clusters was certainly a result of use of the stringent BLASTp threshold. Some small clusters could be formed due to predicting frameshift at a gene overlap made from fission of a single gene into a gene pair with conserved co-location in several species [128].

Also, a number of clusters with up to several dozen fs-genes with very similar or even identical sequences, originated from several closely related genomes such as complete genomes of 30 *E. coli* strains. Some fs-genes were predicted in several copies in the same genome (e.g. genes for transposases).

3.2.3 Clusters identified as programmed frameshift clusters

A cluster of fs-genes with characteristic conserved nucleotide sequence motifs located uniformly close to predicted frameshift positions (Table 5), was classified as a cluster of fs-genes with a programmed frameshift. The conserved motifs were identified by alignment of the frameshift box sequences using the Gibbs Sampler method (see Methods). This approach, as expected, detected several known families of genes with programmed frameshifts; the corresponding conserved motifs were identified.

Many known "slippery" sequences include poly-A/T stretches (such as A_AAA_AAG [118, 133] and A_AAA_AAA [132] implicated in PRF or A_n, n >7 [74] and T_n, n >8 [75] involved in PTR). Poly-A/T sequences are prone to frameshifting during translation, transcription or even replication (as DNA polymerase may produce indel errors at poly-A/T stretches [134]).

Among clusters containing at least five fs-genes we found 145 clusters in which at least 50% of the fs-genes contained one of the seven heptamers mentioned above in Methods. With addition of the cluster of *prfB* genes we had 146 programmed frameshift clusters with 4,302 fs-genes that was divided into two groups: i/ clusters of fs-genes with known programmed frameshifts (Table 7) and ii/ new clusters of fs-genes predicted to use programmed frameshifts (Table 8).

3.2.4 Genes with known programmed frameshifts

The recent work by Sharma et al. [23] has extended the Recode database collection of prokaryotic genes with known programmed frameshifts ([126], ≈1,500 entries), the largest record of confirmed recoding events. Programmed frameshifts identified by Sharma et al. [23] were found by an "all against all" search among protein products of disrupted CDS annotated in prokaryotic genomes. Further, the homologous proteins were grouped into clusters without taking into account position and direction of frameshifts. Further tBLASTn searches against the NCBI nr database were used to

Table 6: Correspondance between GeneTack clusters and clusters from Sharma et al. established based on BLASTn search (e-value threshold 10^{-20}).

GeneTack cluster IDs	Largest cluster name	Sharma et al.
667870043;495557484;178902778	HTH_Tnp_IS630 (Transposase)	1
474411093	Release Factor 2	2
188472814	DDE_Tnp_1 (Transposase)	5;10;16;21;38;46;60
667870043	HTH_Tnp_IS630 (Transposase)	6;7;40;62
858558073	HTH_Tnp_IS630 (Transposase)	7
696263973;435865080	DDE_Tnp_IS1 (Transposase)	8
919140783	DDE superfamily endonuclease	9
910763088	DDE_Tnp_1 (Transposase)	10
675840861	HTH_Tnp_1 (Transposase)	11;13;19;31;33;35
241541714	HTH_Tnp_1 (Transposase)	12
777059633;282094684	DDE_Tnp_ISAZ013 (Transposase)	15
665826121	Hypothetical protein	17
992341191	rve	23
888244788	DDE_Tnp_1 (Transposase)	24
255701500	Hypothetical protein	25
869047494	DDE_Tnp_1 (Transposase)	28
928695812	Transposase, IS4 family	36
405503343	Phage integrase / recombinase	43
952432539	rve	44

enrich clusters by disrupted protein coding regions not annotated as such. Over all 49 clusters with 8,032 genes with programmed frameshifts were identified.

To establish correspondence between clusters of fs-genes with programmed frameshifts identified by Sharma et al. [23] and clusters defined in the current work, we ran BLASTn with each of 8,032 genes [23] as a query to search a database composed of 5,632 fs-genes that belonged to GeneTack clusters of fs-genes predicted to contain programmed frameshifts.

Among 146 GeneTack clusters of fs-genes with putative programmed frameshifts, we observed 14 clusters with significant sequence similarity to fs-genes in 31 Sharma et al. clusters. Finding these 14 clusters by the *ab initio* method serves as supportive evidence for effectiveness of the method. Note that 10 out of the 14 clusters (Table 7) belong to a group of the 12 largest clusters from the set of 146 (ranging from 1,699 fs-genes down to 36 fs-genes). Among the 14 clusters, 13 clusters contained transposase genes and one cluster contained genes encoding Release Factor 2 (see Table 6).

3.2.4.1 Genes of transposases

Besides the 13 transposase clusters matching clusters of Sharma et al., we identified six GeneTack DDE_Tnp_1 (Transposase) clusters (three with +1 and three with -1 frameshifts). Only three of them (the largest ones) matched corresponding Sharma et al clusters. Two clusters with a -1 frameshift (with 29 and 6 fs-genes) and one with a +1 frameshift (6 fs-genes) that did not have a match, could be new branches in the family of transposase genes utilizing programmed frameshifting. In total, there were 7 new clusters of transposase genes containing a relatively small number of fs-genes (from 5 to 29).

```
Thermus_thermophilus*              ------GAGGGAG--AAAAAAAAAGCC---TGA------- 22
Chlorobium_tepidum                 ---TCGGCGGACG--AAAAAAAAAAGCT--TGAGCCT--- 30
Prosthecochloris_aestuarii         -----GGCCGCCGGTAAAAAAAAAAGCCCCTGAAC----- 30
Chlorobium_luteolum                ----CAGGAGCCT-CAAAAAAAAAAGCCCCTGATG----- 30
Pelodictyon_phaeoclathratiform     ----CTGCAGGCT-CAAAAAAAAAAGCACCTGACA----- 30
Capnocytophaga_ochracea            ----TGAGGGGGA-TAAAAAAAAAA-----TGATGGACTT 30
Coraliomargarita_akajimensis       ---GGTGCGGACG-AAAAAAAAAAGTCCAGTGAT------ 30
Flavobacterium_johnsoniae          ACTTTGATAGA----AAAAAAAAAGTTGAGCAAT------ 30
Flavobacterium_psychrophilum       CTTTTGATAGA----AAAAAAAAAGCTAAACAAT------ 30
Cytophaga_hutchinsonii             ----TGGAAGACC-TAAAAAAAAAAGTAAATAAAC----- 30
                                         *     *********        *
```

Figure 17: Alignment of poly-A motifs from the 9 fs-genes of DNA polymerase III cluster with frameshift motif from *Thermus thermophilus* for which transcriptional realignment was previously shown.

3.2.4.2 Release Factor 2 (prfB genes)

GeneTack detected 428 frameshifted genes encoding bacterial Release Factor 2 (RF2). Expression of this genes utilizes one of the best known cases of programmed frameshifting [1]. All these genes joined a single cluster, which should be even larger in size since it was estimated that about 70% of all eubacteria utilize programmed frameshifts to regulate expression of RF2 gene [7]. GeneTack may not predict frameshifts in some RF2 genes where the frameshifts are located less than 50nt from the gene start.

3.2.4.3 DnaX genes: the analysis reveals that not all dnaX genes are encoded in a single ORF

Another well-known gene family with programmed frameshifts is the *dnaX* genes encoding the τ and γ subunits of DNA polymerase III. The *E. coli* τ subunit is the full-length gene product while the shorter γ subunit is also synthesized from the *dnaX* gene; the N terminal region of the γ subunit is identical to that of the τ subunit. The C terminal of the γ subunit is encoded in the -1 shifted ORF within the *dnaX* reading frame [58, 60, 59]. In *E. coli* the frameshift occurs during translation (PRF mechanism) at the frameshift motif A_AAA_AAG [61], while in *Thermus thermophilus* the same outcome is accomplished with PTR at a stretch of 9 A's [62]. Out of 17 fs-genes in the cluster 12 have poly-A stretches in their frameshift boxes: 9 of 10 consecutive A's and then 9 A's (*Chloroherpeton thalassium*), 8 A's (*Chlorobium chlorochromatii*) and 7 A's (*Chlorobium phaeobacteroides*). That the 10 A's sequences

align well with the frameshift motif from *T. thermophilus* (see Fig. 17) suggests prevalence of the PTR mechanism in the cluster. *C. thalassium dnaX* has a long poly-AT motif AATAAAAAAAAA while both fs-genes with 8 and 7 A's are present in the *Chlorobium* genus where the *dnaX* gene of two *Chlorobium* species exhibit 10 A motifs. Therefore, the *dnaX* genes with shorter poly-A motifs are likely to use PTR as well.

The Recode database [126] contains 7 records for the *dnaX* gene from four genera: *Escherichia*, *Neisseria*, *Salmonella* and *Vibrio* (cases of PRF). The full length protein in these genera is synthesized by standard translation (without a frameshift) while the frameshifting yields a shorter product. GeneTack did not predict frameshifts in the *dnaX* genes annotated in Recode. In most frame transition cases, the specific protein coding region pattern of nucleotide frequencies transits from one frame to another. Still in a few cases, particularly in some *dnaX* genes, the three periodic frequency pattern of the protein encoding sequence remains strong in the original frame. In such cases GeneTack may not recognize a possible switch to a second frame.

The GeneTack *dnaX* cluster contains 17 fs-genes from 12 different genera in addition to the four genera annotated in Recode. None of these 17 genes have a programmed frameshift annotated in GenBank. The *dnaX* genes containing deletions at the genomic level that are corrected via programmed frameshifting, have not previously been described.

Table 7: The largest GeneTack programmed frameshift clusters that correspond to known cases of programmed frameshifting. **Cluster ID** – unique identifier of a cluster; **Function** – expected gene function derived for the corresponding fs-proteins from Pfam domains and BLASTp hits against the NCBI nr database; **Size** – number of fs-genes in the cluster (FS), number of different genera (G); **D** – frameshift direction (+1 or -1); **BR** – possible biological role (PTR – programmed transcriptional realignment, PRF – programmed ribosomal frameshifting, TC – translational coupling); **FS coord** – median value of the relative frameshift coordinate for all frameshifts in the cluster; **SD** – standard deviation of the relative frameshift coordinate; **Heptamer** – overrepresented heptamer (the fraction of the cluster's fs-genes that contain the heptamer is shown in parentheses), contrary to Table 5 we specify consensus sequence rather than regular expression pattern; **Frameshift site Logo** – Logo of the frameshift site (see text for details); **Sharma et al clusters** – ID(s) of the corresponding Sharma et al clusters.

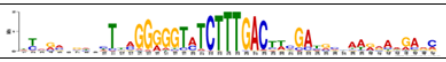
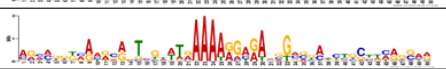
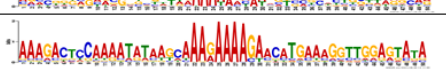
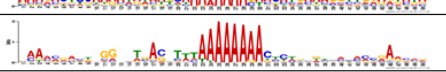
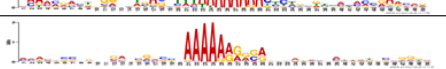
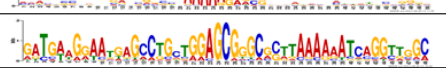
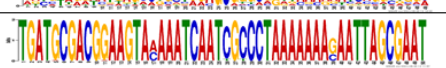
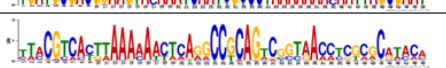
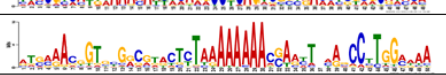
Cluster ID	Function	Size (D, BR)	FS coord (SD)	Heptamer	Frameshift site Logo	Sharma et al clusters	Other References
474411093	Release Factor 2	428 FS, 138 G (+1, PRF)	0.05 (0.11)	C.TT_T.GA_C (86%)		2	(62)
239165634	DNA polymerase III	17 FS, 12 G (-1, PTR)	0.66 (0.09)	A_AA.A_AA.A (53%)			(39,48)
675840861	HTH_Tnp_1 (Transposase)	1699 FS, 106 G (-1, PTR)	0.26 (0.09)	AAAAAAG (49%)		11; 13; 19; 31; 33; 35	(13)
241541714	HTH_Tnp_1 (Transposase)	51 FS, 12 G (+1, PTR)	0.23 (0.05)	AAAAAAA (51%)		12	(13)
667870043	HTH_Tnp_IS630 (Transposase)	495 FS, 20 G (-1, PTR)	0.39 (0.09)	AAAAAAA (75%)		6; 7; 40; 62	(13)
858558073	HTH_Tnp_IS630 (Transposase)	185 FS, 28 G (+1, PTR)	0.37 (0.11)	AAAAAAA (85%)		7	(13)
188472814	DDE_Tnp_1 (Transposase)	384 FS, 37 G (-1, PTR)	0.44 (0.11)	AAAAAAG (28%)		5; 10; 16; 21; 38; 46; 60	(13)
888244788	DDE_Tnp_1 (Transposase)	108 FS, 5 G (+1, PTR)	0.32 (0.09)	AAAAAAT (69%)		24	(13)
910763088	DDE_Tnp_1 (Transposase)	36 FS, 1 G (+1, PTR)	0.45 (0.11)	AAAAAAA (100%)		10	(13)
696263973	DDE_Tnp_IS1 (Transposase)	230 FS, 8 G (-1, PTR)	0.33 (0.11)	AAAAAAC (63%)		8	(13)
919140783	DDE superfamily endonuclease	43 FS, 10 G (-1, PTR)	0.39 (0.09)	AAAAAAA (74%)		9	(13)

Table 8: Programmed frameshift clusters predicted by GeneTack that were selected for experimental verification. **Experimental results** – summary of the results shown on Fig. 18 and Fig. 19 (FS – number of tested fs-genes that showed ribosomal frameshifting; TC – number of tested fs-genes that showed translational coupling)

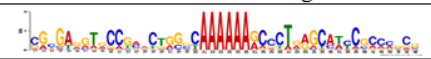
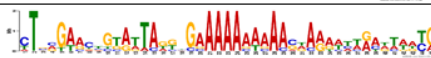
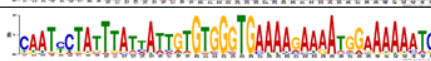
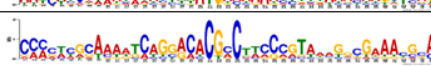
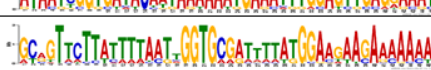
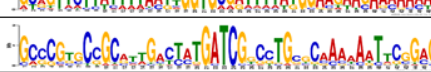
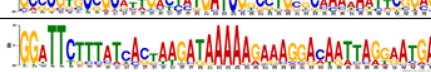
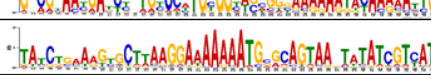
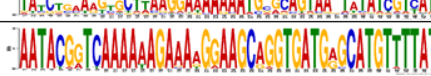
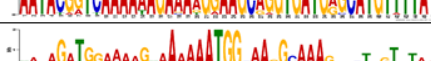
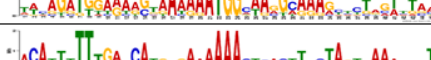
Cluster ID	Function	Size	FS coord	Heptamer	Frameshift site Logo	Experimental results
131733585	Magnesium chelatase	23 FS, 6 G (-1, PRF)	0.60 (0.16)	A_AA.A_AA.G (61%)		3/3 FS (17%, 60%, 41%); 0/3 II
782478235	DUF111	8 FS, 4 G (-1, PRF)	0.66 (0.07)	A_AA.A_AA.A (100%)		2/2 FS (39%, 34%); 1/2 II
621432021	DUF772	14 FS, 2 G (-1, PRF)	0.46 (0.02)	A_AA.A_AA.G (93%)		2/2 FS (24%, 9%); 0/2 II
447662180	Spore germination protein	19 FS, 4 G (-1, PRF)	0.51 (0.12)	G_AA.A_AA.A (84%)		2/2 FS (13%, 4%); 1/2 II
862991913	Phage tail assembly chaperone	41 FS, 5 G (-1, PRF)	0.52 (0.1)	A_AA.A_AA.G (93%)		2/2 FS (7%, 6%); 0/2 II
430699271	Cyclic-nucleotide phosphodiesterase	20 FS, 3 G (-1, PRF)	0.53 (0.08)	A_AA.A_AA.G (50%)		1/2 FS (6%); 0/2 II
720147899	phaP protein / Dehydratase (maoC family)	18 FS, 1 G (-1, PRF)	0.41 (0.01)	A_AA.A_AA.G (94%)		1/2 FS (6%); 0/2 II
181800409	Tetraacyldisaccharide kinase, acyltransferase	16 FS, 4 G (+1, PRF+II)	0.53 (0.1)	G.AA_A.AA_A (88%)		1/2 FS (7%); 2/2 II
931215581	Bac_DNA_binding | Formyl_trans_N	9 FS, 5 G (+1, PRF+II)	0.71 (0.09)	C.TA_A.AA_A (56%)		2/2 FS (6%, 3%) 1/2 II
786465964	ATP-gua_Ptrans | UVR	14 FS, 2 G (+1, PRF+II)	0.33 (0.01)	C.TA_A.AA_A (50%)		1/2 FS (8%); 1/2 II
522343807	DNA glycosylase / Dephospho-CoA kinase	21 FS, 2 G (-1, PRF+II)	0.55 (0.08)	G_AA.A_AA.A (67%)		1/2 FS (low); 1/2 II
392008946	Aminotran_1_2 | Dala_lig_C | Dala_Dala_lig_N | GntR	14 FS, 3 G (+1, PRF+II)	0.60 (0.02)	A.TA_A.AA_A (71%)		1/2 FS (6%); 1/2 II (high)
309851863	ABC transporter	19 FS, 3 G (+1, PRF+II)	0.79 (0.03)	A.AA_A.AA_A (79%)		2/2 FS (5%, low); 2/2 II (high)
310905921	DMRL_synthase | NusB	18 FS, 7 G (-1, II)	0.49 (0.09)	C_AA.A_AA.A (56%)		1/2 FS (1%) 2/2 II (high)
884136395	Ribosomal RNA methyltransferase	23 FS, 5 G (-1, TC)	0.66 (0.04)	G_AT.A_AA.A (74%)		0/2 FS; 1/2 II (high)
970108792	Preprotein translocase subunit SecA	13 FS, 8 G (+1, II)	0.87 (0.13)	A.AA_A.AA_T (62%)		1/2 FS (3%); 2/2 II (high)
984773919	Thymidylate kinase	100 FS, 25 G (+1, II)	0.52 (0.12)	G.AA_A.AA_A (57%)		0/3 FS; 1/3 II
645374543	Fumarylacetoacetase / Homogentisate 1,2-dioxygenase	18 FS, 1G (+1, II)	0.44 (0.002)	A.AA_A.AA_G (72%)		0/2 FS; 2/2 II
523977875	MATE efflux family protein (transporter)	11 FS, 4 G (-1, II)	0.70 (0.19)	A_AA.A_AA.G (61%)		0/2 FS; 2/2 II
655521599	Epimerase | URO-D	7 FS, 3 G (-1, II)	0.49 (0.07)	A_AA.A_AA.A (100%)		0/1 FS; 1/1 II

Figure 18: Western blot analysis and quantification of frameshift products.

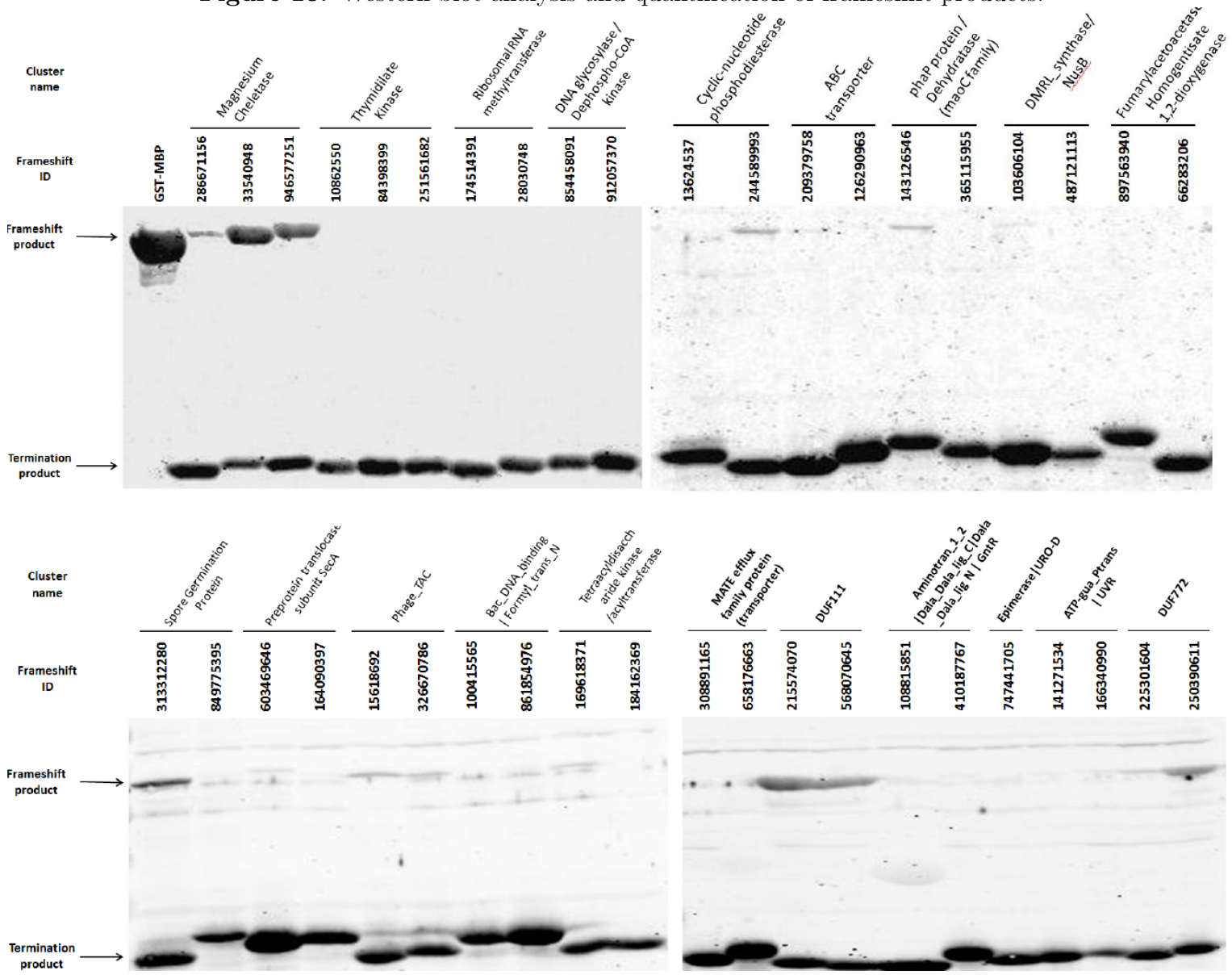


Figure 19: Anti-his western blot analysis of frameshift products and internal initiation (translational coupling) products.

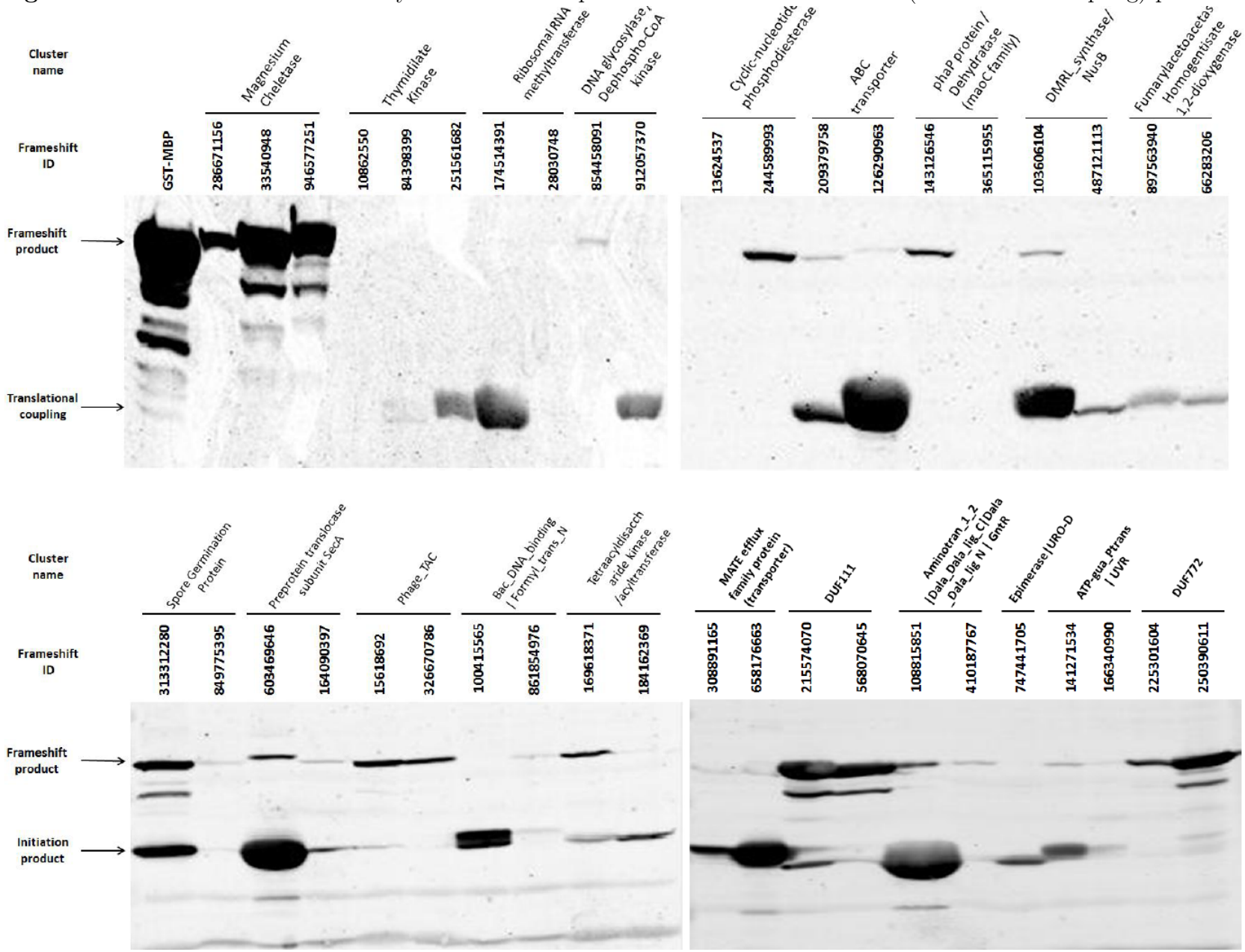


Table 9: Inserts cloned in between GST and MBP genes that showed highest frameshifting efficiency. **FS %** – frameshifting efficiency detected in experiments.

Cluster name [Cluster_ID]	Frameshift ID	INSERT SEQUENCE	FS %
Magnesium chelatase [131733585]	33540948	AAA_AAG_GAA_AAT_AAT_GAC_GT[A_AAA_AAA_AAA_A]CA_AGA_ATA_AAT_TAA_ATA_ATG_AAT_CAA_ATA_ATA_ATG_AGA_ATA_ATA_AT	63% ±4.36
	946577251	GAG_CAA_AAA_AAT_GAT_GAC_GT[A_AAA_AAA_AAC_A]TG_ATG_AAA_TAA_GAA_ATG_AGT_TTG_AAG_AAG_AAA_ATG_AGG_ATT_CAA_AT	40% ±1.73
	286671156	AAG_GTG_GGA_GCG_CCC_GCC_AC[A_AAA_AAA]_GCC_TGA_GCC_CCC_GCG_GCC_CCG_CGC_CAC_AGG_CAA_GGG_CTG_CGG_GGG_	10% ±6.25
DUF111 [782478235]	215574070	ATA_CTC_AGA_ACT_GTA_TTA_GGT_GAA_AAA_AAA_AAG_AAG_AAT_TAA_TTT_ATG_AAA_TTA_GTG_CAA_ATA_TTG_ATG_ATA_TGT_CTT_CAG_ AAA_TCT	39% ±5.32
	568070645	AAT_ATA_GTA_AGA_GCA_ATC_ATA_GGA_AAA_AAA_AAC_TAA_ACT_TGA_GTA_GCA_AAT_ACT_TTG_AGA_TAT_TTG_CTA_ACG_TTG_ATG_ATA	34% ±5.91
DUF772 [621432021]	250390611	AAA_GAT_AGA_ATT_AAT_CAT_AAT_AAA_AAG_CCT_CTA_AAA_AAA_AGA_CTA_AAA_TAG_CTG_AAA_CTA_AGG_AAA_TAA_AAG_TAA_GTA_CAA_ CTG_ATC_CAG_ACA_GTG	24% ±5.01
	225301604	GCT_GCG_GTA_ATC_GAA_GAT_CGT_GAG_GTA_CAT_GGA_AAA_AAG_AAT_TAA_AAC_CTA_GAA_AGG_AAA_GTG_ATA_CCC_CTA_CCA_AAA_AAA_ CTC_GTA_TAA	9% ±4.39
Spore Germination Protein [447662180]	313312280	CTA_TTT_ATT_ATT_GTG_TGG_GTG_AAA_AAA_AAA_TGG_AAA_AAA_CCA_AAA_GAA_AAT_TAA_TAC_TTC_TTA_TTT_GCA_TCA_GTG_TCT_TCA_ GTC_TAA_CTG_GTT_GTT	13% ±3.09
	849775395	CGC_TTG_TTC_GGA_TTC_AAT_CTA_CGT_CAT_TCG_TTG_TAT_TGG_ATT_ATA_CCC_GTG_ATC_TAT_GTA_GCC_TCC_TTG_TCG_CTG_CTT_TCC_ AGA_CAG_CAG_ATG_AGT_CGG_ATG_ACG_ACT_ATT_TAC_TCG_CAT_ATA_ACT_CTG_TAC_ATT_ATT_TAC_GTG_TAC_CCC_TAT_TTT_TTG_TAC	4% ±1.56
Phage_TAC [862991913]	15618692	GCT_GAT_GCA_GAG_TCG_GCC_AGA_AAA_AAG_TAG_CCC_GCC_CGG_AAA_TTC_GCT_TTC_TGA_TGC_GAC_TTG_CGC_TCC_GTC_TGG	7% ±0.79
	326670786	GGA_ATG_AGT_CAG_GAA_GAA_GCG_GGA_AAG_CCG_TAA_AGC_AGC_CGC_TGA_CCT_TCT_TTC_TGC_TGT_CAC_TGG_CGC_TCC_GGC_TGG	6% ±0.61
Cyclic-nucleotide phosphodiesterase [430699271]	13624537	TAC_CAG_TTA_ATC_CCG_ATC_AA[T_TTA_AAG_AAA_AAA]_TTG_AGA_AAG_CTG_ATG_GCA_CTC_GCG_AGC_ACG_TCT_TTT_ATA_CCC_AAG	
	244589993	GCT_CAT_ATG_CAT_AAG_CTT_GT[A_AAA_AAG]_AAG_TTG_TAA_ATG_GTG_TCA_TTA_TCA_CGG_AAC_CAG_ATA_AAT_ATG_GAA	6% ±2.08
phaP protein / Dehydratase (maoC family) [720147899]	143126546	CTT_CAA_AAA_CAA_TTA_GAT_GA[T_TTT_TTG_ACG_GAG_TTC_AAG_TCT_ACA_CAA_CTG_GAA_CTT_GTA_AAA_AAG]_TTC_GAG_GAA_ AAC_TCC_AAA_AAT_CTA_TTT_ACT_TCC_ATC_AAA_TAA_GAA_	6% ±1.16
	365115955	AAG_TCC_AAA_CAA_CTA_GAA_CT[C_GCA_AAG_CAG_TTC_GAG_GAA_AAC_TCA_AAA_AAT]_CTA_TTT_ACT_TCC_ATC_AAG_TAA_GAA_ AAA_TGT_GGC_AAC_TAA_CTG_CAG_	
Tetraacyldisaccharide kinase /acyltransferase [181800409]	169618371	ATA_AAC_CAT_CCT_GAT_TTA_TTA_AAT_GAA_AAA_ATT_TTT_AAA_AAA_GCT_TAG_ATA_TTT_AAT_TGA_GTA_TTT_TAT_CGT_TGT_TAT_ATT_ TCT_GAA_AGT_AAT	7% ±2.45
	184162369	ACA_AAC_AAA_CTA_ATT_AAA_TTA_AAT_GAA_AAA_AAT_TAA_ATA_TTT_TTT_TGA_ATT_TTT_AAT_TAT_ATC_TTC_TCT_TTT_TAT_TAT_TTA	
Bac_DNA_binding | Formyl_trans_N [931215581]	100415565	CGA_AAC_AAT_AAG_GCA_GTT_TTA_GAT_GAG_CAA_GAA_CTT_CCA_GAA_TCT_GGT_TAT_GCA_AAC_GAC_TAA_GCC_TTT_CGC_GGT_AGG_CGC_ ACT_GTT_AGC_GAT_TTC_GCT_TTC_ATG_TGA	6% ±4.09
	861854976	ACT_ATG_GAA_GCA_GAT_TAT_GCA_GTT_CTT_GAC_GAA_ACA_AAA_CTT_CCT_GCA_CAC_GGC_GCG_CAG_TAC_CCA_CAA_TAA_AAA_AAT_AAG_ TGC_CCT_GCT_ATT_ATG_TGT_GGG_CAC_TTT_ACT_TCC_AAC	3% ±1.44
ATP-gua_Ptrans | UVR [786465964]	141271534	CGT_GAT_CAG_ATT_AAT_CAG_CTA_AAA_AAT_CAG_AAT_ACT_ACC_GAT_GCT_CCC_TAA_TCA_TAT_TCT_TAC_TGC_TAT_CGC_AAC_GAT_CAA_ GCA_TTC_TTT_GAG_AAC	8% ±5.41
	166340990	CGA_GAC_CAG_ATT_AAT_CAT_TTA_AAA_AAT_CAG_AAT_TCG_CAT_GAT_TCT_TCC_CAA_TGA_CTT_ACT_TCT_TAA_TTT_TGC_TAG_TAA_GAA_ AGA_CGC_CCC_TCC_TAC_AAA	

3.2.5 New genes that may utilize programmed frameshifting

The remaining 134 GeneTack clusters may also contain genes with programmed frameshifts of previously unknown type. To experimentally test some of these predictons, we selected fs-genes from the 20 (out of the 134) clusters (Table 8). To test the frameshifting efficiency of the candidates the sequences of putative frameshift site plus 18 nt upstream and 45 nt downstream were derived from the frameshift vicinity of the corresponding fs-genes[2] – see Table 9.

Putative frameshift-relevant sequences were cloned in vector pJ307 (see Methods). This vector has a strong promoter, pTAC, with a lac operator, the glutathione S-transferase (GST) gene lacking a terminator and fused in-frame to a maltose binding protein (MBP) gene with a PSPXI-BamH1 cloning site between GST and MBP. The plasmid separately encodes the LacIq repressor so that expression from the pTAC promoter is inducible by addition of IPTG. The cassettes of putative frameshift-relevant sequences were inserted at the cloning site and framed such that the putative frameshifting would yield fusion protein, and the termination product would be a measure of ribosomes that failed to frameshift. The ratio of frameshift and non-frameshift-derived products was determined to estimate frameshift efficiency. (Fig. 18).

In another experiment translational coupling was also measured. This involved a His-tag encoding sequence at the 3' end of the MBP gene with quantification by Western blots with His-tag specific antibody. The results (Fig. 19) complemented those with anti-GST Western blots for frameshift identification.

[2]The URL http://topaz.gatech.edu/GeneTack/cgi/fs_view.cgi?id=FRAMESHIFT_ID can be used to retrieve information about a frameshift

3.2.5.1 Confirmed new cases of efficient Recoding

Genes from four clusters showed frameshifting efficiency higher than 10%. These clusters are magnesium chelatase (frameshifting efficiency up to 63%), DUF111 (up to 39%), DUF772 (up to 24%) and Spore germination protein (up to 13%).

Genes for magnesium chelatase form a cluster of 23 predicted recoded genes from 6 different genera with putative programmed -1 frameshifting (from both bacteria: *Pseudomonas*, *Burkholderia*, *Delftia* and *Herpetosiphon*; and archaea: *Methanocaldococcus* and *Methanococcus*). Cassettes from three of these were tested and all showed significant levels of frameshifting (63%, 40% and 10%). Interestingly, the lengths of the poly-A runs in the cassettes correlate with the frameshift efficiency: A_AAA_AAA_AAA_A (63% – 11A's), A_AAA_AAA_AA (40% – 9A's) and A_AAA_AAA (10% – 7A's) – see Table 9.

The predicted fs-genes for magnesium chelatase are annotated in GenBank as two adjacent genes each about 1,000 nucleotide long. The upstream part is annotated as a gene for magnesium chelatase while the downstream part is annotated as either a hypothetical gene or a gene for von Willebrand factor type A. However, a BLASTp search against NCBI nr database reveals several magnesium chelatase proteins (from *Chloroflexus aggregans*, *Rubrobacter xylanophilus* and others) made by a fusion of the two parts, an indication that the fusion protein and the proteins produced by the recoding are likely to carry similar function.

The clusters DUF111 and DUF772 were named after the Pfam domain found in the fs-proteins from these clusters (DUF stands for "Domain of Unknown Function"). In GenBank the regions corresponding to the fs-genes from the DUF111 cluster are annotated as two hypothetical proteins and regions for the DUF772 cluster as two separate transposases. For both clusters BLASTp search in the NCBI nr database gave a number of hits that are fusions of the two proteins.

Fs-genes from the Spore germination protein cluster are annotated as two separate

genes encoding "Spore germination protein". There is no BLASTp fusion hits in the nr database, however, our experimental results show that the production of a fusion product via programmed frameshifting is possible in the cell.

3.2.5.2 Clusters that show less efficient Recoding

Frameshifting efficiency less than 10% was observed in three clusters: phage tail assembly chaperone (7%), cyclic-nucleotide phosphodiesterase (6%) and phaP protein / dehydratase (6%).

Conceptual translation of the fs-genes from the phage tail assembly chaperone cluster (41 fs-genes) had significant similarity to a protein from *Enterobacteria phage HK97*; we presume that these fs-genes are of viral origin and use Recoding. Notably, expression of a number of viral genes utilizes programmed frameshifting [135, 136].

Protein products of cyclic-nucleotide phosphodiesterase fs-gene have hits to fused proteins in the nr database suggesting similar function for products of Recoding genes and fused genes akin to the magnesium chelatase gene family.

3.2.5.3 Translational coupling

Tested constructs from seven clusters did not show any level of programmed frameshifting. However, initiation of translation was observed that should result in the synthesis of a downstream ORF product. Such cases may represent instances of conserved translational coupling where initiation of downstream ORF translation depends on termination of the upstream ORF translation and such co-regulation contributes to the fitness. The clusters classified as translational coupling include Thymidylate kinase, Ribosomal RNA methyltransferase and Fumarylacetoacetase / Homogentisate 1,2-dioxygenase.

Also, there were six clusters for which both programmed frameshifting and translation coupling was observed. Indeed it is possible that both events occur in the cell [94]. Notably, strong translational coupling signal was observed in two clusters out

of these six (Fig. 19).

3.2.6 Other large clusters of fs-genes

Among 40 clusters that contained 100 and more fs-genes (Table 10), only 8 clusters were classified immediately as clusters of fs-genes with programmed frameshifts. Still, after additional analysis, two more large clusters were classified as such.

Notably, one more transposase gene cluster containing 112 fs-genes did not contain any typical programmed frameshift heptamers near the positions of their predicted frameshifts. However, a conserved TTA_TTN sequence could constitute a slippery site while these fs-genes belonging to a family routinely using Recoding for their gene expression.

Another large cluster of 105 kinase/phosphatase fs-genes with a conserved CAT_TTT motif was identified as a programmed frameshift cluster of fs-genes potentially using both PRF and PTR.

In the remaining 30 clusters we have evidence suggestive of phase variation or translational coupling.

3.2.6.1 Phase variation clusters

First, we collected a set of 38 genes with known phase variation produced by the SSM mechanism [96]. Protein products of these genes were used in a BLASTp search (with E-value 10^{-10}) against the database of all the fs-proteins. Hits for 14 queries were detected in 13 clusters with 5 or more members (Table 11). The 13 clusters were likely to be clusters of genes with conserved phase variation. (this criteria was not used to find phase variation clusters, instead we used %S and %AT from the Table 10).

Next, we attempted to detect short sequence repeats (see Methods) in the 50 nt vicinity of a frameshift in an fs-gene from a large cluster. Since poly-AT is a slippery sequence for DNA polymerase (as well as for RNA polymerase and ribosomes), a

Table 10: The largest clusters containing 100 or more fs-genes. **Size** – number of fs-genes in the cluster, **#G** – number of different genera in the cluster; **D** – frameshift direction; **%AT** – fraction of fs-genes with 7+ nt poly-AT stretch located near predicted frameshift; **%R** – fraction of fs-genes with tandem repeats located near predicted frameshifts; **%S** – fraction of fs-genes with ORF2 start codon ATG ([1]GTG) located within 10nt ([2]20nt) from the ORF1 stop codon; **%B** – fraction of fs-proteins validated by BLASTp against NCBI nr database; **BR** - biological role (PF – programmed frameshifting, PV – phase variation, TC – translational coupling); **?** – putative prediction of biological role; *experimentally verified.

Cluster ID	Cluster name	Size	#G	D	%AT	%R	%S	%B	BR
474411093	Release Factor 2	428	138	+1	49%	4%	1%	2%	PF*
675840861	HT_Tnp_1 (Transposase)	1699	106	-1	72%	7%	3%	75%	PF
188472814	DDE_Tnp_1 (Transposase)	384	37	-1	67%	4%	2%	85%	PF
888244788	DDE_Tnp_1 (Transposase)	108	5	+1	80%	0%	0%	95%	PF
667870043	HTH_Tnp_IS630 (Transposase)	495	20	-1	98%	1%	0%	96%	PF
858558073	HTH_Tnp_IS630 (Transposase)	185	28	+1	100%	5%	0%	86%	PF
696263973	DDE_Tnp_IS1 (Transposase)	230	8	-1	90%	0%	0%	72%	PF
784826247	Transposase IS911/IS222	112	5	-1	6%	0%	0%	0%	PF?
752989859	Kinase / Phosphatase	105	23	+1	67%	1%	16%	0%	PF?
279791230	HATPase_c, HisKA	594	148	+1	57%	5%	35%	13%	PV
487884579	HATPase_c, HisKA	292	98	+1	36%	18%	31%	35%	PV
107592512	HATPase_c, HisKA	162	51	-1	36%	4%	56%	5%	PV?
437298609	BPD transporter	238	79	+1	34%	9%	34%	5%	PV
672517721	BPD transporter	149	46	+1	41%	14%	42%	26%	PV
953823467	BPD transporter	100	22	-1	5%	17%	10%	0%	PV
6376240	tRNA synthetase	215	81	+1	60%	7%	36%	1%	PV
138502135	Aminotransferase	175	88	+1	38%	13%	30%	13%	PV
354349696	Secretion system	140	51	+1	41%	6%	18%	9%	PV
322052632	Fucose synthase / Dehydratase	139	78	+1	44%	9%	37%	4%	PV
631171255	PqiA membrane protein	126	38	+1	71%	4%	17%	5%	PV
222950006	ABC transporter	436	116	+1	46%	7%	47%	59%	TC
785097185	ABC transporter	298	66	+1	62%	4%	48%	63%	TC
208900412	ABC transporter	293	97	+1	24%	4%	61%	69%	TC
624178257	ABC transporter	289	102	-1	49%	8%	45%	64%	TC
79330857	ABC transporter	280	97	+1	35%	9%	86%	1%	TC
104388297	ABC transporter	146	61	-1	21%	18%	64%	11%	TC
22890314	ABC transporter	144	49	+1	24%	3%	65%	69%	TC
471276212	ABC transporter	126	48	+1	25%	2%	60%	33%	TC
548076848	Flagella	139	34	+1	48%	3%	71%	0%	TC
585180489	Flagella	111	36	+1	8%	11%	75%	0%	TC
181132644	Flagella	118	46	+1	36%	10%	70%[2]	0%	TC?
847934252	Polyketide cyclase	132	38	+1	46%	4%	81%[2]	0%	TC
697472870	Biotin carboxylase	128	44	+1	73%	5%	75%	2%	TC
876288400	Hydrolase / Epimerase	121	27	+1	28%	2%	79%	0%	TC
458305551	Polyphosphate kinase	113	35	+1	27%	8%	63%	2%	TC
237996460	Mur ligase	112	33	+1	78%	3%	83%[1,2]	90%	TC
717516549	Epimerase	111	65	-1	44%	13%	61%	2%	TC
539781944	Oxidoreductase	109	51	-1	39%	1%	80%	0%	TC
515287573	Recombination factor RarA	104	52	+1	55%	5%	74%[2]	4%	TC?
984773919	Thymidylate kinase	100	25	+1	93%	7%	3%	3%	TC*

Table 11: GeneTack clusters with members representing known cases of phase variation. **Query gene name (Organism)** – the name of a gene with known phase variation; **# hits** – number of fs-proteins found by BLASTp search (how many of them belong to clusters is specified in brackets); **Main cluster** – name of the fs-cluster with the largest number of hits; **Size** – size of the cluster (number of fs-proteins found by the BLASTp search is specified in brackets).

Query gene name (Organism)	Function	# hits	Main cluster	Size
DNA methylase (*S. pneumoniae*)	DNA mod	106 (82)	DNA methylase	59 (59)
Mod (*Helicobacter pylori*)	DNA mod	38 (27)	Methyltransferase	6 (6)
FlhA (*Campylobacter coli*)	Flagella	143 (130)	Bac_export_2, FHIPEP	72 (72)
FliP (*Helicobacter pylori*)	Flagella	134 (116)	FliO, FliP	48 (48)
SpxB (*S. pneumoniae*)	Metabolism	170 (134)	TPP_enzyme	56 (56)
HifB (*Haemophilus influenzae*)	Pilus	67 (53)	Pili_assembly	21 (21)
pilE (*Neisseria meningitidis*)	Pilus	13 (7)	Pilin	5 (5)
PilS (*Neisseria meningitidis*)	Pilus	13 (7)	Pilin	5 (5)
LgtA (*Neisseria meningitidis*)	Transferase	100 (34)	Glycos_transf_2	8 (7)
LgtC (*Neisseria meningitidis*)	Transferase	30 (25)	Glyco_transf_8	17 (17)
lgtD (*Neisseria meningitidis*)	Transferase	31 (10)	Glycos_transf_2	5 (5)
wlaN (*Campylobacter jejuni*)	Transferase	96 (34)	Glycos_transf_2	5 (5)
pgtA (*Neisseria gonorrhoeae*)	Transferase	36 (18)	Glycos_transf_1	8 (8)
bvgS (*Bordetella pertussis*)	Regulation	250 (132)	HATPase_c, HisKA	83 (59)

stretch of poly-AT could cause phase variation. The fraction of a cluster's fs-genes that have a poly-AT stretch (with minimal length 7) near a predicted frameshift is shown in the %AT column (Table 10).

Finally, we used the Tandem Repeats Finder program [137] to identify other type of repeats (such as poly-G or poly-GC). The program parameters were set to report stretches of the same nucleotide as a repeat (with minimal length 7). The fraction of a cluster's fs-genes with tandem repeats (other than Poly-A and poly-T) near the predicted frameshifts is shown in the column %R (Table 10).

The %AT and %R features indicate characteristic properties of phase variation clusters. A large cluster is classified as phase variation cluster if %AT and/or %R is higher than %S which is related to characteristic property of translational coupling clusters (see below).

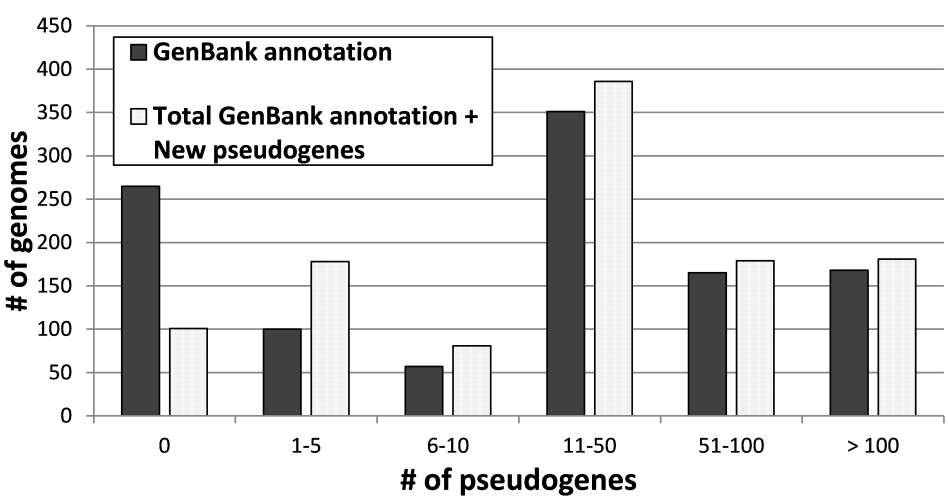


Figure 20: Distribution of the number of annotated and predicted pseudogenes among prokaryotic genomes. Black bars were obtained for pseudogenes annotated in GenBank. The white bars show the updated distribution with 4,806 pseudogenes identified in this work added to the annotated pseudogenes. The largest change in the distribution has been observed in the genomes with less than 10 pseudogenes annotated in GenBank.

3.2.6.2 Translational coupling clusters

Notably, there were 137 clusters of sequences encoding proteins with ABC transporter function having 5 or more members (4,560 cases of phase transition in total); 8 clusters contained more than 100 frameshifts (Table 10). We classified the ABC transporter clusters as candidate clusters gene pairs with translational coupling (see Methods). Earlier, translational coupling was experimentally shown to occur in ABC transporters e.g. drrAB genes from *Streptomyces peucetius* [138]. The protein products of the drrAB genes have similarity to proteins in the ABC transporter cluster containing 36 fs-genes.

Interestingly, the *p78* gene from the ABC transporter operon in *Mycoplasma fermentans* was characterized in another publication [97] as a gene with phase variation. The protein product of this *p78* gene did not have any match in our data.

3.2.7 Pseudogene clusters

In our set of 1,106 genomes, the total number of GenBank annotated pseudogenes was 59,318, with numbers of annotated pseudogenes varying significantly between genomes. No single pseudogene was annotated in 265 genomes while several genome annotations had over a thousand pseudogenes; for example 1,116 out of 2,770 protein coding genes in the parasitic bacteria *Mycobacterium leprae* genome (NC_011896) are annotated as pseudogenes. In spite of the natural dependence of the number of pseudogenes on the evolutionary path of the species in which they reside, a significant part of the difference in the numbers of pseudogenes among genomes might be related to variability in the accuracy of pseudogene annotation.

Comparison of every case of frame transition predicted by GeneTack with genome annotation revealed that 18,619 fs-genes were annotated as pseudogenes, i.e. GeneTack identified 31% of all annotated pseudogenes. Among predicted annotated pseudogenes 7,186 belonged to 3,329 clusters and other 11,433 were singletons (Fig. 21). Thus more than 50% of the predicted pseudogenes did not belong to clusters, which is not surprising taking into account that pseudogenes degrade rapidly [139].

To zoom in on the 3,329 clusters that contained at least one annotated pseudogene we excluded clusters that contain fs-genes from three or more different genera because of the presumption that almost all bacterial pseudogenes are of relatively recent origin [140]. Also we excluded clusters of fs-genes with evolutionary conserved programmed frameshift sites in the frameshift boxes (they might be clusters of fs-genes with Recoding). The remaining 2,810 clusters with at least one annotated pseudogene (Fig. 21) contained 10,290 fs-genes with 5,484 fs-genes annotated as pseudogenes. The remaining 4,806 fs-genes in these clusters with sequence similarity to annotated pseudogenes from the same cluster and frameshifts in the same position as in homologous annotated pseudogenes, were assumed to be pseudogenes as well. Notably, many of the new pseudogenes appeared in genomes with no GenBank annotated pseudogenes

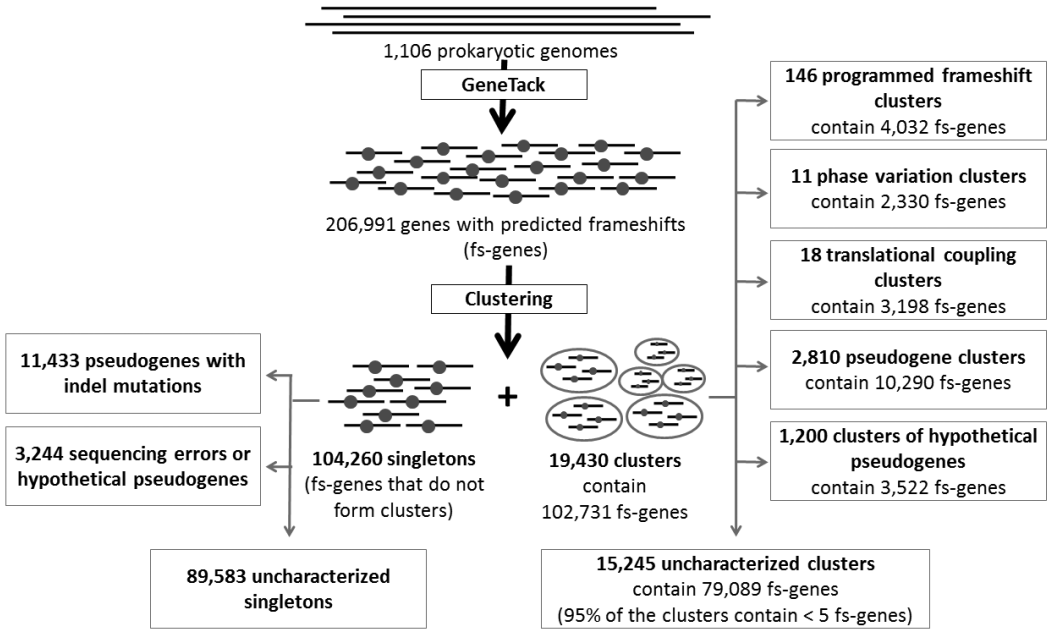


Figure 21: Classification of predicted frameshifts was done by using features specified in Table 12. One of the most important properties of a predicted fs-gene was its membership in a cluster. Singletons are likely to be result of indel mutation or sequencing error, while clustered fs-genes could represent programmed frameshifts, phase variation and translational coupling, as well as pseudogene clusters or clusters of genes with indel mutations.

(see Figure 20).

There were clusters that did not contain annotated pseudogenes but possessed properties typical for pseudogene clusters. Still, these fs-genes need experiential testing to check that their truncated protein products are nonfunctional. We considered these clusters as being comprised of hypothetical pseudogenes. Namely there were 1,200 such clusters containing 3,522 fs-genes in total (Fig. 21). The fs-genes in the clusters did not have features typical for Recoding genes; they contained fs-genes from no more than two different genera while more than 50% of the cluster's frameshifts were validated by BLASTp (Table 12).

Table 12: Features (the first column) used to classify predicted frameshifts into Types (the Type names are given in the top two rows). **H-pseudo** – hypothetical pseudogene; **n/r** – the feature is not required; **n/a** – the feature is not applicable; *a cluster must contain at least one annotated pseudogene; **>50% of cluster fs-genes must be validated by BLASTp; ***manual verification includes functional analysis of the fs-proteins and literature survey.

	Cluster type					Singleton type	
	Programmed frameshift	**Phase Variation**	**Translational Coupling**	**Pseudo gene**	**H-pseudo gene**	**Pseudo gene**	**H-pseudo / Error**
Cluster contains 5 or more fs-genes	YES	YES	YES	n/r	n/r	n/a	n/a
Conserved frameshift site	YES	n/r	n/r	NO	NO	n/a	n/a
Cluster with small (≤2) number of genera	n/r	n/r	n/r	YES	YES	n/a	n/a
GenBank annotation of a pseudogene*	n/r	n/r	n/r	YES	NO	YES	NO
Tandem repeat near frameshift position	n/r	YES	n/r	n/r	n/r	n/r	n/r
ORF2 start is located close to ORF1 stop	n/r	n/r	YES	n/r	n/r	n/r	n/r
BLASTp validation**	n/r	n/r	n/r	n/r	YES	n/r	YES
Pfam validation	n/r	n/r	n/r	n/r	n/r	n/r	YES
Manual verification***	n/r	YES	YES	n/r	n/r	n/r	n/r

3.2.8 Singletons: authentic indel mutations or sequencing errors?

More than 50% of all the predicted fs-genes (104,260) that did not cluster made a set of singletons. The frame transitions in singletons might be caused by sequencing errors or recent indel mutations. In addition, frame transitions in singletons may

represent pairs of adjacent genes. In general, the distribution of relative frameshift coordinates for singletons is more flat than that for all frameshifts (Fig. 22); thus sequencing errors and indel mutations are more frequent among singletons. Still an enrichment of singletons in its middle part of the distribution is indirect evidence for the presence of gene pairs among singletons.

In GenBank 11,433 predicted singletons are annotated as pseudogenes. Unfortunately, there is no method to distinguish a singleton due to a sequencing error from a singleton with indel mutations, other than by resequencing.

Additionally, we could confirm predicted fs-genes by BLASTp and Pfam. Out of 104,260 fs-genes that did not cluster, there were 3,244 fs-genes confirmed by both BLASTp and Pfam. To identify their true nature, these sequences with frame transitions would need to be resequenced.

3.2.9 Distribution of relative frameshift coordinates in fs-genes

The distribution of frameshift coordinates normalized to the whole length of a predicted fs-gene had a characteristic shape on (0,1) interval (Fig. 22). The distribution was computed for the whole set of frameshifts including i/ frameshifts at random location of an fs-gene (sequencing errors and indel mutations); ii/ frameshifts predicted at overlaps of adjacent genes; iii/ programmed frameshifts, etc. The positions of random frameshifts are assumed to follow a uniform distribution. To model the coordinate of a GeneTack prediction at an overlap of a gene pair we used the formula

$$r = \frac{x + t}{x + y - z} \tag{4}$$

where x, y are lengths of adjacent genes X and Y, t is the error in detecting the frameshift position and z is the length of gene overlap (positive if X and Y overlap and negative if the gene Y start codon is located downstream of gene X). The distribution of values t was determined from GeneTack predictions for artificial fs-genes in *E. coli*

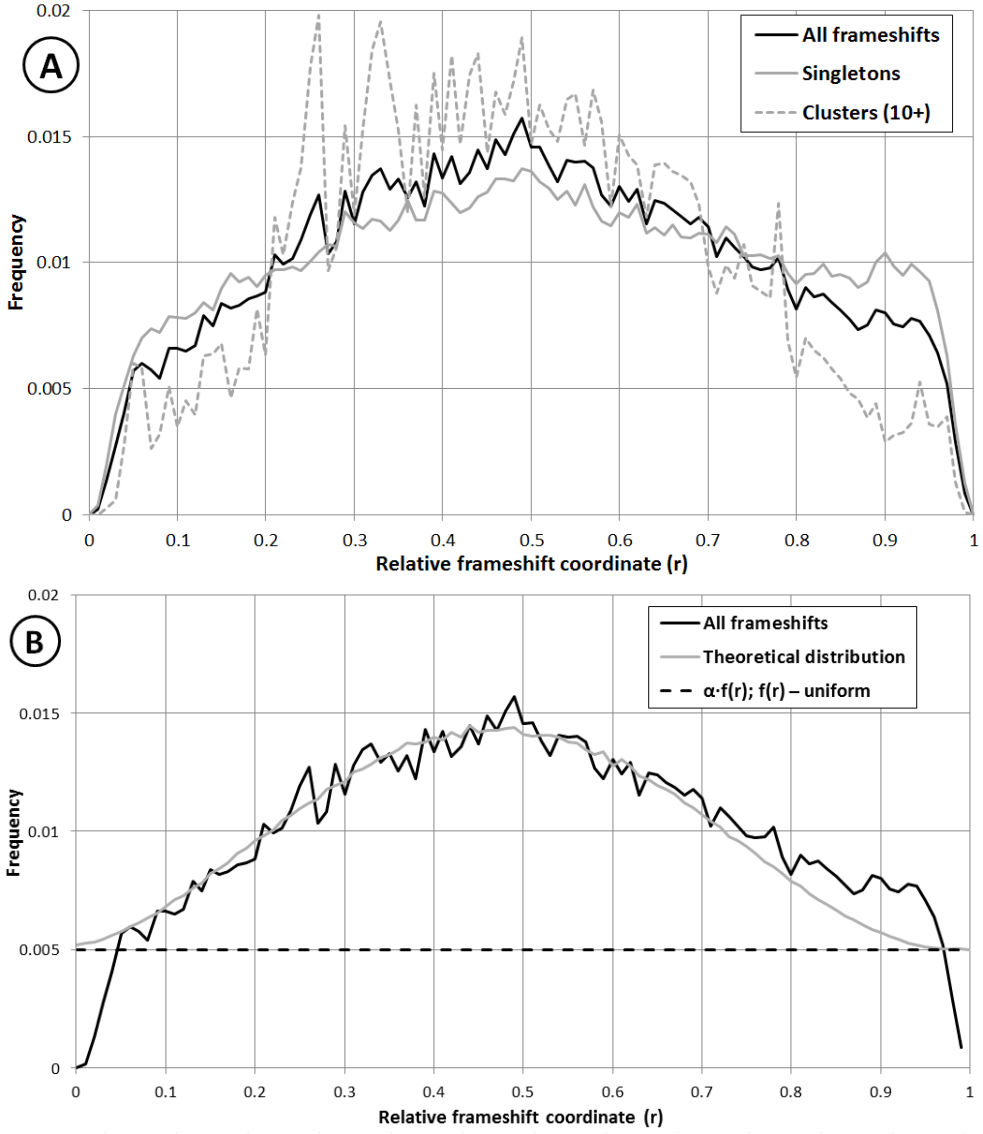


Figure 22: (A) Empirical distributions of frameshift coordinates relative to fs-gene lengths for 1/ all the predicted frameshifts (206,991 fs-genes), 2/ singletons (104,260 fs-genes) and 3/ frameshifts in clusters containing 10 or more members (47,278 fs-genes in total). (B) Distribution of relative coordinates of all the predicted frameshifts (A1) is shown along with the theoretical distribution combining a uniform distribution of coordinates of random frameshifts and distribution of false positive predictions (the $(x-t)/(x+y+z)$ distribution – see text). The random frameshifts correspond to indel mutations and sequencing errors while the false positives are predicted for adjacent genes (with overlapping ORFs). The theoretical curve has good fit to the observed distribution, with the value of parameter $\alpha = 0.005$. See text for more details.

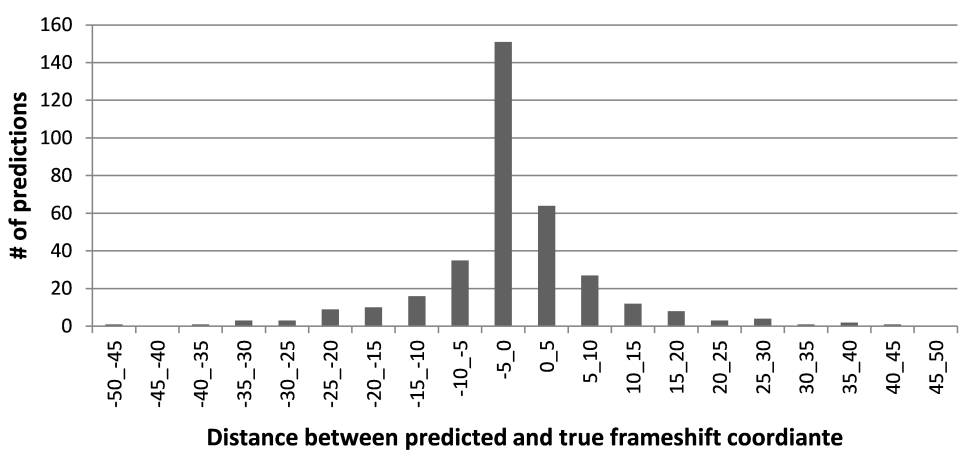


Figure 23: Distribution of frequencies of GeneTack magnitudes of errors in predicting frameshift positions. We have applied GeneTack to 400 *E. coli* genes (longer than 1,000 nt) with a single frameshift created at random in a position separated by at least 150 nt from the gene border. The program successfully predicted 351 frameshifts (the remaining 49 frameshifts were predicted as adjacent genes). This distribution shows that in 83% of cases a prediction is located within 5 nt from the true frameshift position.

(Fig. 23); distribution of z values was derived from all the predicted fs-genes. The X and Y gene lengths, x and y respectively, were assumed to vary between 20 and 3000 codons and follow gamma distribution [121] with the value of the scale parameter equal to 100 codons. Distributions of x, y, t, and z were used for computational modeling of the r distribution $f(r)$.

Also, an observed empirical distribution of the relative frameshift coordinate θ could be represented as the weighted sum of uniform distribution U with density 1 and the $f(r)$ distribution.

$$\theta = \alpha U + (1 - \alpha) f(r) \tag{5}$$

With $\alpha = 0.005$ the theoretical distribution follows closely the observed distribution of relative frameshift coordinates (Fig. 22B).

The empirical θ distribution also shows elevation of frequency of predicted frameshifts at fs-gene ends. This feature could be related to mutations at gene ends that, even

though they lead to premature stop codons, do not affect protein function and remain in the population.

It should be noted, that the distribution approaches zero early at both ends since frameshifts predicted within 50 nt distance from fs-gene borders were filtered out.

The dashed line outlining the "uniform base" of the distribution provides a rough division between random frameshifts – sequencing errors and indel mutations and frame transitions predicted at programmed frameshift sites (with peaks related to specific large clusters of genes with Recoding) as well as at gene overlaps with, or without, evolutionary conservation .

3.3 Materials and Methods

3.3.1 Translation of predicted fs-genes; BLASTp and Pfam confirmations

Two overlapping ORFs in the same strand predicted by GeneTack [127] to reflect a single gene with a frameshift are in the present work referred to as fs-genes. They were united with respect to the predicted position of the frameshift and its direction (+1, -1). The extended ORF was conceptually translated into the fs-protein. When, in a few cases, several frameshifts were predicted in a single gene, several overlapping fs-genes were generated by making predicted additional frameshift positions within the fs-gene borders. Consequently, there was a one-to-one correspondence between a predicted frameshift, an fs-gene and an fs-protein.

The GeneTack false discovery rate (FDR) determined earlier on a set of 17 prokaryotic genomes is about 32% [127]. Taking into account the relatively high FDR, we used two complementary methods to confirm GeneTack predictions. The first one was the BLASTp search for a protein in the NCBI nr database whose alignment score to the fs-protein had an E-value lower than 10^{-10}. Moreover, the sequence alignment to a database protein had to cover at least a 10 amino acids fragment of the fs-protein centered at the predicted frameshift position (Fig. 16A). Contrary to this

outcome, an fs-protein query could produce two sets of BLASTp hits disconnected at the frameshift position (Fig. 16B). This type of outcome was considered as an indication that the conceptual translation of predicted fs-gene involved translation of a pair of overlapping genes; therefore the prediction was filtered out. The former inference, assumed to confirm predicted frameshift, has to take into account the possibility of gene fusion and fission. Some BLASTp validated frameshifts may still be false positives, e.g. a frameshift predicted between adjacent genes whose homologs are fused in another genome [141].

To make the validation more stringent we searched for conserved domains in fs-proteins. We sought an alignment to a Pfam domain (with E-value better than 10^{-3}) that would cover at least a 20 amino acids fragment of the fs-protein query centered at the predicted frameshift position.

We assumed that a conserved domain could not be divided between two fused genes; thus, the Pfam confirmation would exclude the possibility of gene fission and so indicate a correctly predicted fs-gene. Note that given the stringency of the validation procedure, not all correctly predicted frameshifts were expected to be confirmed by BLASTp and Pfam.

Finally, some correctly predicted fs-proteins might not be confirmed by BLASTp due to incompleteness of the nr database. Therefore, an fs-protein with no BLASTp validation was not regarded as a false positive.

3.3.2 Ribosome binding site (RBS) of the downstream ORF

For frameshifts caused by pseudogenizaton mutations, as well as for those that appear in a genomic sequence due to sequencing errors, one would not expect to see an RBS motif near the predicted frameshift position other than by chance. The opposite is true if a frameshift is predicted between adjacent genes each having an RBS site and notably, some programmed frameshifts could have stimulatory sequences of the

Shine-Dalgarno type. The gene prediction program GeneMarkS [119] provides an input to GeneTack and computes an RBS score for every predicted gene; in our experience of working with a number of genomes, the RBS scores range between -11 and 8 (larger score corresponds to stronger RBS). Normally, GeneMarkS predicts two genes instead of one gene with a frameshift. Since the downstream "gene" is not a real gene, its "RBS score" is expected to be low. The GeneTack algorithm has a filter removing fs-genes with high RBS scores for the downstream ORF. The cutoff value 2.2 was chosen for analyzing a single sequence [127]. For clusters of fs-genes uniformly elevated RBS scores could be informative of false positive predictions. For 10,434 frameshifts confirmed by both BLASTp and Pfam, an average value of the RBS scores of downstream ORFs was -1, while the average value of the RBS scores for all the remaining downstream ORFs was -0.14. Some clusters with significant RBS scores could represent fs-genes with Recoding. However, due to the overlap in some instances being connected with translational coupling, we were not able to immediately classify clusters of fs-genes with elevated average RBS score of downstream ORF as clusters of pairs of separate genes with a frame transition.

3.3.3 Clustering

All 206,991 fs-proteins (with or without BLASTp and Pfam confirmation) have been grouped into clusters based on sequence similarity and frameshift position conservation in close species as well as frame transition direction (+1 or -1). The clustering was done as follows.

First, in the database of all fs-proteins an "all-against-all" BLASTp search was performed with a stringent E-value threshold 10^{-50} chosen to avoid inclusion of non-homologous proteins in the clusters and facilitate detection of conserved DNA motifs related to programmed frameshifts.

Next, a graph was built with 206,991 fs-proteins as nodes. Two nodes were connected by an edge if: (i) the BLASTp derived pairwise alignment had positions of both frameshifts inside the alignment block (and separated from the alignment border by at least 10 amino acids); (ii) both predicted frameshifts were of the same direction (+1 or -1); and (iii) the distance d between the two frameshift positions was $\leq$50 amino acids. All connected components of the graph with two or more nodes were called clusters (GeneTack clusters).

The fs-genes that did not cluster are likely cases of artifactual frameshifts caused by sequencing errors that penetrated several layers of sequence quality control.

We attempted to classify clusters of homologous fs-genes as i/ clusters of pseudogenes or hypothetical pseudogenes; ii/ fs-genes with programmed frameshifts; iii/ fs-genes with phase variation iv/ fs-genes with translational coupling; as well as v/ fs-genes related to overlapping homologous gene pairs (false positive clusters).

3.3.4 Functional characterization of the GeneTack clusters

In a given GeneTack cluster of homologous fs-proteins we expected to see fs-proteins with similar function. If in a given cluster more than 50% of the fs-proteins contained the same Pfam domain we assigned the domain name to the cluster. Clusters of multi-domain proteins received a "multi-function name" with more frequent domains listed first.

A cluster of fs-proteins with no detected Pfam domain had no function derived name assigned (just a cluster ID) unless a name could be derived from the cluster fs-proteins BLASTp majority hits to functionally characterized proteins.

3.3.5 Identification of clusters of fs-genes with non-standard mechanisms of transcription and translation

Transcriptional realignment and ribosomal frameshifting occur at specific sequences, which efficiency is often being augmented by additional cis-elements. Due to the

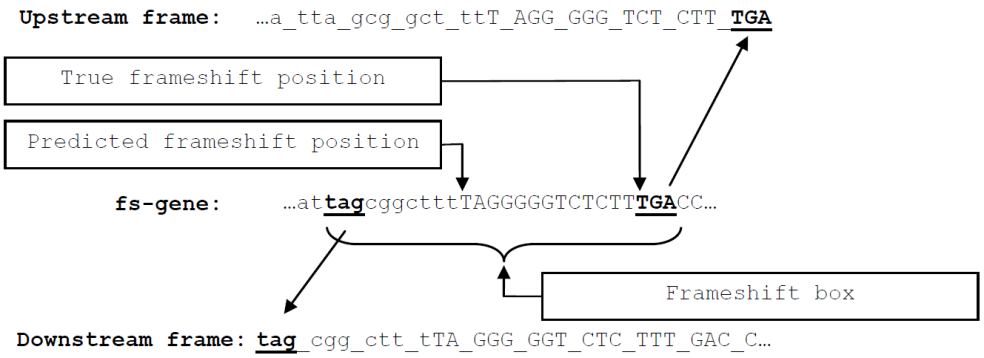


Figure 24: An example of a frameshift box. Predicted frameshift is flanked by two stop codons in different frames (frame 2 TAG stop codon upstream of the frameshift and the frame 1 TGA stop codon downstream of the frameshift). The true frameshift position is always located inside the region between the two stop codons. We call this region "frameshift box".

limited repertoire of shift- and slip- prone sequences, they evolve under purifying selection. In case of PTR, the specific sequences often appear to be homopolymers of eight nt or longer, or combinations of two shorter homoploymers. A PRF event commonly involves rearrangements of tRNAs interacting with two codons in the original frame. The sequence accommodating two codons (tRNAs) interacting to produce -1 and +1 shifts should be at least a heptamer [142]. Thus, it is expected that genuine instances of programmed non-triplet decoding should often contain heptameric sequences evolving under purifying selection. Identification of short conserved sequences of seven nucleotides and longer can be used as supportive evidence for the presence of programmed non-triplet decoding.

To precisely delineate specific sequences prone to PTR or PRF in a particular cluster, we build a multiple alignment of the "frameshift boxes" sequences surrounding the predicted frameshift positions. A frameshift box is a sequence bounded by two stop codons, one at the 5' end for the downstream ORF and the other at the 3' end of the upstream ORF (Fig. 24). Both predicted and true frameshift positions should have occurred within the frameshift box. If the distance between the two stop codons was larger than 100 nt (a frequently case in high GC genomes), the frameshift

box was reduced to the 100 nt long vicinity of the predicted frameshift. Several efficient algorithms and software tools were developed earlier to find conserved motifs in a set of sequences (e.g. the Gibbs Sampler [143] and MEME [144]). However, in case of PRF not only motif per se but also the phase of motif with respect to the reading frame, set by the initiation codon of the upstream ORF, is important. For example, in the *prfB* gene encoding Release Factor 2, the consensus frameshift motif is YTT_TRA_C; here the triplet TRA is a stop codon. Therefore, in search for PRF motifs (contrary to finding PTR motifs) the DNA sequence alphabet needs to be extended by the additional symbol, the underscore indicating the frame of upstream ORF. Frameshift box sequences phased by underscores were used in a custom version of Gibbs Sampler algorithm to produce motifs with given triplet phase. The consensus of motif sequences (a phased motif) was used to characterize the frameshift site in a given cluster.

To initially identify motifs prone to +1 and -1 programmed frameshifts, we searched for framed heptamers that occur in the frameshift boxes of a given cluster more often than other (N_NNN_NNN for -1 frameshifts and NNN_NNN_N for +1). Clusters containing between five and 100 fs-genes with average sequence identity of the frameshift box $\leq$80% were selected (1017 "-1" clusters and 1380 "+1" clusters). The starting positions of motifs were chosen randomly and the Gibbs Sampler was run 100 times searching for N_NNN_NNN motifs in -1 clusters and NNN_NNN_N motifs in +1 clusters.

For large clusters (with 100 or more fs-genes), positions of the most over-represented heptamers were used as start points for the first Gibbs Sampler iteration. Consensus sequences for alignements found by the Gibbs Sampler were recorded (framed heptamers). When motif positions for the first iteration were chosen randomly, consensus heptamers found in different Gibbs Sampler run could vary. The number of times a heptamer X appeared as a Gibbs Sampler (GS) consensus for a particular cluster

was recorded. The score of a heptamer X was computed as follows:

$$Score(X) = \sum_{clusters} ClusterSize \times \frac{Number\,of\,times\,X\,is\,found}{Number\,of\,GS\,runs} \qquad (6)$$

Consensus heptamers containing a start codon for a downstream frame (AT_G for "+1" and A_TG for "-1") indicated that frameshifts were predicted at overlaps of evolutionary conserved gene pairs.

Table 13: Heptamers with maximum score that were used to select the 7 A-rich motifs (plus the *prfB* motif) that could cause programmed frameshifting

Rank	+1 heptamer	Score
1	GTG_CGC_G	594.3
2	**CTT_TGA_C**	**428**
3	GAC_GAG_G	326.04
4	CTG_GAA_A	299.57
5	TGG_CGC_G	280.16
6	**AAA_AAA_A**	**247.87**
7	AAA_GAG_G	195.07
8	**AAA_AAA_T**	**193.47**
9	CTT_TCC_A	175.86
10	GTC_ATC_G	129.79

Rank	-1 heptamer	Score
1	**A_AAA_AAG**	**2105.72**
2	**T_AAA_AAA**	**496.83**
3	T_CAA_TTA	289
4	**A_AAA_AAC**	**231.05**
5	C_CTG_CCG	198.45
6	**A_AAA_AAA**	**79.84**
7	G_CGC_GGC	72.87
8	G_GAG_GCA	69.74
...	...	...
30	**G_AAA_AAA**	**26.53**

It is known that poly-A motifs frequently appear in frameshift sites. Among the consensus heptamers found in our analysis there were seven A-rich heptamers (AAA_AAA_A, AAA_AAA_T, A_AAA_AAG, T_AAA_AAA, A_AAA_AAC, A_AAA_AAA and G_AAA_AAA) – see Table 13.

Programmed frameshift (Recoding) sequence sites are considered as "singular genomic elements" [3] to emphasize that their occurrence in coding regions causes locally beneficial effects: the sites are conserved at specific genomic locations across several species, but are avoided at other locations. This consideration motivated several authors to analyze occurrences of frameshift-prone sequences within protein coding genes. Shah et al. in their analysis of heptanucleotides frequencies in *S. cerevisiae* genes, observed that known frameshift- prone sequences (CTT_AGT_T [77, 145] and

CTT_AGG_C [78]) rank among the least represented heptanucleotides [79]. An analysis of the *E. coli* genome showed that the frameshift prone motifs, A_AAA_AAG [58, 60] and CCC_TGA [146], are indeed underrepresented (especially in highly expressed genes), however, not infrequent [118]. We confirmed this observation for *H. influenzae* and *V. cholerae* genomes (the sets of highly expressed genes were taken from [147]). Among the heptamers selected above only CTT_TGA_C was avoided in highly expressed genes. The other four poly-AT hexamers were present in highly expressed genes, but the their ranking computed for highly expressed genes was always lower than the ranking computed for the full set of genes.

To build a logo of the conserved motif at the frameshift site [148], we used a nucleotide frequency matrix obtained from the Gibbs Sampler alignment (frame information was omitted). The alignments of detected short motifs were extended 20 nucleotides upstream and downstream to include possible stimulatory sequences. The extended logos were built from the frequency matrices produced by extended alignments (Tables 7 and 8).

Notably, finding a conserved motif does not guarantee that the fs-gene cluster in question contains genes with programmed frameshifts. Evolutionary conserved sequences could occur in regions of overlaps of homologous gene pairs. Therefore, for cluster classification we used the several features described in Table 12.

3.3.6 A new measure of motif periodicity

Known biologically meaningful frameshift motifs contain number of conserved nucleotides (for example, the prfB motif with a consensus GGGGGXXTCTTTGAC or the IS motifs that have stretches of conserved A's). We observed that some strong motifs found in the frameshift vicinity had strong periodicity, i.e. such that exhibited a pair of conserved nucleotides (in the first two codon positions) followed by one not conserved nucleotide (in the third codon position). Such periodicity corresponds to a

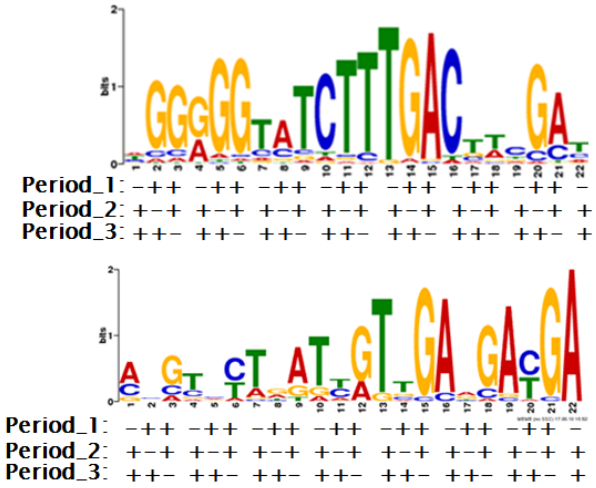


Figure 25: Masks used to calculate motif periodicity of a motif

particular reading frame; it suggests that a relatively strong motif has good E-value because of properties of genetic code in an alignment of orthologous proteins and not because this region has a conserved frameshift motif.

In order to discriminate between periodic and non-periodic motifs the following approach was used (see Fig. 25). Every position in a motif has its own information content (IC) calculated as:

$$IC(i) = 2 + \sum_{\alpha} p_i(\alpha) \log_2(p_i(\alpha)) \tag{7}$$

where $p_i(\alpha)$ is frequency (probability) of a letter α at position i. The information content of a motif position is used to define a total height of a letter height in the Logo diagram ([148]).

We applied three masks to a motif profile. Each mask consists of two 1 and one -1 and corresponds to a particular frame. The length of the mask (total number of 1 and -1) is equal to the motif length, so for every mask (frame) the following score was calculated:

$$score(mask_n) = \sum mask_n(i) IC(i) \tag{8}$$

where $mask(i)$ is either 1 or -1 (depending on the mask). The logic of this approach is that if motif has periodicity, one of three $score(mask_n)$ is much higher than for other

two masks, when the mask 1's correspond to big values of IC and -1's correspond to low values of IC. The three numbers $score(mask_n)$ are converted into a single periodicity score (PS) that is normalized to the motif length:

$$PS = \frac{\max(score(mask_n) : n = \{1,2,3\}) - \min(score(mask_n) : n = \{1,2,3\})}{length(motif)} \quad (9)$$

Larger values of PS correspond to more periodic motifs.

3.3.7 Inferring a type of frameshifting mechanism

It is hard to determine the nature of events that occur at poly-A/T frameshfit sites since reading frame shifts may occur during both transcription and translation (for example in transposase and in *dnaX* gene decoding). Still we have sequence dependent hypotheses on the mechanism of putative programmed frameshifts that occur in fs-gene clusters (Table 7).

Hypotheses regarding the frameshift mechanism utilized were based on the following considerations. The direction of ribosomal frameshifting (+1 or -1) is generally conserved (for example, all Release Factor 2 genes utilize +1 frameshifting). In contrast, our data shows that in genes of transposases that use transcriptional realignment, the direction of the shift in reading frame can be either or both +1 and -1; in this case homologous fs-genes form two separate clusters; with +1 and -1 frameshifts. For example, for HTH_Tnp_IS630 (Transposase), the "-1" cluster contains 495 fs-genes and "+1" cluster contains 185 genes (Table 7). Both clusters had the same overrepresented hexamer AAA_AAA and contained homologous sequences. In either cluster only one of the directions (the one predicted by GeneTack) leads to synthesis of a full length protein. We expect that an fs-gene with a programmed frameshift motif that may cause both +1 and -1 shifts (such as poly-A) could produce three types of products: two distinct truncated proteins (if the programmed frameshift did not occur or a frameshift occurred in a "wrong" direction) and full length functional protein (if the programmed frameshift occurred in the "right" direction).

Our approach in computationally assigning programmed shift mechanisms was mainly based on whether the frameshifts in the cluster are directional or not (Table 7). Namely, if there are several clusters with the same function, but different frameshift direction, we consider them as PTR, otherwise PRF is suggested.

We attempted to find RNA secondary structure conservation in the vicinity of the predicted frameshift site. The presence of conserved structure downstream from the frameshift site would support the PRF mechanism. Unfortunately this approach was not very helpful because conserved secondary structures, quite common in mRNA, were found in gene sequences that undergo both ribosomal frameshifting and transcriptional realignment.

In some known cases of programmed frameshifting the true mechanism is still being debated. Notably, PRF was assumed in the pioneer papers about programmed frameshifts in transposase gene expression [47], however, subsequent work has shown that PTR likely occurs in the expression of quite a number of these genes [16]. Pertinently, different frameshift mechanisms are used in different prokaryotic species to express homologous *dnaX* genes [62, 61].

3.3.8 Clusters of sequences with frame transitions determined by phase variation and translational coupling

Phase variation, a reversible and inheritable change of bacterial phenotype is often considered as a random process evolved to facilitate immune evasion of a host. Phase variation has been studied mainly in bacterial pathogens; however, it may occur in non-pathogens as well [95]. The majority of proteins encoded by genes known to be relevant to phase variation, are exposed to the environment. Examples include proteins involved in capsule, fimbriae, pili, flagella as well as surface proteins: transporters, receptors and porins. Notably, many of the large clusters contain fs-genes for cell surface and secretory proteins. However, phase variation has also been associated with DNA modification and metabolism associated genes [96].

Among molecular mechanisms of phase variation (homologous recombination, inversion of DNA elements, etc.), slipped strand mispairing (SSM) seems to be the most common. During replication SSM may occur at repeat units (such as short sequence repeats, microsatellites or variable number tandem repeats). The repeat unit could be as simple as a homopolymer sequence (e.g. poly-A in the *p78* gene of *M. fermentas* [97] or poly-C/poly-G in the type III methyltransferases genes [98]) or a repeat of more complex subunits (for example AGTC is repeated over 30 times in the mod gene of *H. influenzae*). Insertion or deletion of a repeat unit upon replication creates a frameshift mutation to turn the gene on or off.

Candidate clusters of fs-genes related to phase variation are identified due to the presence of characteristic repeats close to the frameshift position in genes encoding fs-proteins with functions earlier associated with the phase variation mechanism.

Translational coupling, also known as translational re-initiation, occurs in polycistronic mRNAs when the start of the downstream gene is located close to the end of the upstream gene (in the same or a different reading frame). After completing translation of the upstream gene, the ribosome moves several nucleotides backward or forward to the start codon of the next gene to begin translation. This mechanism helps maintain a specific ratio between the concentrations of the expressed proteins.

To find clusters related to translational coupling, we analyzed the vicinity of the upstream ORF (ORF1) stop codon limiting the frameshift box of the fs-gene ORF pair. If a downstream ORF (ORF2) start codon occurred within 10nt (20nt for some clusters) distance from the ORF1 stop codon, the predicted frame transition would indicate a junction of two genes (in different frames) expressed via translational coupling. Observation of phylogenetic conservation of co-location of ORF1 stop and ORF2 start codons would further support the translational coupling hypothesis. For a cluster with 100 or more fs-genes, the fraction of fs-genes with an ORF2 start within 10 nt from the ORF1 stop was calculated (%S). The prediction of the biological process

in Table 10 (phase variation or translational coupling) was based on the comparison of %S with %R and %AT values. We assumed that a high fraction of fs-genes with evolutionary conserved short gene overlap or short distance between adjacent genes (%S column), defines a candidate cluster of fs-genes with translational coupling.

3.3.9 Experimental verification of predicted programmed frameshifting

Bacterial strains. The *Escherichia coli* strain DH5α and MG1655Δ*lacIZ* were used for plasmid propagation and western blot respectively. Strains were grown in Luria-Bertani (LB) plus or minus isopropyl-β,d-thiogalactopyranoside (IPTG).

Plasmid construction. The vector pJ307 was derived from the GST-MBP-His fusion vector (pGMH57) by ligating annealed oligonucleotides (5' GATCAGCTC-GAGCACTAGTCCATGGGGATCCAAG 3' and 5' AATTCTTGGATCCCCATG-GACTAGTGCTCGAGCT 3' into pGMH57 between BamHI-EcoRI restriction sites of pGHM57 [149]. 20 inserts were constructed by PCR amplification of complementary oligonucleotides to give a full length sequence containing 5' XhoI and 3' BglII restriction sites. These were restriction digested and then ligated into the vector pJ307, digested by compatible restriction enzymes PspXI and BamHI (present in the new cloning site of pJ307), so that the *MBP* gene was in an alternative frame (+1 or -1) relative to *GST* or in-frame for positive control. Table 9 contains the full length sequences of the inserts.

Western Analysis. Overnight cultures of strains expressing the appropriate plasmid were diluted 1:100 in LB Broth, grown for two hours at 37°C, and then induced with 100 mM IPTG for an additional two hours at 37°C. Crude extracts obtained by culture centrifugation and re-suspending the bacterial pellet in Laemmli sample buffer. Proteins were separated on 10% SDS PAGE gels and transferred to nitrocellulose membranes (Protran). Immunoblots were incubated at 4°C overnight

in 5% milk/PBS-Tween containing a 1:500 dilution of rabbit anti-GST or 1:2000 dilution of rabbit anti-HIS. Immunoreactive bands were detected on membranes after incubation with appropriate fluorescently labeled secondary antibodies using a LI-COR Odyssey®Infrared Imaging Scanner (LI-COR Biosciences). The amounts of termination and frameshift product were quantified by ImageQuant. The frameshifting efficiency was estimated as the ratio of the amount of frameshift product to the total amount of the termination plus frameshift products.

3.4 Discussion

Here we presented the general strategy for characterization of genes containing reading frame transitions (fs-genes) and the analysis of such genes from the completed bacterial and archaeal genomes. As a result of this analysis we characterized 5,632 genes that can be organized in 146 clusters as candidates recoded genes. Using reporter genetic constructs based on the sequences from 20 of these clusters we confirmed that the dataset is enriched with *bona fide* recoding cases with genes from four clusters showing high frameshifting efficiency.

In our classification of fs-genes and their clusters we considered a number of features (Table 12). Despite this many gene clusters remained uncharacterized as we were unable to ambiguously determine the nature of frame transitions in them. It is possible that the difficulty reflects evolutionary relationship of genes in the cluster. Conservation of overlapping ORFs indicates functional relationship between their products [150]. Therefore such ORFs are likely to be co-regulated, when different mechanisms of co-regulation are utilized in the genes from the same cluster, it is difficult to unambiguously assign the nature of the frame transition. In some examples that have been tested experimentally we observed, both products of frameshifting and initiation at the downstream ORF, suggesting that translational coupling and recoding mechanisms could be interchangeable.

The majority of predicted recoding candidates are insertion sequence elements. This is consistent with previous observations of frequent use of recoding mechanisms in the mobile elements [45, 135].

Genes with frameshifts present a difficulty for standard annotation procedures. Often it is very difficult to discriminate between frameshifts due to deleterious indel mutations and recoding mechanisms. This difficulty leads to errors and incorrect annotations. Even among genes with well proven and established mechanisms of recoding, such annotation errors are frequent. For example 17 out of 428 fs-genes from the release factor 2 cluster were annotated as pesudogenes. In total, out of 5,632 fs-genes that were classified as recoding candidates, 721 were annotated as pseudogenes. It is likely that a large portion of these genes is annotated erroneously.

Due to the lack of universal methods for identification of recoding instances and classification of frameshifted genes, it is likely that erroneous annotations will continue to prevail. Nonetheless we hope that the current analysis and the web-based tools developed in the course of this work is a leap toward to solving the problem of annotation of genes with frameshifts.

Chapter IV

COMPARATIVE GENOMICS ANALYSIS OF EUKARYOTIC MRNAS WITH FRAMESHIFTS

4.1 Introduction

One of the dogmas of modern molecular biology is that protein coding eukaryotic mRNA contains one ORF only. Analysis of a newly sequenced mRNA is focused on the identification of start/stop of a single gene (if it exists) and exon boundaries. Because of this, alternative reading frames are frequently remain not annotated.

The goal of the present study is to identify mRNAs with two or more coding regions in different frames that in the current annotation belong to either the 5' or 3' untranslated region (UTR). Presence of additional coding regions inside mRNAs could be result of frame shifting alternative splicing, indel mutations inside exons or cases of programmed ribosomal frameshifting. Moreover, dual coding regions are also of interest of this study.

Up to one-third of alternative splicing variants contain premature termination codons (PTCs) [100, 101]. Although such mRNAs are degraded by Nonsense-mediated decay (NMD), a number of frameshifts were predicted in PTC-containing mRNAs. We refer to these frameshifts as alternative splicing frameshifts (AS FS).

Indel mutations inside exons are another reason for PTCs formation. These mutations may occur during DNA replication. Indel mutations also lead to shift of reading frame.

Genes utilizing programmed ribosomal frameshifting (PRF) also have frame transition to another frame however it does not lead to PTCs and mRNA degradation by NMD. PRF occurs when ribosome encounters specific signals embedded in the

Table 14: Examples of known eukaryotic dual coding genes. **Genes** – pair of genes in two different frames that share the dual-coding region; **DC (nt)** – length of the dual-coding region (in nucleotides); **Ori** – origin of the dual coding region: AS (alternative splicing isoform), IG (internal gene); **Refs** – references.

Genes	DC (nt)	Ori	Species	Refs
XLαs/ALEX (GNAS1 locus)	1068	IG	Mammals	[154]
Galectin-3/galig	318	IG	*H. sapiens*	[155]
INK4α/ARF	317	AS	Mammals	[156]
4E-BP3/MASK	172	AS	Mammals	[157]
CREB1a/CREB1c	41	AS	*Aplysia*	[158]
IGF-1Ea/IGF-1Ec (MGF)	26	AS	*H. sapiens*	[159]
LRTOMT	$\approx$200	AS	Mammals	[160]

same mRNA and dynamically shifts reading to another frame. Programmed ribosomal frameshifting is a type of recoding – the term that is used to refer to cases of nonstandard decoding [151]. Genes utilizing PRF are present in all domains of life. In eukaryotes PRFs are most frequently used in retrotransposons [52, 152, 53], but there are several known examples of eukaryotic cellular programmed frameshifted genes including ornithine decarboxylase antizyme [65], human *PEG10* gene [82] and its mouse homolog *Edr* [83], human paraneoplastic *Ma3* gene [84] and yeast *EST3* [77] and *ABP140* genes [153].

Dual coding is a phenomenon when the same stretch of DNA encodes two protein sequences in different frames. Due to the codon codependency of the overlapping frames dual coding is generally thought to prevail in pathogenic organisms under pressure to maintain a compact genome [161] and only recent results pointed to its possible importance in mammals in general and in the human genome in particular. The examples of eukaryotic dual coding genes are listed in Table 14. Although dual coding in prokaryotes and viruses evolved to compact their genome, the described phenomenon in eukaryotes participates in the extension of complexity and plasticity. The most striking example is the expression of ALEX protein from the GNAS1 locus where a single mRNA simultaneously produces the alpha subunit of G-protein from

the main reading frame, and a completely different protein, ALEX, from the 1068 nt long ORF in the +1 frame. Taking into account its genomic organization the ALEX can be called internal gene of longer XLαs gene. Another example of internal gene is the human *galig* that is located inside *Galectin-3* gene (318 nt long ORF in the alternative frame). In other known cases of dual coding there is just a fragment of dual coding sequence shared by two ORFs and each product is translated from its own mRNA produced by alternative splicing. Our analysis revealed several potentially new cases of eukaryotic internal genes.

There were a number of computational studies aiming to specifically predict eukaryotic frame shifting alternative splicing isoforms [101], indel mutations [162], programmed frameshifting [84, 41] and dual coding genes [163, 164, 165]. The approach used in the present study, the GeneTack program [127], allowed us to identify instances of all these events together. We are not aware of any similar computational study that would analyze all available mRNAs for more than 100 eukaryotic species. It should be noted that GeneTack was previously applied for frameshift prediction in eukaryotic EST sequences from *Puccinia triticina* [166].

4.2 Materials and methods

4.2.1 Sequence data

Intronless mRNA sequences were downloaded from RefSeq and GenBank databases. The number of available mRNAs significantly varies for different species and 91 genera each containing at least 15,000 mRNAs have been selected. In total there were 2,972,967 sequences related to the chosen genera that were used to train the HMM models (sequences containing non-ACGT characters were disregarded). Out of them the 1,165,799 mRNAs with annotated CDS were selected for GeneTack run. These sequences had at least one of the following properties: (i) the 3' UTR is longer than 50 nt; (ii) the 5' UTR is longer than 50 nt or (iii) the length of the annotated CDS

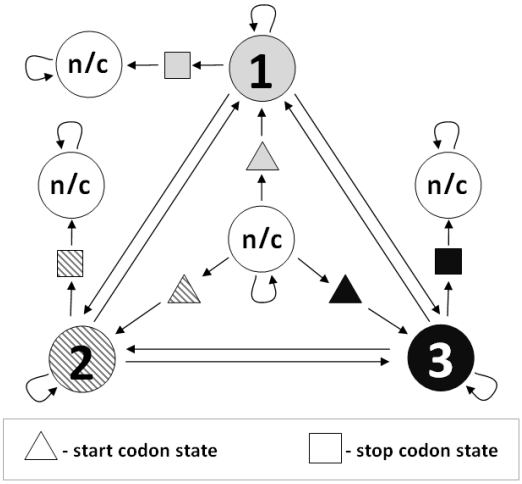


Figure 26: GeneTack HMM that was used to predict frameshift in eukaryotic mRNAs. The HMM is based on the assumption that eukaryotic mRNAs contain one gene only.

is not divisible by 3.

Poly-A tails (3 or more consecutive A's) at the 3' end of all mRNA have been trimmed. Duplicated mRNAs (identical sequences from the same species) have been removed as well.

4.2.2 HMM structure and parameters

It is known that eukaryotic mRNA contains one coding sequence only, so the original GeneTack HMM designed for frameshift predicted in prokaryotic genomes was modified. Namely, the ability to predict several adjacent genes (overlapping or not) in the same sequence was removed. The eukaryotic version of the HMM used in this work allows prediction of one gene only (with any number of internal frameshifts) flanked by optional UTRs (see Figure 26).

Emission probabilities for the HMM model were estimated from a collection of mRNA sequences using self-training program GeneMark.hmm [121]. In order to increase amount of sequences for training we combined all the mRNAs from the same genera (for example, all mRNAs of different *Drosophila* species were combined together) and grouped the sequences by GC% into 5% bins. For every bin containing

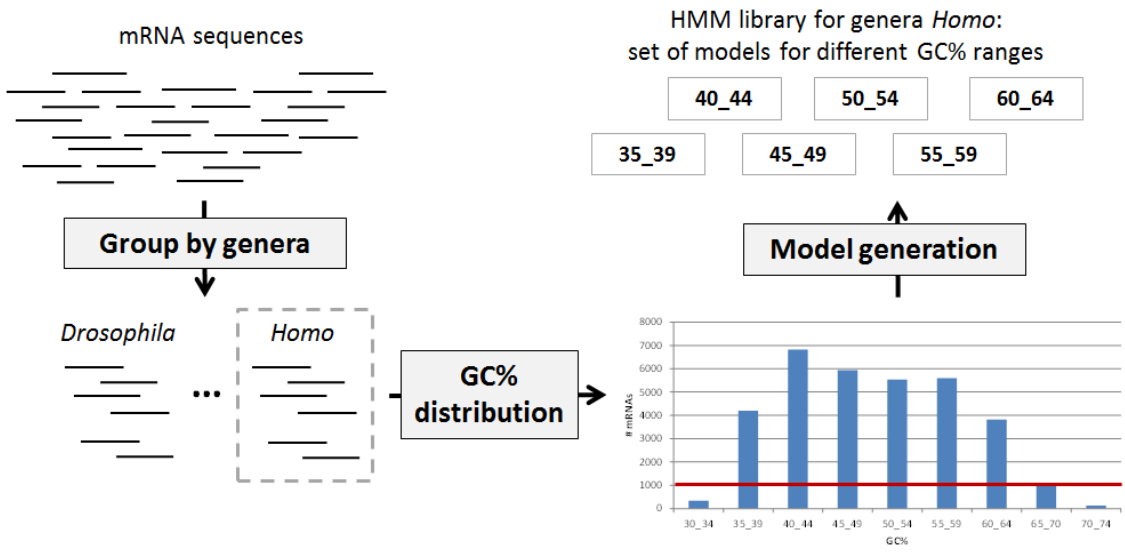


Figure 27: Genera-specific models for GeneTack were generated from 5% GC% ranges, each containing at least 1,000 mRNAs. In the example below, for the mRNA with GC% lower than 35 the 35_39 model would be used and for sequences with GC% higher than 64 the 60_64 model would be used.

at least 1,000 mRNAs an HMM model was generated (Fig. 27). Therefore a library of HMM models was generated for each of the 91 selected genera.

For frameshift prediction the model was chosen in the following way. For each input mRNA the library corresponding to its genus was used. Next, GC% of the input mRNA sequence was calculated (the poly-A tail was trimmed for all sequences and does not affect GC%) and the model for the corresponding GC% range was chosen. If the model for the GC% of the input mRNA does not exist (because there were not enough sequences in this GC% range), the model for the nearest available GC% range was used.

We have shown that combining mRNAs from different species of the same genus as well as use of 5% bins (rather than 1%) do not decrease quality of frameshift prediction, but allowed us to generate models for larger number of eukaryotic species.

4.2.3 Preparation of the test set with artificial frameshifts

1,000 human mRNAs with annotated CDS longer 1,000 nt and containing 5' and 3' UTRs longer 50 nt were randomly selected from the pool of all available human mRNAs. The sequences were selected with uniform distribution of GC% (the lowest GC% was 35% and the highest was 74%). Inside every CDS a frameshift was simulated by inserting a single nucleotide at a random position not closer than 50 nt from gene border. The locations of the simulated frameshifts have been recorded and were later used to discriminate true positive from false positive predictions. Namely a predicted frameshift was considered as true positive if the distance to the simulated location was closer than 50 nt.

4.2.4 Filters

Two filters were applied to the predicted frameshifts. We noticed that many frameshifts (up to 10) could be predicted in some mRNA sequences. Such a high number of predicted frameshifts probably reflected the unusual codon frequencies used in the mRNA rather than cases of authentic reading frame transitions. mRNAs with 4 or more predicted frameshifts were filtered out in our analysis.

Another filter was similar to one of the filters in prokaryotic GeneTack-GM program. Frameshifts predicted closer than 50 nt to the fs-gene border were discarded.

4.2.5 Clustering

All the genes containing predicted frameshifts (fs-genes) were conceptually translated into proteins (fs-proteins). Based on "all-against-all" BLASTp search results (with E-value threshold 10^{-10}) the fs-proteins were grouped into clusters taking into account sequence similarity, frameshift location and direction. The same clustering approach was used to cluster prokaryotic fs-genes (see above).

4.2.6 Exon mapping

Information about exon locations for 15 species with maximum number of mRNAs in our data were downloaded from Genome Browser website[1] and matched with the mRNAs where frameshifts were predicted. In total, 12,159 mRNAs had exon annotation.

4.3 Results

4.3.1 Testing quality of frameshift prediction in eukaryotic mRNAs

To measure the frameshift prediction accuracy of the eukaryotic GeneTack HMM we prepared two test sets: (i) 1,000 human mRNAs with artificially simulated frameshifts and (ii) mRNAs containing real annotated PRFs (such as antizyme).

On the first test set the best GeneTack performance in terms of average sensitivity and specificity was observed with the frameshift probability 10^{-7} (the direct transition between any two coding states). With this probability the sensitivity was 84.1% and specificity was 87.2%.

The second test set consisted of 7 human mRNA sequences corresponding to known genes utilizing PRF: four ornithine decarboxylase antizyme (NM_001134939, NM_002537, NM_004152, NM_016178) and three PEG10 (NM_001172437, NM_001184961, NM_015068) sequences. The locations of the programmed frameshifts are annotated in these sequences and were used to estimate accuracy of the frameshift prediction. Frameshifts were predicted in all the sequences: in 6 cases the distance between prediction and the true location was less than 50 nt and in 1 case (ornithine decarboxylase antizyme 2) the distance was 55 nt.

These results indicate that the eukaryotic version of GeneTack is able to efficiently predict frameshifts in eukaryotic mRNA sequences.

[1] http://genome.ucsc.edu/cgi-bin/hgTables

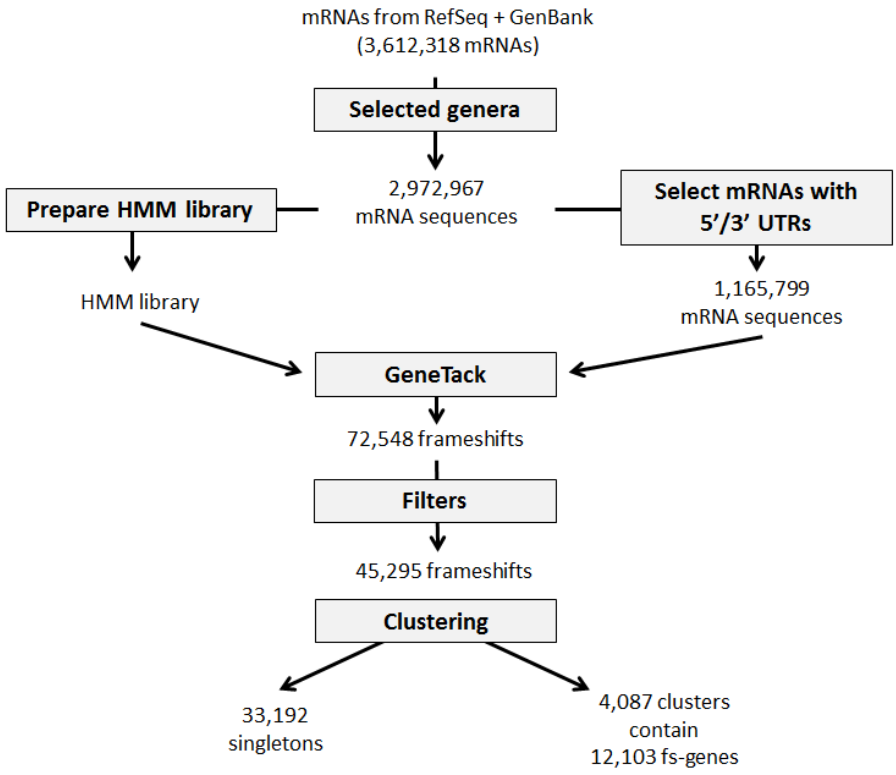


Figure 28: Pipeline of the work (see text for more details).

4.3.2 Predicting frameshifts in the eukaryotic mRNAs

GeneTack was applied to the selected 1,165,799 mRNAs. 45,295 frameshifts were predicted in 37,307 mRNAs, so only 3.2% of all mRNAs contained genes with frameshifts (fs-genes). All the predicted fs-genes were conceptually translated into fs-proteins and 12,103 of them were clustered in 4,087 clusters (Fig. 28). Most of the clusters contained just a few frameshifts: there were 2,608 clusters containing two fs-genes only and 3,837 (94% of all clusters) contained 5 or less fs-genes.

4.3.3 Rediscovery of known programmed frameshifting events

First we attempted to find known cases of programmed frameshifting among the obtained clusters. This was done in two rounds. First, 207 protein sequences that correspond to eukaryotic programmed frameshifting events were downloaded from the RECODE database [126]. BLASTp search with e-value threshold of 10^{-10} was

Table 15: GeneTack clusters that correspond to known cases of programmed frameshifting. **Cluster ID** – unique identifier of a cluster; **Name** – cluster name; **Size** – number of fs-genes in the cluster; **Species** – number of different genera; **D** – frameshift direction (+1 or -1); **FS Site** – frameshift site; **Ref** – references.

Cluster ID	Name	Size	Species	D	FS site	Ref
347704463	Antizyme 1/2/4	19	13	+1	TCC_TGA	[65]
711658885	Antizyme 3	11	10	+1	TCC_TGA	[65]
624530279	Antizyme (*Drosophila*)	5	1	+1	TCC_TGA	[65]
579332711	PEG10/Edr gene	11	7	-1	G_GGA_AAC	[82, 83]
891480314	Ma3 antigen	5	3	-1	G_GGA_AAC	[84]

done in order to find fs-proteins that are similar to the selected RECODE sequences. Subsequent analysis identified 4 clusters with 5 or more fs-proteins that had RECODE hits (Table 15). 29 singletons also had Recode hits: 15 of them corresponded to mobile elements, 8 had hits to the protein kinase Ndr from *Euplotes octocarinatus* [167], 3 fs-proteins corresponded to antizyme and 3 – to the product of *ABP140* gene [78].

An additional search for programmed frameshift clusters was based on analysis of the frameshift vicinity. We attempted to find clusters with conservation of known frameshift sites (that included X_XXY_YYZ motif causing -1 frameshifting). Manual analysis of the clusters with 5 or more fs-genes with conserved frameshift site revealed a cluster of five paraneoplastic antigen *Ma3* that is absent in the Recode database (see Table 15).

4.3.4 Frame shifting alternative splicing isoforms and indel mutations

The frameshifts predicted in frame shifting alternative splicing mRNAs variants are located close to exon-exon junctions. On the other hand the frameshifts corresponding to the indel mutations are randomly distributed along the exon. If an indel mutation located far enough from the exon-exon junction it can be distinguished from the alternative splicing frameshifts. The approach is based on analysis of stop codons between the exon 5' end and the predicted frameshift. Namely if stop codons are

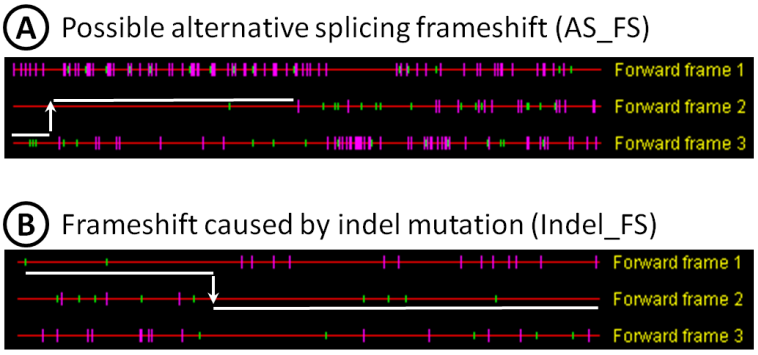


Figure 29: Discrimination between a frameshift caused by alternative splicing (AS_FS) and indel mutation inside exon (Indel_FS). Three-frame translation of exons where frameshifts were predicted is shown (the images were obtained using http://arbl.cvmbs.colostate.edu/molkit/translate/). White lines indicate the translation path predicted by GeneTack with arrows indicating frameshift locations. Purple and green dashes indicate stop and start codons respectively. (A) The translation of a terminal exon started in frame 3 while the correct frame is 2. -1 frameshift (from frame 3 to frame 2) is predicted near the beginning of the exon suggesting that this is likely to be frame shifting AS isoform (note that there is no stop codons in frame 2 upstream of the frameshift position); (B) An example of an exon with indel mutation. The translation of the exon started in the correct frame 1, because there are stop codons in frame 2 and 3 upstream of the predicted frameshift position. +1 frameshift predicted in the middle part of the exon corresponds to indel mutation.

observed in two frames other than the reading frame, the frameshift cannot be result of alternative splicing (Fig. 29).

The distance from the predicted frameshifts and the closest exon-exon junction is the main indicator of the AS_FS. The further predicted frameshift from the exon boundary, the more chances to see stop codon in both alternative frames and reject the hypothesis that the frameshift is a result of alternative splicing (Fig. 30).

Many large clusters were classified as result of alternative splicing indicating that homologous frame shifting alternative isoforms are produced in a number of different species (Table 16). Among the large clusters we also identified two cases of indel mutations and three clusters containing instances of both indel and alternative splicing frameshifts.

Table 16: Largest clusters containing 10 or more fs-genes from at least 5 different species. **Cluster ID** – unique identifier of a cluster; **Name** – cluster name; **Size** – number of fs-genes in the cluster; **#S** – number of different species; **D** – frameshift direction; **%E** – fraction of cluster's fs-genes that have exon annotation; **EJC** – exon junction distance (average distance in nucleotides from the frameshift to the nearest exon-exon junction); **Type** – possible biological nature of the cluster (DC – cases of possible dual coding, AS – alternative splicing, Indel – indel mutation inside exon, AS & Indel – mixture of AS and Indel frameshifts, FP – GeneTack false positive)

Cluster ID	Cluster name	Size	#S	D	%E	EJD	Type
347704463	Antizyme 1/2/4	19	13	+1	26%	115	Known PRF
711658885	Antizyme 3	11	10	+1	27%	37	Known PRF
579332711	PEG10/Edr gene	11	7	-1	45%	1122	Known PRF
165060982	Protein FAM98A	19	15	+1	37%	426	DC
620965413	Collagen alpha	19	10	-1	32%	40	DC
275483014	Helicase SRCAP	12	10	+1	17%	162	DC?
383732599	Helicase SRCAP	14	10	-1	36%	228	DC?
573284995	Protein phosphatase 1	12	9	+1	42%	43	DC?
637024743	Keratin 2	13	8	+1	54%	34	DC?
961071914	Zinc finger protein	164	15	-1	60%	583	AS & Indel
374934679	Zinc finger protein	57	7	+1	47%	417	AS & Indel
586305929	X-box binding protein	27	17	-1	41%	236	AS & Indel
564038905	RNA-binding protein EWS	36	16	+1	28%	33	AS
259815531	dystrobrevin alpha	24	7	+1	38%	8	AS
359078667	Calcium-transporting ATPase	23	12	+1	26%	6	AS
420611633	Transcription factor 7	23	7	-1	70%	29	AS
313759418	Molybdopterin synthase subunit	17	14	+1	18%	68	AS
771141998	Extracellular sulfatase SULF-1	17	11	+1	53%	38	AS
589659706	Mental retardation protein 1	15	9	+1	20%	1	AS
408865523	Ral GTPase-activating protein	15	7	+1	47%	36	AS
492939429	Histidine-rich glycoprotein	14	9	+1	29%	24	AS
570576076	Pkinase	14	7	-1	43%	22	AS
583991734	Pkinase	11	6	+1			AS
525426412	Tumor protein p63	13	5	-1	54%	4	AS
664710746	GNAI polypeptide	11	6	-1	36%	25	AS
142597958	Metabotropic glutamate receptor	11	6	+1	36%	31	AS
375223683	P2X_receptor	14	5	-1	50%	7	AS
358564776	Phosphatidylinositol transfer	10	10	-1	20%	54	AS
813971855	MHC, class I	10	5	+1	80%	22	AS
480481244	Zinc finger protein 44	12	11	-1	33%	56	AS
299671469	Zinc finger protein	141	15	+1	59%	825	Indel
559722434	Pro-neuregulin-1	12	6	+1	25%	65	Indel
380612508	Fork_head	10	6	-1	50%	537	FP
658903509	Eyes absent homolog 1	16	5	+1	38%	80	FP
704349068	Zinc finger protein	13	7	+1	54%	189	FP
705569377	Guanylate cyclase subunit	12	5	+1	33%	583	FP
886555733	Ankyrin 2	18	8	+1	56%	11	FP
193751720	Potential cation channel	55	13	+1	35%	228	FP

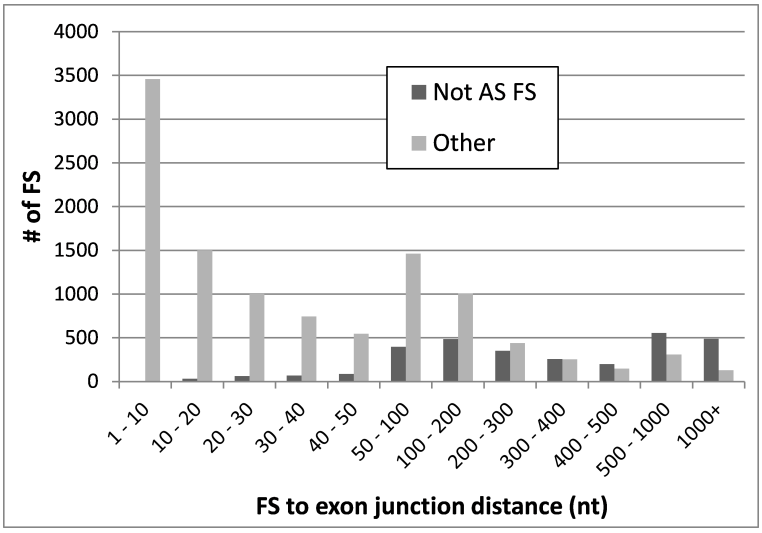


Figure 30: Distribution of the distance between predicted frameshift and the closest exon-exon junction for all mRNAs with known exon locations. The frameshifts predicted near exon-exon junctions are more likely to be caused by alternative splicing.

As was mentioned above many alternative mRNA isoforms with premature termination codons are apparent targets for NMD. Even though the majority of the predicted frameshifts are related to the alternative splicing isoforms, the fraction of such mRNAs in the pool of all the analyzed sequences is very small (about 3%). This indicates that the NMD is an efficient process and the sequenced frame shifting alternative isoforms could be recently produced mRNAs that have not been subject to NMD quality control yet.

4.3.5 Dual-coding mRNA sequences

Dual-coding regions of mRNA sequences have very unusual codon frequencies and relatively long ORFs in two different frames. We observed that such dual coding fragments are prone to frameshift prediction by GeneTack because the same region contains coding potential in two different frames. Analysis of the 38 largest clusters revealed 5 cases of potential internal genes (type of dual coding) that include two "Helicase SRCAP" clusters (containing +1 and -1 frameshifts) corresponding to the same mRNAs (Table 16).

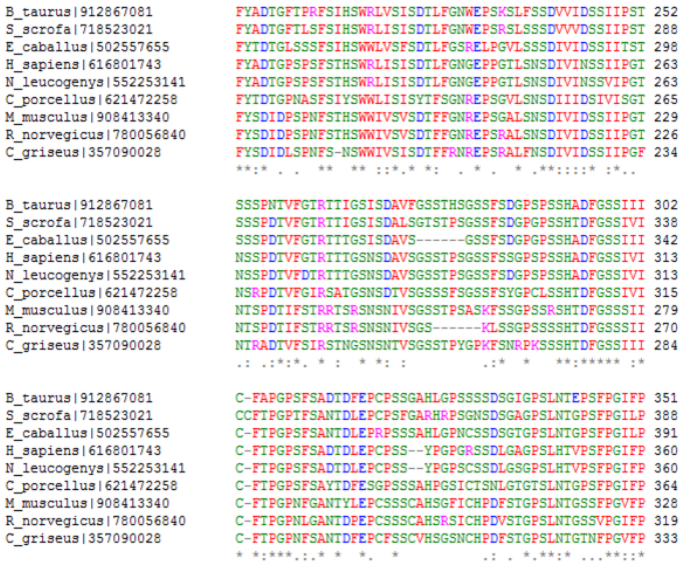


Figure 31: An alignment of the alternative frame translations derived from the SRCAP cluster (cluster ID 275483014) – one of the dual coding candidates.

In each of these clusters frameshifts were predicted where an ORF in the alternative frame was inside longer ORF in the main frame. The fact that GeneTack predicts a frameshift and shifts to alternative reading frame indicates that the ORF2 has coding potential higher than in ORF1. Alignments of the ORF2 translations revealed the conservation on the protein level across number of eukaryotic species (for example see Fig. 31).

To further confirm dual coding hypothesis for the candidates we analyzed K_a/K_s values [168] of the alternative reading frames. The K_a/K_s is the ratio of the number of non-synonymous substitutions to the number of synonymous substitutions in the pair of orthologous genes. The smaller the value of K_a/K_s, the stronger is the stabilizing evolutionary pressure on the gene that does not allow to freely change amino acids. Genes with K_a/K_s values less than 1 are usually important functional genes that are under evolutionary selection.

To discriminate dual coding candidates from the usual, single coding eukaryotic genes we first analyzed a homologous genes derived from HomoloGene database [169]. The database contains over 20,000 groups of homologous genes from a number of

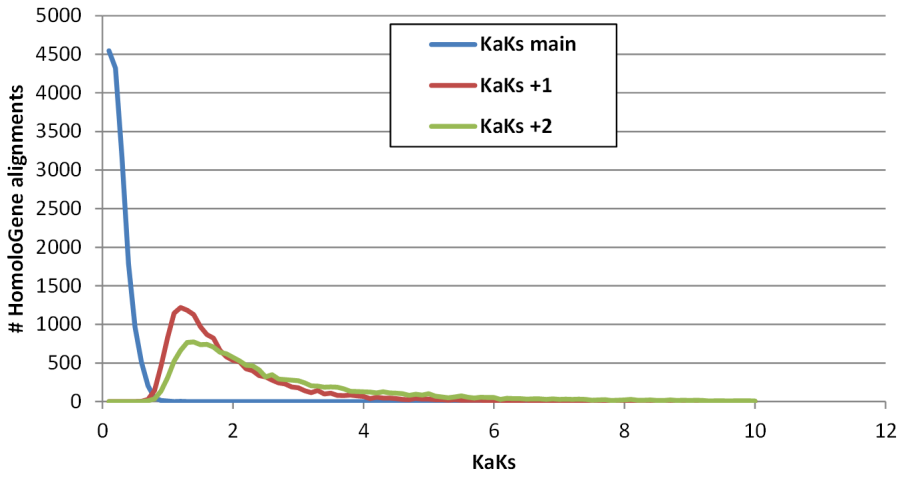


Figure 32: Distribution of the K_a/K_s values calculated for the main frame and for the two alternative frames for 15,576 pairwise alignments of human-mouse homologous genes from HomoloGene database. The distribution demonstrates that K_a/K_s values can efficiently discriminate main (coding) frame from alternative frames.

eukaryotic species. For our analysis 15,576 HomoloGene groups were selected. Each group contained one human and one mouse gene. All the human and mouse genes were translated in the main frame and pairwise protein alignments were built for each homologous group using MUSCLUE program [170]. The obtained protein alignments were converted into nucleotide alignments which were used to calculate the K_a/K_s values for the main frame and for the two alternative frames (+1 and +2). When calculating K_a/K_s for the alternative frames the "nonsense mutations" that lead to stop codons were considered as non-synonymous mutations; codons that included gaps were ignored.

As we expected, the K_a/K_s values for the main frame were usually less than 1 (indicating stabilizing selection) while for the alternative frames the values were higher than 1. Alternative reading frame in dual coding genes should posses two properties: (i) K_a/K_s values indicating stabilizing selection and (ii) existence of a relatively long ORF. Plots on Fig. 33 were built to check the dual coding hypothesis for the candidate fs-gene clusters. Based on the plots we concluded that the SRCAP cluster is the strongest dual coding candidate. We also noticed several HomoloGene

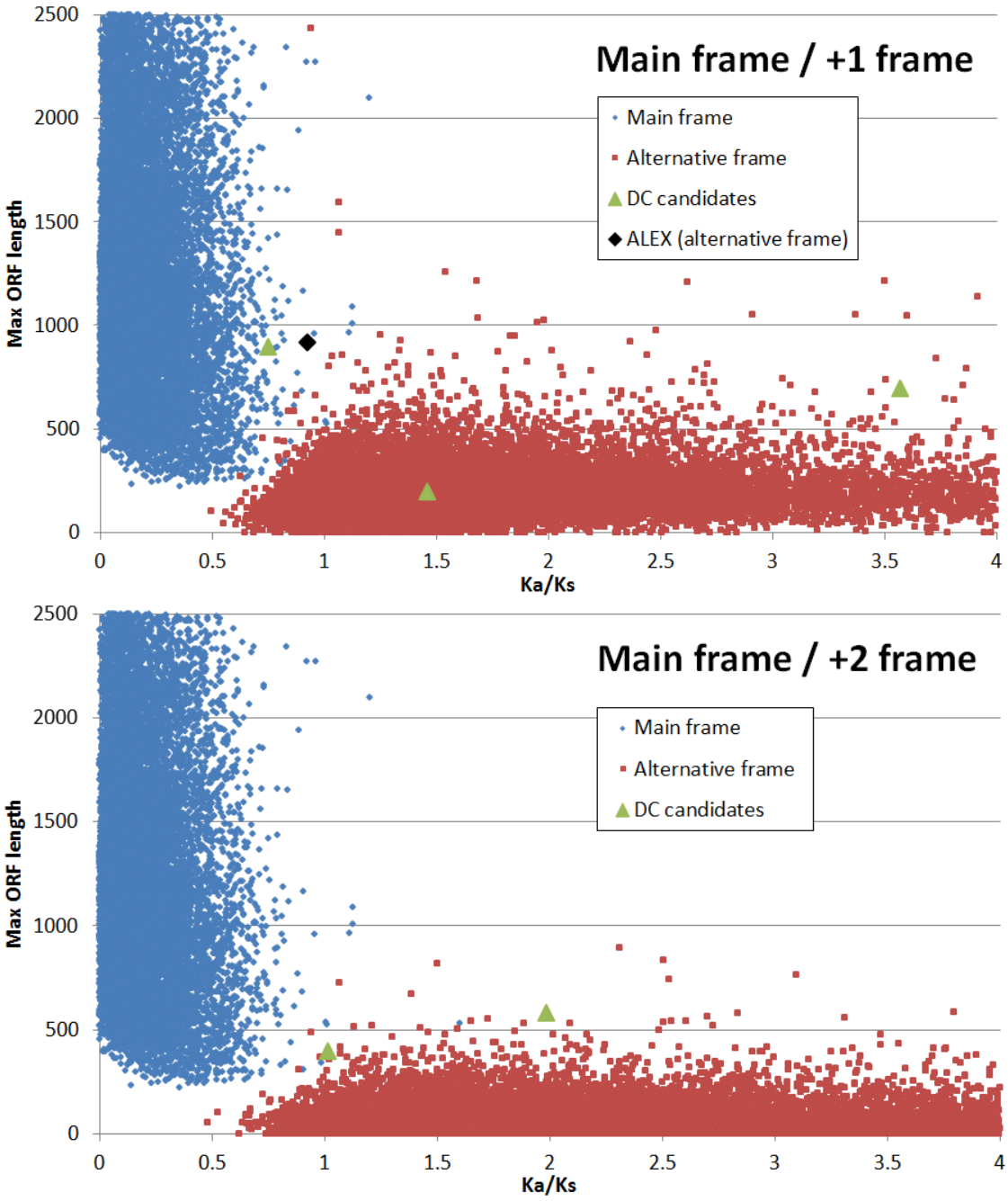


Figure 33: Comparison of K_a/K_s values and ORF lengths between the main frame and alternative frames obtained from pairwise alignments of homologous human-mouse genes from HomoloGene database. A single pairwise alignment produces two dots on each plot – one dot for the main frame (blue) and one – for an alternative frame (red). For a pair of human-mouse genes the longest ORF was found in a given frame (main, +1 or +2) and the average length between human and mouse longest ORFs was used for this frame as a Y value. Black dot on the top plot corresponds to the alternative frame of a confirmed case of dual-coding – the ALEX protein. This dot can be used as a reference point to find other coding candidates. The green dots correspond to the alternative frames of the dual coding candidates found among fs-gene clusters in normal, single coding genes.

groups that could also be cases of dual coding.

It should be noted that none of the proteins obtained from the identified alternative reading frames have reliable hits in the NCBI nr database suggesting that these could be new proteins produced in eukaryotic cells. Experimental verification is needed to support or reject our predictions.

4.4 Discussion

Here we presented the analysis of frameshift prediction in mRNAs from 91 eukaryotic genera. As a result of this analysis we predicted 45,295 frameshifts in 37,307 mRNAs (some mRNAs contain several predicted frameshifts). Taking into account GeneTack False Discovery Rate the expected fraction of false positive predictions is 13%. Up to 50% of the predicted frameshifts are related to the frame shifting alternative splicing isoforms. A number frameshifts are result of indel mutations, but the exact percentage of these events is hard to estimate. Finally there are few examples of programmed ribosomal frameshifting and dual coding genes.

We have clustered the 12,301 fs-proteins in 4,087 clusters. The set of 38 largest clusters containing 10 or more fs-genes from at least 5 different species was manually analyzed. Alternative splicing was the most common nature of the clusters. This is consistent with previous observations of the variety of frame shifting alternative isoforms [100]. Existence of clusters of AS isoforms indicates that the production of mRNAs with premature termination codons from the same gene is common among many species.

Analysis of the frameshift vicinity and comparison with RECODE database allowed us to classify 5 clusters as known cases of programmed frameshifting with three of them belonging to the list of 38 largest clusters.

We have identified 5 clusters of mRNAs that have dual coding regions encoding

internal gene. Alignment of the proteins obtained from the alternative frame translations showed conservation between number of species suggesting potential expression and functional importance of proteins encoded in the alternative reading frames.

It should be noted that many of the predictions were not classified and may be of interest to other researches. Web based interface to browse and search the mRNAs with predicted frameshifts is available on the GeneTack web page.

Chapter V

GENETACK DATABASE: GENES WITH FRAMESHIFTS IN PROKARYOTIC GENOMES AND EUKARYOTIC MRNA SEQUENCES

5.1 Introduction

Database annotations of prokaryotic genomes and eukaryotic mRNA sequences pay relatively low attention to frame transitions that disrupt protein coding genes. Earlier we have developed an algorithm and software program GeneTack for *ab initio* frameshift finding in intronless genes [127]. Here, we describe a database containing genes with frameshifts (fs-genes) predicted by GeneTack.

Frameshifts predicted by GeneTack correspond to reading frame transitions. The transition could be caused by many reasons, among them sequencing errors [15], indel mutations [14], programmed frameshifting events [7, 6, 4], phase variation [95], overlapping of adjacent genes [138], dual-coding regions [154], eukaryotic alternative splicing [100].

While sequencing errors are artifacts of sequencing technologies, authentic indel mutations correspond to real sequence features. These mutations usually lead to gene pseudogenization; still some pseudogenization remain conserved in evolution if the transcript (not truncated contrary to the protein product) carry some regulatory function [171].

In case of phase variation reversible indel mutations occur at high frequencies at specific sites. They generate a population of bacterial pathogens with heterogeneous sequences of phase variant gene thus increasing population fitness, since it may help some bacteria to escape immune response of a host [96]. Phase variation results in

reversible and inheritable variation of bacterial phenotype.

Programmed frameshifting occurs either during translation (programmed ribosomal frameshifting – PRF) or transcription (programmed transcriptional realignment – PTR). PRF and PTR violate standard triplet decoding allowing for a single protein to be produced from two overlapping ORFs. Hence, GeneTack predicts frame transition between these ORFs. PRF and PTR occur at sites with specific sequence patterns conserved in evolution since programmed frameshifting regulates gene expression. Programmed frameshifting usually results in synthesis of two protein products (standard and frameshift) that share the same N-terminal sequence but possess different C-terminal parts. Among chromosomal genes, the best studied examples are bacterial *prfB* gene encoding Release Factor 2 [1] and eukaryotic genes encoding Ornithine decarboxylase antizyme [65]. PRF is abundant in viruses [5], bacteriophages and transposons [135, 23]. The largest available collection of known PRF genes is available in the Recode database [126].

A frameshift could be predicted when two adjacent sequences (CDSs) that carry genetic code in different frames are located close to each other or overlap. Notably, a co-location of some of CDS pairs could be evolutionary conserved if expression of the two genes is linked by translational coupling mechanism. Such gene pairs predicted as fs-genes are present in the GeneTack database as well.

The eukaryotic part of the database was built using known mRNA sequences; a large number of predicted fs-genes was found in alternative spliced transcripts containing premature termination codons (PTCs) [100, 172]. This fact is not surprising taking into account that in mammals up to one-third of alternative splicing (AS) events produce PTC-containing splice variants [101, 173].

The database contains fs-genes that represent possible dual coding in eukaryotic mRNAs. Dual coding allows the same stretch of DNA to encode two protein sequences in different frames [174]. Multiple instances of dual coding in human genome

were detected by analysis of ribosomal profiling data obtained from HeLa cells [175]. Several instances of dual coding are well studied, such as the *xbp1* gene encoding x-box binding protein 1. The products of initial rounds of *xbp1* mRNA translation facilitate endonuclease mediated excision of a 26 nt fragment of its own mRNA. As a result, mRNA downstream of excision appears in a different frame [176] and different protein product is synthesized from the same mRNA at the later rounds of translation. Another well studied example is expression of the ALEX protein from the GNAS1 locus where a single mRNA is able to simultaneously produce two protein products from different reading frames of the same mRNA [154]. Similarly, tumor suppressor proteins P16(INK4a) and P14(ARF) are produced from the same gene, where the same sequence appears in alternative frames in two alternative transcripts [177]. Due to the codon codependency of overlapping frames [178] dual coding regions have unusual codon frequencies that make them prone to frameshift prediction by GeneTack.

The GeneTack database contains all types of frame transition events (prokaryotic and eukaryotic); ≈20% of entries have been characterized in terms of the probable nature of predicted frame transition.

To help explore the nature of predicted fs-genes they were grouped into clusters of orthologous fs-genes based on sequence similarity, conservation of frameshift direction (-1, +1) and location. We characterized the fs-genes that formed the largest clusters based on comparative genomics analysis (Antonov et al, paper in preparation). Although the nature of more than 80% of the predicted frameshifts was not revealed, (at least 1.5% have a strong evidence to be sequencing errors, while up to 54% could be related to sequencing errors) this database will be useful for improving annotation of new genomes, re-annotation of old ones as well as for stimulating experimental studies leading to identification of new programmed events and other cases of frame transitions under evolutionary selection.

Table 17: Statistics on eukaryotic and prokaryotic sections of the GeneTack database

	Prokaryotes	Eukaryotes
Number of species analyzed	1,006	100
Total number of predicted frameshifts	206,991	45,295
Total number of clusters	19,430	4,087
Number of fs-genes in clusters	102,731	12,103
Number of singleton fs-genes	104,260	33,192
Clusters with <5 fs-genes	14,441	3,701
Programmed frameshift clusters	239	5
Indel mutations clusters	4,010	2
Clusters of PTC-containing splice variants	n/a	21

5.2 Database statistics and usage

The data are stored in a local MySQL database queried by CGI scripts embedded in the web interface. The database also includes some pre-built data, such as Sequence LOGOs [148] of conserved motifs observed in overlapping ORFs for all the clusters.

The database consists of two sections – prokaryotic and eukaryotic. Notably, the method of frameshift prediction was slightly different in prokaryotic genomic DNA and eukaryotic mRNA. For prokaryotes, genes in a complete genome sequence were predicted by GeneMarkS [119] the self-training program that derived parameters both for itself as well as for GeneTack. A single statistical model was generated for each prokaryotic genome and use in GeneTack.

Eukaryotic genes with frameshifts were identified in mature mRNA sequences. Several HMM models were generated for each eukaryotic genus. Each model was generated by a self-training algorithm, a version of GeneMarkS, from a set of mRNAs with a close GC% content. All the eukaryotic and prokaryotic models are available at the GeneTack web page; a database user can choose an appropriate pre-built model for a query sequence.

Currently the database contains fs-genes from 1,106 prokaryotic and 100 eukaryotic species (see Table 17). Since the length of prokaryotic genomes as well as the

total size of available eukaryotic mRNAs vary for different species, the number of predicted fs-genes also varies. For example, in 115,001 human mRNAs we predicted 8,700 frameshifts while only 839 frameshifts were predicted in 32,155 mRNA sequences of *Rattus norvegicus*. Conceptual translation of predicted fs-genes produced a database of fs-proteins used for clustering. All over, 50% of prokaryotic and 27% of eukaryotic fs-genes formed clusters while other fs-genes were singletons.

The database home page is the user's entry point. The user can perform sequence similarity search by BLASTp for a query sequence of interest, search for a cluster name using a query string, or browse prokaryotic or eukaryotic clusters of fs-genes. Majority of the clusters were named using names of Pfam domain detected in the cluster of fs-proteins. However, several clusters (e.g. known cases of programmed frameshifting) were manually renamed to reflect gene and protein names. Thus, Release Factor 2 cluster can be found by using the gene name "prfB" as a query.

To allow search against the GeneTack database of fs-proteins two BLASTp databases (containing either prokaryotic or eukaryotic fs-proteins) were built. The BLASTp hit may reveal the nature of a frameshift mechanism in a novel sequence.

Finally, a user can browse sections of either of the two databases in the following ways. First, a particular species can be selected from a list of species. For a given species a list of all the predicted fs-genes is available (see Figure 34). The list provides information about every frameshift such as its direction and genomic coordinates. More detailed information about an fs-gene can be accessed by clicking on the fs-gene ID. A page with frameshift details provides the following information: the species name, the frameshift coordinate (in the prokaryotic genome or the eukaryotic mRNA), the frameshift direction (+1 or -1), the coordinates of the fs-gene, its length and the length of encoded protein. The initial fs-gene sequence (with a frameshift), the corrected fs-gene sequence and the sequence of conceptually translated protein

#	FS_ID	Coord	D	GeneL	GeneR	S	F	G	P	BLASTp	Pfam	COF	RBS
1	297796143	35380	-1	34781	36162	-	6877	780	260	-- / --	-- / --	63855303	0.48
2	453258176	72934	+1	72229	75453	-	4454	1584	528	8 / 0	-- / --	654703201	-0.73
3	256999994	74521	-1	72229	75453	-	2867	930	310	-- / --	-- / --	953823467	1.16
4	365072851	93162	+1	91413	98403	+	9294	1746	582	422 / 0	-- / --	237996460	1.39
5	606254630	97072	-1	91413	98403	+	13204	1074	358	-- / --	-- / --	448938455	0.92
6	928154345	115717	+1	114522	117051	-	3016	1332	444	-- / --	-- / --	354349696	1.56
7	146205650	129463	+1	129394	131161	-	2152	1695	565	-- / --	-- / --	125091330	
8	499384488	156422	+1	156334	156883	-	5683	459	153	-- / --	-11 / 2e-06	750652363	
9	687563479	159271	-1	159175	160094	-	2834	819	273	-- / --	4 / 0	612498387	

Figure 34: The GeneTack database entries: fs-genes predicted in genome of *Escherichia coli str. K-12 substr. DH10B*. **FS_ID** – unique fs-gene identificator, **Coord** – frameshift coordinate in the input sequence, **D** – frameshift direction (+1 or -1), **GeneL** – coordinate of left border of fs-gene (gene start for '+' strand, gene end for '-' strand), **GeneR** – coordinate of right border of fs-gene (gene end for '+' strand, gene start for '-' strand), **S** – the fs-gene strand, **F** – frameshift coordinate in fragment (the sequence used as input to GeneTack), **G** – frameshift coordinate in fs-gene, **P** – frameshift coordinate in fs-protein, **BLASTp** – information on the BLASTp hit covering frameshift position in the fs-protein, **Pfam** – information on the Pfam domain covering frameshift position in the fs-protein, **COF** – cluster ID (if available), **RBS** – RBS score of the downstream gene defined by GeneMarkS.

product are available as well. Additional information for a frameshift includes reference to the BLASTp/Pfam hit if it did occur to cover predicted frameshift position in the fs-protein. Link to the corresponding cluster is provided if the fs-gene belongs to the cluster. It should be noted that an fs-gene can belong to one cluster only.

Another way of browsing the database is by using a probable type of fs-gene. Some of the predicted fs-genes and the clusters of the fs-genes were grouped together based on their types. Each group of the clusters (for example, all prokaryotic programmed frameshift clusters) can be seen as a list on a single web page with general information about each cluster.

The type of a cluster was predicted using a range of cluster's features. To identify programmed frameshift clusters, sequences in the vicinities of the frameshifts were analyzed in order to find a conserved motif that would resemble a frameshift site. Pseudogene clusters must have BLASTp hits in nr database indicating that predicted frameshift is a result of an indel mutation. Elevated frequency of tandem

repeats near predicted frameshifts was chosen as a characteristic property of a phase variation clusters. On the other hand, conserved start codons for downstream ORF2 are expected in the vicinity of the frameshifts in translational coupling clusters.

There are a number of large clusters for which the nature was not predicted but they may be of interest to research community. To provide access to these clusters additional groups were introduced: clusters with 100+ and 50-100 fs-genes (in case of prokaryotes) and 10+ fs-genes (in case of eukaryotes), so clusters could be retrieved by size.

Additionally, during the search for prokaryotic programmed frameshift clusters we have analyzed the frameshift vicinities and grouped clusters by the most overrepresented heptamer. The heptamers include special symbols (underscores) to indicate the reading frame of the upstream ORF1.

The cluster details page contains the same information as the fs-gene details page except that the information is provided for all the cluster's fs-genes together, e.g. a multi-fasta file where all the fs-gene or the fs-protein sequences are provided instead of a single sequence. The cluster information page may also include figures visualizing frequencies of nucleotides in conserved motifs (Sequence LOGOs) located close to the frameshift position as well as the distributions of frameshift coordinates and the fs-gene lengths (see Figure 35). Sequence LOGOs were generated with the MEME software package [144].

5.3 Tools for frameshift prediction

Besides the database the GeneTack server contains a number of tools for frameshift identification in nucleotide sequences. There are four main programs – GeneTack-GM [127], GeneTack-Prok [127], GeneTack-Euk (Antonov et al, paper in preparation) and MetaGeneTack (Tang et al, paper submitted).

GeneTack-GM is a combination of frameshift prediction program GeneTack and a

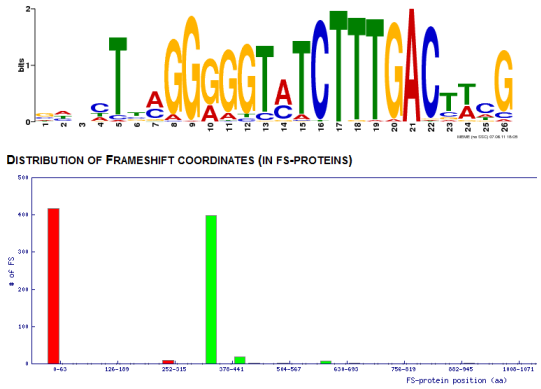


Figure 35: (A) Logo of the conserved motif and (B) distribution of coordinates of frameshifts in 428 fs-genes of Release Factor 2 collected in a cluster (ID 474411093) [1]. Red bars in (B) correspond to frameshift positions and green bars show the total length of fs-proteins. The small green bars indicate existence of subgroups of longer fs-proteins.

self-training gene prediction program GeneMarkS [119]. GeneTack-GM could be used to predict frameshfits in long prokaryotic sequences (longer 300 kB). The model parameters are automatically generated by a self-training program GeneMarkS. GeneTack-GM also includes a number of filters to remove false positive predictions.

GeneTack-Prok and GeneTack-Euk can be used to analyze shorter prokaryotic and eukaryotic sequences with length insufficient for self-training. Eukaryotic sequences must be intronless, e.g. mRNAs or ESTs can be used. Both programs feature a number of pre-built species specific models. A user should choose the one that corresponds to the input sequence. No filters are applied to the frameshifts predicted by these two programs.

GeneTack cannot be directly applied to short metagenomic sequences because it requires a species-specific statistical model. Yet another *ab initio* frameshift finder, MetaGeneTack, can be used in this case (Tang et al, paper submitted). MetaGeneTack uses heuristic models [179] and applies several additional filters for removing false positive predictions.

5.4 Application of the tools and database

The GeneTack tools predict frameshifts in all types of sequences. Using one of the tools a user can find candidate genes with frameshfits in a new prokaryotic genome, contig or metagenome or explore a single protein coding mRNA for a presence of frameshifts. The predicted fs-genes are automatically translated into fs-proteins that could be used as queries against GeneTack database. Hits to large clusters will show phylogenetic conservation of the frameshift. An association with a large cluster can be used to argue that the predicted frameshift is not a result of sequencing error. Moreover, if the type of the cluster is known (e.g. programmed frameshift) it is likely that the input sequence has a frameshift of the same type as well.

5.5 Availability

The interface to GeneTack database is at http://topaz.gatech.edu/GeneTack/db.html. All data are available for download as flat files (sequences in fasta format), and also as a set of MySQL relational database files. Each fs-gene as well as each fs-gene cluster has a unique identification number (ID). The genes or clusters are accessible via URLs: http://topaz.gatech.edu/GeneTack/cgi/fs_view.cgi?id=FS_ID (for fs-genes) or cof_view.cgi?id=CLUSTER_ID (for clusters).

REFERENCES

[1] W. J. Craigen and C. T. Caskey, "Expression of peptide chain release factor 2 requires high-efficiency frameshift," *Nature*, vol. 322, no. 6076, pp. 273–5, 1986.

[2] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, "Ncbi reference sequences: current status, policy and new initiatives," *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D32–6, 2009.

[3] J. F. Atkins and R. F. Gesteland, *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. Springer, 1st ed., 2010.

[4] J. Dinman, "Mechanisms and implications of programmed translational frameshifting," *Wiley Interdisciplinary Reviews: RNA*, vol. 3, no. 5, pp. 661–73, 2012.

[5] A. Firth and I. Brierley, "Non-canonical translation in rna viruses," *Journal of General Virology*, vol. 93, no. Pt 7, pp. 1385–1409, 2012.

[6] O. Namy, J. P. Rousset, S. Napthine, and I. Brierley, "Reprogrammed genetic decoding in cellular gene expression," *Mol Cell*, vol. 13, no. 2, pp. 157–68, 2004.

[7] P. V. Baranov, R. F. Gesteland, and J. F. Atkins, "Recoding: translational bifurcations in gene expression," *Gene*, vol. 286, no. 2, pp. 187–201, 2002.

[8] W. H. Lin and E. Kussell, "Evolutionary pressures on simple sequence repeats in prokaryotic coding regions," *Nucleic Acids Res*, vol. 40, no. 6, pp. 2399–413, 2012.

[9] Y. Kashi and D. G. King, "Simple sequence repeats as advantageous mutators in evolution," *Trends Genet*, vol. 22, no. 5, pp. 253–9, 2006.

[10] M. Ronaghi, "Pyrosequencing sheds light on dna sequencing," *Genome Res*, vol. 11, no. 1, pp. 3–11, 2001.

[11] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-dna interactions," *Science*, vol. 316, no. 5830, pp. 1497–502, 2007.

[12] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church, "Accurate multiplex polony sequencing of an evolved bacterial genome," *Science*, vol. 309, no. 5741, pp. 1728–32, 2005.

[13] S. J. Laken, G. M. Petersen, S. B. Gruber, C. Oddoux, H. Ostrer, F. M. Giardiello, S. R. Hamilton, H. Hampel, A. Markowitz, D. Klimstra, S. Jhanwar, S. Winawer, K. Offit, M. C. Luce, K. W. Kinzler, and B. Vogelstein, "Familial colorectal cancer in ashkenazim due to a hypermutable tract in apc," *Nat Genet*, vol. 17, no. 1, pp. 79–83, 1997.

[14] C. Deshayes, E. Perrodou, S. Gallien, D. Euphrasie, C. Schaeffer, A. Van-Dorsselaer, O. Poch, O. Lecompte, and J. M. Reyrat, "Interrupted coding sequences in mycobacterium smegmatis: authentic mutations or sequencing errors?," *Genome Biol*, vol. 8, no. 2, p. R20, 2007.

[15] C. Medigue, M. Rose, A. Viari, and A. Danchin, "Detecting and analyzing dna sequencing errors: toward a higher quality of the bacillus subtilis genome sequence," *Genome Res.*, vol. 9, no. 11, pp. 1116–27, 1999.

[16] P. V. Baranov, A. W. Hammer, J. Zhou, R. F. Gesteland, and J. F. Atkins, "Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in is element gene expression," *Genome Biol*, vol. 6, no. 3, p. R25, 2005.

[17] J. J. Wernegreen, S. N. Kauppinen, and P. H. Degnan, "Slip into something more functional: selection maintains ancient frameshifts in homopolymeric sequences," *Mol Biol Evol*, vol. 27, no. 4, pp. 833–9, 2010.

[18] W. A. Decatur and M. J. Fournier, "Rna-guided nucleotide modification of ribosomal and other rnas," *J Biol Chem*, vol. 278, no. 2, pp. 695–8, 2003.

[19] E. M. Byrne, G. J. Connell, and L. Simpson, "Guide rna-directed uridine insertion rna editing in vitro," *EMBO J*, vol. 15, no. 23, pp. 6758–65, 1996.

[20] B. E. Wulff, M. Sakurai, and K. Nishikura, "Elucidating the inosinome: global approaches to adenosine-to-inosine rna editing," *Nat Rev Genet*, vol. 12, no. 2, pp. 81–5, 2011.

[21] A. Kiran, G. Loughran, J. J. O'Mahony, and P. V. Baranov, "Identification of a-to-i rna editing: dotting the i's in the human transcriptome," *Biochemistry (Mosc)*, vol. 76, no. 8, pp. 915–23, 2011.

[22] S. Maas, "Posttranscriptional recoding by rna editing," *Adv Protein Chem Struct Biol*, vol. 86, pp. 193–224, 2012.

[23] V. Sharma, A. Firth, I. Antonov, O. Fayet, J. Atkins, M. Borodovsky, and P. Baranov, "A pilot study of bacterial genes with disrupted orfs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment," *Molecular Biology and Evolution*, vol. 28, no. 11, pp. 3195–3211, 2011.

[24] R. B. Weiss, D. M. Dunn, J. F. Atkins, and R. F. Gesteland, "Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting," *Cold Spring Harb Symp Quant Biol*, vol. 52, pp. 687–93, 1987.

[25] R. B. Weiss, D. M. Dunn, J. F. Atkins, and R. F. Gesteland, "Ribosomal frameshifting from -2 to +50 nucleotides," *Prog Nucleic Acid Res Mol Biol*, vol. 39, pp. 159–83, 1990.

[26] J. F. Atkins, R. B. Weiss, and R. F. Gesteland, "Ribosome gymnastics–degree of difficulty 9.5, style 10.0," *Cell*, vol. 62, no. 3, pp. 413–23, 1990.

[27] J. F. Atkins and G. R. Bjork, "A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight p-site realignment," *Microbiol Mol Biol Rev*, vol. 73, no. 1, pp. 178–210, 2009.

[28] P. V. Baranov, B. Vestergaard, T. Hamelryck, R. F. Gesteland, J. Nyborg, and J. F. Atkins, "Diverse bacterial genomes encode an operon of two genes, one of which is an unusual class-i release factor that potentially recognizes atypical mrna signals other than normal stop codons," *Biol Direct*, vol. 1, p. 28, 2006.

[29] I. Brierley, R. J. Gilbert, and S. Pennell, "Rna pseudoknots and the regulation of protein synthesis," *Biochem Soc Trans*, vol. 36, no. Pt 4, pp. 684–9, 2008.

[30] D. P. Giedroc and P. V. Cornish, "Frameshifting rna pseudoknots: structure and mechanism," *Virus Res*, vol. 139, no. 2, pp. 193–208, 2009.

[31] A. Devaraj and K. Fredrick, "Short spacing between the shine-dalgarno sequence and p codon destabilizes codon-anticodon pairing in the p site to promote +1 programmed frameshifting," *Mol Microbiol*, vol. 78, no. 6, pp. 1500–9, 2010.

[32] B. Larsen, N. M. Wills, R. F. Gesteland, and J. F. Atkins, "rrna-mrna base pairing stimulates a programmed -1 ribosomal frameshift," *J Bacteriol*, vol. 176, no. 22, pp. 6842–51, 1994.

[33] O. L. Gurvich, S. J. Nasvall, P. V. Baranov, G. R. Bjork, and J. F. Atkins, "Two groups of phenylalanine biosynthetic operon leader peptides genes: a high level of apparently incidental frameshifting in decoding escherichia coli phel," *Nucleic Acids Res*, vol. 39, no. 8, pp. 3079–92, 2011.

[34] B. Larsen, J. Peden, S. Matsufuji, T. Matsufuji, K. Brady, R. Maldonado, N. M. Wills, O. Fayet, J. F. Atkins, and R. F. Gesteland, "Upstream stimulators for recoding," *Biochem Cell Biol*, vol. 73, no. 11-12, pp. 1123–9, 1995.

[35] F. Iseni, F. Baudin, D. Garcin, J. B. Marq, R. W. Ruigrok, and D. Kolakofsky, "Chemical modification of nucleotide bases and mrna editing depend on hexamer or nucleoprotein phase in sendai virus nucleocapsids," *RNA*, vol. 8, no. 8, pp. 1056–67, 2002.

[36] I. Ferrer, G. Santpere, and F. W. van Leeuwen, "Argyrophilic grain disease," *Brain*, vol. 131, no. Pt 6, pp. 1416–32, 2008.

[37] J. Turnbough, C. L., "Regulation of gene expression by reiterative transcription," *Curr Opin Microbiol*, vol. 14, no. 2, pp. 142–7, 2011.

[38] M. Chamberlin and P. Berg, "Deoxyribo ucleic acid-directed synthesis of ribonucleic acid by an enzyme from escherichia coli," *Proc Natl Acad Sci U S A*, vol. 48, pp. 81–94, 1962.

[39] L. A. Wagner, R. B. Weiss, R. Driscoll, D. S. Dunn, and R. F. Gesteland, "Transcriptional slippage occurs during elongation at runs of adenine or thymine in escherichia coli," *Nucleic Acids Res*, vol. 18, no. 12, pp. 3529–35, 1990.

[40] M. Bekaert, J. F. Atkins, and P. V. Baranov, "Arfa: a program for annotating bacterial release factor genes, including prediction of programmed ribosomal frameshifting," *Bioinformatics*, vol. 22, no. 20, pp. 2463–5, 2006.

[41] M. Bekaert, I. P. Ivanov, J. F. Atkins, and P. V. Baranov, "Ornithine decarboxylase antizyme finder (oaf): fast and reliable detection of antizymes with frameshifts in mrnas," *BMC Bioinformatics*, vol. 9, p. 178, 2008.

[42] C. Theis, J. Reeder, and R. Giegerich, "Knotinframe: prediction of -1 ribosomal frameshift events," *Nucleic Acids Res*, vol. 36, no. 18, pp. 6013–6020, 2008.

[43] P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler, "Isfinder: the reference centre for bacterial insertion sequences," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D32–6, 2006.

[44] Y. Sekine and E. Ohtsubo, "Frameshifting is required for production of the transposase encoded by insertion sequence 1," *Proc Natl Acad Sci U S A*, vol. 86, no. 12, pp. 4609–13, 1989.

[45] M. Chandler and O. Fayet, "Translational frameshifting in the control of transposition in bacteria," *Mol Microbiol*, vol. 7, no. 4, pp. 497–503, 1993.

[46] P. Polard, M. F. Prere, M. Chandler, and O. Fayet, "Programmed translational frameshifting and initiation at an auu codon in gene expression of bacterial insertion sequence is911," *J Mol Biol*, vol. 222, no. 3, pp. 465–77, 1991.

[47] Y. Sekine, N. Eisaki, and E. Ohtsubo, "Translational control in production of transposase and in transposition of insertion sequence is3," *J Mol Biol*, vol. 235, no. 5, pp. 1406–20, 1994.

[48] K. Vogele, E. Schwartz, C. Welz, E. Schiltz, and B. Rak, "High-level ribosomal frameshifting directs the synthesis of is150 gene products," *Nucleic Acids Res*, vol. 19, no. 16, pp. 4377–85, 1991.

[49] C. Loot, C. Turlan, P. Rousseau, B. Ton-Hoang, and M. Chandler, "A target specificity switch in is911 transposition: the role of the orfa protein," *EMBO J*, vol. 21, no. 15, pp. 4172–82, 2002.

[50] B. Ton-Hoang, P. Polard, L. Haren, C. Turlan, and M. Chandler, "Is911 transposon circles give rise to linear forms that can undergo integration in vitro," *Mol Microbiol*, vol. 32, no. 3, pp. 617–27, 1999.

[51] M. Shehu-Xhilaga, S. M. Crowe, and J. Mak, "Maintenance of the gag/gag-pol ratio is important for human immunodeficiency virus type 1 rna dimerization and viral infectivity," *J Virol*, vol. 75, no. 4, pp. 1834–41, 2001.

[52] M. F. Belcourt and P. J. Farabaugh, "Ribosomal frameshifting in the yeast retrotransposon ty: trnas induce slippage on a 7 nucleotide minimal site," *Cell*, vol. 62, no. 2, pp. 339–52, 1990.

[53] P. J. Farabaugh, H. Zhao, and A. Vimaladithan, "A novel programed frameshift expresses the pol3 gene of retrotransposon ty3 of yeast: frameshifting without trna slippage," *Cell*, vol. 74, no. 1, pp. 93–103, 1993.

[54] J. Brandt, A. M. Veith, and J. N. Volff, "A family of neofunctionalized ty3/gypsy retrotransposon genes in mammalian genomes," *Cytogenet Genome Res*, vol. 110, no. 1-4, pp. 307–17, 2005.

[55] X. Gao, E. R. Havecker, P. V. Baranov, J. F. Atkins, and D. F. Voytas, "Translational recoding signals between gag and pol in diverse ltr retrotransposons," *RNA*, vol. 9, no. 12, pp. 1422–30, 2003.

[56] E. Scolnick, R. Tompkins, T. Caskey, and M. Nirenberg, "Release factors differing in specificity for terminator codons," *Proc Natl Acad Sci U S A*, vol. 61, no. 2, pp. 768–74, 1968.

[57] M. R. Capecchi and H. A. Klein, "Release factors mediating termination of complete proteins," *Nature*, vol. 226, no. 5250, pp. 1029–33, 1970.

[58] A. L. Blinkowa and J. R. Walker, "Programmed ribosomal frameshifting generates the escherichia coli dna polymerase iii gamma subunit from within the tau subunit reading frame," *Nucleic Acids Res*, vol. 18, no. 7, pp. 1725–9, 1990.

[59] A. Blinkova, M. F. Burkart, T. D. Owens, and J. R. Walker, "Conservation of the escherichia coli dnax programmed ribosomal frameshift signal in salmonella typhimurium," *J Bacteriol*, vol. 179, no. 13, pp. 4438–42, 1997.

[60] Z. Tsuchihashi and A. Kornberg, "Translational frameshifting generates the gamma subunit of dna polymerase iii holoenzyme," *Proc Natl Acad Sci USA*, vol. 87, no. 7, pp. 2516–20, 1990.

[61] A. M. Flower and C. S. McHenry, "The gamma subunit of dna polymerase iii holoenzyme of escherichia coli is produced by ribosomal frameshifting," *Proc Natl Acad Sci U S A*, vol. 87, no. 10, pp. 3713–7, 1990.

[62] B. Larsen, N. M. Wills, C. Nelson, J. F. Atkins, and R. F. Gesteland, "Nonlinearity in genetic decoding: homologous dna replicase genes use alternatives of transcriptional slippage or translational frameshifting," *Proc Natl Acad Sci USA*, vol. 97, no. 4, pp. 1683–8, 2000.

[63] B. Larsen, R. F. Gesteland, and J. F. Atkins, "Structural probing and mutagenic analysis of the stem-loop required for escherichia coli dnax ribosomal frameshifting: programmed efficiency of 50%," *J Mol Biol*, vol. 271, no. 1, pp. 47–60, 1997.

[64] S. Matsufuji, T. Matsufuji, Y. Miyazaki, Y. Murakami, J. F. Atkins, R. F. Gesteland, and S. Hayashi, "Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme," *Cell*, vol. 80, no. 1, pp. 51–60, 1995.

[65] I. P. Ivanov and J. F. Atkins, "Ribosomal frameshifting in decoding antizyme mrnas from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation," *Nucleic Acids Res*, vol. 35, no. 6, pp. 1842–58, 2007.

[66] I. P. Ivanov, K. Simin, A. Letsou, J. F. Atkins, and R. F. Gesteland, "The drosophila gene for antizyme requires ribosomal frameshifting for expression and contains an intronic gene for snrnp sm d3 on the opposite strand," *Mol Cell Biol*, vol. 18, no. 3, pp. 1553–61, 1998.

[67] I. P. Ivanov, S. Matsufuji, Y. Murakami, R. F. Gesteland, and J. F. Atkins, "Conservation of polyamine regulation by translational frameshifting from yeast to mammals," *EMBO J*, vol. 19, no. 8, pp. 1907–17, 2000.

[68] I. P. Ivanov, R. F. Gesteland, and J. F. Atkins, "Antizyme expression: a subversion of triplet decoding, which is remarkably conserved by evolution, is a sensor for an autoregulatory circuit," *Nucleic Acids Res*, vol. 28, no. 17, pp. 3185–96, 2000.

[69] I. P. Ivanov, R. F. Gesteland, and J. F. Atkins, "Evolutionary specialization of recoding: frameshifting in the expression of s. cerevisiae antizyme mrna is via an atypical antizyme shift site but is still +1," *RNA*, vol. 12, no. 3, pp. 332–7, 2006.

[70] M. K. Chattopadhyay, Y. Murakami, and S. Matsufuji, "Antizyme regulates the degradation of ornithine decarboxylase in fission yeast schizosaccharomyces pombe. study in the spe2 knockout strains," *J Biol Chem*, vol. 276, no. 24, pp. 21235–41, 2001.

[71] N. Mejlhede, J. F. Atkins, and J. Neuhard, "Ribosomal -1 frameshifting during decoding of bacillus subtilis cdd occurs at the sequence cga aag," *J Bacteriol*, vol. 181, no. 9, pp. 2930–7, 1999.

[72] C. Fu and J. Parker, "A ribosomal frameshifting error during translation of the argi mrna of escherichia coli," *Mol Gen Genet*, vol. 243, no. 4, pp. 434–41, 1994.

[73] R. Schwartz and J. F. Curran, "Analyses of frameshifting at uuu-pyrimidine sites," *Nucleic Acids Res*, vol. 25, no. 10, pp. 2005–11, 1997.

[74] C. Penno, A. Hachani, L. Biskri, P. Sansonetti, A. Allaoui, and C. Parsot, "Transcriptional slippage controls production of type iii secretion apparatus components in shigella flexneri," *Mol Microbiol*, vol. 62, no. 5, pp. 1460–8, 2006.

[75] C. Penno, P. Sansonetti, and C. Parsot, "Frameshifting by transcriptional slippage is involved in production of mxie, the transcription activator regulated by the activity of the type iii secretion apparatus in shigella flexneri," *Mol Microbiol*, vol. 56, no. 1, pp. 204–14, 2005.

[76] B. Cobucci-Ponzano, A. Trincone, A. Giordano, M. Rossi, and M. Moracci, "Identification of an archaeal alpha-l-fucosidase encoded by an interrupted gene. production of a functional enzyme by mutations mimicking programmed -1 frameshifting," *J Biol Chem*, vol. 278, no. 17, pp. 14622–31, 2003.

[77] D. K. Morris and V. Lundblad, "Programmed translational frameshifting in a gene required for yeast telomere replication," *Curr Biol*, vol. 7, no. 12, pp. 969–76, 1997.

[78] T. Asakura, T. Sasaki, F. Nagano, A. Satoh, H. Obaishi, H. Nishioka, H. Imamura, K. Hotta, K. Tanaka, H. Nakanishi, and Y. Takai, "Isolation and characterization of a novel actin filament-binding protein from saccharomyces cerevisiae," *Oncogene*, vol. 16, no. 1, pp. 121–30, 1998.

[79] A. A. Shah, M. C. Giddings, J. B. Parvaz, R. F. Gesteland, J. F. Atkins, and I. P. Ivanov, "Computational identification of putative programmed translational frameshift sites," *Bioinformatics*, vol. 18, no. 8, pp. 1046–53, 2002.

[80] X. Saulquin, E. Scotet, L. Trautmann, M. A. Peyrat, F. Halary, M. Bonneville, and E. Houssaint, "+1 frameshifting as a novel mechanism to generate a cryptic cytotoxic t lymphocyte epitope derived from human interleukin 10," *J Exp Med*, vol. 195, no. 3, pp. 353–8, 2002.

[81] J. D. Dinman, T. Icho, and R. B. Wickner, "A -1 ribosomal frameshift in a double-stranded rna virus of yeast forms a gag-pol fusion protein," *Proc Natl Acad Sci U S A*, vol. 88, no. 1, pp. 174–8, 1991.

[82] H. Lux, H. Flammann, M. Hafner, and A. Lux, "Genetic and molecular analyses of peg10 reveal new aspects of genomic organization, transcription and translation," *PLoS One*, vol. 5, no. 1, p. e8686, 2010.

[83] E. Manktelow, K. Shigemoto, and I. Brierley, "Characterization of the frameshift signal of edr, a mammalian example of programmed -1 ribosomal frameshifting," *Nucleic Acids Res*, vol. 33, no. 5, pp. 1553–63, 2005.

[84] N. M. Wills, B. Moore, A. Hammer, R. F. Gesteland, and J. F. Atkins, "A functional -1 ribosomal frameshift signal in the human paraneoplastic ma3 gene," *J Biol Chem*, vol. 281, no. 11, pp. 7082–8, 2006.

[85] L. L. Major, E. S. Poole, M. E. Dalphin, S. A. Mannering, and W. P. Tate, "Is the in-frame termination signal of the escherichia coli release factor-2 frameshift site weakened by a particularly poor context?," *Nucleic Acids Res*, vol. 24, no. 14, pp. 2673–8, 1996.

[86] S. Mottagui-Tabar and L. A. Isaksson, "The influence of the 5' codon context on translation termination in bacillus subtilis and escherichia coli is similar but different from salmonella typhimurium," *Gene*, vol. 212, no. 2, pp. 189–96, 1998.

[87] J. Shine and L. Dalgarno, "Growth-dependent changes in terminal heterogeneity involving 3'-adenylate of bacterial 16s ribosomal rna," *Nature*, vol. 256, no. 5514, pp. 232–3, 1975.

[88] J. Ma, A. Campbell, and S. Karlin, "Correlations between shine-dalgarno sequences and gene features such as predicted expression levels and operon structures," *J Bacteriol*, vol. 184, no. 20, pp. 5733–45, 2002.

[89] M. J. Johansson, A. Esberg, B. Huang, G. R. Bjork, and A. S. Bystrom, "Eukaryotic wobble uridine modifications promote a functionally redundant decoding system," *Mol Cell Biol*, vol. 28, no. 10, pp. 3301–12, 2008.

[90] A. Vimaladithan, S. Pande, H. Zhao, and P. J. Farabaugh, "Peptidyl-trnas promote translational frameshifting," *Nucleic Acids Symp Ser*, no. 33, pp. 190–3, 1995.

[91] B. Bonetti, L. Fu, J. Moon, and D. M. Bedwell, "The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in saccharomyces cerevisiae," *J Mol Biol*, vol. 251, no. 3, pp. 334–45, 1995.

[92] Z. Nagy and M. Chandler, "Regulation of transposition in bacteria," *Res Microbiol*, vol. 155, no. 5, pp. 387–98, 2004.

[93] G. Duval-Valentin and M. Chandler, "Cotranslational control of dna transposition: a window of opportunity," *Mol Cell*, vol. 44, no. 6, pp. 989–96, 2011.

[94] M. F. Prere, I. Canal, N. M. Wills, J. F. Atkins, and O. Fayet, "The interplay of mrna stimulatory signals required for auu-mediated initiation and programmed -1 ribosomal frameshifting in decoding of transposable element is911," *J Bacteriol*, vol. 193, no. 11, pp. 2735–44, 2011.

[95] M. W. van der Woude, "Re-examining the role and random nature of phase variation," *FEMS Microbiol Lett*, vol. 254, no. 2, pp. 190–7, 2006.

[96] M. W. van der Woude and A. J. Baumler, "Phase and antigenic variation in bacteria," *Clin Microbiol Rev*, vol. 17, no. 3, pp. 581–611, 2004.

[97] P. Theiss and K. S. Wise, "Localized frameshift mutation generates selective, high-frequency phase variation of a surface lipoprotein encoded by a mycoplasma abc transporter operon," *J Bacteriol*, vol. 179, no. 12, pp. 4013–22, 1997.

[98] Y. N. Srikhanta, K. L. Fox, and M. P. Jennings, "The phasevarion: phase variation of type iii dna methyltransferases controls coordinated switching in multiple genes," *Nat Rev Microbiol*, vol. 8, no. 3, pp. 196–206, 2010.

[99] B. Blencowe, "Alternative splicing: new insights from global analyses," *Cell*, vol. 126, no. 1, pp. 37–47, 2006.

[100] C. Zhang, A. R. Krainer, and M. Q. Zhang, "Evolutionary impact of limited splicing fidelity in mammalian genes," *Trends Genet*, vol. 23, no. 10, pp. 484–8, 2007.

[101] B. P. Lewis, R. E. Green, and S. E. Brenner, "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans," *Proc Natl Acad Sci U S A*, vol. 100, no. 1, pp. 189–92, 2003.

[102] L. E. Maquat, "Nonsense-mediated mrna decay: splicing, translation and mrnp dynamics," *Nat Rev Mol Cell Biol*, vol. 5, no. 2, pp. 89–99, 2004.

[103] H. Le Hir, E. Izaurralde, L. E. Maquat, and M. J. Moore, "The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mrna exon-exon junctions," *EMBO J*, vol. 19, no. 24, pp. 6860–9, 2000.

[104] Y. Ishigaki, X. Li, G. Serin, and L. E. Maquat, "Evidence for a pioneer round of mrna translation: mrnas subject to nonsense-mediated decay in mammalian cells are bound by cbp80 and cbp20," *Cell*, vol. 106, no. 5, pp. 607–17, 2001.

[105] J. Posfai and R. J. Roberts, "Finding errors in dna sequences," *Proc Natl Acad Sci U S A*, vol. 89, no. 10, pp. 4698–702, 1992.

[106] J. M. Claverie, "Detecting frame shifts by amino acid sequence comparison," *J Mol Biol*, vol. 234, no. 4, pp. 1140–57, 1993.

[107] W. R. Pearson, T. Wood, Z. Zhang, and W. Miller, "Comparison of dna sequences with protein sequences," *Genomics*, vol. 46, no. 1, pp. 24–36, 1997.

[108] E. Birney, J. D. Thompson, and T. J. Gibson, "Pairwise and searchwise: finding the optimal alignment in a simultaneous comparison of a protein profile against all dna translation frames," *Nucleic Acids Res*, vol. 24, no. 14, pp. 2730–9, 1996.

[109] X. Guan and E. C. Uberbacher, "Alignments of dna and protein sequences containing frameshift errors," *Comput Appl Biosci*, vol. 12, no. 1, pp. 31–40, 1996.

[110] G. A. Fichant and Y. Quentin, "A frameshift error detection algorithm for dna sequencing projects," *Nucleic Acids Res*, vol. 23, no. 15, pp. 2900–8, 1995.

[111] T. Schiex, J. Gouzy, A. Moisan, and Y. Oliveira, "Framed: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences.," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3738–3741, 2003.

[112] A. Kislyuk, A. Lomsadze, A. L. Lapidus, and M. Borodovsky, "Frameshift detection in prokaryotic genomic sequences," *Int J Bioinform Res Appl*, vol. 5, no. 4, pp. 458–77, 2009.

[113] S. Moon, Y. Byun, H.-J. Kim, S. Jeong, and K. Han, "Predicting genes expressed via -1 and +1 frameshifts.," *Nucleic Acids Res.*, vol. 32, no. 16, pp. 4884–4892, 2004.

[114] Y. Byun, S. Moon, and K. Han, "A general computational model for predicting ribosomal frameshifts in genome sequences," *Comput Biol Med*, vol. 37, no. 12, pp. 1796–801, 2007.

[115] A. B. Hammell, R. C. Taylor, S. W. Peltz, and J. D. Dinman, "Identification of putative programmed -1 ribosomal frameshift signals in large dna databases," *Genome Res*, vol. 9, no. 5, pp. 417–27, 1999.

[116] J. L. Jacobs, A. T. Belew, R. Rakauskaite, and J. D. Dinman, "Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of saccharomyces cerevisiae," *Nucleic Acids Res*, vol. 35, no. 1, pp. 165–74, 2007.

[117] P. Y. Liao, Y. S. Choi, and K. H. Lee, "Fsscan: a mechanism-based program to identify +1 ribosomal frameshift hotspots," *Nucleic Acids Res*, vol. 37, no. 21, pp. 7302–11, 2009.

[118] O. L. Gurvich, P. V. Baranov, J. Zhou, A. W. Hammer, R. F. Gesteland, and J. F. Atkins, "Sequences that direct significant levels of frameshifting are frequent in coding regions of escherichia coli," *EMBO J*, vol. 22, no. 21, pp. 5941–50, 2003.

[119] J. Besemer, A. Lomsadze, and M. Borodovsky, "Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions," *Nucleic Acids Research*, vol. 29, no. 12, pp. 2607–18, 2001.

[120] J. Besemer and M. Borodovsky, "Heuristic approach to deriving models for gene finding," *Nucleic Acids Res*, vol. 27, no. 19, pp. 3911–20, 1999.

[121] A. V. Lukashin and M. Borodovsky, "Genemark.hmm: new solutions for gene finding," *Nucleic Acids Res*, vol. 26, no. 4, pp. 1107–15, 1998.

[122] A. M. Shmatkov, A. A. Melikyan, F. L. Chernousko, and M. Borodovsky, "Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes," *Bioinformatics*, vol. 15, no. 11, pp. 874–86, 1999.

[123] T. S. Larsen and A. Krogh, "Easygene–a prokaryotic gene finder that ranks orfs by statistical significance," *BMC Bioinformatics*, vol. 4, p. 21, 2003.

[124] M. Borodovsky and J. McIninch, "Genemark: parallel gene recognition for both dna strands.," *Computers & Chemistry*, vol. 17, no. 19, pp. 123–133, 1993.

[125] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[126] M. Bekaert, A. E. Firth, Y. Zhang, V. N. Gladyshev, J. F. Atkins, and P. V. Baranov, "Recode-2: new design, new search tools, and many more genes," *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D69–74, 2010.

[127] I. Antonov and M. Borodovsky, "Genetack: frameshift identification in protein-coding sequences by the viterbi algorithm," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 3, p. 535, 2010.

[128] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D561–8, 2011.

[129] R. A. Spanjaard and J. van Duin, "Translation of the sequence agg-agg yields 50% ribosomal frameshift," *Proc Natl Acad Sci U S A*, vol. 85, no. 21, pp. 7967–71, 1988.

[130] O. L. Gurvich, P. V. Baranov, R. F. Gesteland, and J. F. Atkins, "Expression levels influence ribosomal frameshifting at the tandem rare arginine codons agg_agg and aga_aga in escherichia coli," *J Bacteriol*, vol. 187, no. 12, pp. 4023–32, 2005.

[131] N. Mejlhede, P. Licznar, M. F. Prere, N. M. Wills, R. F. Gesteland, J. F. Atkins, and O. Fayet, "-1 frameshifting at a cga aag hexanucleotide site is required for transposition of insertion sequence is1222," *J Bacteriol*, vol. 186, no. 10, pp. 3274–7, 2004.

[132] M. H. Mazauric, P. Licznar, M. F. Prere, I. Canal, and O. Fayet, "Apical loop-internal loop rna pseudoknots: a new type of stimulator of -1 translational frameshifting in bacteria," *J Biol Chem*, vol. 283, no. 29, pp. 20421–32, 2008.

[133] G. Wang, D. Rasko, R. Sherburne, and D. Taylor, "Molecular genetic basis for the variable expression of lewis y antigen in helicobacter pylori: analysis of the α (1, 2) fucosyltransferase gene," *Molecular microbiology*, vol. 31, no. 4, pp. 1265–1274, 2002.

[134] J. M. Kirchner, H. Tran, and M. A. Resnick, "A dna polymerase epsilon mutant that specifically causes +1 frameshift mutations within homonucleotide runs in yeast," *Genetics*, vol. 155, no. 4, pp. 1623–32, 2000.

[135] P. Baranov, O. Fayet, R. Hendrix, and J. Atkins, "Recoding in bacteriophages and bacterial is elements," *Trends Genet*, vol. 22, no. 3, pp. 174–181, 2006.

[136] J. Xu, R. W. Hendrix, and R. L. Duda, "Conserved translational frameshift in dsdna bacteriophage tail assembly genes," *Mol Cell*, vol. 16, no. 1, pp. 11–21, 2004.

[137] G. Benson, "Tandem repeats finder: a program to analyze dna sequences," *Nucleic Acids Res*, vol. 27, no. 2, pp. 573–80, 1999.

[138] P. Pradhan, W. Li, and P. Kaur, "Translational coupling controls expression and function of the drrab drug efflux pump," *Journal of Molecular Biology*, vol. 385, no. 3, pp. 831–842, 2009.

[139] C. H. Kuo and H. Ochman, "The extinction dynamics of bacterial pseudogenes," *PLoS Genet*, vol. 6, no. 8, 2010.

[140] E. Lerat and H. Ochman, "Psi-phi: exploring the outer limits of bacterial pseudogenes," *Genome Res*, vol. 14, no. 11, pp. 2273–8, 2004.

[141] K. Suhre and J. M. Claverie, "Fusiondb: a database for in-depth analysis of prokaryotic gene fusion events," *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D273–6, 2004.

[142] P. BARANOV, R. GESTELAND, and J. ATKINS, "P-site trna is a crucial initiator of ribosomal frameshifting," *Rna*, vol. 10, no. 2, pp. 221–230, 2004.

[143] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment," *Science*, vol. 262, no. 5131, pp. 208–14, 1993.

[144] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "Meme suite: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. Web Server issue, pp. W202–8, 2009.

[145] D. Taliaferro and P. J. Farabaugh, "An mrna sequence derived from the yeast est3 gene stimulates programmed +1 translational frameshifting," *RNA*, vol. 13, no. 4, pp. 606–13, 2007.

[146] M. O'Connor, "Imbalance of trna(pro) isoacceptors induces +1 frameshifting at near-cognate codons," *Nucleic Acids Res*, vol. 30, no. 3, pp. 759–65, 2002.

[147] S. Karlin, J. Mrazek, A. Campbell, and D. Kaiser, "Characterizations of highly expressed genes of four fast-growing bacteria," *J Bacteriol*, vol. 183, no. 17, pp. 5025–40, 2001.

[148] T. Schneider and R. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[149] A. Herr, C. Nelson, N. Wills, R. Gesteland, and J. Atkins, "Analysis of the roles of trna structure, ribosomal protein l9, and the bacteriophage t4 gene 60 bypassing signals during ribosome slippage on mrna1," *Journal of molecular biology*, vol. 309, no. 5, pp. 1029–1048, 2001.

[150] Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin, "Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context," *Genome Res*, vol. 11, no. 3, pp. 356–72, 2001.

[151] R. F. Gesteland, R. B. Weiss, and J. F. Atkins, "Recoding: reprogrammed genetic decoding," *Science*, vol. 257, no. 5077, pp. 1640–1, 1992.

[152] J. J. Clare, M. Belcourt, and P. J. Farabaugh, "Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast ty1 transposon," *Proc Natl Acad Sci U S A*, vol. 85, no. 18, pp. 6816–20, 1988.

[153] P. J. Farabaugh, E. Kramer, H. Vallabhaneni, and A. Raman, "Evolution of +1 programmed frameshifting signals and frameshift-regulating trnas in the order saccharomycetales," *J Mol Evol*, vol. 63, no. 4, pp. 545–61, 2006.

[154] M. Klemke, R. H. Kehlenbach, and W. B. Huttner, "Two overlapping reading frames in a single exon encode interacting proteins–a novel way of gene usage," *EMBO J*, vol. 20, no. 14, pp. 3849–60, 2001.

[155] M. Guittaut, S. Charpentier, T. Normand, M. Dubois, J. Raimond, and A. Legrand, "Identification of an internal gene to the human galectin-3 gene with two different overlapping reading frames that do not encode galectin-3," *J Biol Chem*, vol. 276, no. 4, pp. 2652–7, 2001.

[156] N. E. Sharpless, "Ink4a/arf: a multifunctional tumor suppressor locus," *Mutat Res*, vol. 576, no. 1-2, pp. 22–38, 2005.

[157] F. Poulin, A. Brueschke, and N. Sonenberg, "Gene fusion and overlapping reading frames in the mammalian genes for 4e-bp3 and mask," *J Biol Chem*, vol. 278, no. 52, pp. 52290–7, 2003.

[158] D. Bartsch, A. Casadio, K. Karl, P. Serodio, and E. Kandel, "Creb1 encodes a nuclear activator, a repressor, and a cytoplasmic modulator that form a regulatory unit critical for long-term facilitation," *Cell*, vol. 95, no. 2, pp. 211–223, 1998.

[159] M. Hameed, R. Orrell, M. Cobbold, G. Goldspink, and S. Harridge, "Expression of igf-i splice variants in young and old human skeletal muscle after high resistance exercise," *The Journal of physiology*, vol. 547, no. 1, pp. 247–254, 2003.

[160] Z. Ahmed, S. Masmoudi, E. Kalay, I. Belyantseva, M. Mosrati, R. Collin, S. Riazuddin, M. Hmani-Aifa, H. Venselaar, M. Kawar, *et al.*, "Mutations of lrtomt, a fusion gene with alternative reading frames, cause nonsyndromic deafness in humans," *Nature genetics*, vol. 40, no. 11, pp. 1335–1340, 2008.

[161] C. Rancurel, M. Khosravi, A. K. Dunker, P. R. Romero, and D. Karlin, "Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation," *J Virol*, vol. 83, no. 20, pp. 10719–36, 2009.

[162] R. E. Mills, W. S. Pittard, J. M. Mullaney, U. Farooq, T. H. Creasy, A. A. Mahurkar, D. M. Kemeza, D. S. Strassler, C. P. Ponting, C. Webber, and S. E. Devine, "Natural genetic variation caused by small insertions and deletions in the human genome," *Genome Res*, vol. 21, no. 6, pp. 830–9, 2011.

[163] E. Kovacs, P. Tompa, K. Liliom, and L. Kalmar, "Dual coding in alternative reading frames correlates with intrinsic protein disorder," *Proc Natl Acad Sci U S A*, vol. 107, no. 12, pp. 5429–34, 2010.

[164] H. Liang and L. F. Landweber, "A genome-wide study of dual coding regions in human alternatively spliced genes," *Genome Res*, vol. 16, no. 2, pp. 190–6, 2006.

[165] S. Ribrioux, A. Brungger, B. Baumgarten, K. Seuwen, and M. R. John, "Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts," *BMC Genomics*, vol. 9, p. 122, 2008.

[166] J. Xu, R. Linning, J. Fellers, M. Dickinson, W. Zhu, I. Antonov, D. Joly, M. Donaldson, T. Eilam, Y. Anikster, T. Banks, S. Munro, M. Mayo, B. Wynhoven, J. Ali, R. Moore, B. McCallum, M. Borodovsky, B. Saville, and

G. Bakkeren, "Gene discovery in est sequences from the wheat leaf rust fungus puccinia triticina sexual spores, asexual spores and haustoria, compared to other rust and corn smut fungi," *BMC Genomics*, vol. 12, no. 161, 2011.

[167] M. Tan, A. Liang, C. Brunen-Nieweler, and K. Heckmann, "Programmed translational frameshifting is likely required for expressions of genes encoding putative nuclear protein kinases of the ciliate euplotes octocarinatus," *J Eukaryot Microbiol*, vol. 48, no. 5, pp. 575–82, 2001.

[168] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.," *Molecular biology and evolution*, vol. 3, no. 5, pp. 418–426, 1986.

[169] K. Pruitt and D. Maglott, "Refseq and locuslink: Ncbi gene-centered resources," *Nucleic acids research*, vol. 29, no. 1, pp. 137–140, 2001.

[170] R. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[171] A. Khachane and P. Harrison, "Assessing the genomic evidence for conserved transcribed pseudogenes under selection," *BMC Genomics*, vol. 10, no. 1, p. 435, 2009.

[172] N. McGlincy and C. Smith, "Alternative splicing resulting in nonsense-mediated mrna decay: what is the meaning of nonsense?," *Trends in biochemical sciences*, vol. 33, no. 8, pp. 385–393, 2008.

[173] Q. Pan, A. Saltzman, Y. Kim, C. Misquitta, O. Shai, L. Maquat, B. Frey, and B. Blencowe, "Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mrna decay to control gene expression," *Genes & development*, vol. 20, no. 2, p. 153, 2006.

[174] W. Chung, S. Wadhawan, R. Szklarczyk, S. Pond, and A. Nekrutenko, "A first look at arfome: dual-coding genes in mammalian genomes," *PLoS Computational Biology*, vol. 3, no. 5, p. e91, 2007.

[175] A. M. Michel, K. Roy Choudhury, A. E. Firth, N. T. Ingolia, J. F. Atkins, and P. V. Baranov, "Observation of dually decoded regions of the human genome using ribosome profiling data," *Genome Res*, p. doi:10.1101/gr.133249.111, 2012.

[176] K. Yanagitani, Y. Kimata, H. Kadokura, and K. Kohno, "Translational pausing ensures membrane targeting and cytoplasmic splicing of xbp1u mrna," *Science*, vol. 331, no. 6017, p. 586, 2011.

[177] D. Ouelle, F. Zindy, R. Ashmun, and C. Sherr, "Alternative reading frames of the ink4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest," *Cell*, vol. 83, no. 6, pp. 993–1000, 1995.

[178] A. Nekrutenko, S. Wadhawan, P. Goetting-Minesky, and K. Makova, "Oscillating evolution of a mammalian locus with overlapping reading frames: an xlαs/alex relay," *PLoS Genetics*, vol. 1, no. 2, p. e18, 2005.

[179] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences," *Nucleic Acids Research*, vol. 38, no. 12, pp. e132–e132, 2010.

VITA

Ivan Antonov was born in Bogoroditsk, Tula region, USSR on July 22, 1986 to Valentin Antonov and Marina Antonova. He grew up in Ozery, Moscow region where he graduated from the High School №3 in 2003. Ivan received his M.Sc. in Bioengineering and Bioinformatics from Moscow State University, Russia in 2008. From February 2005 to June 2008 he worked at GeneGo Inc. developing software for systems biology and drug discovery. After graduating from the Moscow State University in 2008 Ivan moved to Atlanta, Georgia to join Dr. Borodovsky's lab and started working on his PhD.