Prediction of Ribosomal -1 Frameshifts in the *Escherichia coli* K12 Genome

Sanghoon Moon, Yanga Byun, and Kyungsook Han*

School of Computer Science and Engineering, Inha University,
Inchon 402-751, Korea
jiap@inhaian.net, quska@inhaian.net, khan@inha.ac.kr

Abstract. Ribosomal frameshifting at a particular site can yield two protein products from one coding sequence or one protein product from two overlapping open reading frames. Many organisms are known to utilize ribosomal frameshifting to express a minority of genes. However, finding ribosomal frameshift sites by a computational method is difficult because frameshift signals are diverse and dependent on the organisms and environments. There are few computer programs available for public use to identify frameshift sites from genomic sequences. We have developed a web-based application program called FSFinder2 for predicting frameshift sites of general type. We tested FSFinder2 on the *Escherichia coli* K12 genome to detect potential -1 frameshifting genes. From the genome sequence, we identified 18,401 frameshift sites following the X XXY YYZ motif. 11,530 frameshift sites out of the 18,401 sites include secondary structures. Comparison with the GenBank annotation produced 11 potential frameshift sites, including 3 known frameshift sites. The program is useful for analyzing frameshifts of various types and for discovering new genes expressed by frameshifts.

1 Introduction

Ribosomes in general terminate translation at three kinds of stop codons (UAG, UGA and UAA), but some ribosomes continue to decode after the stop codons. This alternative translational event is called 'recoding'. Recoding events include frameshifting, read-through and bypassing [1-3]. In frameshifting, ribosomes shift reading frame by one or more nucleotides at a specific mRNA signal between overlapping genes [4]. Frameshifts are classified into different types depending on the number of nucleotides shifted and the shifting direction. The most common type is a -1 frameshift, in which the ribosome slips a single nucleotide in the upstream direction. -1 frameshifting requires a frameshift cassette that consists of a slippery site, a stimulatory RNA structure and a spacer. +1 frameshifts are much less common than -1 frameshifts, but have been observed in diverse organisms [1].

No program exists to predict general types of frameshift. In addition, existing computational models predict too many false positives. In previous work we developed a program called FSFinder (Frameshift Signal Finder) for predicting -1 and +1

* Correspondence author.

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNBI 4115, pp. 612 – 621, 2006.
© Springer-Verlag Berlin Heidelberg 2006

frameshift sites [5]. That program is written in Microsoft C# and is executable on Windows systems only. To remove these limitations and to handle frameshifts of general type, we developed a new web-based application called FSFinder2. Users can predict frameshift sites of any type online from any web browser and operating system.

In previous experimental results of testing FSFinder2 on ~190 genomic and partial DNA sequences showed that it predicted frameshift sites efficiently and with greater sensitivity and specificity than other programs, because it focused on the overlapping regions of ORFs and prioritized candidate signals (For -1 frameshifts, sensitivity was 0.88 and specificity 0.97; for +1 frameshifts, sensitivity was 0.91 and specificity 0.94) [5-7].

Using the web service of the FSFinder2, we analyzed the *Escherichia coli* (*E. coli*) K12 genome sequence to find the -1 frameshifting genes with high probability. From the *E. coli* K12 genome sequence, we found 18,401 frameshift sites after the X XXY YYZ motif. Among these sequences, 11,530 frameshift sites included secondary structure such as pseudoknots or stem-loops. Using the GenBank description we found 312 overlapping regions of two genes with more than 1 base. Using FSFinder we found 309 overlapping regions with more than 30 bases. After removing redundant ones, we obtained 195 overlapping regions and found 66 potential frameshift sites in the 195 overlapping regions. Among these sites, 11 sites including 3 known frameshift sites were considered significant based on the gene length, shape and the length of overlapping region. We believe that at least 4 new frameshift sites are highly likely to be frameshift sites.

2 Analyzing Method

2.1 Finding Frameshift Motifs

The cassettes of -1 frameshift consist of three parts: slippery site, spacer and secondary structure. The slippery site is usually a heptameric sequence in the form X XXY YYZ (in the incoming 0-frame), where X, Y and Z can be same nucleotides [5-7]. The spacer is a short sequence of 5 to 11 nucleotides separating slippery site and the downstream secondary structure. The downstream structure is usually a pseudoknot or simple stem-loop, as shown in Fig. 1.

For analyzing frameshift sites in the *E. coli* genome sequence, we detected heptameric sequence with a secondary structure. In previous work by others [5-7], frameshift sites have two constraints. In slippery site X XXY YYZ, X is any nucleotide, Y is A or U, and Z is either A, U or C. However, in our work, any kind of nucleotide with secondary structure can be located in the slippery sequence.

Fig. 2 shows the parameters for the FSFinder web service. The web service and web application of FSFinder2 were implemented using XML, XSLT and JavaScript. If the user sends a query to the FSFinder2 server after setting parameters or defining a new model, all the computations are done on the server side. After computation, the server sends the results to the user.

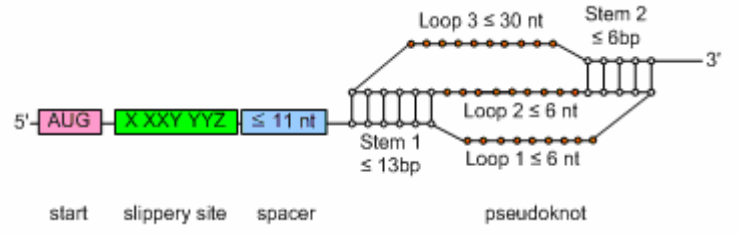

Fig. 1. A programmed -1 ribosomal frameshift signal with H-type pseudoknot. Stem 1 has 11 base pairs, stem 2 has 6 base pairs, and both loops of the pseudoknot have 6 nucleotides. In particular, any nucleotide can be located in the slippery site not the same as previous work.

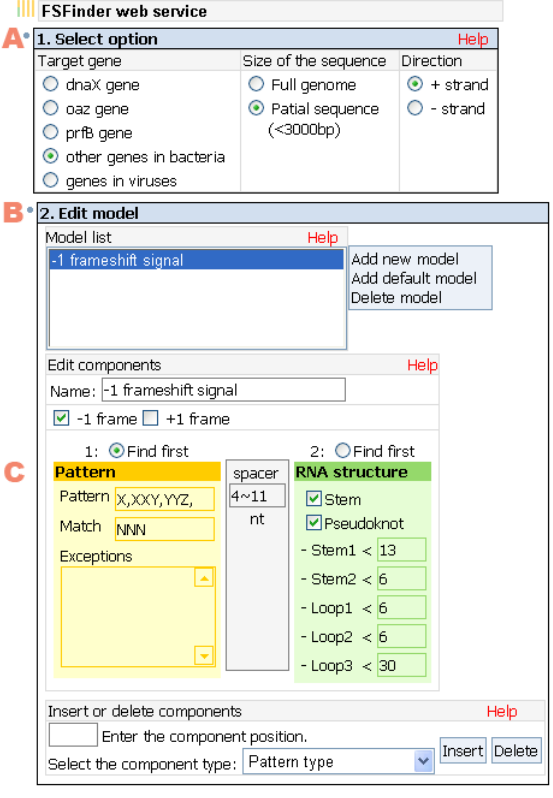

Fig. 2. The input page of the FSFinder web service. (A) The select option lets the user choose the type of genes expressed via frameshifts, the size of the sequence and its direction. (B, C) The user can define a new model by specifying its components and their locations. To analyze *E. coli* genome, we set the selection option as other genes in bacteria, partial sequence and +1 strand. Because we focused on -1 frameshift model, we made a model which is the general motif X XXY YYZ of the -1 frameshift. There was no limitation of frameshift site, thus any kind of nucleotide can be located in the frameshift site. For downstream secondary structure, the length of the stems and loops set by default.

```
<FSFinder_run xmlns="http://wilab.inha.ac.kr/WSFSFinder/">
  <FSFinder_Input xmlns:xsd=http://www.w3.org/2001/XMLSchema
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <Sequence_information>
      <Target_gene>other genes in bacteria</Target_gene>
      <Sequence_size>Patial</Sequence_size>
      <Sequence_direction>Plus strand</Sequence_direction>
      <Sequence>input sequence</Sequence>
      <Sequence_name>E. coli K-12</Sequence_name>
    </Sequence_information>
    <FSFinder_Model_List>
      <FSFinder_Model Find_first="0" Overlap_with="-1">
        <Model_name>-1 frameshift signal</Model_name>
        <Components>
          <Pattern_type Spacer="0">
            <Pattern>X,XXY,YYZ,</Pattern>
            <Pattern_match>NNN</Pattern_match>
          </Pattern_type>
          <RNA_structure_type Spacer="4~11" Structure=
            "StemLoop;Pseudoknot">
            <Stem1_size_max>13</Stem1_size_max>
            <Stem2_size_max>6</Stem2_size_max>
            <Loop1_size_max>6</Loop1_size_max>
            <Loop2_size_max>6</Loop2_size_max>
            <Loop3_size_max>30</Loop3_size_max>
          </RNA_structure_type>
        </Components>
      </FSFinder_Model>
    </FSFinder_Model_List>
  </FSFinder_Input>
</FSFinder_run>
```

Fig. 3. XML schema with the parameters shown in Fig. 2. If the user sets the parameters of a model in the web page of the FSFinder web service, the parameters are converted to XML. After the request of a web service is sent to the FSFinder server, all the computations are done on the server side.

To analyze *E. coli* genome, we set the selection option as other genes in bacteria, partial sequence and + strand (Fig. 2A). Because we focused on -1 frameshift sites (Fig. 2B), we defined a signal that fits all kinds of motif X XXY YYZ of the -1 frameshift. There was no limitation of frameshift site, thus any kind of nucleotide can be located in the frameshift site. Thus we set the match type as NNN (N is any nucleotide) and no exception of arrangement. That is, frameshift site can occur even at A AAA AAA or U UUU UUU. Default values were used for the lengths of the stems and loops in the downstream secondary structure (Fig. 2C). Fig. 3 shows the XML schema for the parameters of FSFinder2. When the user sets the parameters of a

model in the web page of the FSFinder web service, the parameters are converted to XML. After the request of a web service is sent to the FSFinder server, all the computations are done on the server side.

2.2 Finding Overlapping Regions of Genes

Fig. 4 shows the frameshift site of the genome of SARS corona virus (NC_004718). ORF1a (265..13398) and ORF1b (13398..21485) partially overlap each other. U UUA AAC is the slippery sequence and the secondary structure is a pseudoknot [8]. In the GenBank description, the overlapping region of ORF1a and ORF1b has only one 1 nucleotide. ORF1b starts from codon AAA instead of a regular start codon. The start codon of ORF2 (red triangle) exists outside the overlapping region. Frameshifting can occur even in overlapping regions with 1 nucleotide. To solve these problems, finding an overlapping region is divided into two processes: finding motifs and finding overlapping regions.

2.2.1 Finding Overlapping Regions Using FSFinder

The shape of an overlapping region is considered to find an overlapping region. As shown in Fig. 5, the overlapping region is extended from stop codon 1 of open reading frame 2 to stop codon 2 of open reading frame 1 (green color). The minimum length of both genes should be longer than 100 nucleotides, and the length of the overlapping region should be longer than 30 nucleotides. In addition, the location of the start codon of ORF2 does not matter. As in the SARS corona virus, the start codon is not always located in the overlapping region of frameshifting genes.

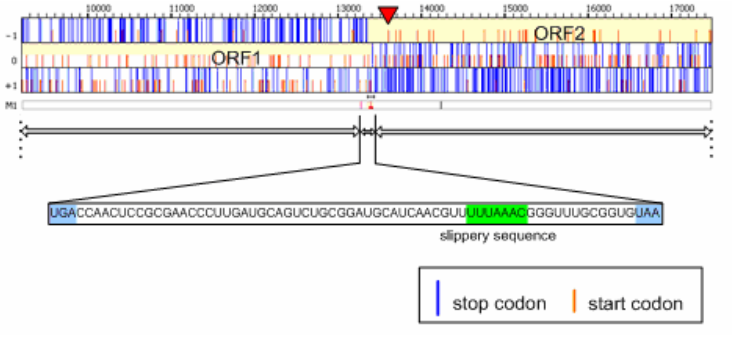

Fig. 4. The frameshift site of the SARS corona virus (the slippery sequence UUUAAAC shown in green background)

2.2.2 Finding Overlapping Regions Using the GenBank Description

Additional method to find overlapping regions is to use the description of the GenBank file. If more than 1 nucleotide is overlapped, we consider two open reading frames as a candidate of partially overlapping genes. Finally, the overlapping regions found both by FSFinder and by the GenBank description are used for further analysis.

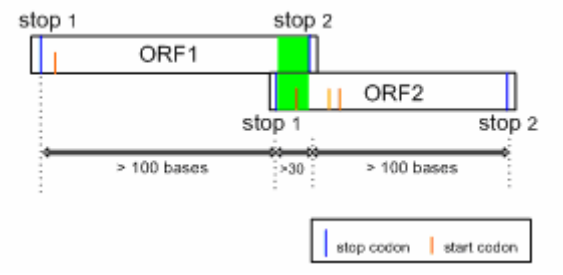

Fig. 5. Finding overlapping regions by FSFinder. Unlike general overlapping regions, FSFinder extends the overlapping region from stop codon 1 of open reading frame 2 to stop codon 2 of open reading frame 1 (green color). The length of both ORFs should be longer than 100 nucleotides and the length of the overlapping region should be longer than 30 nucleotides.

3 Results and Discussion

The *E. coli* K12 genome sequence (NC_000913) was obtained from GenBank. As shown in Fig. 6, all the heptameric sequences that follow X XXY YYZ motif were

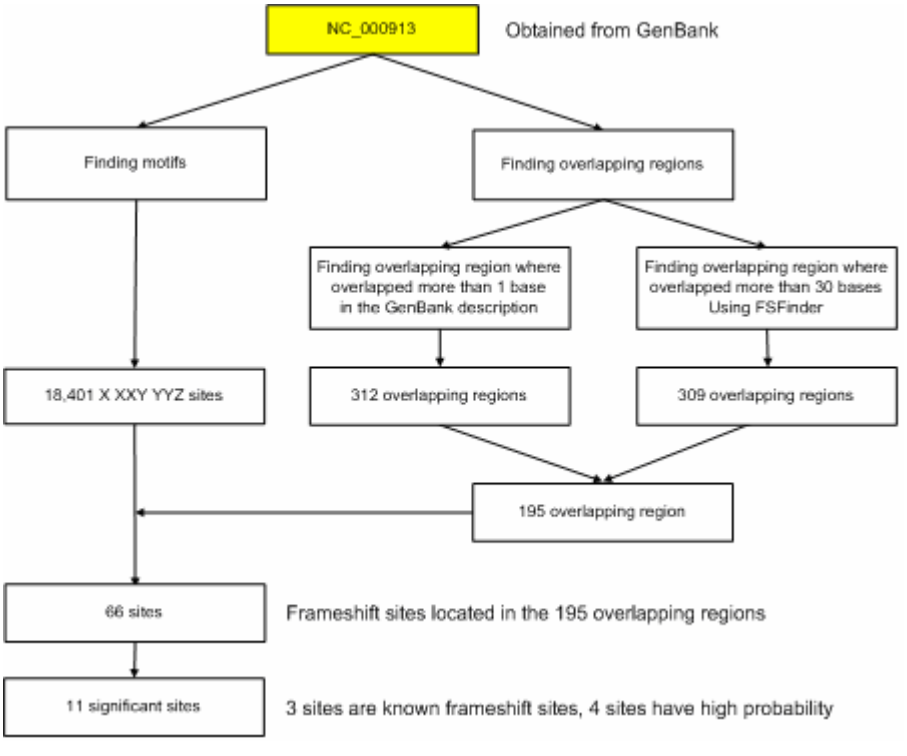

Fig. 6. The prediction process of frameshift sites from the *Escherichia coli* K 12 genome sequence

examined. As a result of this process 18,401 sites were found. Whether these sites are located in the overlapping region or not was not considered when finding these sites.

For overlapping regions of ORFs, we found 312 candidate regions that were partially overlapped more than 1 base according to the GenBank description and 309 candidate regions that were overlapped more than 30 bases using the FSFinder web service. After removing redundant sites, we obtained 195 overlapping regions. From these regions, heptameric sequences that were not located in the overlapping region were filtered out. Just 66 sites remained in the overlapping regions. From the gene length, shape and the length of the overlapping region, 11 sites including 3 known frameshift sites were identified as significant candidates.

This process consists of two sub-processes: finding motifs and finding overlapping regions. The *E. coli* K 12 genome sequence (NC_000913) was obtained from the GenBank. To find motifs, all the heptameric sequences that follow X XXY YYZ motif were found. 18,401 sites were detected. According to the GenBank description, there were 312 candidate overlapping regions that were partially overlapped more than 1 base in. The FSFinder web service found 309 candidate overlapping regions that were overlapped more than 30 bases. After removing redundant sites, we obtained 195 overlapping regions. After further removing heptameric sequences that were not located in the overlapping region, 66 sites remained in the overlapping regions. Considering the gene length, shape and length of the overlapping region, 11 frameshift sites including 3 known sites were considered significant candidates.

Table 1 shows the names of overlapping genes, the locations of the slippery sites, slippery sequences and number of overlapped nucleotides from our analysis. The three genes, yi21_1-yi22_1, yi21_5-yi22_5, yi21_6-yi22_6, are known genes expressed via -1

Table 1. The predicted frameshift sites in the E. coli K 12 genome sequence. The first threegenes marked with * symbol are those with known frameshift sites. PK represents a pseudoknot.

Gene name	Slippery site	Slippery sequence + secondary structure	# bases in the overlapping region found by FSFinder	# bases in the overlapping region based on GenBank
i21_1..yi22_1*	380892	A AAA AAG + PK	61	43
i21_5..yi22_5*	3184526	A AAA AAG + PK	61	43
i21_6..yi22_6*	4496612	A AAA AAG + PK	61	43
insA_6..insB_6	3581757	A AAA AAC +PK	103	82
entB..entA	627723	A AAC CCG + PK	76	1
tehA..tehB	1499529	G GGA AAA + PK	154	4
xdhB..xdhC	3001496	G GGG GGA + stem	37	4
mhpD..mhpF	372127	C CCA AAA only	61	4
fliM..fliN	2019082	U UUA AAU only	40	4
atoS..atoC	2319873	G GGA AAU only	85	4
yijC..yijD	4159773	G GGA AAU only	46	1

frameshifts. These three genes and insA_6-insB_6 are insertion sequences. We believe that entB-entA, tehA-tehB, and xdhB-xdhC have a high probability using frameshift events. All these seven genes have either pseudoknot or stem loop as a downstream secondary structure. The rest four genes, mhpD-mhpF, fliM-fliN, atoS-atoC and yijC-yijD, have no downstream secondary structures and have a lower probability of frameshifting than the other 7 genes.

There exist previous works similar to our approach. Hammell *et al.* [7] studied -1 ribosomal frameshift signals in the large databases. Using their well-established model, they found that -1 frameshifts occur with frequencies from two- to six-fold greater than random. They considered the nucleotides of a frameshift site, spacer, and pseudoknot. However, they focused on the -1 frameshifts only. Bekaert *et al.* [6] performed a computational method similar to Hammell's approach. From their model, they designed a model for -1 eukaryotic frameshifting. But these two programs are not available for public use.

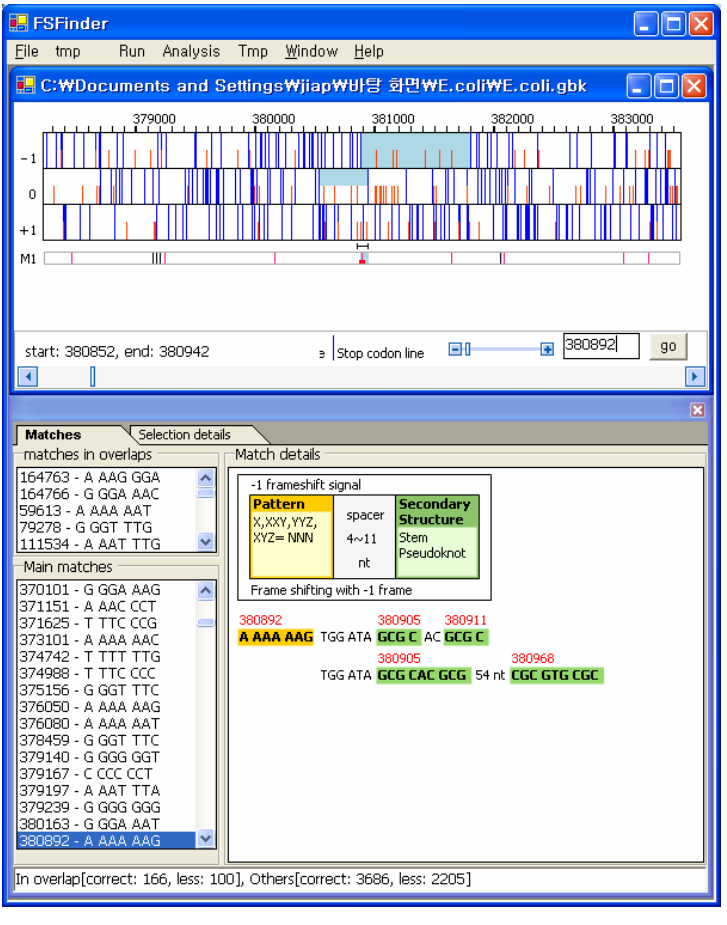

Fig. 7. Frameshift site of overlapping genes i21_1..yi22. FSFinder2 found the sequences of the frameshift cassettes and locations of the slippery site.

The FreqAnalysis [9] is an application program available for public use. FreqAnalysis was implemented in the Java language and can find putative translational frameshift from probabilistic calculation. In fact, Shah *et al.* [9] found putative frameshift sites from the *Sacharomyces cerevisiae* ORFs. But FreqAnalysis does not consider the X XXY YYZ frameshift motif. Thus it is hard to compare the results of FreqAnalysis with ours directly.

Fig. 7 shows the result of the FSFinder2. Two rectangles filled with sky blue color in the upper window represent i21_1 and yi22_1, respectively. The i21_1 gene located in 0 frame and the yi22_1 gene located in -1 frame are partially overlapped. When translational frameshift occurs, translation continues past the stop codon at 380,938 until the stop codon at 381,801.

4 Conclusion

Ribosomal frameshifting is unusual event which is known to affect producing heterogeneous proteins, auto-regulation. Prediction of frameshift sites is very difficult. On the other hand if prediction is possible, this is very useful to understand of biological phenomenon. And unveiling of unknown protein production mechanisms can be realized. For this valuable advantage, we develop a web application and serve web service for prediction of frameshift sites.

Using the FSFinder, we analyzed *Escherichia coli* K 12 genome sequence to detect potential -1 frameshifting genes. From the E. coli K 12 genome sequence, we have got 18,401 frameshift sites followed X XXY YYZ motif. Among these sequences, 11,530 frameshift sites included secondary structure. Comparing with GenBank description, we have got 11 sequences including 3 known frameshift sites. Among these sequences, we believe that at least 4 sequences have a high probability to use frameshift event and all these 4 sites have a downstream secondary structure such as pseudoknot and stem loops. Other 4 gene do not have downstream structure, but we believe that these sequences have less probabilities than above 7 genes but significant.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) under grant R01-2003-000-10461-0 and in part by the Brain Korea 21 Project.

References

1. Farabaugh, P.J.: Programmed Translational Frameshifting. Ann. Rev. Genetics 30 (1996) 507-528
2. Gesteland, R.F., Atkins, J.F.: Recoding: Dynamic Reprogramming of Translation. Annu. Rev. Biochem. 65 (1996) 741-768
3. Herr. A.J., Gesteland, R.F., Atkins, J.F.: One Protein From Two Open Reading Frames: Mechanism of a 50 ntTranslational Bypass. EMBO J. 19 (2000) 2671-2680
4. Baranov, P.V., Gesteland, R.F., Atkins, J. F.: Recoding: Translational Bifurcations in Gene Expression. Gene 286 (2002) 187-201

5. Moon, S., Byun, Y., Kim, H.-J., Jeong, S., Han, K.: Predicting Genes Expressed via -1 and +1 Frameshifts. Nucleic Acids Research 32 (2004) 4884-4892
6. Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J., Froidevaux, C., Hatin, Rousset, J., Termier, M.: Towards a Computational Model for -1 Eukaryotic Frameshifting Sites. Bioinformatics 19 (2003) 327-335
7. Hammell, A.B., Taylor, R.C., Peltz, S.W., Dinman, J.D.: Identification of Putative Programmed -1 Ribosomal Frameshift Signals in Large DNA Databases. Genome Research 9 (1999) 417-427
8. Ramos, F.D., Carrasco, M., Doyle, T., Brierley, I.: Programmed -1 Ribosomal Frameshifting in the SARS Coronavirus. Biochemical Society Transactions 32 (2004) 1081-1083
9. Shah, A.A., Giddings, M.C., Parvaz, J.B., Gesteland, R.F., Atkins, J.F., Ivanov, I.P.: Computational Identification of Putative Programmed Translational Frameshift Sites. Bioinformatics 18 (2002) 1046-1053