Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
Computational Biology and Chemistry 31 (2007) 298–302
www.elsevier.com/locate/compbiolchem

Database note

FSDB: A frameshift signal database

Sanghoon Moon, Yanga Byun, Kyungsook Han *

School of Computer Science and Engineering, Inha University, Inchon 402-751, Republic of Korea

Received 12 April 2007; accepted 7 May 2007

Abstract

Programmed frameshifting is a recoding event in which a ribosome shifts reading frame by one or more nucleotides at a specific mRNA signal between overlapping genes. Programmed frameshifting is involved in the expression of many genes in a wide range of organisms, especially in viruses and bacteria. The mechanism of programmed frameshifting is not fully understood despite many studies, and there are few databases available for detailed information on programmed frameshifting. We have developed a database called FSDB (Frameshift Signal Database), which is a comprehensive compilation of experimentally known or computationally predicted data about programmed ribosomal frameshifting. FSDB provides a graphical view of frameshift signals and the genes using programmed frameshifting for their expression. It also allows the user himself/herself to find programmed frameshift sites in genomic sequences using a program called FSFinder (http://wilab.inha.ac.kr/fsfinder2). We believe FSDB will be a valuable resource for scientists studying programmed ribosomal frameshifting. FSDB is freely accessible at http://wilab.inha.ac.kr/fsdb/.

Keywords: Programmed ribosomal frameshift; Recoding; Translation; Frameshift signal

1. Introduction

In the translation process some genes make two proteins from the same mRNA sequence by selecting alternative translation at specific regions. This event is called 'recoding' (Gesteland and Atkins, 1996; Baranov et al., 2002). Programmed frameshifting is one type of recoding event in which a ribosome shifts reading frame by one or more nucleotides at a specific mRNA signal between overlapping genes (Baranov et al., 2002). It is known to play an important role in autogenous control. Frameshifts are classified according to the number of nucleotides shifted and the direction of shift. Most known programmed frameshifts are −1 or +1 frameshifts (slippage of one base either forward or backward).

Programmed frameshifting is involved in the expression of certain genes in a wide range of organisms, especially in viruses and bacteria. For example, infectious viruses such as HIV, human herpes virus, and human coronavirus are known to utilize programmed frameshifting. Many factors appear to be involved in programmed frameshifting, but the mechanism of programmed frameshifting is not fully understood despite a lot of biochemical approaches and some computational approaches. Previous works have identified several slippery sequences and stimulatory structures as major factors of programmed frameshifting. Thus, there should be a compilation to summarize these factors and to provide useful information for users.

In contrast with many databases in other researches, there are few databases providing detailed information on programmed frameshifting. RECODE (Baranov et al., 2001, 2003) has been the only database that provides information about recoding events based on the scientific literature and personal communications. It covers three kinds of recoding event: programmed frameshifting, readthrough, and bypassing. RECODE offers useful information such as gene name, *cis*-elements, *trans*-elements, product and references, but does not provide complete details of frameshifting. For example, it is not easy to figure out the overall structure of a frameshift signal from the text-based information of RECODE. The locations of the slippery sequences in several RECODE entries differ from those in GenBank, causing confusion to users who want to locate a frameshift site in the sequence.

Very recently Jacobs et al. (2007) released a database called PRFdb for −1 frameshifts in *Saccharomyces cerevisiae*. However, PRFdb is limited to text-based information for −1 frameshifts in one organism. It provides the location of genes,

* Corresponding author. Tel.: +82 32 860 7388; fax: +82 32 863 4386.
E-mail address: khan@inha.ac.kr (K. Han).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.05.004

but does not provide the detailed information such as the location of the slippery sequence and spacer.

We have developed a database called FSDB (Frameshift Signal Database), which is focused on ribosomal frameshift events. It is a comprehensive compilation of experimentally known or computationally predicted data about programmed ribosomal frameshifting. In particular it provides structural information on frameshift factors. The entries in FSDB were obtained from three sources: RECODE (Baranov et al., 2001, 2003), PseudoBase (van Batenburg et al., 2000) and FSFinder (Moon et al., 2004). FSDB also provides all the parameters of a frameshift model that allows the user to identify frameshift sites using FSFinder.

Currently FSDB has 253 entries. The pseudoknot structure data for 22 entries were obtained from PseudoBase, 122 were from RECODE, and 109 were predicted by FSFinder and then confirmed by GenBank annotations. The structural data for all the entries except the 22 entries from PseudoBase were predicted by pknotsRG (Reeder and Giegerich, 2004) and visualized by PseudoViewer (Byun and Han, 2006; Han and Byun, 2003). The graphical views of frameshift signals, including drawings of pseudoknots and secondary structures, should help workers to understand ribosomal frameshifting more easily and quickly, which in turn should help to identify new programmed ribosomal frameshifts.

2. The database description

As shown in Table 1, the FSDB database has 253 entries. All the data available in FSDB can be saved in extensible markup language (XML) format and downloaded for later analysis. At present it contains data on frameshifts of −1 and +1 type, and more complex types will be included in the future.

2.1. Contents of the database

The contents of FSDB are divided into two: overall information and graphical views. The contents can be outlined as follows. More details of the overall information and graphical views are given in the section on the user interface.

Overall information
- FSDB ID: unique identifier of an FSDB entry.
- Type of frameshift: −1 or +1 frameshift.
- Data type: experimental if the slippery sequence was determined by an experimental method, and predicted if it is predicted by a program.
- Kingdom of organism: viruses, prokaryota or eukaryotes.
- Definition: organism, sequence length and sequence type.
- Resources: relevant references to the entry.
- Nucleic acid sequence: nucleotide sequence from GenBank.
- Amino acid sequence: amino acid sequence of the protein product and CDS.

Graphical view
- Three open reading frames.
- Slippery sequence, location of slippery sequence, and sequence of secondary structure.
- Secondary structure drawing.
- Components of the frameshift model.
- Target gene: *prfB*, *oaz*, *dnaX*, other genes in bacteria or in viruses.
- Sequence type: either omplete genome or partial sequence.
- Direction: + strand or − strand.
- Components of the user-defined frameshift model.

2.2. Database entry

The FSDB database can be searched by a combination of five selections: data type, frameshift type, organism type, slippery sequence, and key word. Data type is either experimental or predicted. Experimental data are experimentally verified data obtained from PseudoBase and RECODE. Predicted data are: (1) data not experimentally verified, obtained from RECODE and (2) data predicted computationally by FSFinder and confirmed by GenBank annotations.

Currently FSDB provides information on two frameshift types: −1 and +1 frameshifts. −1 and +1 frameshifts are the most common frameshift events. We have simplified organism type to three categories: viruses, prokaryota and eukaryotes. Bacteriophages and prophages are included in viruses. Both bacteria and bacterial insertion sequences are included in prokaryota, and eukaryotic transposable elements are included in eukaryota. We provide a detailed classification of organisms based on the GenBank definition. For example, the abbreviated words BCT, PHG and VRL represent bacterial, bacteriophage and virus nucleotide sequences, respectively.

The identifier used in FSDB is represented in the form of EFkOn, where E is the data type of a slippery sequence (E for

Table 1
Current statistics of FSDB

Type	Organisms						Total
	Viruses		Prokaryota		Eukaryota		
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted	
−1 frameshifting	38	75	7	6	3	13	142
+1 frameshifting	1	0	2	83	12	13	111
Total	114		98		41		253
	Experimental data: 63				Predicted data: 190		

the experimental data and P for the prediction data), F is the frameshift type (P for + frameshift and M for − frameshift), k is the number of nucleotides shifted, O is the organism type (V for viruses, P for prokaryota, and E for eukaryota), and *n* is a four-digit number. For example, EM1E0001 refers to the experimental data of a slippery sequence for −1 frameshifts in eukaryota.

2.3. User interface

The user can search the database by selecting the data type, frameshift type, organism, slippery sequence and key word. Each menu of the data type, frameshift type, organism and slippery sequence lists possible choices. By typing one or more key words separated by space in the 'Keyword' field, the user can search entries that include the key words in the GenBank definition (Fig. 1A). For example, typing 'PHG' or 'phg' allows the user to find all bacteriophages in the virus category of FSDB. All the entries, or those matched to the user selection, are listed in the left pane of the window (Fig. 1B). If the user clicks a specific entry from the listed entries, detailed information about the entry is displayed in the view area of the window (Fig. 1C–G).

The entry details are in two parts: overall information (Fig. 1C) and graphical view (Fig. 1D–G). FSDB ID, frameshift type, data type and kingdom are shown at the top of the window. The resource of an entry displays relevant references or 'predicted by FSFinder'. The nucleic acid sequence field is linked to GenBank. The user can download the sequence file in FASTA or GenBank format from FSDB and use it as input to FSFinder. The amino acid sequence field consists of gene and coding sequence (CDS), providing information about the gene and CDS that utilize the ribosomal frameshift. That information was obtained from GenBank. If there is no information on gene or CDS in the GenBank annotation, the amino acid sequence field remains empty.

The graphical view visualizes the frameshift signal. Fig. 1D shows three open reading frames with start codons and stop codons marked in each frame. The light yellow region represents the open reading frames. The overlapping region of the open reading frames (Fig. 1E) is blown up in Fig. 1F, which displays the nucleotide sequence of the region. The frameshift site and stimulatory structures are highlighted in yellow and green, respectively. The exact locations of the frameshift site and spacer sequence are also displayed. Fig. 1G shows a frameshift model of the entry. The parameters of the frameshift model can be used directly to locate the frameshift entity by the user himself/herself using FSFinder. When running FSFinder, frames may have to be alternated to get the same result as the entry.

Like the data type of slipper sequences, secondary structure data of frameshift signals are of two types: experimental or predicted. Out of the total 30 experimental secondary structure data, 22 structure data were obtained from PseudoBase and the remaining 8 structure data were obtained from literatures (Blinkova et al., 1997; Mejlhede et al., 2004; Dulude et al., 2002; Ivanov et al., 2004; Baranov et al., 2002).

Except the 30 experimental structure data, remaining secondary structures were predicted by pknotsRG. When predicting a stimulatory secondary structure, the longest possible subsequence was used for the structure, and the spacer length was determined by FSFinder. Both predicted and experimental secondary structures were visualized by PseudoViewer.

3. Resources

3.1. RECODE

One hundred and twenty-two entries in FSDB were extracted from 185 entries in the RECODE database (97 −1 frameshift data and 88 +1 frameshift data). Sixty-three of the 185 entries were excluded for a variety of reasons. First, some RECODE entries have no GenBank accession number and so no sequence data associated with them. Second, the same GenBank accession number may appear in more than one RECODE entry. For example, in +1 frameshifting, *Botryotinia fuckeliana* has the same accession number (AF291578.1) as *Schizosaccharomyces japonicus*. Third, some RECODE entries are redundant with the PseudoBase data. Fourth, some slippery sequences are too simple and are found too many times in a sequence. For example, the slippery sequence of the *argI* gene in *Escherichia coli* is UUUC, and there are too many UUUC occurrences in *E. coli*.

The frameshift sites in most organisms are found using the default parameters of FSFinder. However, if the organism uses a specific slippery sequence, frameshift sites can be found with the user-defined model of FSFinder. All the data are classified into experimental data and predicted data, as mentioned before.

3.2. PseudoBase

Twenty-two experimental data of secondary structures in FSDB were obtained from the PseudoBase database, and visualized by PseudoViewer. When there is discrepancy in structure data from different resources, we used the structure data in PseudoBase.

3.3. FSFinder

Every entry in FSDB was added to FSDB by the following procedure.

1. Find a region that contains the components (slippery sequence, spacer, stimulatory structure, SD-like sequences, etc.) of basic frameshift signals using FSFinder.
2. Compare the slippery sequence in the region with known slippery sequences and find an overlapping open reading frame (ORF) that includes the slippery sequence.
3. Determine that the frameshift candidate is associated with an actual gene if it satisfies any of the followings:
   A. There is a GenBank annotation that the candidate causes frameshifting.
   B. The GenBank annotation does not classify the frameshift candidate as a hypothetical gene and the overlapping ORF is identical to a gene (such as *gap-pol*) in GenBank.
   C. The GenBank annotation classifies the frameshift candidate as a hypothetical gene, but the ORF translates into a protein product confirmed by InterproScan or BLAST.

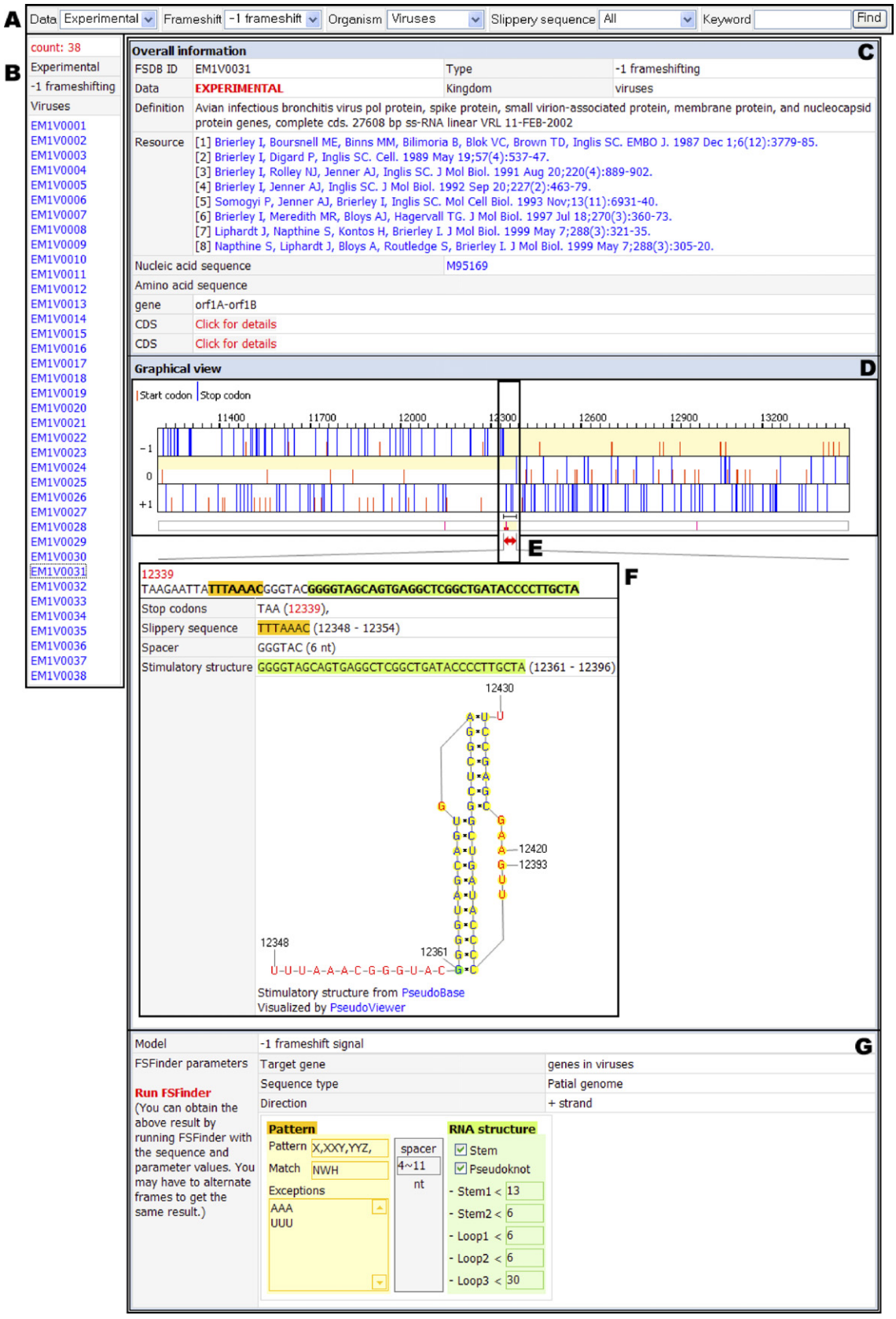

Fig. 1. The user interface of the FSDB. (A) Selection menu. (B) List of FSDB entries that fit the combination of selections in A. (C) Overall information about an entry. (D–G) Graphical view of the entry. (D) Three open reading frames with start codons and stop codons marked in each frame. (E) The overlapping region of the open reading frames in which a frameshift signal is found. (F) Components of the frameshift signal, including slippery sequence, spacer, and stimulatory structure. (G) Frameshift model of the entry. The parameters of the frameshift model can be used directly to locate the frameshift entity using FSFinder. (For interpretation of the references to colour in the text, the reader is referred to the web version of the article.)

After analyzing 1 968 virus genome sequences (April 2006) and 1 161 prokaryotic genome sequences (August 2006) of NCBI, we added 109 entries to FSDB as prediction data. Using the research results by Bekaert et al. (2006), we also corrected the wrong annotations of 'shifty' RF2 genes in GenBank and added 25 entries to FSDB.

4. Comparison with other programmed ribosomal frameshift databases

As mentioned above, RECODE has been the only database that provides information about translational recoding events including frameshifting. It offers useful information such as gene name, *cis*-elements, *trans*-elements, product, and references. However, the RECODE data is in text form and is not easy to understand. Moreover, the location of the slippery sequence in several RECODE entries differs from that in GenBank, causing confusion to a user who wants to locate a frameshift site in the GenBank sequences.

Very recently Jacobs et al. (2007) released a database called PRFdb for −1 frameshifts in *Saccharomyces cerevisiae*. However, PRFdb is limited to text-based information for −1 frameshifts in one organism. It provides the location of genes, but does not provide the detailed information such as the location of the slippery sequence and spacer.

FSDB is complementary to RECODE. It uses a graphical view to represent programmed ribosomal frameshifts, and stimulatory structures as well as open reading frames are visualized. In addition to the nucleotide sequence of a gene and the amino acid sequence of the protein product, all components of the frameshift signals are displayed, which permits much easier understanding than the text-only representation. Since the location of the frameshift site is calculated from the GenBank annotation, the user can immediately use the GenBank file for further analysis.

5. Conclusion

Programmed ribosomal frameshifting has been studied for a long time, but there are few relevant databases and programs dealing with both −1 and +1 programmed frameshift. We have developed a database called FSDB which is a compilation of experimentally known or computationally predicted data about programmed ribosomal frameshifting. FSDB provides graphical views of the frameshift signals, the genes utilizing frameshifting for their expression, and the protein products resulting from frameshifting. It also allows the user to find frameshift sites himself/herself from genome sequences using the program FSFinder. All the data available in the database can be saved in XML format and downloaded for later analysis. FSDB promises to be a valuable resource for biologists studying programmed ribosomal frameshifting and infectious diseases associated with programmed ribosomal frameshifting.

Acknowledgements

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

References

Baranov, P.V., Gurvich, O.L., Fayet, O., Prere, M.F., Miller, W.A., Gesteland, R.F., Atkins, J.F., Giddings, M.C., 2001. RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucl. Acids Res. 29, 264–267.

Baranov, P.V., Gesteland, R.F., Atkins, J.F., 2002. Recoding: translational bifurcations in gene expression. Gene 286, 187–201.

Baranov, P.V., Gurvich, O.L., Hammer, A.W., Gesteland, R.F., Atkins, J.F., 2003. RECODE. Nucl. Acids Res. 31, 87–89.

Bekaert, M., Atkins, J.F., Baranov, P.V., 2006. ARFA: a program for annotating bacterial release factor genes, including prediction of programmed ribosomal frameshifting. Bioinformatics 22, 2463–2465.

Blinkova, A., Burkart, M.F., Owens, T.D., Walker, J.R., 1997. Conservation of the *Escherichia coli* dnaX programmed ribosomal frameshift signal in *Salmonella typhimurium*. J. Bacteriol. 179, 4438–4442.

Byun, Y., Han, K., 2006. PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. Nucl. Acids Res. 34, W416–W422.

Dulude, D., Baril, M., Brakier-Gingras, L., 2002. Characterization of the frameshift stimulatory signal controlling a programmed −1 ribosomal frameshift in the human immunodeficiency virus type 1. Nucl. Acids Res. 30, 5094–5102.

Gesteland, R.F., Atkins, J.F., 1996. Recoding: dynamic reprogramming of translation. Annu. Rev. Biochem. 65, 741–768.

Han, K., Byun, Y., 2003. PseudoViewer2: visualization of RNA pseudoknots of any type. Nucl. Acids Res. 31, 3432–3440.

Ivanov, I.P., Anderson, C.B., Gesteland, R.F., Atkins, J.F., 2004. Identification of a new antizyme mRNA +1 frameshifting stimulatory pseudoknot in a subset of diverse invertebrates and its apparent absence in intermediate species. J. Mol. Biol. 339, 495–504.

Jacobs, J.L., Belew, A.T., Rakauskaite, R., Dinman, J.D., 2007. Identification of functional, endogenous programmed −1 ribosomal frameshift signals in the genome of *Saccaromyces cerevisiae*. Nucl. Acids Res. 35, 165–174.

Mejlhede, N., Licznar, P., Prere, M., Wills, N.M., Gesteland, R.F., Atkins, J.F., Fayet, O., 2004. −1 frameshifting at CGA AAG hexanucleotide site is required for transposition of insertion sequence IS 1222. J. Bacteriol. 186, 3274–3277.

Moon, S., Byun, Y., Kim, H.-J., Jeong, S., Han, K., 2004. Predicting genes expressed via −1 and +1 frameshifts. Nucl. Acids Res. 32, 4884–4892.

Reeder, J., Giegerich, R., 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynimics. BMC Bioinform. 5, 104.

van Batenburg, F.H.D., Gultyaev, A.P., Pleij, C.W.A., Ng, J., Oliehoek, J., 2000. PseudoBase: a database with RNA pseudoknots. Nucl. Acids Res. 28, 201–204.