GeneTack database: genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences

Ivan Antonov¹, Pavel Baranov ² and Mark Borodovsky¹, ³, ⁴, *

¹School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, ²Biochemistry Department, University College Cork, Cork, Ireland, ³Department of Molecular and Biological Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia and ⁴Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

ABSTRACT

Database annotations of prokaryotic genomes and eukaryotic mRNA sequences pay relatively low attention to frame transitions that disrupt protein-coding genes. Frame transitions (frameshifts) could be caused by sequencing errors or indel mutations inside protein-coding regions. Other observed frameshifts are related to recoding events (that evolved to control expression of some genes). Earlier, we have developed an algorithm and software program GeneTack for ab initio frameshift finding in intronless genes. Here, we describe a database (freely available at http://topaz.gatech.edu/GeneTack/db.html) containing genes with frameshifts (fs-genes) predicted by GeneTack. The database includes 206 991 fs-genes from 1106 complete prokaryotic genomes and 45 295 frameshifts predicted in mRNA sequences from 100 eukaryotic genomes. The whole set of fs-genes was grouped into clusters based on sequence similarity between fs-proteins (conceptually translated fs-genes), conservation of the frameshift position and frameshift direction (−1, +1). The fs-genes can be retrieved by similarity search to a given query sequence via a web interface, by fs-gene cluster browsing, etc. Clusters of fs-genes are characterized with respect to their likely origin, such as pseudogenization, phase variation, etc. The largest clusters contain fs-genes with programed frameshifts (related to recoding events).

INTRODUCTION

Frameshifts predicted by ab initio program GeneTack (1) correspond to reading frame transitions. The transition could be caused by many reasons, among them sequencing errors (2), indel mutations (3), programed frameshifting events (4–6), phase variation (7), overlapping of adjacent genes (8), dual-coding regions (9) and eukaryotic alternative splicing (10).

Although sequencing errors are artifacts of sequencing technologies, authentic indel mutations are features of real sequences. 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 function (11).

In case of phase variation, reversible indel mutations occur at high frequencies at specific sites. They generate a population of bacteria with heterogeneous sequences of phase variant gene, thus increasing population fitness, for example it may help some bacterial pathogens to escape immune response of a host (12). Phase variation results in reversible and inheritable variation of bacterial phenotype.

Programed frameshifting occurs either during translation (PRF, programed ribosomal frameshifting) or transcription (PTR, programed transcriptional realignment). PRF and PTR violate standard triplet decoding allowing for a single protein to be produced from two overlapping open reading frames (ORFs). Hence, GeneTack predicts frame transition between these ORFs. PRF and PTR occur at sites with specific sequence patterns conserved in evolution because programed frameshifting is required for gene expression. Programed 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 (13) and eukaryotic genes encoding ornithine decarboxylase antizyme (14). PRF is abundant in viruses (15), bacteriophages and transposons (16,17). The largest available collection of known PRF genes is available in the Recode database (18).

*To whom correspondence should be addressed. Tel: +1 404 894 8432; Fax: +1 404 894 4243; Email: borodovsky@gatech.edu

© The Author(s) 2012. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.
A frameshift could be predicted when two coding sequences (CDSs) that are in different frames and located close to each other or overlap. Notably, a co-location of some of the 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.

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) (10, 19). 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 (20, 21).

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 (22). Multiple instances of dual coding in human genome were detected by the analysis of ribosomal profiling data obtained from HeLa cells (23). 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 (24) and different protein product is synthesized from the same mRNA at the later rounds of translation. Another well-studied example (9) is expression from the GNAS1 locus (coding for Guanine Nucleotide binding protein Alpha Subunit 1) an alternative product is synthesized from the same mRNA at the later rounds of translation. Another well-studied example (9) is expression from the GNAS1 locus (coding for Guanine Nucleotide binding protein Alpha Subunit 1) an alternative

Table 1. Statistics on eukaryotic and prokaryotic sections of the GeneTack database

<table>
<thead>
<tr>
<th>Database statistics</th>
<th>Prokaryotes</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of species analyzed</td>
<td>1106</td>
<td>100</td>
</tr>
<tr>
<td>Total number of predicted frameshifts</td>
<td>206991</td>
<td>45295</td>
</tr>
<tr>
<td>Total number of clusters</td>
<td>19430</td>
<td>4087</td>
</tr>
<tr>
<td>Number of fs-genes in all clusters</td>
<td>102731</td>
<td>12103</td>
</tr>
<tr>
<td>Number of singleton fs-genes</td>
<td>104260</td>
<td>33192</td>
</tr>
<tr>
<td>Number of clusters with less than five fs-genes</td>
<td>14441</td>
<td>3701</td>
</tr>
<tr>
<td>Number of programmed frameshift clusters</td>
<td>146</td>
<td>5</td>
</tr>
<tr>
<td>Number of indel mutation clusters</td>
<td>4010</td>
<td>2</td>
</tr>
<tr>
<td>Number of clusters of PTC-containing splice variants</td>
<td>n/a</td>
<td>21</td>
</tr>
</tbody>
</table>

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 (27) 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 (28), 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 used 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 percent 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 1106 prokaryotic and 100 eukaryotic species (Table 1). 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 115001 human mRNAs, we predicted 8700 frameshifts, whereas 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, whereas other fs-genes were singletons.
The database home page is the user's entry point. The user can browse prokaryotic or eukaryotic clusters of fs-genes, perform sequence similarity search by BLASTp for a query sequence of interest or search for fs-genes or clusters using a query string. The query string could be the fs-gene/cluster identification number (ID) or cluster name.

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 programed framesshifting) 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 (Figure 1). 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 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 programed 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 gene features. To identify programed frameshift clusters, sequences in the vicinities of the frameshifting were analyzed in order to find a conserved motif that would resemble a frameshift site. Protein products from 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 programed frameshift clusters, we have analyzed the frameshift vicinities and grouped clusters by the most over-represented 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 clusters' 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...

Table 1. The GeneTack database entries: fs-genes predicted in the genome of \textit{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.

Figure 1. The GeneTack database entries: fs-genes predicted in the genome of \textit{Escherichia coli} str. K-12 substr. DH10B.
coordinates and the fs-gene lengths (Figure 2). Sequence LOGOs were generated with the MEME software package (29).

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 (1), GeneTack-Prok (1), GeneTack-Euk (Antonov I et al., manuscript in preparation) and MetaGeneTack (Tang S et al., accepted for publication to Bioinformatics).

GeneTack-GM is a combination of frameshift prediction program GeneTack and a self-training gene prediction program GeneMarkS (28). GeneTack-GM could be used to predict frameshifts in long prokaryotic sequences (longer than 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 expression sequence tags (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 \textit{ab initio} frameshift finder, MetaGeneTack, can be used in this case (Tang S et al., accepted for publication). MetaGeneTack uses heuristic models (30) and applies several additional filters for removing false-positive predictions.

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 frameshifts 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. programed frameshift) it is likely that the input sequence has a frameshift of the same type as well.

AVAILABILITY

The interface to GeneTack database is at http://topaz.genet.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 ID. The genes or clusters are accessible through URLs: http://topaz.genet.edu/GeneTack/cgi/fs_view.cgi?id=FS_ID (for fs-genes) or cof_view.cgi?id=CLUSTER_ID (for clusters).

ACKNOWLEDGEMENTS

We would like to thank Alexandre Lomsadze for valuable help.

FUNDING

Funding for open access charge: The work of I.A. and M.B. was supported in part by the USA National Institute of Health grant [HG000783 to M.B.]; the work of P.V.B. was supported in part by Wellcome Trust grant [094423 to P.V.B.].

Conflict of interest statement. None declared.

REFERENCES


