BIOINFORMATICS

Vol. 14 no. 4 1998
Pages 367–371

Frame: detection of genomic sequencing errors

Nigel P. Brown[1,2,3], Chris Sander[1,2] and Peer Bork[2,3]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, CB10 1SD, UK , [2]European Molecular Biology Laboratory (EMBL-HD), Biocomputing Unit, D-69012 Heidelberg, and [3]Max Delbrück Centrum für Molekulare Medizin (MDC), Berlin-Buch, D-13122 Berlin, Germany

Received on May 19, 1997; revised and accepted on June 16, 1997

Abstract

Motivation: *The underlying error rate for genomic sequencing sometimes results in the introduction of artificial frameshifts and in-frame stop codons into putative protein encoding genes. Severe errors are then introduced into the inferred transcripts through mis-translation or premature termination.*

Results: *We describe a system for screening segments of DNA for frameshift and in-frame stop errors in coding regions. The method is based on homology matching using blastx to compare all six reading frames of the query nucleotide sequence against selected protein sequence databases. Fragments of protein matching neighbouring regions of the query DNA are united and extended laterally to define candidate open reading frames, within which, frameshifts and stops are identified. Suitable targets include prokaryotic or other intron-free genomic sequence and complementary DNAs. As an example of its use, we report here two frameshifted ORFs that deviate from the original TIGR sequence annotations for the recently released Helicobacter pylori genome.*

Availability: *The tool is accessible via the URL http://www.sander.ebi.ac.uk/frame/.*

Contact: *brown@ebi.ac.uk.*

Introduction

Estimates of error rates in genomic sequencing vary from ~0.1% up to as high as 3–4% of nucleotides (States, 1992; Beck, 1993). Of the various errors, the introduction of frameshifts and stop codons in putative genes are the most severe because of their effect at the protein level: mistranslation and premature termination of inferred transcripts, respectively. Frameshift errors were estimated to affect ~2.5% of prokaryotic open reading frames (ORFs) in GenBank (Posfai and Roberts, 1992).

This tool provides a simple and rapid screen for frameshift and internal stop errors in coding regions in intron-free DNA sequences supplied via a Web interface. The underlying method is a simple form of ORF assignment by homology, with sequencing error indications as a by-product, and is similar to some other techniques (Posfai and Roberts, 1992; Gish and States, 1993).

It operates by scanning a query nucleotide sequence against databases of protein sequences and effectively hybridizing similar fragments of protein onto the query in any of its six reading frames. Potential frameshifts are inferred from changes of frame between consecutive hybridized fragments, while internal stop errors are found by inspection of the in-frame codons of the apparent ORF subsuming these fragments. Multiple protein database sequences that show similarity to a given ORF in the query (and may derive from different organisms) provide cumulative evidence for such inferences and focus the user's attention on their location. Summaries of putative ORFs, sequencing errors, and the associated evidence from matched database proteins, are displayed. When a possible error is found, the user can then apply a specialised dynamic programming alignment method (e.g., Guan and Uberbacher, 1996; Huang and Zhang, 1996; Birney *et al.*, 1996), to refine the error analysis based on the list of database hits, and/or inspect the original sequencing data directly.

ORF assembly and frameshift detection

A query nucleotide sequence is searched against selected protein sequence databases using blastx (Altschul *et al.*, 1990; Gish and States, 1993). Long sequences are searched in 20 kilobase pair (kbp) segments with an overlap of 2 kbp to ensure adequate coverage of ORFs bridging segment termini. Blastx output is collated and parsed to extract scoring information and to assign a unique identifier to each hit for subsequent retrieval. Replicate blastx hits from the segment overlaps and nested hits are removed, and the remaining blastx hits are grouped by database protein identifier for analysis.

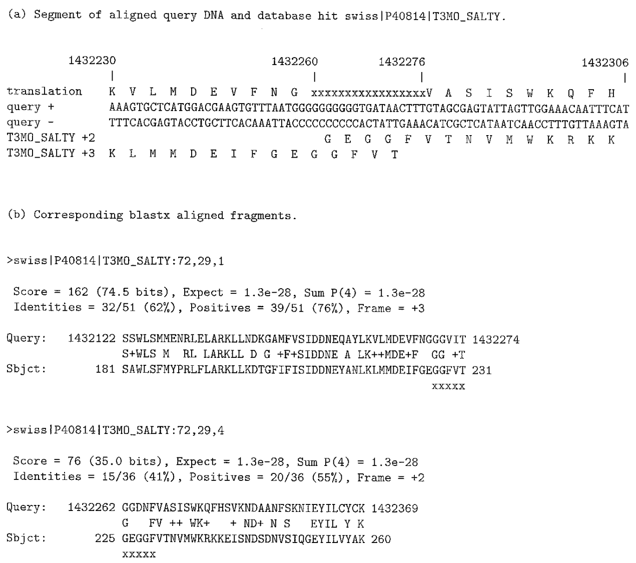
Potential ORF locations along the query sequence are assigned for each protein according to these criteria: several blastx hits (1) lie on the same strand of the query; (2) satisfy preset blastx score thresholds; (3) maintain the same ordering along both query and matching protein sequences; (4) map to the same region of the query sequence.

For each database protein and for each strand orientation (criterion 1), a two-dimensional sparse array is constructed with axes defining the individual hit orderings along the query sequence and along the matching protein sequence, respectively. The corresponding list of hits is sorted by descending blastx score and ascending *P* value. The best hit is removed from the list and forms a 'nucleus' for ORF assembly provided it satisfies a threshold blastx score (criterion 2: nucleation cut-off). Further hits are removed from the list and added to the ORF assembly if they lie diagonally (criterion 3) through the two-dimensional array from the nucleation hit, and they satisfy a second blastx score (criterion 2: assembly cut-off), and they lie within a limiting number of nucleotides from the last diagonal hit (criterion 4: gap cut-off). Extension around a nucleus terminates when criteria 2 or 4 are violated at both growing ends, or there are no further hits to examine.

The bounds of each ORF assembly are further extended downstream to the first in-frame (with respect to the last blastx hit) stop codon, and upstream to the first potential in-frame start codon, as defined by the selected genetic code for the query sequence. Extension is prematurely terminated at either end of a linear query sequence. Alternatively, a 'wraparound' capability continues extension past the apparent endpoints for circular sequences.

The generation of further assemblies continues until the hit list is empty or there are no more nucleation candidates. Each assembly is scored by its cumulative blastx score derived from the component blastx hits with simple averaging of scores for overlapping regions, and by the lowest *P* value of the component hits. A scan of each ORF assembly then determines any frame conflicts between neighbouring blastx hits and in-frame internal stop errors in un-frameshifted regions. Frameshifted regions can be 'sequential' or 'overlapping'. The approximate boundaries of a frameshifted region are given by the inner endpoints of two sequential discrete blastx hits, or by the common extent of two overlapping hits. The frameshift region boundaries are adjusted to consume partial codons at either end. An example of an overlapping frameshift is given in Figure 1.

ORF assemblies terminating at a common stop position are grouped to define an ORF representing a set of 'analogue' (putative homologue) proteins. Two such groupings are presently supported: (1) a 'minimal ORF' has boundaries given by the common stop and by the most distal of the start positions in the analogue set, while (2) a 'maximal ORF' has boundaries given by the downstream stop and by the first potential start codon after an upstream in-frame stop encompassing the analogue set. Each ORF is annotated with the protein identifier of the best scoring analogue therein as a crude indicator of function.



```
(a) Segment of aligned query DNA and database hit swiss|P40814|T3MO_SALTY.

           1432230                       1432260         1432276                       1432306
                 |                             |               |                             |
translation      K  V  L  M  D  E  V  F  N  G  xxxxxxxxxxxxxxxxV  A  S  I  S  W  K  Q  F  H
query +          AAAGTGCTCATGGACGAAGTGTTTAATGGGGGGGGGGTGATAACTTTGTAGCGAGTATTAGTTGGAAACAATTTCAT
query -          TTTCACGAGTACCTGCTTCACAAATTACCCCCCCCCCACTATTGAAACATCGCTCATAATCAACCTTTGTTAAAGTA
T3MO_SALTY +2                                    G  E  G  G  F  V  T  N  V  M  W  K  R  K  K
T3MO_SALTY +3    K  L  M  M  D  E  I  F  G  E  G  G  F  V  T


(b) Corresponding blastx aligned fragments.


>swiss|P40814|T3MO_SALTY:72,29,1

 Score = 162 (74.5 bits), Expect = 1.3e-28, Sum P(4) = 1.3e-28
 Identities = 32/51 (62%), Positives = 39/51 (76%), Frame = +3

Query:   1432122 SSWLSMMENRLELARKLLNDKGAMFVSIDDNEQAYLKVLMDEVFNGGGVIT 1432274
                 S+WLS M  RL LARKLL D G +F+SIDDNE A LK++MDE+F  GG +T
Sbjct:       181 SAWLSFMYPRLFLARKLLKDTGFIFISIDDNEYANLKLMMDEIFGEGGFVT 231
                                                               xxxxx


>swiss|P40814|T3MO_SALTY:72,29,4

 Score = 76 (35.0 bits), Expect = 1.3e-28, Sum P(4) = 1.3e-28
 Identities = 15/36 (41%), Positives = 20/36 (55%), Frame = +2

Query:   1432262 GGDNFVASISWKQFHSVKNDAANFSKNIEYILCYCK 1432369
                 G   FV ++ WK+   + ND+ N S   EYIL Y K
Sbjct:       225 GEGGFVTNVMWKRKKEISNDSDNVSIQGEYILVYAK 260
                 xxxxx
```

Fig. 1. Detail of a single overlapping frameshifted ORF on the plus strand from *Helicobacter pylori*. (**a**) Thirty base pairs either side of the frameshifted region are shown for both the plus and minus DNA strands, numbered with respect to the plus strand as published by TIGR. Beneath the DNA sequences lie the amino acid sequences for the best matching database protein, swiss|P40814|T3M0_SALTY, which yielded two blastx hits in reading frames +2 and +3 in this region. Above the DNA is the translation of the query with untranslated bases in the frameshifted region shown as 'x'. (**b**) The corresponding raw blastx fragments around the frameshifted region, which is marked underneath with 'x'.

Web interface

The system interface is provided by two Web tools: a query sequence submission tool, and a results viewer for both precomputed whole genomes and user-supplied queries.

The query submission tool allows the user to: select the protein sequence databases to be searched (currently SWISS–PROT and TREMBL; Bairoch and Apweiler, 1997); select the genetic code of the query sequence; select the query sequence topology (linear, circular); specify that optional blastx filters for low sequence complexity should be used; specify fragment assembly controls (nucleation, assembly, and gap cut-offs); submit a sequence, either using cut&paste, or by local filename if in Netscape. Acceptable formats include plain, fasta and GCG. Input sequence length is currently restricted to 50 kbp, although longer sequences, including whole chromosomes, can be processed by arrangement.

Submission of a completed form initiates a job running under a new ‘project’ on the remote server and generates a project-status page for the user. Under Netscape, this page will automatically and periodically reload to report any change in job status. Alternatively, the user may reload the page manually, perhaps via a previously saved bookmark if in another session. On successful job termination, the updated project-status page will contain a live link to the results and an additional form allowing alternative trials with altered assembly parameters under the same project. Projects that are not accessed within a time limit (currently 7 days) are automatically deleted.

Results (pre-computed or from a user query) are viewed through a simple table hierarchy: (i) inferred minimal and maximal ORFs of either strand ordered by stop location along the query sequence, with intervening unassigned gap regions, counts of analogue proteins, and counts of errors; (ii) for each ORF, a list of protein analogues ordered by score, with error summaries; (iii) for each analogue, a location ordered list of component blastx fragments and error information.

Implementation

The software is written in Perl, version 5 (Wall *et al.*, 1996), and currently runs on SGI workstations although it should be portable to any POSIX compliant UNIX system with minimal effort. Extensive use is made of Perl5 class libraries modelling sequences and sequence fragments (codon, blast hit, fragment assembly, ORF, etc. (Brown, unpublished)).

The UNIX ‘make’ utility is used to coordinate blastx runs and subsequent processing steps based on an automatically generated job-specific ‘makefile’. Sequence format conversion is performed using the ‘readseq’ program (Gilbert, 1990). Byte offsets into raw blastx output are held in ‘ndbm’ (a UNIX database subroutine library) files giving random access lookup of individual blastx hits by a unique key.

ORF and frameshift assignment results are output as relational tables following the Perl RDB (Hobbs, 1993) format and are further processed using those tools to construct various views within the web viewers. Cross-referencing of protein identifiers in the web viewers to their entries in the respective protein sequence databases is through the SRS system (Etzold *et al.*, 1996). Web interfaces are built around the CGI.pm (Stein, 1996) Perl5 class library and follow the HTML Level 3 standard, with the exceptions of the Netscape file upload and meta tag refresh extensions.

Results

The method was applied to the whole *Helicobacter pylori* genome (Tomb *et al.*, 1997), strain 26695, recently released by TIGR (release Aug 6, 1997), using SWISS-PROT 34 as the search database. Two strong candidates for frameshifted ORFs are presented in Table 1.

The first example conflicts with HP0684 (+ 734056–734370) and HP0685 (+ 734285–734800). Both overlapping TIGR ORFs are annotated as ‘flagellar biosynthesis protein (fliP) {Bacillus subtilis}’. The frameshifted maximal ORF starts in an unassigned region after HP0683 just upstream of HP0684 and co-terminates with HP0685. Compared to, e.g., *Escherichia coli*, fliP, at 245 amino acids (aa), both TIGR ORFs are quite short, producing transcripts of 105 and 172 aa, respectively, while the frameshifted ORF is a better length match at 261 aa. Conserved patterns in the fliP family are present either side of the putative frameshift, but absent in one or other case from the TIGR ORFs.

In the second example, an ORF homologous to several bacterial restriction enzymes conflicts with HP1369 (+1430856–1432277, ‘type III restriction enzyme M protein (mod) {Salmonella choleraesuis}’ and HP1370 (+ 1432418–1433281, ‘type II restriction enzyme M protein (mod) {Haemophilus influenzae}’. The maximal ORF co-starts with HP1369 and co-terminates with HP1370. The approximate region of the frameshift begins near the downstream end of HP1369 allowing extension of this ORF into HP1370, which has the same functional assignment. Figure 1 details the region around the frameshift together with the blastx fragments used to assemble this ORF. Transcript lengths would be 808 aa for the maximal ORF, 406 aa for the minimal ORF, 474 aa for HP1369, and 288 for HP1370. These compare with 651 aa for the best match T3MO_SALTY. The most likely explanation seems to be a single ORF terminating at 1433281, with a start codon lying intermediate between the two extremes given by the minimal and maximal ORFs.

Discussion

The user must interpret potential errors in the query sequence with care. The method cannot infer the presence of an ORF where there is no significant similarity to any database protein sequence. ORF assignments are only for minimal and maximal ORFs as described above. The method assumes the correctness of the downstream stop, but has no way of inferring the correct start.

In particular, no information concerning neighbouring minimal ORFs, or other information such as potential promoter sites, is used to resolve conflicts or extend an ORF into unassigned regions of the query sequence. The software does not function well when introns are present, since spurious internal stops are likely to be indicated in the intronic regions, and frameshifts will be wrongly inferred around non-symmetrical introns.

Table 1.

ORF locations

			maximal ORF			minimal ORF		
best match	function	sense	lower	upper	aa	lower	upper	aa
P33133|FLIP_ECOLI	Flagellar biosynthetic protein fliP	+	734020	734800	260	734242	734800	186
P40814|T3MO_SALTY	Type III restriction-modification system stylti enzyme MOD	+	1430856	1433281	808	1432062	1433281	406

ORF properties

best match	score	frameshift	support	affects
P33133|FLIP_ECOLI	465.15	olp(+1, +2, 734311, 734320)	6	HP0684, HP0685
P40814|T3MO_SALTY	343.54	olp(+3, +2, 1432260, 1432276)	3	HP1369, HP1370

Candidate frameshifted ORFs in the *Helicobacter pylori* genome, strain 26695 (TIGR release Aug 6, 1997) from screen against SWISS-PROT 34. Both maximal and minimal ORFs are shown, representing the largest and smallest encompassing region of chromosome containing the assembled blastx fragments, bounded by suitable start and stop codons. Notes: (best match) most similar database protein, by blastx score; (function) SWISS-PROT description of best match; (sense) located on plus or minus DNA strand; (lower, upper) boundaries of ORF in plus strand sense counting from 1 and excluding the stop codon; (aa) length of likely transcript as number of amino acid residues; (score) is a cumulative value computed from fragment blastx scores with averaging over overlapped regions; (frameshift) gives type of frameshift region 'seq'–sequential, or 'olp'–overlapping, and frame changes and lower and upper boundaries along plus strand; (support) number of matching database proteins supporting decision; (affects) TIGR identifiers of ORFs affected.

Certain characteristics of blastx and of the reference protein databases must be borne in mind. Blastx occasionally reports hits that are translated past a valid stop and this will cause an incorrect stop to be selected downstream of the actual one, as well as an apparent internal stop error. A short double frameshift hidden within a single reported blastx hit will never be detected by this approach. Further, frameshifts lying central to an ORF are more likely to be found than peripheral ones, because of the need for two blastx fragments bridging the error. An apparent frameshift might be due to an incorrect database sequence (Bork and Bairoch, 1996): here, the corroboratory effect of multiple database hits reporting the same or similar errors in the query can be of help.

There are some biological sources of misreported errors. Distinct neighbouring genes in the query can match a gene fusion product in the database, suggesting an apparent single ORF with an internal stop. Inferences of ORF locations and sequencing errors could also be perturbed in cases where ribosomal frameshifting or post-transcriptional modification occur, or where selenocysteine (stop encoded) codons are present.

Nevertheless, the method requires no training (e.g. neural network recognition of coding regions) on test sequences and builds upon tools already familiar to the community (blastx, WWW). Suitable applications include error checking of prokaryotic or other intron-free genomic sequence and cDNAs (complementary DNAs) with the prerequisite that protein homologues (irrespective of whether their function is known or not) exist in the public databases—an increasingly likely prospect.

The URL http://www.sander.ebi.ac.uk/frame/ gives access to the entry level page offering the user a user query submission form and system documentation.

Acknowledgements

This software was developed at the European Molecular Biology Laboratory, Heidelberg, during 1995–1996 with a Deutsche Forschungsgemeinschaft grant administered via the Max Delbrück Center, Berlin. It was further developed and is presently maintained at the European Bioinformatics Institute, Cambridge.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Apweiler,R. (1997) The SWISS–PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.*, **25**, 31–36.

Beck,S. (1993) Accuracy of DNA sequencing: should the sequence quality be monitored? *J. DNA Sequenc. Mapp.*, **4**, 215–217.

Birney,E., Thompson,J.D. and Gibson,T.J. (1996) PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.*, **24**, 2730–2739.

Bork,P. and Bairoch,A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet.*, **12**, 425–427.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Gilbert,D.G. (1990) *ReadSeq (Version 1Feb93).* Biology Department, Indiana University, Bloomington, IN 47405, USA. ftp://ftp.bio.indiana.edu/molbio/readseq/readseq.shar

Gish,W. and States,D.J. (1993) Identification of protein coding regions by database similarity search. *Nature Genet.*, **3**, 266–272.

Guan,X. and Uberbacher,E.C. (1996) Alignments of DNA and protein sequences containing frameshift errors. *Comput. Applic. Biosci.*, **12**, 31–40.

Hobbs,W.V. (1993) *RDB: A Relational Database Management System (Version 2.5k).* RAND Corp., USA. ftp://unix.hensa.ac.uk/mirrors/perl-CPAN/modules/dbperl/scripts/rdb/

Huang,X. and Zhang,J. (1996) Methods for comparing a DNA sequence with a protein sequence. *Comput. Applic. Biosci.*, **12**, 497–506.

Posfai,J. and Roberts,R.J. (1992) Finding errors in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 4698–4702.

States,D.J. (1992) Molecular sequence accuracy: analysing imperfect data. *Trends Genet.*, **8**, 52–55.

Stein,L.D. (1996) *CGI—Simple Common Gateway Interface class (Version 2.23).* Whitehead Institute MIT Genome Center, Cambridge, MA 02142, USA. http://www-genome.wi.mit.edu/ftp/pub/software/WWW/cgi-docs.html

Tomb,J.-F., *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.

Wall,L., Christiansen,T. and Schwartz,R.L. (1996) *Programming Perl*, 2nd edn. Nutshell Handbooks, O’Reilly & Associates, Inc.