Short Technical Report

JavaScript DNA Translator: DNA-Aligned Protein Translations

BioTechniques 33:1318-1320 (December 2002)

William L. Perry III
Lilly Research Laboratories,
Indianapolis, IN, USA

ABSTRACT

There are many instances in molecular biology when it is necessary to identify ORFs in a DNA sequence. While programs exist for displaying protein translations in multiple ORFs in alignment with a DNA sequence, they are often expensive, exist as add-ons to software that must be purchased, or are only compatible with a particular operating system. JavaScript Translator is a shareware application written in JavaScript, a scripting language interpreted by the Netscape® Communicator and Internet Explorer Web browsers, which makes it compatible with several different operating systems. While the program uses a familiar Web page interface, it requires no connection to the Internet since calculations are performed on the user's own computer. The program analyzes one or multiple DNA sequences and generates translations in up to six reading frames aligned to a DNA sequence, in addition to displaying translations as separate sequences in FASTA format. ORFs within a reading frame can also be displayed as separate sequences. Flexible formatting options are provided, including the ability to hide ORFs below a minimum size specified by the user. The program is available free of charge at the BioTechniques Software Library (www.BioTechniques.com).

INTRODUCTION

Today, it is relatively easy to search and retrieve DNA sequences from public databases over the Internet using a Web browser such as Microsoft® Internet Explorer or Netscape® Communicator. Sequence analysis is also possible through Web page interfaces to other computers or through the use of JavaScript programs embedded in Web pages that run on the user's own computer. While the latter appears to the user as a typical Web page, no connection to the Internet is necessary, and proprietary sequences are not compromised. Since JavaScript programs are interpreted by a Web browser and not by an operating system, one program can be compatible with several computer platforms including PC-compatibles, Macintosh® computers, and computers that run UNIX®.

Molecular biologists often need to identify ORFs in DNA sequences as a starting point for cloning experiments, *in vitro* mutagenesis, or identifying genes encoded by genomic DNA. The potential to use a Web browser to both retrieve DNA sequences from public or private databases and to analyze these sequences is attractive. Several tools exist for generating protein translations aligned to a DNA sequence using a Web browser interface including: (a) the ExPASy translation tool (<http://www.expasy.ch/tools/dna.html>); (b) the Molecular Tool Kit (<http://arbl.cvmb.colostate.edu/molkit/translate>); (c) the BCM Search Launcher (<http://searchlauncher.bcm.tmc.edu/>); (d) the DNA Sequence Translator (<http://biocommons.bcc.washington.edu/services/psoftware/dnatranslator/index.html>); (e) the DNA to Protein Translator (<http://bio.lundberg.gu.se/edu/translate.html>); (f) the Translate applet of the Sequence Ma-

nipulation Suite (<http://www.ualberta.ca/~stothard/javascript/>) (4); (g) the Map program of Genetics Computer Group software using the SeqWeb interface (subscription software) (1); and (h) the Translation tool of Incyte Genomics interface to LifeSeq® Gold (subscription software; Incyte Genomics, Palo Alto, CA, USA). However, these programs have drawbacks. Some produce alignments of a DNA sequence with a protein translation in only one reading frame at a time (a, b, d, and f) or with only the three forward or three reverse frames at a time (c). A few programs produce FASTA formatted protein sequences in addition to the alignments to DNA (a, c, e, and h), but only one identifies ORFs of a user-selected size within these translations (h). Only one program lets the user vary the number of DNA base pairs to display per line (b), and none of the programs let the user choose to display only translations of methionine-to-stop-codon ORFs of a certain size in alignments with a DNA sequence. Furthermore, only one of these programs performs translations on the user's machine (f); the rest send sequences to another computer for processing, potentially compromising novel or proprietary sequences. Only one program can process more than one different sequence at a time (h), and two programs are only available as a part of a software package and institutional licensing (g–h), limiting their use by the scientific community. The limitations of these existing tools prompted the writing of JavaScript DNA Translator.

This JavaScript application generates protein translations of one or multiple DNA sequences in up to six ORFs, both aligned to the DNA sequence and as

MATERIALS AND METHODS

program was subsequently tested on Internet Explorer 5.1.4 for the Macintosh, Netscape Communicator 6.2.2 for the Macintosh, Netscape Communicator 6.2.3 for Windows and Netscape Communicator 4.75 for UNIX running on a Sun[®] workstation.

RESULTS

JavaScript DNA Translator can be obtained at the BioTechniques Software Library (www.Biotechniques.com) and can be saved directly to the user's machine. To begin, the program must be loaded into a JavaScript-compatible Web browser. For Internet Explorer or Netscape Communicator, the user can open the file like a Web page from the program menu or by dragging the file from its current folder into an open browser window.

Input Sequence Format

The sequence to be translated can be typed or pasted into the second window of the sequence input page (Figure 1). Sequences may be either a DNA-only

sequence (without annotation) or one or more FASTA format sequences. If FASTA format sequences are detected then text preceding the first ">" character will be ignored. Sequence titles of FASTA files will be taken from the first line of the sequence and may be up to 300 characters. DNA-only sequences may be annotated using the "Sequence Title" input box and may contain line breaks. Sequences to be translated may contain specific (A, G, C, or T) or degenerate codes (R, Y, S, M, K, H, B, V, D, or N) for nucleic acids, as recommended by the Nomenclature Committee of the International Union of Biochemistry (3), as either lowercase or uppercase letters. Other letters will appear as lowercase in the DNA sequence output and will result in codons being displayed as "???" or "X" characters in the three- and one-letter translations, respectively. Additional punctuation, line feeds, carriage returns, or numbers will be removed from sequences upon reformatting, and the sequence to be translated will be changed to uppercase except as noted above. This allows sequences in GenBank® format (that contain spaces and numbers) to be used as input if only the

sequence region of the file is copied to the program's sequence window. Single sequences of either type may also be reverse complemented by clicking on this button in the input window (Figure 1). The title of a FASTA sequence will be moved to the sequence title window, and the input sequence is replaced with its reverse complement.

The length of the input sequence that can be analyzed at one time is limited by the size of the input buffer and the speed of one's computer. Netscape Communicator 4.7 for Windows truncates sequences pasted into the sequence input window to about 25 kb, although this is not a problem for Internet Explorer 5.5 and above or Netscape Communicator 6.2.3 for Windows. Sequences of 105, 80, 40, and 20 kb were translated in 17.5 min, 9.5 min, 2 min 15 s, and 47 s, respectively, on a 733 MHz PC-compatible computer running Windows 2000. Processing 50 cDNA sequences with an average size of 1 kb on the same machine took 2 min and 20 s. Processing was fastest when other applications and browser windows were closed.

Display Options for Translations Aligned with a DNA Sequence

Translations in up to six reading frames (three forward and three reverse) are displayed aligned to the DNA sequence in a new browser window (Figure 1). Methionines and stop codons are color-coded to make ORFs easier to analyze. From the sequence input page, the user can choose between one- or three-letter abbreviations for amino acids, which reading frames to translate, and the number of bases to translate per line from drop-down selection boxes (Figure 1). While numbered dsDNA sequences are generated by default, ssDNA sequences and/or unnumbered sequences can be selected by unchecking the appropriate boxes on the sequence input page. The user may also select to display only ORFs beginning with methionine codons of a minimum length (between 5 and 1000 amino acids) in the selected reading frames. This is particularly helpful for analyzing cDNA sequences in which one large ORF is expected. However, the program cannot distinguish between the actual translation initiation

site and upstream in-frame ATGs. The longest ORF will be displayed.

Options for Displaying Sequence Translations in FASTA Format

Following the alignment of the DNA sequence with protein translations in the selected reading frame(s), the user also has the option of displaying ORFs in FASTA format (Figure 1, left panel). One checkbox on the sequence input page allows the user to display the complete translation of the selected ORFs. Stop codons are displayed as red asterisks, methionines are displayed as blue "M"s, and unknown amino acids (caused by inclusion of unknown "N" bases or degenerate bases in the input sequence) are displayed as "X"s. The complete translations are also highlighted to distinguish them from partial-frame translations. The user may also choose to display ORFs within the reading frame of a minimum selectable size (between 5 and 1000 amino acids). The ORFs can be limited further to those that start with methionines by checking the appropriate box on the sequence input page (Figure 1). The translated frame and the ORF number are appended to the end of the sequence name. If multiple FASTA files are given as input then the translation results appear in the same translation window separated by horizontal lines.

DISCUSSION

JavaScript DNA Translator is an easy-to-use application that should be widely compatible with different computer platforms and Web browsers. Output from the program can be printed directly, saved as an HTML file for later use, or copied as "formatted text" to another application such as a word processor. The ability to generate alignments to sequences of different line lengths makes printing translation results in the landscape orientation or in a smaller font size to conserve space practical. This program has been used to translate genomic sequences that were subsequently edited to only retain translations over the exons of relevant genes. The consequences of exon skipping or alternate splice site usage on the protein coding

potential of mRNA could then be predicted. This program has also been very useful in generating translations of cDNA sequences in which a single ORF is displayed above the DNA sequence using the "MET-to-Stop" and minimum ORF length options. Figures in this format are often used for publication. To the best of my knowledge, the ability to translate multiple FASTA format sequences at once and to display ORFs of a selected size (both aligned to DNA and as separate sequences) are unique features of the JavaScript DNA Translator among free software available to the scientific community. Although it was possible to translate sequences in excess of 100 kb using this program, processing speed and memory limitations make the program more practical for analyzing sequences of 50 kb or less.

JavaScript DNA Translator should make it a little easier for molecular biologists to perform sequence analysis in the same applications used for database searching and sequence retrieval.

ACKNOWLEDGMENTS

The author would like to thank John Calley and Qingqin Li for thoughtful reading of this manuscript and testing of the program in different operating systems.

REFERENCES

1. **Dolz, R.** 1994. GCG: displaying restriction sites and possible translations in a DNA sequence. *Methods Mol. Biol.* 24:47-55.
2. **Holzner, S.** 1998. JavaScript Complete, Appendix A, p. 505-520. McGraw-Hill, New York.
3. **Nomenclature Committee of the International Union of Biochemistry (NC-IUB).** 1886. Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *J. Biol. Chem.* 261:13-17.
4. **Stothard, P.** 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* 28:1102-1104.

Received 31 May 2002; accepted 25 July 2002.

Address correspondence to:

Dr. William L. Perry III
Lilly Research Laboratories
Lilly Corporate Center
Indianapolis, IN 46285, USA
e-mail: bperry@lilly.com