Virtual Ribosome - a comprehensive DNA translation tool with support for integration of sequence feature annotation

Wernersson, Rasmus

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gkl252

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W385–W388***
doi:10.1093/nar/gkl252

Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation

Rasmus Wernersson*

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received February 14, 2006; Revised March 1, 2006; Accepted March 20, 2006

ABSTRACT

Virtual Ribosome is a DNA translation tool with two areas of focus. (i) Providing a strong translation tool in its own right, with an integrated ORF finder, full support for the IUPAC degenerate DNA alphabet and all translation tables defined by the NCBI taxonomy group, including the use of alternative start codons. (ii) Integration of sequences feature annotation—in particular, native support for working with files containing intron/exon structure annotation. The software is available for both download and online use at http://www.cbs.dtu.dk/services/VirtualRibosome/.

INTRODUCTION

A large number of software packages for translating DNA sequences already exist, as services on the World Wide Web [e.g. the Expasy Translate Tool (http://www.expasy.ch/tools/dna.html)], as command-line tools [e.g. the GCG package (1)] and as user-friendly graphical applications [e.g. DNA strider (a personal favorite) (2) and ApE (http://www.biology.utah.edu/jorgensen/wayned/ape/)]. However, many of these fine tools do not support translating sequences containing degenerate nucleotides, have no or limited support for alternative translation tables (including alternative initiation codons) and in general have problems handling special case situations. The software described here aims at addressing these issues and providing a comprehensive solution for translation. The software is build on the experience gained from writing and maintaining the RevTrans server (3).

Another part of the rationale for creating Virtual Ribosome is to create an easy and consistent way to map the underlying intron/exon structure of a gene onto its protein product. This makes it easy to build datasets that can be used for analyzing how the underlying exon structure is reflected in the protein [e.g. how exon modules maps onto the 3D structure of the protein, see the FeatureMap3D server (4) elsewhere in this issue].

SOFTWARE FEATURES

Support for the degenerate nucleotide alphabet

The software has full support for the IUPAC alphabet (Table 1) for degenerate nucleotides. For example, the codon **TCN** correctly translates to **S** (serine) and not **X** (unknown) as often seen in other translators.

Support for a wide range of translation tables

Full support for all translation tables defined by the NCBI taxonomy group (5) (see the list below). The command-line version of the software also has support for

Table 1. IUPAC alphabet of degenerate nucleotides

Letter	Description	Bases represented
A	Adenine	A
T	Thymine	T
G	Guanine	G
C	Cytosine	C
Y	pYrimidine	C T
R	puRine	A G
S	Strong	G C
W	Weak	A T
K	Keto	T G
M	aMino	A C
B	Not A	C G T
D	Not C	A G T
H	Not G	A C T
V	Not T/U	A C G
N	aNy	A C G T

*Tel: +45 45252485; Fax: + 45 45931585; Email: raz@cbs.dtu.dk

```
AAs    = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M------------MMMM---------------M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Figure 1.

reading an arbitrary translation table defined by the user.

```
 [1] Standard Genetic Code
 [2] Vertebrate (Mitochondrial)
 [3] Yeast (Mitochondrial)
 [4] Mold, Protozoan, Coelenterate (Mitochondrial)
     and Mycoplasma/Spiroplasma
 [5] Invertebrate Mitochondrial
 [6] Ciliate, Dasycladacean and Hexamita (Nuclear)
 [9] Echinoderm and Flatworm (Mitochondrial)
[10] Euplotid (Nuclear)
[11] Bacterial and Plant Plastid
[12] Alternative Yeast (Nuclear)
[13] Ascidian Mitochondrial
[14] Alternative Flatworm (Mitochondrial)
[15] Blepharisma (Nuclear)
[16] Chlorophycean (Mitochondrial)
[21] Trematode (Mitochondrial)
[22] Scenedesmus obliquus (Mitochondrial)
[23] Thraustochytrium (Mitochondrial)
```

Start and Stop codons

Virtual Ribosome also uses the table of alternative translation initiation codons de.ned in the translation tables mentioned above. Figure 1 is the definition for translation table 11 (the Bacterial and Plant plastid code).

In this case, the codons **TTG**, **CTG**, **ATT**, **ATC**, **ATA**, **ATG** and **GTG** are all allowed as a start codon, and all of them will translate to methionine if used as a start codon. [For a recent report on the use of **GTG** as a methione coding start-codon, please see (6)]. The use of alternative methionine codons at the first position can be disabled using the 'all internal' option (useful for working with sequence fragments).

In addition, the software has support for either terminating the translation at the first encountered Stop codon, or reading through the entire sequence annotating stop codons with '*'.

Reading frames and ORF finder

The reading frame used for translation can be selected by the user, as a single reading frame (1, 2, 3, −1, −2, −3) or as a set of reading frames (all, positive, negative). Following translation the protein sequences are available for download, and a visualization, in which all possible Start and Stop codons are highlighted, is presented to the user. The example below shows how the result is visualized if a single reading frame has been selected.

```
   M   V   L   S   A   A   D   K   G   N   V   K   A   A   W   G   K   V   G   *
5' ATGGTGCTGTCTGCCGCCGACAAGGGCAATGTCAAGGCCGCCTGGGGCAAGGTTGGCTAA 60
   >>>...))).....................................................***
```

The 'strict' Start codons (always coding for methionine) are annotated with '>>>', the 'alternative' Start codons (only coding for methionine at the start position) are annotated with ')))' and Stop codons are annotated with '***'.

If multiple reading frames are selected the results are stacked as shown in the example below. Notice how the Start codon 'arrows' are reversed on the minus strand to indicate the direction of translation.

```
     G   A   V   C   R   R   Q   G   G   Q   C   Q   G   R   L   G   Q   G   W   L
   W   C   C   L   P   P   T   R   A   M   S   R   P   P   G   A   R   L   A
   M   V   L   S   A   A   D   K   G   N   V   K   A   A   W   G   K   V   G   *
5' ATGGTGCTGTCTGCCGCCGACAAGGGCAATGTCAAGGCCGCCTGGGGCAAGGTTGGCTAA 60
   >>>...))).)))..............>>>..........)))........))).*.***
   .....................(((...(((..***(............(((......***.
3' TACCACGACAGACGGCGGCTGTTCCCGTTACAGTTCCGGCGGACCCCGTTCCAACCGATT 60
   H   H   Q   R   G   G   V   L   A   I   D   L   G   G   P   A   L   N   A   L
     T   S   D   A   A   S   L   P   L   T   L   A   A   Q   P   L   T   P   *
     P   A   T   Q   R   R   C   P   C   H   *   P   R   R   P   C   P   Q   S
```

Virtual Ribosome has the option of working as an ORF (open reading frame) finder. When this option is used all specified reading frames are scanned for ORFs and the longest ORF is reported. The rules for defining an ORF can be adjusted to (i) only open an ORF at 'strict' Start codons, (ii) open an ORF at any Start codon and (iii) open an ORF at any codon except Stop (useful for working with small DNA fragments). The position of the ORF within the DNA sequences is visualized as shown in the following example.

```
>Seq1_rframe3_ORF
Reading frame: 3
             M   F   L   S   F   P   T   T   K   T   Y   F   P   H
5' CCCTGGAGAGGATGTTCCTGAGCTTCCCCACCACCAAGACCTACTTCCCCCACTAAGAGG 60
   ..)))......>>>...))).................................***....
```

Intron/exon annotation

Besides working on the standard FASTA format files (sequence only), Virtual Ribosome natively understands the TAB file format for containing both sequence and sequence feature annotation described in (7). Briefly, each line in the TAB format file describes one sequence (DNA or peptide) in four fields, separated by tabs: Name, Sequence, Annotation and Comment. The Annotation field is a string of exactly the same length as the Sequence field. Each position in the annotation string describes the nature of the corresponding position in the sequence string using a single-letter code. TAB files containing intron/exon structure can easily be generated by the FeatureExtract server (7), or by submitting a GenBank file directly to Virtual Ribosome. If a GenBank file is submitted, CDS sections (including information about intron/exon structure) are extracted to the TAB format before translation, by running the FeatureExtract software in the background with default parameters.

If a GenBank or TAB file is supplied as input, only the exonic parts of DNA sequences is used for the translation. Furthermore, the underlying exon structure will be rejected in the translated sequence (also in the TAB format). By default, each amino acid will be annotated with a number indicating the exon that encoded this particular amino acid (see example below).

```
Name: 'J00044, Goat adult alpha-ii-globin gene'
MVLSAADKSNVKAAWGKVGSNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
11111111111111111111111111111112222222222222222222222222222 60
EKVAAALTKAVGHLDDLPGTLSDLSDLHAHKLRVDPVNFKLLSHSLLVTLACHHPSDFTP 120
22222222222222222222222222222222222222222333333333333333333 120
AVHASLDKFLANVSTVLTSKYR* 143
33333333333333333333 143
```

Alternatively, the positions and the phase of the introns can be indicated.

- Phase 0: an intron exists right before the codon encoding the amino acid.
- Phase 1: an intron exists in between positions 1 and 2 of the codon.
- Phase 2: an intron exists in between positions 2 and 3 of the codon.

The following example illustrates the principle.

```
Name: 'J00044, Goat adult alpha-ii-globin gene'
MVLSAADKSNVKAAWGKVGSNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
.............................2............................ 60
EKVAAALTKAVGHLDDLPGTLSDLSDLHAHKLRVDPVNFKLLSHSLLVTLACHHPSDFTP 120
.........................................0................... 120
AVHASLDKFLANVSTVLTSKYR* 143
...................... 143
```

Easy to use interface

The interface to the Virtual Ribosome server has been designed to be intuitive and easy to use. Figure 2 shows the basic part of the interface. Notice that it is possible to submit a

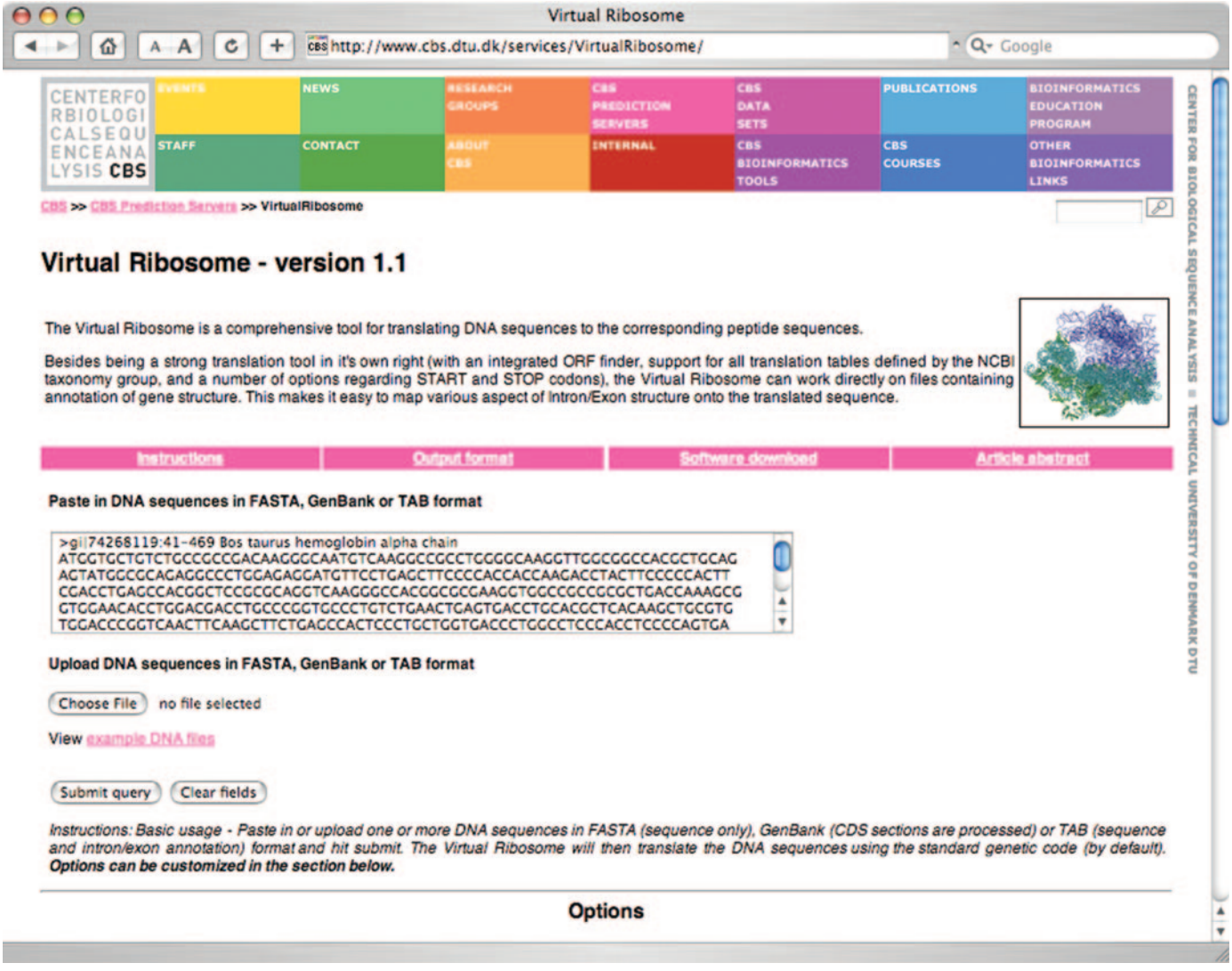

Figure 2. Screenshot of the basic part of the Virtual Ribosome interface.

sequence for translation using the default parameters, without having to scroll through a page of obscure options. The options are grouped into logical sections further down the web page. For each option a short explanation is provided together with a link to a detailed description.

ACKNOWLEDGEMENTS

Special thanks to Ulrik de Licthenberg for comments on the content and layout of the manuscript, and to Henrik Nielsen for inspiration for implementing the 'Intron phase' functionality. This work is supported by a grant from The Danish National Research Foundation and The Danish Research Agency. Funding to pay the Open Access publication charges for this article was provided by The Danish Research Agency.

Conflict of interest statement. None declared.

REFERENCES

1. Dölz,R. (1994) GCG: translation of DNA sequence. *Methods Mol. Biol*., **24**, 129–142.
2. Douglas,S.E. (1994) DNA Strider. A Macintosh program for handling protein and nucleic acid sequences. *Methods Mol. Biol*., **24**, 181–194.
3. Wernersson,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res*., **31**, 3537–3539.
4. Wernersson,R., Rapacki,K., Stærfeldt,H.-H., Sackett,P.W. and Mølgaard,A. (2006) FeatureMap3D: a tool to map protein features and sequence conservation onto homologous structures in the PDB. *Nucleic Acids Res*., **34**, W536–W540.
5. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **28**, 10–14.
6. Abramczyk,D., Tchorzewski,M. and Grankowski,N. (2003) Non-AUG translation initiation of mRNA encoding acidic ribosomal P2A protein in *Candida albicans*. *Yeast*, **20**, 1045–1052.
7. Wernersson,R. (2005) FeatureExtract—extraction of sequence annotation made easy. *Nucleic Acids Res*., **33**, W567–W569.