Nucleic Acids Research, 2002, Vol. 30 No. 15 3449–3453

BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences

Dean Laslett, Bjorn Canback[1,*] and Siv Andersson[1]

Murdoch University, Perth, Western Australia, Australia and [1]Department of Molecular Evolution, University of Uppsala, Norbyvägen 18C, SE-752 36, Uppsala, Sweden

Received March 18, 2002; Revised and Accepted June 11, 2002

ABSTRACT

A computer program, BRUCE, was developed for the identification of transfer-messenger RNA (tmRNA) genes. The program employs heuristic algorithms to search for a tRNA^{Ala}-like secondary structure surrounding a short sequence encoding the tag peptide. In the 57 completely sequenced bacterial genomes where tmRNA genes have been reported previously, BRUCE identified all with no false positives. In addition, BRUCE found 99 of the 100 tmRNAs identified previously in other bacteria, red chloroplasts and cyanelles. The output of the program reports the proposed tRNA secondary structure, the tmRNA gene sequence and the tag peptide.

INTRODUCTION

Transfer-messenger RNAs (tmRNAs) are named for their dual tRNA-like and mRNA-like functions. The role of tmRNAs is to liberate the mRNA from stalled ribosomes (1). This is accomplished by using part of the tmRNA as a reading frame that ends with one or more translation termination signals. The tmRNA codes for a hydrophobic peptide, the proteolysis tag, which is attached to the C-terminus of the incomplete protein enabling recognition by a protease (2). The canonical tmRNA secondary structure consists of a tRNA-like domain at the 5′ and 3′ ends surrounding an internal region consisting of stem–loops and pseudo-knots (Fig. 1). The tRNA domain contains alanyl-tRNA synthetase recognition signals, a TψC-stem (T-stem) and TψC-loop (T-loop), a shortened variable loop (V-loop), and an extended anticodon stem, but the canonical dihydro-uridine stem (D-stem) and dihydro-uridine loop (D-loop) are replaced by a single loop that does not base pair with itself. In some tmRNA genes, for instance in *Bacillus subtilis* (3), the 3′ CCA tail is not encoded within the gene, but is only present on the mature tmRNA molecule, possibly added by a nucleotidyltransferase.

According to the current model (4), the 3′ end is first charged with an alanyl moiety that attaches to the incomplete protein chain before translation recommences. The internal reading frame is terminated with at least one stop codon, which is usually preceded by two or more non-polar amino acid codons (5). Identification of where the reading frame begins, called the resume codon, is more speculative, but a consensus sequence of WATARNYGCNAANNANNA [W: A or T; R: A or G (purines); N: A, C, G or T (any nucleotide); Y: C or T (pyrimidines)] has been noted around the resume codon in most tmRNA genes (5). Recently, tmRNAs have been identified that are encoded in two parts (6), probably as a result of a translocation event. In some of these permuted genes, the T-loop consensus motif subset GTTC has diverged toward GGGC, resulting in an overall consensus motif of GKKC (K: G or T) for all currently identified tmRNA genes.

tmRNA genes have now been identified in all sequenced eubacterial genomes (5). tmRNA genes have also been identified in 'red' chloroplasts (from red algae, colourless algae or diatoms) and cyanelles. No tmRNAs have been identified in archaea, 'green' chloroplasts (from green algae or higher plants) or the nuclear genomes of eukaryotes (5). A tmRNA-like sequence has been identified in the genome of *Reclinomonas americana* mitochondrion, although it is lacking a peptide tag reading frame (6). The presence of proteolysis tags in *Escherichia coli* and *Thermus thermophilus* have been confirmed experimentally (7,8). Recently, a group I intron has been identified in the T-loop of the tmRNA gene for *Clostridium botulinum* (5).

To date, *in silico* searches for tmRNA genes have been carried out manually using the PATSCAN pattern-searching program (9) to identify tRNA^{Ala}-like patterns in a sequence, followed by BLAST searches (10) to establish sequence homology with other tmRNA genes. While effective, a limitation of this approach is that additional manual steps are required to predict the amino acid sequence of the proteolysis tag. The purpose of this study was to develop an algorithm to search *in silico* for tmRNA genes and predict the tag amino acid sequence. The resulting program, BRUCE, was found to successfully detect these genes and predict the tag sequence.

MATERIALS AND METHODS

tmRNA sequences

The most comprehensive source of tmRNA information is the tmRNA website found at http://www.indiana.edu/~tmrna which is curated by Kelly Williams (5). The website is frequently updated and contains information about tmRNA sequences, tag peptides, alignments as well as careful

*To whom correspondence should be addressed. Tel: +46 18 471 4203; Fax: +46 18 471 6404; Email: bjorn.canback@ebc.uu.se

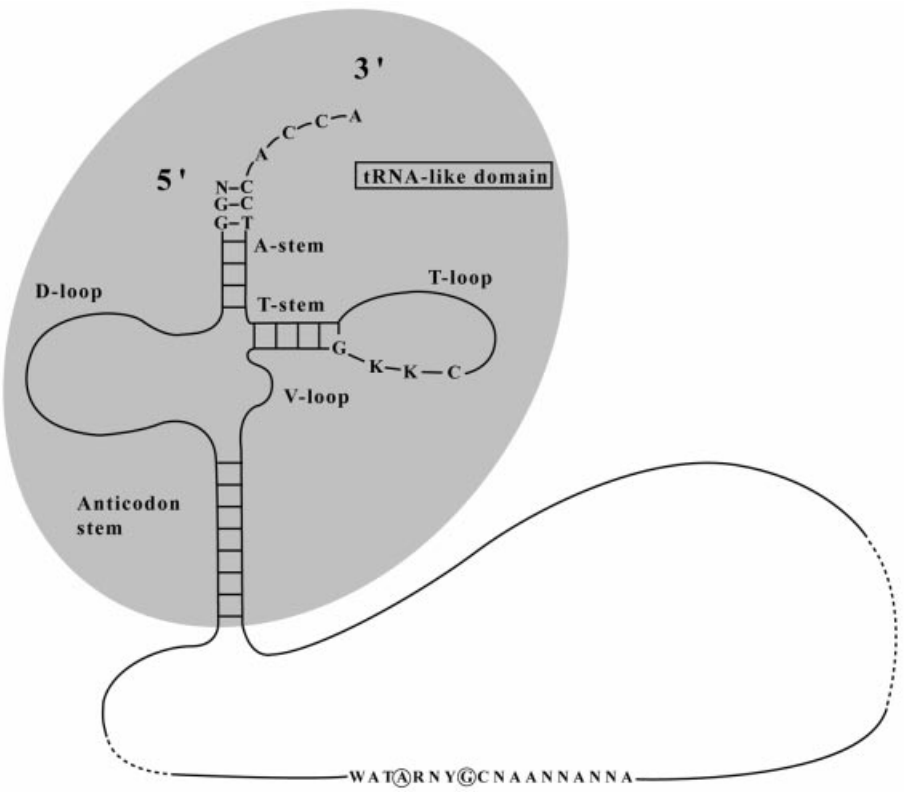

Figure 1. Secondary structure of a typical tmRNA. Lower sequence corresponds to the consensus motif. Dashed lines represent parts of the figure which are not in scale. Adapted from Muto *et al*. (2) and Williams (5).

annotations. Another website, the tmRDB (tmRNA database), found at http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html (11), includes a smaller set of tmRNAs and focuses on secondary structures and three-dimensional models. Annotated tmRNA sequences are also found in the public sequence databases. However, these include only a subset of all the available sequences.

Search algorithm

As all known tmRNA genes contain a tRNA-like domain, the search algorithm assumes that any tmRNA gene within the target sequence will contain a standard T-stem and T-loop secondary structure within its tRNA-like domain (Fig. 1). Hence, the algorithm first searches the target sequence for the T-loop consensus motif subset GKKC. Around each motif, the algorithm attempts to construct a T-loop from 5 to 9 bases long and T-stem from 4 to 5 bp long, with preference for a 5 bp T-stem and a 7 base T-loop. For each candidate T-loop, the sequence from 216 to 427 bases upstream, and from 25 to 384 bases downstream is searched for possible 5′ amino-acyl acceptor arm (A-stem) sequences that can base pair with the putative 3′ A-stem sequence downstream of the T-loop. If a 5′ A-stem is found downstream of the T-stem, this is taken as an indication of a permuted tmRNA gene.

For non-permuted candidates, the sequence is searched from 166 to 322 bases upstream of the T-loop for a resume consensus motif. For permuted candidates, the sequence is searched from 61 to 466 bases downstream of the T-loop for a resume consensus motif. Any box matching 8 or more bases, without gaps, from the consensus motif ATARNYGC-NVVNNMNNA (V: A, C or G; M: A or C) is considered to be a candidate. Extra weighting is assigned to the G and the second A, as they are conserved in most tmRNA sequences in the tmRNA database (5). The sequence is then searched downstream 12–102 bases from the motif for a reading frame terminated by two non-polar amino acid codons and a stop codon. Finally, extra preference is given to a candidate if the 5′ end begins with NGG and if the nucleotides TCCA are present in the 3′ end.

For each tmRNA candidate, the secondary structure of the tRNA-like domain, the position of the resume consensus motif, and the amino acid sequence of the proteolysis tag is predicted. The algorithm is capable of detecting non-permuted tmRNA genes from 235 to 451 bases long, and permuted tmRNA genes from 97 to 457 bases long. However, BRUCE does not predict the secondary structure of the mRNA-like domain, so the precise beginning and end of a permuted tmRNA gene cannot be located. This is because the gene has been cut into two pieces within the mRNA-like domain and swapped. Instead, the gene is assumed to begin 54 bases upstream of the 3′ anticodon stem, and end 145 bases downstream of the tag-reading frame. To enable detection across the origin of counting for circular genomes, the search is wrapped around the beginning and end of the sequence to the maximum allowed length of a tmRNA gene.

Availability of software

BRUCE is written in C. The source code can be downloaded from the website http://artedi.ebc.uu.se/Bjorn/Tmrna/index.html. The website also contains a user interface to the

```
-------------------------
BRUCE v1.0   Dean Laslett
-------------------------

Searching for tmRNA genes
Using standard genetic code for peptide tag prediction
Assuming circular topology, search wraps around ends
Searching both strands


gi|6626251|gb|U00096.1|U00096 Escherichia coli K-12 MG1655 complete genome
4639221 nucleotides in sequence
Mean G+C content = 50.8%

1.

                ca
               c
              a
            g-c
            g-c
            g+t
            g-c
            c-g
            t-a
            g-c     tc
        ctta   cgccc  a
      gt       !!!!!  a
     g         gcggg  c
     a         c    tt
      tt      a
        cga   g
           c-g
           g-c
           g+t
           g+t
           a-t
           t-a
           t-a
           t+g



     tmRNA (tRNA domain)
     67 bases, %GC = 58.2
     Sequence [2753614,2753976]


tmRNA Sequence in gi|6626251|gb|U00096.1|U00096 Escherichia coli K-12 MG1655
complete genome

1   .   10    .   20    .   30    .   40    .   50
ggggctgattctggattcgacgggatttGCGAAACCCAAGGTGCATGCCG
AGGGGCGGTTGGCCTCGTAAAAAGCCGCAAAAAATAGTCgcaaacgacga
aaactacgctttagcagcttaataaCCTGCTTAGAGCCCTCTCTCCCTAG
CCTCCGCTCTTAGGACGGGGATCAAGAGAGGTCAAACCCAAAAGAGATCG
CGTGGAAGCCCTGCCTGGGGTTGAAGCGTTAAAACTTAATCAGGCTAGTT
TGTTAGTGGCGTGTCCGTCCGCAGCTGGCAAGCGAATGTAAAGACTGACT
AAGCATGTAGTACCGAGGATGTAGgaatttcggacgcgggttcaactccc
gccagctccacca
```

Figure 2. Output of the computer program BRUCE from a search on the complete genome sequence of *Escherichia coli* K12.

program allowing the user to upload a sequence and run the program on a server. BRUCE accepts as input a file with one or more nucleotide sequences in FASTA format. By default BRUCE assumes that each sequence has a circular topology (search wraps around ends), that both strands should be searched, that progress of the search is not reported, and that the tRNA domain secondary structure should be output (Fig. 2). These settings can be individually changed to linear topology (no wrapping), search the sense strand only, report search progress and no output of tRNA secondary structure. An abbreviated output format is also available. For each sequence in the input file, only the sequence name and tab delimited information about each gene in the sequence is given. It should be noted that when searching through files consisting of one or more short sequences containing single tmRNA genes, BRUCE will report two tmRNA genes for each sequence; one non-permuted, and one permuted, unless linear topology is specified.

Test sequences

The speed, sensitivity and selectivity of BRUCE were tested against three different sequence sets. The first set comprised

all completely sequenced prokaryotic genomes and eukaryotic chromosomes. These were downloaded from NCBI (http://www.ncbi.nlm.nih.gov). Of the prokaryotic genomes, 57 were from bacteria and 12 from archaea. The second set contained 100 putative full-length tmRNA genes from other genomes, downloaded from the tmRNA website (http://www.indiana.edu/~tmrna) (5). The putative tmRNA gene from the *R.americana* mitochondrion was not included as it does not contain any coding sequence for the peptide tag. Approximately 15% of the tmRNA genes in this set are permuted. The third set was constructed using randomly generated sequences. This set includes seven 100 Mb sequences with a G + C content of 20, 30, 40, 50, 60, 70 and 80%, respectively.

RESULTS

To enable searches for tmRNA genes by automatic methods, we developed a computer program, BRUCE, which uses a heuristic algorithm to search for $tRNA^{Ala}$-like secondary structures surrounding a short sequence encoding the tag peptide. The performance of BRUCE was tested on three different data sets: tmRNA genes identified at the tmRNA website, tmRNA genes identified in completely sequenced genomes, and randomly generated nucleotide sequences.

BRUCE identified tmRNA genes in all full-length sequences at the tmRNA website with the exception of the sequence from *Dehalococcoides ethenogenes*. When a single nucleotide from the T-stem region of this sequence was removed, BRUCE identified a tmRNA gene with a tag sequence of GERELVLAG, agreeing with the tag sequence from the tmRNA website. In fact, all proteolysis tags predicted by BRUCE agreed completely with those from the tmRNA website, except for *Aquifex aeolicus*. In this case, BRUCE predicted a tag sequence of AKTAPEAELALAA, 3 amino acids longer than the previously predicted sequence, APEAELALAA.

BRUCE also correctly detected all 57 putative tmRNA genes (5) in the bacterial genomes sequenced so far without reporting any false positives. One of these genomes, *Chlamydophila pneumoniae* strain J138, is not included in the tmRNA website (5). However, BRUCE detected a tmRNA gene identical in sequence to those from *C.pneumoniae* strains AR39 and CWL029. For all genomes other than the alpha-proteo bacteria, BRUCE and the tmRNA website reported the same nucleotide sequences from the 5′ end to the CCA triplet at the 3′ end, except that the 3′ end often differed in the very last nucleotides, probably due to post-transcriptional modification of the mRNA (3). For the alpha-proteo bacteria BRUCE reported 5′ ends that were 11–16 nucleotides longer. No tmRNAs were detected in any of the sequenced archeal genomes or eukaryotic chromosomes, supporting previous findings that bacterial-like tmRNA genes are not present in these lineages.

Finally, we generated 700 Mb of random sequences with G + C content values ranging from 20 to 80%, which were searched for the presence of tmRNA genes. BRUCE predicted a total of only five putative tmRNAs, all of which were found in sequences with a G + C content of 70% or higher. Thus, the frequency of false positive tmRNA prediction is less than one per 100 Mb of sequence data.

The execution speed of BRUCE was tested on a computer with a Pentium 1 GHz processor. A search through the smallest sequenced genome so far, *Mycoplasma genitalium*, with a size of 0.58 Mb, was processed in 2.9 s. The largest bacterial genome sequenced so far, *Mesorhizobium loti*, with a size of 7.6 Mb, was processed in 147 s. The execution time was found to increase with the G + C content of the sequence data. For example, the 100 Mb randomly generated sequence with a G + C content of 20% was processed with an average speed of 3.0 s/Mb. For comparison, the random sequence data with a G + C content of 50% was processed with an average speed of 8.5 s/Mb, and the one with a G + C content of 80% with an average speed of 94 s/Mb. The increasing search time reflects the fact that tRNA genes are G + C rich, and therefore many more putative tRNA-like structures are present in sequences with high G + C content.

DISCUSSION

We have here developed a computer program for automatic detection of tmRNA genes based on sequences that have previously been predicted by manual methods to correspond to tmRNA genes. The tests performed so far are promising. Only one tmRNA sequence out of 151, that of *D.ethenogenes*, was not detected by BRUCE. An extra nucleotide appears to have been inserted into the T-arm of this sequence. As the whole genome sequence is preliminary (http://www.tigr.org), the apparent insertion may be an effect of low sequence quality. These results demonstrate the high stringency criteria used by BRUCE for detection of tmRNA genes and suggest that tmRNA genes that deviate from the consensus tRNA domain secondary structure may not be recognized.

Indeed, no false positives were predicted in the completely sequenced bacterial genomes analysed here. Furthermore, in the randomly generated sequences with a G + C content of 70% or higher, the number of predicted tmRNAs was only one in 40 Mb on average. Considering that only eight of the 57 sequenced genomes have a G + C content of 60% or higher, the average number of false positives should be considerably lower when searching through complete genomes.

As we have used previous predictions as our training set, the accuracy of BRUCE in detecting tmRNA genes is dependent on the quality and accuracy of the tmRNA sequences found at the tmRNA website (5). BRUCE and the tmRNA website should therefore be regarded as complementary, rather than independent sources of tmRNA sequence information. To our knowledge only two tmRNA sequences have been experimentally verified to correspond to tmRNA genes (7,8). As more sequences receive experimental verification, the accuracy of BRUCE will become more firmly established. Only a minority of bacterial genomes contain annotations for tmRNA genes in the public sequence databases. We believe that with its ease of use, ability to predict the tag sequence in a single step and the free availability of a web interface, BRUCE will become a valuable tool for global analyses of tmRNA sequences as well as for more comprehensive annotation of bacterial genomes.

ACKNOWLEDGEMENTS

Thanks to Håkan Svensson for providing useful scripts, to Ola Lundström for help with Figure 1 and to Kat Taylor for sharing valuable information.

REFERENCES

1. Gillet,R. and Felden,B. (2001) Emerging views on tmRNA-mediated protein tagging and ribosome rescue. *Mol. Microbiol.*, **42**, 879–885.
2. Muto,A., Ushida,C. and Himeno, H. (1998) A bacterial RNA that functions as both a tRNA and an mRNA. *Trends Biochem. Sci.*, **23**, 25–29.
3. Ushida,C., Himeno,H., Watanabe,T. and Muto,A. (1994) tRNA-like structures in 10Sa RNAs of *Mycoplasma capricolum* and *Bacillus subtilis*. *Nucleic Acids Res.*, **22**, 3392–3396.
4. Keiler,K.C., Waller,P.R. and Sauer,R.T. (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, **271**, 990–993.
5. Williams,K.P. (2002) The tmRNA Website: invasion by an intron. *Nucleic Acids Res.*, **30**, 179–182.
6. Keiler,K.C., Shapiro,L. and Williams,K.P. (2000) tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: a two-piece tmRNA functions in *Caulobacter*. *Proc. Natl Acad. Sci. USA*, **97**, 7778–7783.
7. Tu,G.F., Reid,G.E., Zhang,J.G., Moritz,R.L. and Simpson,R.J. (1995) C-terminal extension of truncated recombinant proteins in *Escherichia coli* with a 10Sa RNA decapeptide. *J. Biol. Chem.*, **270**, 9322–9326.
8. Op De Bekke,A., Kiefmann,M., Kremerskothen,J., Vornlocher,H.P., Sprinzl,M. and Brosius,J. (1998) The 10Sa RNA gene of *Thermus thermophilus*. *DNA Seq.*, **9**, 31–35.
9. Dsouza,M., Larsen,N. and Overbeek,R. (1997) Searching for patterns in genomic data. *Trends Genet.*, **13**, 497–498.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Knudsen,B., Wower,J., Zwieb,C. and Gorodkin,J. (2001) tmRDB (tmRNA database). *Nucleic Acids Res.* **29**, 171–172.