Sequence analysis

ARWEN, a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences

Dean Laslett[1] and Björn Canbäck[2*]

[1]Murdoch University, Perth, Western Australia, Australia and [2]Björn Canbäck Bioinformatik, Vindögatan 66, 257 33 Rydebäck, Sweden

ABSTRACT

Motivation: Mitochondrial genomes encode their own transfer RNAs (tRNAs). These are often degenerate in sequence and structure compared to tRNAs in their bacterial ancestors. This is one of the reasons why current tRNA gene predictor programs perform poorly identifying mitochondrial tRNA genes. As a consequence there is a need for a new program with the specific aim of predicting these tRNAs.

Results: In this study we present the software ARWEN that identifies tRNA genes in metazoan mitochondrial nucleotide sequences. ARWEN detects close to 100% of previously annotated genes.

Availability: An online version, software for download and test results are available at www.acgt.se/online.html.

Contact: bcanback@acgt.se

1 INTRODUCTION

The mitochondrial transfer RNA (tRNA) genes of many metazoan (including mammalian) species exhibit less conformity to the canonical cloverleaf secondary tRNA structure, and less homology to recognized tRNA consensus motifs, than cytosolic or prokaryotic tRNA genes, to the extent that these mitochondrial genes are referred to as "bizarre" (Helm et. al., 2000). Smaller than usual dihydrouridine (D) stems and loops; smaller than usual TφC (T) stems and loops; changes in the number of connector bases between the acceptor (A) stem and D-stem, and D-stem and anticodon (C) stem; and elongated C-stems, have been reported, but perhaps the most astounding feature in some of these tRNA genes is the complete absence of either the whole D-arm, or the whole T-arm and variable (V) loop. These are replaced by short sequences, called the D replacement loop and TV replacement loop respectively. Furthermore, mitochondria often use a genetic code that differs from the universal genetic code.

For these reasons, conventional tRNA detection programs perform poorly. For example, the ARAGORN tRNA detection program (Laslett and Canback, 2004) detects only 3 of the 22 tRNA genes in the *Homo sapiens* mitochondrion, and tRNAscan-SE (Lowe and Eddy, 1997) using standard settings detects only 1 gene.

The purpose of this study is to develop a heuristic algorithm to search *in silico* for metazoan tRNA genes. The software should have a detection rate close to 100% even if this results in a number of falsely predicted tRNAs since these can be easily removed when analyzing the genome sequence. The program should be user friendly with a limited number of (user) parameter settings, produce results that are easy to interpret, and a website should be available for the user to perform on-line analysis. The ARWEN program successfully fulfills all of these requirements.

* To whom correspondence should be addressed.

2 METHODS

2.1 Search algorithm

ARWEN employs a heuristic algorithm that searches for hairpin structures with a 5 to 6 base-pair (bp) stem and a 6 to 8 base loop that could be a candidate C-arm. For every candidate C-arm, the upstream sequence is searched for possible D-arm structures (2 to 5 bp stem and loop of 3 to 17 bases), and the downstream sequence for possible T-arm structures (2 to 7 bp stem and loop of 3 to 31 bases). Both upstream and downstream sequences are assessed for base pairing interactions that could indicate the presence of an A-stem (5 to 8 bp). ARWEN then attempts to combine these structures into a complete tRNA gene containing at least three out of four of these structures. Unlike ARAGORN, ARWEN does not allow for the presence of introns in the C-loop.

Three different algorithms are used, one for each type of tRNA (D replacement loop, TV replacement loop and standard cloverleaf), to assign a score to a candidate sequence. Because a standard cloverleaf tRNA can also form a D or TV replacement loop tRNA by opening one hairpin, the initial score for the cloverleaf type is set higher to favour the formation of a full cloverleaf if possible.

The presence of TA in the spacer sequence between A-stem and D-arm, CT<nnn>AA in the C-arm (where <nnn> denotes the anticodon sequence), and GTTC homology in the T-arm (when present) are given extra points respectively. However, the importance of stem base pairing interactions and tertiary structure interactions is increased compared to the ARAGORN program, and the importance of consensus sequence homology is reduced.

In all stems, GC bonds are considered to be the most thermodynamically stable, and are not penalised. AT bonds, GT bonds, and non-bonding base pairs are penalised in order of increasing magnitude. The score for each stem is further modified by stem termination and stem opening base combinations (the last base pairs at either end of the stem and one base beyond), tandem repeats and nearest neighbour interactions within the stem, and lengths of stem and loop for the T-arm that differ from the canonical values (5 bp stem and 7 base loop). Sequences with a high (> 50%) or low (< 10%) overall GC content are also penalised.

The three different algorithms then use 40, 32, and 65 different partial combinations respectively of stem base-pairing, tertiary

interactions and consensus sequence motifs to further modify the score.

If the final score from each algorithm is greater than a cutoff threshold value then the candidate sequence is accepted as a possible tRNA gene. Since D-loop replacement tRNA genes were given the lowest initial score, they have the lowest cutoff value, followed by TV loop replacement genes, and then full cloverleaf genes.

The precise scoring magnitudes and cut-off thresholds within each algorithm were calibrated manually to achieve maximum sensitivity using a training set of 125 annotated metazoan mitochondrial genomes with a total sequence length of 1.98 Megabases. The chosen calibration generated 39 false negatives and 370 false positives from 2696 annotated tRNA genes, giving a sensitivity of 98.5% and a selectivity of 186.6 per Megabase, or almost 3 false positives per genome on average. The typical selectivity for the ARAGORN program was reported as less than 0.0035 per Megabase (Laslett and Canback, 2004). Hence ARWEN is not suitable for use on longer cytoplasmic genomes.

2.2 Testing the Algorithm

The ARWEN algorithm described here is tested against a number of datasets:

(1) a set of mammalian mitochondrial tRNA gene sequences from the Mamit-tRNA database (at http://mamit-trna.u-strasbg.fr/). The set contains 678 tRNA gene sequences from 31 mammalian mitochondrial genomes arranged into 22 groups according to amino acid specificity. One incomplete sequence (tRNA-SerGCT from *Didelphis virginiana*) was removed, and two sequences with errors in the C-loop (tRNA-PheGAA from *Canis familiaris* and tRNA-LysTTT from *Oryctolagus cuniculus*) were corrected, using sequences from the original complete mitochondrial genomes accessed from GenBank, to leave a final set of 677 tRNA sequences. This set includes 30 tRNA-Ser with D-replacement loops.

(2) a set of metazoan mitochondrial tRNA gene sequences from the OGRe database (Jameson et. al. 2003) at http://www.bioinf.man.ac.uk/ogre. This set was extensively corrected using sequences from the original complete mitochondrial genomes. The set contained 10346 tRNA gene sequences.

(3) a set of 23 mitochondrial genomes with RefSeq annotations (Pruitt et al. 2005) which were randomly chosen by using stratified selection to allow representatives from distant lineages with few members. RefSeq is a comprehensive set of sequences which have been manually curated by NCBI staff. The 23 genomes contain 469 tRNAs according to the RefSeq annotations. Detection rates and rate of false positives are compared with those produced by tRNAscan-SE (version 1.23). For the mammalian genomes, ARWEN was invoked by the -mtmam switch (search for mammalian tRNAs) and the -gcmam switch (use the mammalian mitochondrial genetic code). For all other genomes, ARWEN was invoked with the -gcmet switch (use a composite metazoan mitochondrial genetic code). tRNAscan-SE was invoked by the -O option (search for organellar tRNAs) and the -g option (use a specified alternate genetic code). None of the 23 genomes was included in the training set.

3 RESULTS

When running ARWEN on sequences from the Mamit-tRNA database set, a detection sensitivity of 100% was achieved (tRNAscan-SE: 93.6%). A 99.1% detection sensitivity was achieved for the corrected OGRe database set (tRNAscan-SE: 82.8%).

When comparing ARWEN with tRNAscan-SE against the set of 23 mitochondrial genomes containing 469 tRNAs according to RefSeq annotations, ARWEN achieved a 99.4% detection sensitivity. The corresponding value for tRNAscan-SE was 72.1%. ARWEN completed scanning the 23 genomes almost 30 times faster than tRNAscan-SE (total run-time 3.29 minutes on an Intel Centrino Duo 1.6 Ghz CPU, 512 Mb memory running cygwin 1.5.24-2 under windows XP-SP2; compared to 96.09 minutes total run-time for tRNAscan-SE). The number of reported false positives when using ARWEN was 115. The corresponding value for tRNAscan-SE was 8. When the threshold values in ARWEN were raised until the total number of false positives was also 8 (to give an overall selectivity equivalent to tRNAscan-SE), then the overall sensitivity was 81.7%, almost 10 percentage points above tRNAscan-SE.

Test results are summarized in table 1. For more detail, please see supplementary material found at www.acgt.se/online.html.

4 IMPLEMENTATION

ARWEN is written in C. The source code can be downloaded from the website. The website also contains a user interface to the program allowing the user to upload a sequence and run the program on a server. Aggregate sequence lengths up to 2Mb are allowed. ARWEN accepts as input a file with one or more nucleotide sequences in FASTA format. By default, ARWEN assumes that each sequence has a circular topology (search wraps around ends), that both strands should be searched, and that the progress of the search is not reported. These settings can be individually changed to linear topology (no wrapping), search of the sense strand only, and report of search progress. For each candidate mitochondrial tRNA, secondary structure, anti-codon position, and amino acid iso-acceptor species are predicted (Figure 1). If there has been a deviation in the genetic code for a particular anticodon triplet within kingdom metazoa, then more than one possible iso-acceptor species is reported. An abbreviated output format is also available. In this case, for each sequence in the input file, only the sequence name and tab delimited information about each gene detected in the sequence are given.

5 DISCUSSION

The results for the Mamit tRNA gene database and the OGRe database indicate that ARWEN identifies a mitochondrial tRNA in almost every case. ARWEN achieves a detection rate close to 100% for these sequences.

However, a high detection rate may lead to a high level of falsely predicted tRNAs (false positives). By testing ARWEN and tRNAscan-SE on whole-genome sequences a direct comparison of

detection rates and number of false positives is possible. The results presented in table 1 show that, with the options used, ARWEN is superior to tRNAscan-SE in detecting true tRNAs. ARWEN only misses 3 tRNAs while tRNAscan-SE misses 131 tRNAs or 28% of the total number of 469. The missed tRNAs are not evenly distributed among the genomes. In 6 of the genomes tRNAscan-SE detects less than 50% of the annotated tRNAs. In the most extreme case, that of the genome of *Leptotrombidium pallidum*, tRNAscan-SE detects none of the 21 annotated tRNA genes. When ARWEN and tRNAscan-SE are used together, no tRNAs are missed in any of the genomes tested.

On the other hand, ARWEN predicts 115 false positives while tRNAscan-SE only reports 8. Again the distribution of the false positives predicted by ARWEN is not even. 64% of the falsely predicted tRNAs are derived from 6 genomes. Clearly, some genomes have sequence properties that fool the selection filters in the algorithms. Also, with the algorithm used in ARWEN, selectivity will be expected to degrade for genomes with an extraordinary high G+C content, leading to detection of more false positives. Taken together, it becomes evident that the two software programs are written with different aims. While ARWEN identifies nearly all tRNAs, tRNAscan-SE almost never mis-predicts a tRNA. Combined use of the two will lead to a great improvement in annotation of tRNAs in metazoan mitochondrial sequences. Typically these genomes have been annotated by first screening for tRNAs by using tRNAscan-SE and then manual identification of missing tRNAs by comparing the genome under investigation with that of a related organism. From an annotator's point of view, it is better to have too much than too little. Many falsely predicted tRNAs can be easily removed since they are positioned in protein encoding sequences.

It should also be noted that RefSeq annotations in at least a few cases seems to be heavily dependent on predictions from tRNAscan-SE. One such case is the genome of *Axinella corrugata* where RefSeq annotations are identical to the output produced by tRNAscan-SE. If so, the test results will be biased in favour of tRNAscan-SE.

We have here developed a computer program for detection of metazoan mitochondrial tRNA genes. We see several advantages of releasing a new algorithm. (I) The sensitivity is 100% for most mammalian mitochondrial genomes sequenced so far, and greater than tRNAscan-SE for most metazoan mitochondrial sequences. (II) ARAGORN reports the tRNA secondary structure in an intuitive way, as a cloverleaf diagram. tRNAscan-SE also reports secondary structure, however the linear representation of the secondary structure is not as easy to interpret. (III) ARWEN may be run with no options and still produce the desired results. For an inexperienced user tRNAscan-SE may be more difficult to use, especially when using appropriate translation tables.

ACKNOWLEDGEMENTS

The authors wish to thank the Murdoch University Guild of Students and School of Mathematics for their generous provision of access to Unix computers.

REFERENCES

Helm *et. al.* (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA*, **6,** 1356-79.

Jameson,D. *et al.* (2003) OGRe: a relational database for comparative analysis of mitochondrial genomes. *Nucleic Acids Res.*, **31,** 202-06.

Laslett,D. and Canback,B. (2004) ARAGORN, a program for the detection of transfer RNA and transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32,** 11-6.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25,** 955-964.

Pruitt,K.D. *et al.* (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33,** D501-D504.

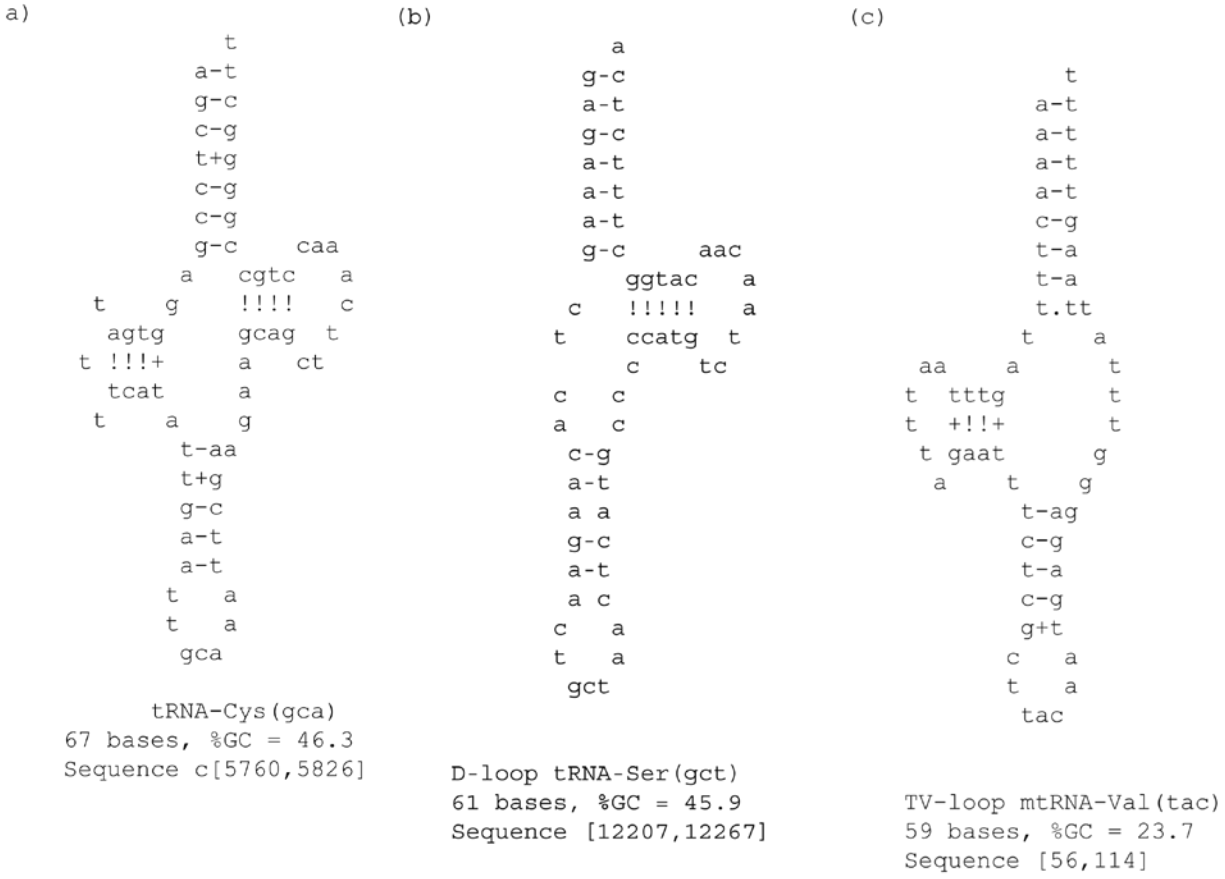

Fig. 1. Example output of the computer program ARWEN from searches on *Homo sapiens* and *Ascaris suum* mitochondrial genomes. (a) Standard cloverleaf structure tRNA-Cys from *H. sapiens* mitochondrion. (b) D-replacement tRNA-Ser from *H. sapiens* mitochondrion. (c) TV-replacement tRNA-Val from *A. suum* mitochondrion.

Table 1. ARWEN and tRNAscan-SE test results for 23 randomly chosen metazoan mitochondrial genomes with RefSeq annotations.

	Reported number of tRNAs			tRNAs not detected			False positives		
Organism	**ARWEN**	**tRNAscan-SE**	**RefSeq[1]**	**ARWEN**	**tRNAscan-SE**	**Both[2]**	**ARWEN**	**tRNAscan-SE**	**Both[3]**
Placozoa									
Trichoplax adhaerens	45	24	22	0	0	0	23	2	1
Porifera									
Axinella corrugata	38	25	25	0	0	0	13	0	0
Cnidaria									
Ricordea florida	5	2	2	0	0	0	3	0	0
Briareum asbestinum	7	1	1	0	0	0	6	0	0
Acoelomata									
Echinococcus multilocularis	24	2	22	0	20	0	2	0	0
Schistosoma mansoni	23	8	23	1	15	0	1	0	0
Pseudocoelomata									
Ancylostoma duodenale	26	8	22	0	15	0	4	1	0
Ascaris suum	24	2	22	0	21	0	2	1	0
Mollusca									
Crassostrea virginica	27	18	23	0	5	0	4	0	0
Arthropoda									
Artemia franciscana	25	10	22	0	12	0	3	0	0
Leptotrombidium pallidum	28	0	21	0	21	0	7	0	0
Heterodoxus macropus	36	18	22	0	5	0	14	1	0
Echinodermata									
Acanthaster planci	24	20	22	0	2	0	2	0	0
Chordata									
Arenaria interpres	24	21	22	0	1	0	2	0	0
Chauliodus sloani	34	24	22	0	1	0	12	3	3
Coreoleuciscus splendidus	25	21	22	0	1	0	3	0	0
Cottus reinii	22	20	22	0	2	0	0	0	0
Dallia pectoralis	22	21	22	0	1	0	0	0	0
Elephas maximus	23	21	22	0	1	0	1	0	0
Onychodactylus fischeri	24	21	22	0	1	0	2	0	0
Porichthys myriaster	24	19	22	1	3	0	3	0	0
Saccopharynx lavenbergi	28	20	22	0	2	0	6	0	0
Semnopithecus entellus	23	20	22	1	2	0	2	0	0
TOTAL	**581**	**346**	**469**	**3**	**131**	**0**	**115**	**8**	**4**

1) Number of tRNAs according to RefSeq annotations.
2) Numbers of tRNAs not detected when using both ARWEN and tRNAscan-SE.
3) Numbers of false positives in common when using both ARWEN and tRNAscan-SE.