Annotating animal mitochondrial tRNAs: an experimental evaluation of four methods

Stacia K. Wyman [*] Jeffrey L. Boore [†]

Abstract

We present the results of an experimental analysis of four methods for identifying animal mitochondrial tRNAs tested on a dataset consisting of 6,600 tRNAs extracted from the entire set of complete animal mitochondrial genomes in GenBank. Methods were evaluated based on false negatives (missing actual tRNAs) and false positives (falsely identifying tRNAs). An exploration of training covariation models to identify animal mitochondrial tRNAs is also presented.

Keywords: *tRNAs, annotation, secondary structure, covariation model*

Identifying and annotating genes is currently a time consuming and error fraught process and, with the input of high throughput genome centers, is becoming a rate-limiting step in the production of complete mitochondrial genome sequences. Clearly, a more automated and accurate method must be developed to streamline this process. In so doing, we may also be able to use this system as a model on which to base methods of finding other types of structured RNA molecules.

Transfer RNAs are approximately 70 nucleotides in length and are necessary components to a cell's protein synthesis machinery. They fold into a complex shape including both single-stranded regions and helices based on internal nucleotide pairings. This can be represented in schematic form as a cloverleaf with four stems (see Figure 1). In animal mitochondrial (mt) genomes, tRNAs make up 22 of the 37 genes and yet no program exists which can automate the identification process and many animal mitochondrial genomes remain unannotated due to the difficulty of identifying tRNAs. Methods based on conservation of nucleotide sequence exist [4, 5, 8, 10, 1], but are not suitable for animal mt tRNAs. This is because selection operates on functional tRNAs based on maintenance of base-pairing (secondary) structure rather than conservation of nucleotide sequence. Because animal mitochondrial tRNAs have almost no conservation of sequence at the nucleotide level, methods must focus on covariation of basepairing in the secondary structure.

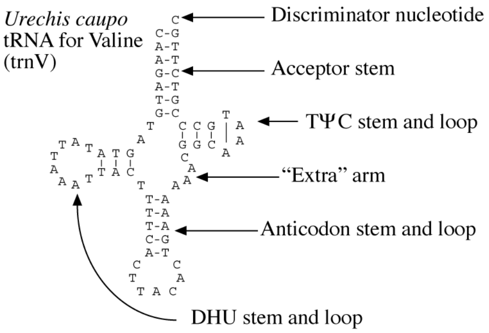

Figure 1. Schematic representation of a typical tRNA encoded by an animal mitochondrial genome. Nucleotides paired by hydrogen bonds are indicated by dashes.

We present an analysis of the performance of four existing methods for identifying animal mt tRNAs as well as an in-depth exploration of ways to train covariation models. The methods were rigorously tested by running each program on the 300 complete animal mt genomes in GenBank [6] containing 6,600 tRNAs. The programs which we tested were: COVE [3], tRNAscan-SE [8], RNAMotif [9], and our own method.

Methods Identifying tRNAs using covariation models (a generalization of hidden Markov models) was first proposed in 1994 simultaneously by Eddy and Durbin [3] and Sakakibarra *et al.* [11] and was implemented by Eddy and Durbin in the COVE software package. A covariation model (CM) is created in COVE by training the model with a set of sequences. The set of sequences may be previously aligned or not. This is well-suited to the animal mt tRNA problem since we are not required (and, in fact, would not be able to) align the training sequences based on primary structure. Once the model has been created, a candidate RNA sequence is aligned to the CM using a three-dimensional dynamic programming algorithm, and the score is then calculated based on the probability of the alignment with the model. For COVE, we created a covariation model specifically trained to identify animal mitochondrial tRNAs. The model was trained with 1,432 tRNAs from 65 complete ani-

[*] `stacia@cs.utexas.edu`. Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712. Supported by NSF IGERT 0114387.

[†] `jlboore@lbl.gov`. DOE Joint Genome Institute 2800 Mitchell Drive Walnut Creek, CA 94598.

mal mt genomes taken from the set of 300 genomes. COVE was then run on the datasets with and without the training sequences.

tRNAscan-SE, probably *the* most popular program today for identifying tRNAs, is primarily targeted towards non-organellar tRNAs. It uses two methods as a prepass for identifying candidate sequences based on sequence content and then passes the candidates to COVE as a subroutine. For organellar tRNAs, it bypasses the prepass step (because the sequences that they search for don't exist in organellar genomes) and gives the sequences directly to the covariation model. This process, without the preprocessing, is equivalent to running the COVE program on the tRNA CM included in the software package. For tRNAscan-SE, we used the tRNA CM that the program came with. This was trained on a dataset of 1,415 aligned tRNAs from the 1993 Sprinzl database [12].

RNAMotif is a descendent of past motif-based programs [5, 7], but has a more powerful descriptor language than its predecessor, RNAMOT [5], and it has a global scoring scheme. The user creates a descriptor for a molecule based on stem and loop motifs and scoring rules. We created an RNAMotif descriptor file for animal mt tRNAs (see our web site for the descriptor [2]).

Finally, the fourth method we tested was one we implemented as part of an organellar genome annotation package. It is a pattern-matching algorithm which combines structural motif searching with an integrated scoring system designed specifically for animal mt tRNAs.

Evaluation Each program was tested on the 300 genome dataset and was evaluated based on its ability to identify the 22 tRNAs (one for each amino acid and two for both serine (tRNA-Ser) and leucine (tRNA-Leu)) in each genome and counting the number of false negatives (FN) and false positives (FP). A false negative occurs when a program fails to identify an actual tRNA, and a false positive occurs when a program identifies a sequence as a tRNA when it is not one. The false negatives for each genome were counted and plotted in Figure 2. It shows, for each of the four methods, for each number of false negatives, how many genomes missed that many tRNAs. The false negatives for each of the 22 tRNAs also were counted individually and are presented in Figure 3 with the FN for the two tRNA-Ser and tRNA-Leu combined. This figure shows for each tRNA, how many genomes missed it for each method.

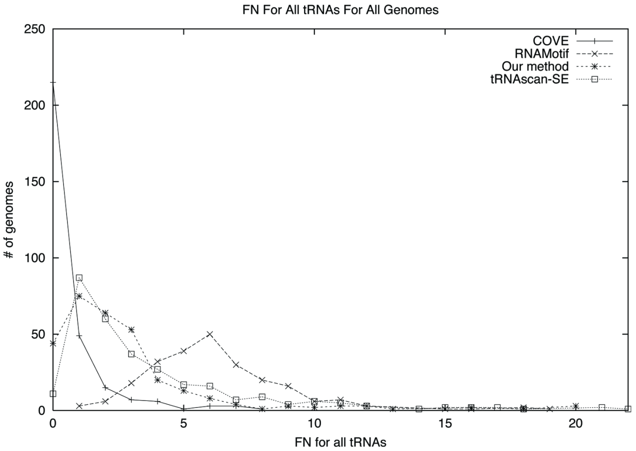

Figure 2. For each value of FN, the number of genomes with that FN is plotted.

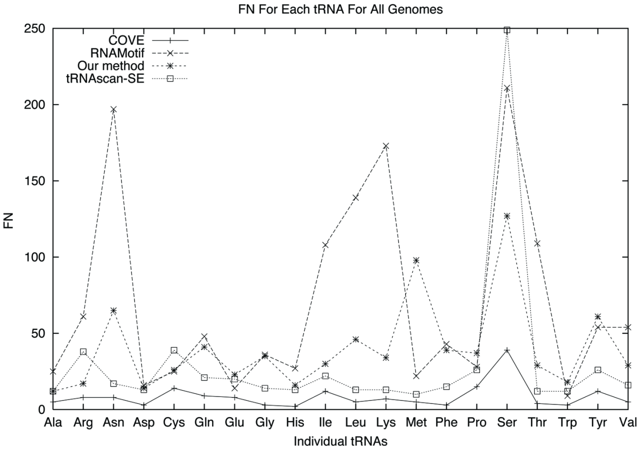

Figure 3. For each tRNA, the number of genomes that missed that tRNA is plotted.

Results and Discussion As can be seen in Figure 2, the best-performing method was, by far, COVE. The COVE results are for all 300 genomes, including the training set. COVE found all 22 of the tRNAs for 215 out of 300 genomes. This is compared to just 11 for tRNAscan-SE, 44 for our method and *none* for RNAMotif. Even as the best performer, however, the CM trained for animal mitochondrial tRNAs still missed some tRNAs and in the worst case genome, missing 8. Figure 3 shows that COVE, unlike the other methods, performed uniformly well. The spike on the FN for COVE on serine is primarily because the FN for the two serines is combined. Figure 3 illustrates that COVE is much less sensitive to degenerate cases of animal mt tRNA secondary structure than the other methods. Although very successful for non-organellar genomes [8], the tRNAscan-SE CM, which had been trained on non-organellar genomes, does not perform nearly as well on animal mitochondrial tRNAs (Figure 2). tRNAscan-SE only found all 22 of the tRNAs in 11 genomes (compared to 215 for COVE) and even missed *every* tRNA for one of the genomes. This can also be seen in Figure 3 where tRNAscan-SE missed tRNA-Ser (which is often missing the D arm) for most of the genomes.

The least successful of the methods was RNAMotif, with over half of the genomes missing more than 25% of the tRNAs and not finding all of the tRNAs in any of the genomes. To be fair, RNAMotif is meant to be a general purpose program for identifying RNA molecules. This becomes a problem because the descriptor for mitochondrial tRNAs must be very general because individual nucleotides are not constrained, also allowing for a huge number of matches. For the graph in Figure 2, if a candidate with the correct coordinates was found in the top 20 answers for each tRNA, it was counted. This is a more generous way of counting than used with the other methods and yet RNAMotif still did not perform very well. One of the drawbacks to this method is its "all or nothing" approach. The descriptor language does have some nice regular expression features, but it does not allow for boolean expressions. As an example, one would want to allow for a missing D arm in a tRNA by matching a stem and loop strongly or not at all. Our own method found all 22 of the tRNAs for 44 of the genomes, and usually only missed 1 or 2 tRNAs. When our method identified the correct tRNA, it was usually the top scoring tRNA or within the top 3 for each tRNA. Although our method performs reasonably well compared to tRNAscan-SE and RNAMotif, it doesn't perform nearly as well as COVE.

With respect to false positives, the both COVE and tRNAscan-SE had very few and while this is indeed a feature, it also doesn't present the user with "second choices" if the folding is not to their satisfaction. The number of false positives for our method is not reported because the user can choose how many of the best-scoring candidates for each tRNA should be returned. The number of false positives reported by RNAMotif is so large as to make the method impractical, sometimes giving hundreds of false positives per tRNA.

Further Training of Covariation Models Encouraged by the strong performance of COVE, we attempted to improve upon the method by trying different approaches to training the model. Despite coming up with several strategies for training, we were surprised to find that none of the new CMs that we trained performed better than the original (trained with 1,432 animal mt tRNAs). We tried increasing the number of training sequences up to 3,300 (half of all the tRNAs in the dataset) and it found all of the tRNAs in 221 genomes versus 215 in the 1,432 CM but this was not uniform over different FN values. For individual tRNAs, the 3,300 CM only missed 8 tRNA-Pro versus 15 in the 1,432 CM but it missed 20 tRNA-Cys when the 1,432 CM only missed 14.

We then tried to construct a CM which would do a better job identifying non-canonical foldings by training it solely on tRNA-Ser (often one of the two tRNA-Ser is missing an arm). Despite using just the tRNA-Ser sequences to train the model, it still missed 52 tRNA-Sers compared to 41 for the original model. We tried expanding this idea to targeting each tRNA family individually, but just as with the CMs targeted for tRNA-Ser, there was not a discernible improvement in performance.

Conclusion Here we have presented an analysis of existing tRNA identification methods and evaluated their performance with respect to animal mt genomes. We have shown that COVE is the most effective and promising method. We have shown that tRNAscan-SE does not perform well for organellar tRNAs because of the way it was trained. We showed that COVE is the most robust method with respect to identifying non-canonical foldings.

Acknowledgments Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, in the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC03-76SF00098.

References

[1] B. Billoud, M. Kontic, and A. Viari. Palingol: a declarative programming language to describe nucleic acids' secondary sructures and to a scan sequence database. *Nucleic Acids Research*, 24:1395–1403, 1996.

[2] `http://www.cs.utexas.edu/users/stacia/trna_desc.txt`.

[3] S. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.

[4] G. Fichant and C. Burks. Identifying potential tRNA genes in genomic DNA sequences. *Journal of Molecular Biology*, 220:659–671, 1991.

[5] D. Gautheret, F. Major, and R. Cedergren. Pattern searching/alignment with RNA primary and secondary structures: An effective descriptor for tRNA. *Comput. Applic. Biosci.*, 6:325–331, 1990.

[6] `http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/mztax_short.html`.

[7] A. Laferriere, D. Gautheret, and R. Cedergren. An RNA pattern matching program with enhanced performance and portability. *Comput. Applic. Biosci.*, 10:211–212, 1994.

[8] T. Lowe and S. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25:955–964, 1997.

[9] T. Macke, D. Ecker, R. Gutell, D. Gautheret, D. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research*, 29:4724–4735, 2001.

[10] A. Pavesi, F. Conterlo, A. Bolchi, G. Dieci, and S. Ottonello. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transriptional control regions. *Nucleic Acids Research*, 22:1247–1256, 1994.

[11] Y. Sakakibarra, M. Brown, R. Hughey, I. Mian, K. Sjolander, R. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22:5112–5120, 1994.

[12] S. Steinberg, A. Misch, and M. Sprinzl. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 21:3011–3015, 1993.