Annotating animal mitochondrial tRNAs: A new scoring scheme and an empirical evaluation of four methods

Stacia K. Wyman
Dept. of Computer Sciences
Univ. of Texas, Austin, TX 78712
(512)471-8854 fax: (512)636-2596
email: stacia@cs.utexas.edu

Jeffrey L. Boore
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598

Abstract

Identification of transfer RNAs in animal mitochondrial genomes is important for many areas of genome analysis including phylogenetic reconstruction, understanding inheritance of disease, and identifying forensic materials. Animal mitochondrial tRNAs differ from the canonical tRNAs in both their secondary structure and level of conservation of nucleotide sequence and therefore, conventional tRNA or general RNA searching software cannot be used for identification and custom methods are required. Here we present the results of an experimental analysis of four different methods tested on a large dataset consisting of 5,720 tRNAs extracted from the entire set of complete animal mitochondrial genomes in GenBank[9]. Methods were evaluated based on number of false negatives and false positives. Additionally, we present a new scoring scheme customized for animal mitochondrial tRNAs.

Keywords: genome annotation, tRNA identification, secondary structure, covariation models

1 Introduction

Genome annotation is a critical aspect of whole genome analysis and is the precursor to many kinds of biological analyses. With the advent of high throughput sequencing, we are presented with a wealth of whole genomes which must be analyzed and out-of-date methods which do not scale to the problems at hand. Annotating whole genomes involves identification of protein-coding genes for which excellent methods exist based on search by sequence similarity[1,17], but also ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) must be identified. In animal mitochondrial (mt) genomes, tRNAs make up 22 of the 37 genes and yet no program exists which can automate the identification process. Methods developed for protein-coding genes don't work for animal mt tRNAs (or most other tRNAs) because selection operates on functional tRNAs based on maintenance of base-pairing structure rather than conservation of nucleotide sequence. Transfer RNAs sharing the same function may appear unrelated based on primary sequence similarity and their close relationship can only be seen once their secondary structure is known.

Many general programs have been developed for identifying RNA molecules[7,8,12,16] but they often focus on features to accommodate identification of general RNA molecules (like psuedoknots) or are based on a combination of secondary structure and primary sequence searching techniques. This is in part because animal mt tRNAs have a non-canonical secondary structure, but also because much of the existing software is targeted towards identifying single RNA molecules within a fragment of DNA. 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.

Here we present an analysis of the performance of four existing methods in identifying animal mt tRNAs. The methods have been rigorously tested by running each program on the 260 complete animal mt genomes in GenBank[9] containing 5,720 tRNAs. The programs were chosen after

preliminary testing showed them to potentially be successful in identifying animal mitochondrial tRNAs. The programs we tested are COVE[5], tRNAscan-SE[12], RNAMotif[13], and our own method. Eddy and Durbin's COVE software is based on probabilistic covariance models (CM) constructed from aligned sequences. This method has a solid theoretical framework and is quite tractable for the small size of animal mt genomes (approx. 15000 nucleotides). Lowe and Eddy's tRNAscan-SE, probably the most popular program today for identifying tRNAs, is a hierarchical method based on a combination of three methods. Macke *et al.*'s is a descendent of RNAMOT which also uses descriptors of structural motifs, but a more powerful descriptor language for capturing secondary structure information as well as a global scoring mechanism. Finally, we tested our own method which implements a custom animal mitochondrial tRNA scoring scheme integrated into a structural motif-based search algorithm.

2 Biological Background

Transfer RNAs (see Söll and RajBhandary[20]) 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. These parts are illustrated in Figure 1.

Each tRNA is enzymatically charged with one particular amino acid according to features internal to the tRNA. The amino acid is chemically linked to the discriminator nucleotide. The tRNA then delivers this amino acid to the growing peptide chain on the ribosome. The order of entry of tRNAs into the ribosome is specified by the messenger RNA (mRNA). The mRNA is a chain of nucleotides (generally hundreds or thousands of nucleotides in length) that threads through the ribosome by triplets, with each triplet (i.e. "codon") binding to the three complementary nucleotides at the base of the tRNA (the anticodon; in the case of Figure 1 these are TAC). Thus, the order of nucleotides in the mRNA specifies the sequencing of tRNAs into the ribosome and, consequently, the order of amino acids in the growing peptide chain.

Of course, these tRNAs are encoded by genes, which are commonly identified by the potential of their sequences to form these cloverleaf-like structures and by certain well-conserved nucleotide positions. This can be challenging to do reliably, and attempts fail both by missing tRNA genes that are poorly conserved or aberrantly structured as well as by generating false positives. These problems are especially acute for the tRNAs that are encoded by animal mitochondrial (mt) genomes, which are especially variable both in sequence and structure. For example, the name of the "TψC" arm derives from the "universal" presence of the three nucleotide sequence thymine-pseudouracil-cytosine; this is not present in mt tRNAs. Mitochondrial tRNAs are also occasionally missing some paired arms or are otherwise varying in structure (see, for example, Wolstenholme et al. 1987[23]).

The comparison of complete mt genome sequences is becoming increasingly important for reconstructing the evolutionary relationships of organisms[3,4,14], for studying population structure and history [18], including those of humans[10], for identifying forensic materials[15], and for understanding the inheritance of certain human diseases [22]. 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.

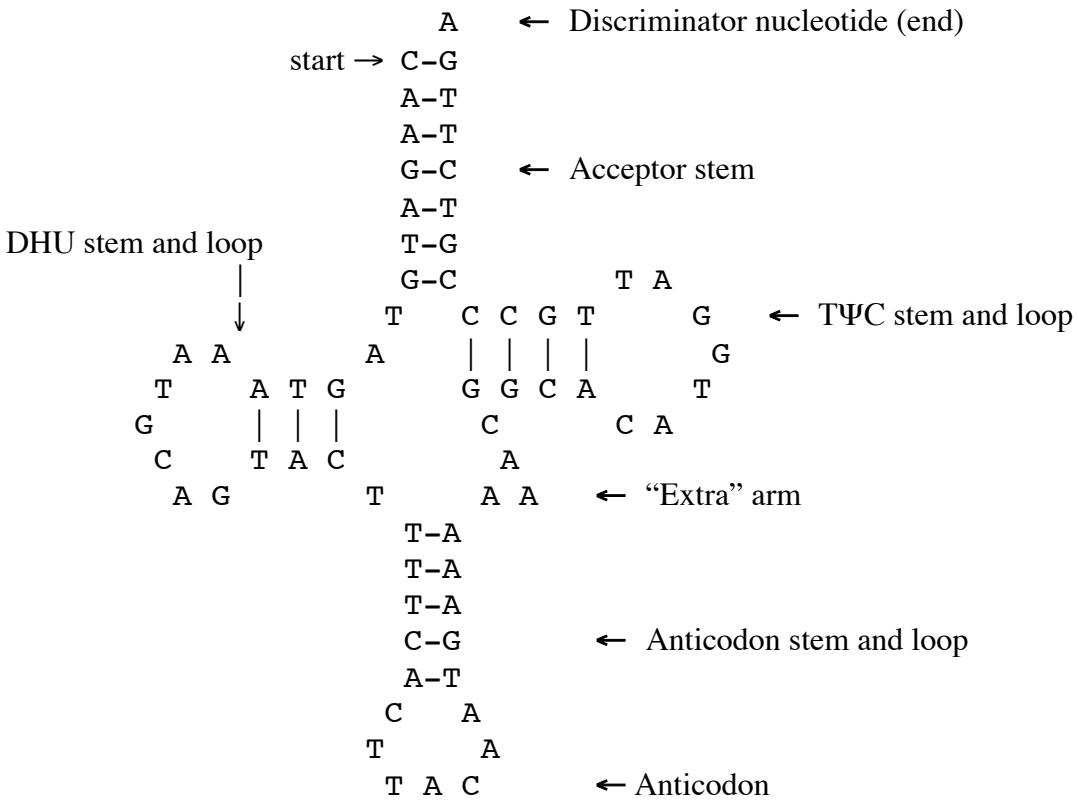

Figure 1: Schematic representation of a typical tRNA encoded by an animal mitochondrial genome. Nucleotides paired by hydrogen bonds are indicated by dashes. The tRNA is folded from a single string of ribonucleotides starting with "CAAGATG", reading counterclockwise around the structure ("TAGTAAATGCAGTACTTTTC ACTTACAATGAAAAAACGGCACATGGATTGCC") and ending with "CGTCTTGA".

3 Methods

Methods for identifying tRNAs typically fall into three basic categories: covariation analysis based on a generalization of hidden markov models (HMMs)[5,19], motif-finding algorithms which include structural as well as nucleotide sequence motifs[6,7,8,11], and minimum-energy based algorithms[24]. We tested four of these programs for identifying tRNAs in animal mt genomes. The four programs — COVE, tRNAscan-SE, RNAMotif, and our own method — are discussed briefly below.

COVE Identifying tRNAs using HMMs was first proposed in 1994 simultaneously by Eddy and Durbin[5] and Sakakibarra *et al.*[19] and was implemented by Eddy and Durbin in the COVE software package. A covariation model is created in COVE by training it with a set of sequences and adjusting the parameters and structure of the model so that high probabilities are assigned to the training 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 calculated based on the probability of the alignment.

Although this method initially showed promise and was very accurate, it was prohibitively time-consuming to run. Lowe[12] estimated that searching the human genome with a tRNA covariance model would take about nine and a half CPU years. However, animal mt genomes, at 15,000 bp, are a very reasonable size and indeed, COVE typically took about one and a half minutes to identify all the tRNAs in a whole animal mt genome.

tRNAscan-SE tRNAscan-SE 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 covariance model. This process, without the preprocessing, is equivalent to running the COVE program on the tRNA CM included in the software package.

RNAMotif RNAMotif is a descendent of past motif-based programs[8,11,2], but has a more powerful descriptor language than its predecessor, RNAMOT[8] and a global scoring scheme. The user creates a descriptor for a molecule based on stem and loop motifs, including any specific information regarding nucleotide values or numbers of mismatches in the the stem. The descriptor also includes a scoring section in which the user can query assignments made to the motifs by the search and from this, can compute a score. RNAMotif first creates a tree representation from the descriptor file. It then does a depth first search through the tree, trying to match the input sequence. The successful candidates are then passed to the scoring routine, where they are evaluated and optionally accepted or rejected based on rules in the score section.

Our method Our method, which is still in the preliminary stages of development, was initially implemented as part of a whole genome annotation package for organellar genomes. It is a pattern-matching algorithm which combines structural motif searching with an integrated scoring system designed specifically for animal mt tRNAs. Its search relies almost exclusively on secondary structure for identification. It searches one-by-one for each tRNA's anticodons so that the user is given the best-scoring candidates for *each* tRNA. This way, none of the tRNAs are missed entirely and the best-scoring possibilities are returned to the user. The program first identifies the anticodon arm (see Figure 1) and then searches to either side for a basepairing arm of length seven which is the

acceptor stem, awarding points for each positive pattern that is matched. The remaining sequence between the anticodon and acceptor stems is then folded to maximize the score. The best-scoring candidates are then saved and reported to the user.

4 Experimental Setup

Our dataset consists of 260 complete animal mitochondrial genomes from GenBank with their accompanying annotation of 5,720 tRNAs. This is the complete list of all animal mitochondrial genomes at the time of this writing.

In testing the COVE method, a covariance model specifically trained to identify animal mitochondrial tRNAs was created. The model was trained with 1,432 tRNAs from 65 complete animal mt genomes taken from the set of 260 genomes. COVE was then run on the datasets with and without the training sequences. The results reported here are the ones on the whole dataset so that they can be more easily compared to the other methods which are tested on the whole dataset. The difference in the result was that there were about one third less genomes for which every tRNA was found (refer to Figure 2). This is to be expected because the CM should correctly identify the tRNAs it was created with. For tRNAscan-SE, we used the CM that the program came with which was trained on a dataset of 1415 aligned tRNAs from the 1993 Sprinzl database[21]. Although this model was not trained on organellar tRNAs, when a user queries the tRNAscan-SE web site with a mitochondrial sequence, this is the model which is used. For RNAMotif, we created a descriptor file for animal mt tRNAs and went through several iterations of testing and modifying the descriptor until we was convinced it was performing as well as it could. we then ran the output through the pruning routine that comes RNAMotif which is meant to prune out subsets of solutions in which the stem might not be as long as possible.

5 Evaluation

Each program was tested on the 260 genome dataset and evaluated based on 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, in fact, is not. 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 were counted 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.

6 Results and Discussion

As can be seen in Figure 2, the best-performing method by far was COVE. In the figure, the COVE results are for all 260 genomes, including the training set. One would expect it to perform well because it was trained on a subset off the dataset, but even without the subset of genomes which the model was trained on, it performed very well, getting all 22 of the tRNAs correct for 178 out of 260 genomes. This is compared to 8 for tRNAscan-SE, 34 for our method and *none* for RNAMotif. Even as the best performer, however, the CM trained for animal mitochondrial tRNAs still missed some tRNAs, in the worst case, missing 8 of the tRNAs. Figure 3 shows that for COVE, unlike the other methods, one can't say that it performed especially poorly on particular tRNAs or for particular genomes except that there is a spike on the FN for COVE on Serine. The FN for both Serines were combined, but the FN is still high. However, it does appear that COVE is much less

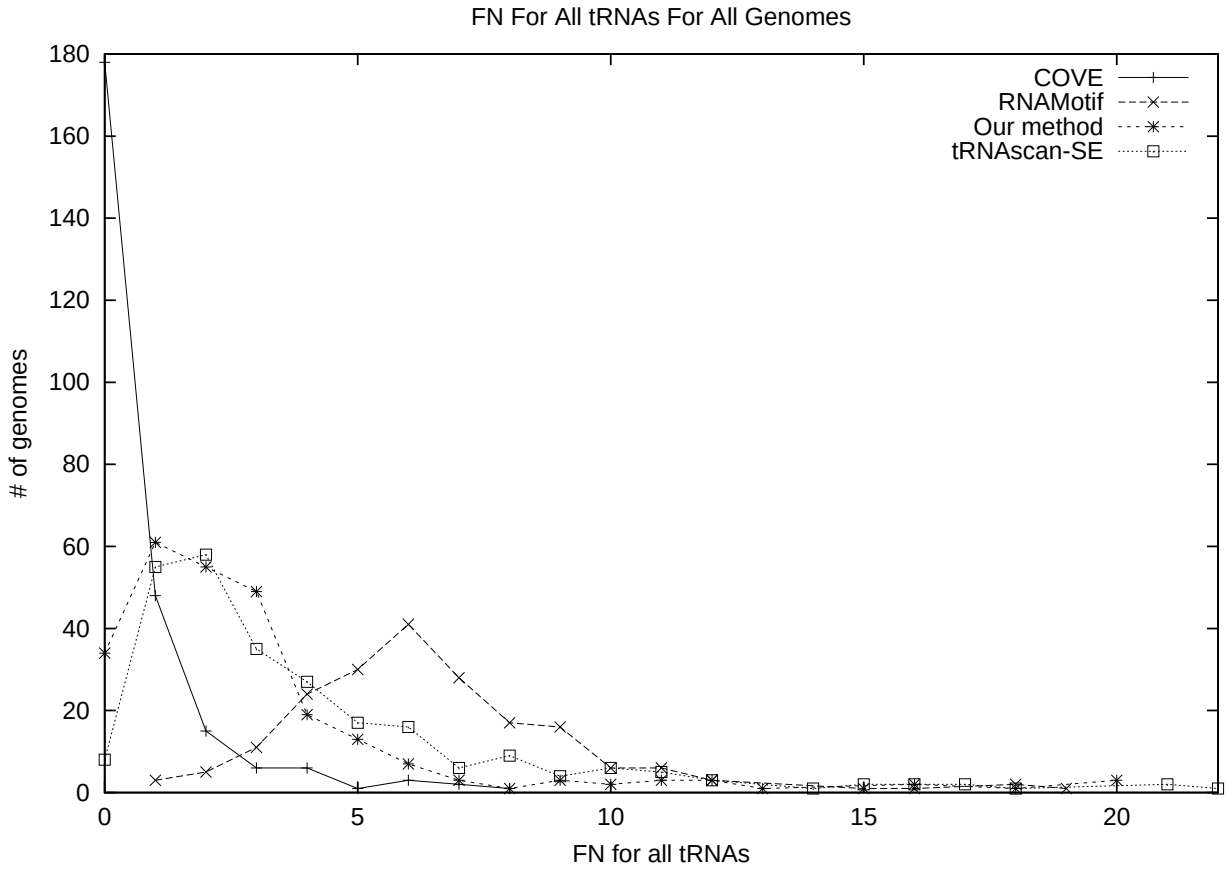

Figure 2: The FN for the four methods. For each number of missed tRNAs (x-axis), the number of genomes with that FN is plotted.

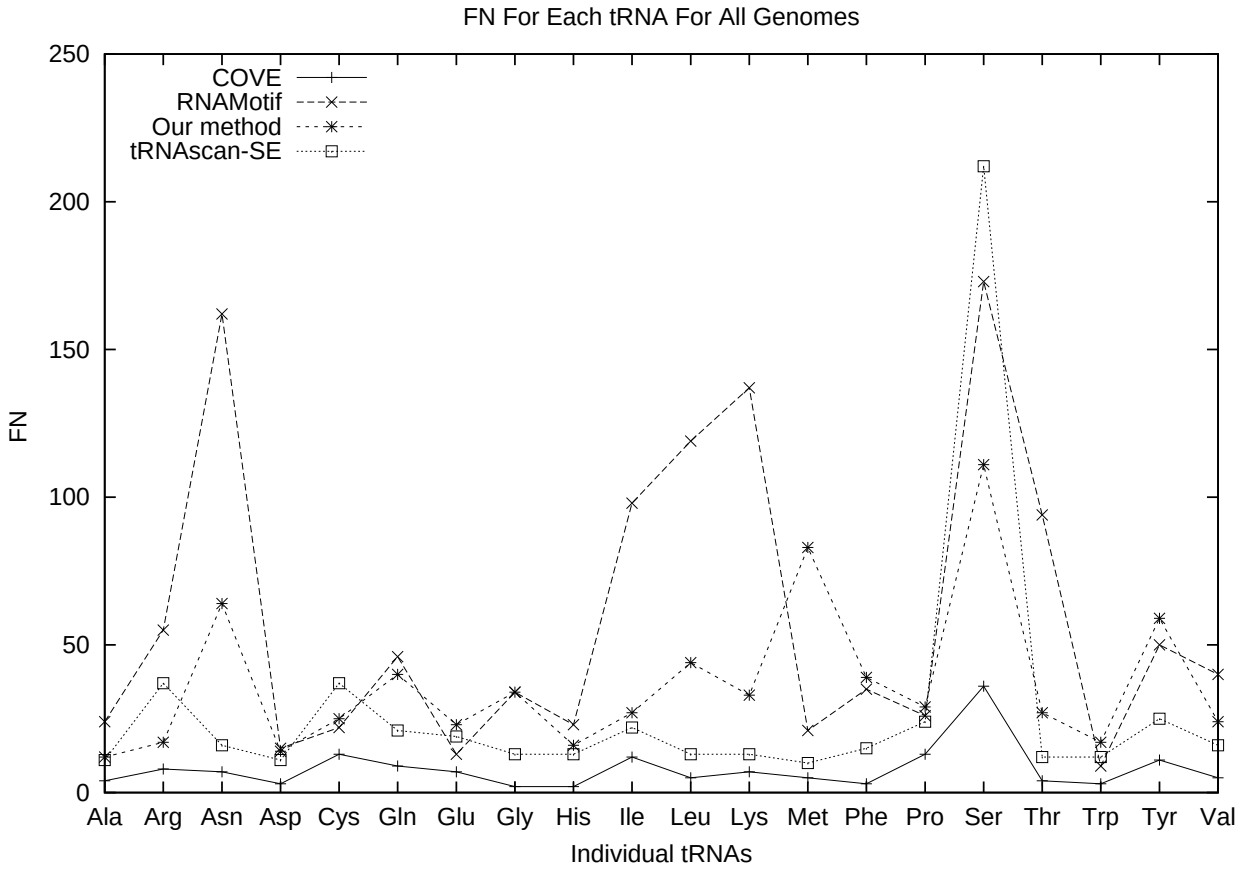

Figure 3: The FN for each tRNA for the four methods. For each tRNA, the number of genomes that missed that tRNA is plotted.

sensitive to degenerate cases of animal mt tRNA secondary structure than the other methods.

tRNAscan-SE is probably *the* most popular method for identifying tRNAs. Although very successful for non-organellar genomes[12], the CM which has been trained on non-organellar genomes does not perform particularly well on animal mitochondrial tRNAs. As can be seen in Figure 2, a CM needs to be appropriately trained for its target molecule. tRNAscan-SE only found all 22 of the tRNAs in 8 genomes and even missed *every* tRNA for two 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. This illustrates how important it is to use care when using probabilistic methods which require training. It is tempting to dismiss a method when it initially doesn't appear to be successful, but they can be extremely rigorous and accurate when properly trained (as in COVE, above).

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. One of the problems with assessing the performance of RNAMotif is the volume of output. The descriptor for mitochondrial tRNAs must be very general because individual nucleotides are not constrained, but this also allows for a huge number of matches. Even when the output is run through the pruning routine that comes with RNAMotif, the deluge of output is more than even the most diligent user could wade through. 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 very generous way of counting (one could not have much confidence in a method where the correct tRNA is ranked eighteenth in a list) and yet it 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 types of 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. We suspect this is also the reason that RNAMotif is the method most sensitive to missing particular tRNAs. In Figure 3, it appears that there a group of tRNAs for which RNAMotif performs exceptionally poorly.

Our method, while still not robust, shows potential. It found all 22 of the tRNAs for 38 of the genomes, and usually only missed 1 or 2. When our method identified the correct tRNA, it was usually the top scoring tRNA or within the top 3 for each tRNA. A feature of our method is that it selects the top-scoring candidates for each tRNA and presents them to the user. The results from our method also illustrate how difficult it can be to develop a system for recognizing a set of tRNAs with such diverse secondary structure and so many exceptions to the canonical tRNA structure.

With respect to false positives, the COVE program 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.

7 Conclusions and Future Directions

Here we have presented an analysis of existing tRNA identification methods and evaluated there performance with respect to animal mt genomes. We have shown that COVE is the most effective and promising method and why other methods are not successful at identifying animal mt tRNAs. COVE is also the most robust method with respect to identifying non-canonical foldings. Future

work will include continued development and improvement of our own method as well as continued investigation of the COVE method. One aspect to investigate further is selection of the training set. The training set of animal mitochondrial genomes was chosen somewhat arbitrarily and it may be that selecting a more phylogenetically representative set of genomes across the animal kingdom would improve the results.

8 Acknowledgments

SKW was supported by National Science Foundation IGERT grant 0114387.

9 References

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
2. 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.
3. J.L. Boore and W.M. Brown. Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.*, 8(6):668–674, 1998.
4. Y. Cao, M. Fujiwara, M. Nikaido, N. Okada, and M. Hasegawa. Interordinal relationships and timesecale of eutherian evolution as inferred from mitochondrial genome data. *Gene*, 259:149–158, 2000.
5. S.R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.
6. N. El-Mabrouk and F. Lisacek. Very fast identification of RNA motifs in genomic DNA. application to tRNA search in the yeast genome. *Journal of Molecular Biology*, 264:46–55, 1996.
7. G.A. Fichant and C. Burks. Identifying potential tRNA genes in genomic DNA sequences. *Journal of Molecular Biology*, 220:659–671, 1991.
8. 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.
9. `http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/mztax_short.html`.
10. M. Ingman, H. Kaessmann, S. Pääbo, and U. Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708–713, 2001.
11. A. Laferriere, D. Gautheret, and R. Cedergren. An RNA pattern matching program with enhanced performance and portability. *Comput. Applic. Biosci.*, 10:211–212, 1994.
12. T.M. Lowe and S.R. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25:955–964, 1997.
13. T.J. Macke, D.J. Ecker, R.R. Gutell, D. Gautheret, D.A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research*, 29:4724–4735, 2001.
14. M. Miya, A. Kawaguchi, and M. Nishida. Mitogenomic exploration of higher teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. *Mol. Biol. Evol.*, 18:1993–2009, 2001.
15. T.J. Parsons and M.D. Coble. Increasing forensic discrimination of mitochondrial DNA testing through the analysis of the entire mitochondrial DNA genome. *Croatian Med. J.*, 42:304–309, 2001.
16. 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.
17. W.R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 183:63–9, 1990.

18. D.M. Rand. The units of selection on mitochondrial DNA. *Ann. Rev. Ecol. Syst.*, 32:415–448, 2001.
19. Y. Sakakibarra, M. Brown, R. Hughey, I.S. Mian, K. Sjolander, R.C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22:5112–5120, 1994.
20. D. Söll and U. RajBhandary, editors. *tRNA: Structure, Biosynthesis, and Function*. American Society for MIcrobiology, Washington, DC, 1995.
21. S. Steinberg, A. Misch, and M. Sprinzl. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 21:3011–3015, 1993.
22. D.C. Wallace. Mitochondrial diseases in man and mouse. *Science*, 283:482–488, 1999.
23. D.R. Wolstenholme, J.L. MacFalane, R. Okimoto, D.O. Clary, and J.A. Wahleithner. Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. *Proc. Natl. Acad. Sci.*, 84:1324–1328, 1987.
24. M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.