tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence

Todd M. Lowe & Sean R. Eddy[1]

Department of Genetics
Washington University School of Medicine
660 S. Euclid, Box 8232
St. Louis, MO 63110
{lowe,eddy}@genetics.wustl.edu

[1]Corresponding author

ABSTRACT

We describe a program, tRNAscan-SE, which identifies 99-100% of transfer RNA genes in DNA sequence while giving less than one false positive per 15 gigabases. Two previously described tRNA detection programs are used as fast, first-pass prefilters to identify candidate tRNAs, which are then analyzed by a highly selective tRNA covariance model. This work represents a practical application of RNA covariance models, which are general, probabilistic secondary structure profiles based on stochastic context-free grammars. tRNAscan-SE searches at approximately 30,000 bp/second. Additional extensions to tRNAscan-SE detect unusual tRNA homologues such as selenocysteine tRNAs, tRNA-derived repetitive elements, and tRNA pseudogenes.

INTRODUCTION

Transfer RNA (tRNA) genes are the single largest gene family. A typical eukaryotic genome contains hundreds of tRNA genes; the human genome contains an estimated 1,300 (1). In a time when complete genomes are being sequenced, one would like to have an accurate means of tRNA gene identification. The tRNA repertoire of an organism affects the codon bias seen in highly expressed protein coding genes. In extreme cases, selective pressure for extremely high or low genomic GC content may have caused loss of a tRNA, producing an unassigned codon (2,3). Suppressor tRNAs are important genetic loci in many model organisms. In addition to authentic tRNA genes, tRNA-derived short interspersed nuclear elements (SINEs) have been identified in rodents and other mammals as likely mobile genetic elements (4,5). Detection and discrimination of these elements from true tRNAs is a desirable feature of tRNA identification methods.

It is commonly believed that the best RNA gene detection methods are custom-written programs that search for one type of RNA gene exclusively (6). Numerous tRNA search programs key on primary sequence patterns and/or secondary structure specific to tRNAs (7-13). Why bother with specialized tRNA-detection software instead of using a fast, commonly available similarity search program such as BLAST (14) or FASTA (15)? Since many functional RNA genes tend to conserve a common base-paired secondary structure better than a consensus primary sequence, the accuracy of RNA similarity searching is much improved by including secondary structure elements. A group of generalized RNA gene search tools look for specific combinations of primary and secondary structure motifs specified by the user (16-21), although tRNA “descriptors” in these pattern-matching languages have typically under-performed custom-written programs.

tRNAscan 1.3 by Fichant & Burks (12) is perhaps the most widely used tRNA detection program. It identifies approximately 97.5% of true tRNA genes and gives 0.37 false positives per million base pairs (Mbp) (12). The algorithm uses a hierarchical, rule-based system in which each potential tRNA must exceed empirically determined similarity thresholds for two intragenic promoters, plus have the ability to form base pairings present in tRNA stem-loop structures. The false positive rate of tRNAscan has been acceptable for small genomes, but for larger eukaryotic genomes it becomes a significant problem. It will produce around 1100 false positive tRNAs for the human genome (0.37 false pos/Mbp for 3000 Mbp); given that there are about 1300 true tRNAs in the genome, almost half of the tRNAs predicted by tRNAscan will be false positives.

Pavesi and colleagues have developed a different tRNA detection algorithm (13) which searches exclusively for linear sequence signals in the form of eukaryotic RNA polymerase III promoters and terminators. The sensitivity and selectivity of this algorithm is roughly comparable to tRNAscan 1.3 in detection of eukaryotic tRNAs. Notably, the Pavesi algorithm identifies tRNAs not detected by tRNAscan 1.3, and vice versa (13). The combined sensitivities of these two programs exceed 99%; however, the combined false positive rate is about five times that of tRNAscan alone.

Eddy & Durbin (22) have developed a general RNA structure similarity search method employing probabilistic RNA structural profiles, or "covariance models". Covariance models are able to capture both primary consensus and secondary structure information through the use of stochastic context-free grammars (SCFG's; 22-24). Much like sequence profiles (25,26), covariance models are constructed from multiple sequence alignments. Sequences are searched against a given covariance model using a three-dimensional dynamic programming algorithm, similar to a Smith-Waterman alignment but including base-pairing terms as well. RNA covariance models have the advantages of high sensitivity, high specificity, and general applicability to any RNA sequence family of interest, obviating the need for custom-written software for each RNA family. However, covariance model dynamic programming algorithms are almost prohibitively CPU-intensive. A tRNA covariance model identifies >99.98% of true tRNAs, with a false positive rate of <0.2/Mbp (22), but searching the human genome with a tRNA covariance model would take about nine and a half CPU-years (based on benchmarks on an SGI Indigo2 R4400/200 CPU, 140 SPECint92).

We describe here a program, tRNAscan-SE, that combines three tRNA search methods to attain the specificity of covariance model analysis with the speed and sensitivities of optimized versions of tRNAscan 1.3 and the Pavesi search algorithm. tRNAscan-SE detects 99-100% of true tRNAs, giving fewer than one false positive per fifteen billion nucleotides of random sequence, at approximately 1,000 to 3,000 times the speed of searching with tRNA covariance models. Additional extensions to tRNAscan-SE allow detection and accurate secondary structure prediction of unusual tRNA species including both prokaryotic and eukaryotic selenocysteine tRNA genes, as well as tRNA-derived repetitive elements and pseudogenes.

METHODS

tRNAscan-SE input consists of DNA or RNA sequences in FASTA format. tRNA predictions are output in tabular, ACeDB, or an extended format including tRNA secondary structure information. tRNAscan-SE does no tRNA detection itself, but instead negotiates the flow of information between three independent tRNA prediction programs, performs some post-processing, and outputs the results (Figure 1).

tRNAscan-SE works in three phases. In the first stage, it runs tRNAscan and the Pavesi algorithm on the input sequence. The first of these two programs is an optimized version of tRNAscan 1.3 (12). The other is an implementation of the Pavesi search algorithm (13) which we call EufindtRNA. Results from both programs are merged into one list of candidate tRNAs. Intron information from tRNAscan 1.3 is discarded because its intron predictions are typically unreliable. Analysis with the tRNA covariance model at a later stage (described below) allows non-ambiguous determination of intron boundaries.

In the second stage, tRNAscan-SE extracts the candidate subsequences and passes these segments to the covariance model search program *covels* (22). Seven flanking nucleotides on both sides of the candidate tRNAs are included in the subsequence in case the tRNA was truncated by the initial prediction. The *covels* search program applies a tRNA covariance model (TRNA2.cm) that was made by structurally aligning 1415 tRNAs from the 1993 Sprinzl database (27). 87 noncanonical "group III" sequences and 509 RNA sequences were removed from the complete 2011 sequence database as described in (22). To improve intron prediction, intron sequences were manually inserted into the Sprinzl alignment for 38 intron-containing tRNAs of known genomic sequence.

Finally, tRNAscan-SE takes predicted tRNAs that have been confirmed with *covels* log odds scores of over 20.0 bits, trims the tRNA bounds to those predicted by *covels*, and runs the covariance model global structure alignment program *coves* (22) to get a secondary structure

prediction. The tRNA isotype is predicted by identifying the anticodon within the *coves* secondary structure output. Introns are identified from this output as runs of five or more consecutive non-consensus nucleotides within the anticodon loop.

tRNAscan-SE uses heuristics to try to distinguish pseudogenes from true tRNAs, primarily on lack of tRNA-like secondary structure. A second tRNA covariance model (TRNA2ns.cm) was created from the same alignment, under the constraint that no secondary structure is conserved (this model is effectively just a sequence profile, or hidden Markov model). By subtracting a tRNA's similarity score to the primary structure-only model from that using the complete tRNA model, a secondary structure-only score is obtained. In Bayesian terms, this difference can be viewed as the evidence for the complete tRNA model, as opposed to a structureless, sequence-only pseudogene model. We observed that tRNAs with low scores for either component of the total score were often pseudogenes. Thus, tRNAs are marked as likely pseudogenes if they have either a score of less than 10 bits for the primary sequence component of the total score, or a score of less than 5 bits for the secondary structure component of the total score. Selenocysteine tRNAs are not checked by these rules since they have atypical primary and secondary structure. Final tRNA predictions are then saved in tabular, ACeDB, or secondary structure output format.

tRNAscan 1.4

tRNAscan-SE uses an optimized version of tRNAscan 1.3 (12) which we refer to as tRNAscan 1.4. The core algorithm is identical to tRNAscan 1.3. tRNAscan versions 1.3 and 1.4 have identical tRNA detection rates except in the case of ambiguous nucleotides occurring within the input sequence. There are implementation errors in tRNAscan 1.3's handling of ambiguous nucleotide codes. tRNAscan 1.4 conservatively calls ambiguous nucleotides as always forming base pairings in stems, and matching the highest scoring choice in consensus promoter matrices. This results in a high false positive rate for sequences containing a large number of ambiguous nucleotides. For our purposes, this is acceptable because the second stage covariance model analysis eliminates false positives. Several command line options were added to tRNAscan 1.4 for convenience in integration with tRNAscan-SE. Additional code changes were made to increase the robustness and speed of the program. These modifications result in roughly a 650-fold increase in search speed and no upper limit on input sequence size.

Implementation of EufindtRNA

EufindtRNA was implemented from the published algorithm by Pavesi and colleagues (13). The step-wise algorithm uses four probabilistic profiles for identifying basic tRNA features: "A box" nucleotide composition, "B box" composition, nucleotide distance between identified A and B boxes, and distance between identified B boxes and RNA polymerase III termination signals (four or more consecutive thymine nucleotides). In a search, an "intermediate" score is obtained by adding scores from identified A and B boxes to the score for the nucleotide distance between them. A final score is obtained by adding the intermediate score to the score for the distance to the nearest termination signal. If the final score is above a specific cutoff, the tRNA identity and location are saved.

Scores from over 30 example tRNAs described in the original publication match our implementation to within 0.1 log odds units. tRNAscan-SE uses a less selective version of the algorithm described above which does not search for transcription termination signals; instead, the intermediate score is used as a final cutoff. Also, the intermediate score cutoff is loosened slightly to -32.10 relative to the intermediate cutoff described in the original algorithm, -31.25. Although the program is designed for eukaryotic tRNA detection, we found EufindtRNA to be effective at

identifying prokaryotic tRNAs if the intermediate cutoff score is further adjusted. tRNAscan-SE has a specific option (-P) for scanning prokaryotic sequences which loosens the intermediate cutoff score to -36.0. Also, as with tRNAscan 1.4, ambiguous nucleotides are automatically assigned the best of the four non-ambiguous nucleotide scores at that position in the scoring matrices.

Selenocysteine tRNA Identification

The primary and secondary structure of selenocysteine tRNAs differ from canonical tRNAs in several respects, most notably an eight base pair acceptor stem, a long variable region arm, and substitutions at several well-conserved base positions. These differences make detection and accurate secondary structure prediction difficult using tRNA search programs geared towards canonical tRNAs. tRNAscan 1.3 fails to detect most selenocysteine tRNAs; the Pavesi algorithm incorporates a separate routine specifically for eukaryotic selenocysteines; and the TRNA2.cm covariance model barely detects selenocysteine tRNAs, giving scores just over the minimum cutoff of 20 bits, and in two cases, below the cutoff. tRNAscan-SE addresses this problem in the first-pass stage using EufindtRNA modifications, and in the second stage using selenocysteine tRNA-specific covariance models.

The first-pass scanner EufindtRNA implements a specialized subroutine described by Pavesi *et al.* (13) for identifying eukaryotic selenocysteine tRNAs (based on a B box score with a value between -2.2 and -3.6, and the motif GGTC(C/T)G(G/T)GGT appearing 36 nucleotides upstream of the B box). To similarly identify prokaryotic selenocysteine tRNAs, a subroutine was added to EufindtRNA which detects tRNAs with B box scores between -2.2 and -4.9, and a conserved sequence motif found in the anticodon loop of all known prokaryotic selenocysteine tRNAs (anticodon in bold): GG(A/T)(C/T)T**TCA**AA(A/T)CC. It is unclear if this motif will generalize well for new selenocysteine tRNAs, but it is conserved among the closely related *Escherichia coli* (Y00299), *Proteus vulgaris* (X14255), *Haemophilus influenzae* (U32753), and *Desulfomicrobium baculatus* (X75790) tRNAs, and in the more distant *Clostridium thermoaceticum* (Z26950) tRNA. After EufindtRNA has identified a candidate selenocysteine tRNA, it is passed to a eukaryotic or prokaryotic selenocysteine-specific covariance model. These two covariance models were developed by aligning selenocysteine tRNAs with inferred secondary structure information. Another program in the covariance model program suite, *coveb*, was used to build covariance models from the structure-annotated RNA sequence alignments. The five prokaryotic tRNAs noted above were used to build the prokaryotic selenocysteine model. Seven selenocysteine tRNAs from *Caenorhabditis elegans, Drosophila melanogaster, Xenopus laevis*, chicken, mouse, bovine, and human were used to build the eukaryotic model.

Databases Tested

tRNA detection rates were assessed primarily by searching two annotated databases: the 1995 release of the Sprinzl tRNA database (retrieved from ftp://ftp.ebi.ac.uk/pub/databases/trna; 27), and a tRNA sequence subset of Genbank (retrieved from the National Center for Biotechnology Information on 9/24/96). Genomic DNA was also searched from *Haemophilus influenzae* (v. 1.0, from the Institute for Genome Research (TIGR) ftp site at ftp://ftp.tigr.org/pub/data), *Mycoplasma genitalium* (rel. 10/9/95, TIGR ftp site), *Methanococcus jannaschii* (retrieved on 8/27/96, TIGR ftp site), *Saccharomyces cerevisiae* (rel. 4/24/96 from ftp://mips.embnet.org/yeast), *Schizosaccharomyces pombe* (completed cosmids retrieved from http://www.sanger.ac.uk/~yeastpub/svw/pombe.html on 9/30/96), *Caenorhabditis elegans* (completed cosmids retrieved 11/13/96 from ftp://ftp.sanger.ac.uk/pub/C.elegans_sequences), and Human (completed cosmids retrieved 8/28/96 from ftp://ftp.sanger.ac.uk/pub/human).

The Sprinzl tRNA database is the most comprehensive tRNA database, containing 2700 entries from a wide variety of organisms (27). It provides a set of trusted "true positives" for evaluating the sensitivity of a detection method. Since tRNAscan-SE was optimized for analyzing bacterial, archaeal, and eukaryotic genomic DNA, the 1144 tRNAs from species in these groups were chosen for analysis, excluding mitochondrial, chloroplast, and viral tRNA sequences. From this set, tRNAs that were used to train the TRNA2.cm covariance model (553 tRNAs in the 1993 release of the database) were removed to increase the independence between training and testing sequence data. Entries were restored to their correct primary sequence by combining the Sprinzl structural alignment with the atypical insertions that are annotated in a separate file. Introns, not present in the Sprinzl sequences or annotation, were not restored. Two prokaryotic sequences (DI1950, DR1420) were removed which would contain introns over 200 bp long had introns been included; none of the current tRNA search programs attempt to detect tRNA genes containing long group I or group II introns.

A broad sample of non-viral, non-organellar Genbank sequences indicating at least one tRNA in their feature tables was also analyzed. *C. elegans* and *S. cerevisiae* sequences were excluded since these genomic sequences were tested separately. The sequences were retrieved using the IRX query system at the National Center for Biotechnology Information (NCBI). Incomplete or synthetic tRNA sequences were removed, yielding a total of 1051 in the set. Genbank sequence annotation was not relied upon as a measure of the true number of tRNAs in the set since annotation quality is highly variable. Instead, tRNA detection by covariance model analysis was used to estimate the total number of tRNAs. Sequences with no tRNAs detected by covariance model analysis were manually examined to determine why annotated tRNAs were not detected, and six believed to be tRNAs were added to the covariance model-detected set. This method gave us a reasonable lower bound on the number of true positives in the Genbank subset.

"Random" Sequence Data

Two types of random sequence databases were created to test false positive rates. The first database is generated by a fifth order Markov chain based on six-mer frequencies within the first 54 Mbp of genomic sequence from the *C. elegans* genome project. Two thousand cosmid-sized sequences, 50 kilobases (Kbp) each, were generated based on these frequencies, totaling 100 Mbp of random sequence which is tRNA-free. The second random database was created to roughly simulate the human genome in size and GC content. Not enough human genomic sequence is available to parameterize a fifth order Markov chain model, so human sequence was simulated based on isochore proportion and %GC content. Ten thousand 300 Kbp sequences were generated, each one with a GC content approximating one of the five isochore types (L1 or L2 = 40% GC, H1 = 45% GC, H2 = 49% GC, H3 = 53% GC; 28). The isochore identities for these random sequences were chosen to approximate the proportion each isochore represents in the human genome (L1 + L2 60%, H1 20%, H2 10%, H3 5%). The remaining 5% of the human genome attributed to ALU-type repeat elements were not included since ALU sequences were tested separately (the absent 5% was distributed proportionally among the other isochore types).

Implementation & Online Analysis

tRNAscan-SE was written in Perl. The implementation of the Pavesi algorithm (13), EufindtRNA, was written in C. A single package of the UNIX-based programs used by tRNAscan-SE is available at http://genome.wustl.edu/eddy/. All analysis times given are for a Silicon Graphics Indigo2 R4400 200 Mhz workstation. A web server is available for on-line tRNA analysis at http://genome.wustl.edu/eddy/tRNAscan-SE/.

RESULTS

A summary of the overall sensitivity, selectivity, and search speed for the four tRNA search programs tested is shown in Table 1. The number of true positives is based on the percentage of tRNAs detected within a test set taken from the Sprinzl tRNA database (see Table 2). The false positive rate is based on analysis of randomly generated sequence data (Table 4). The search speeds for the various programs are shown for a scan of the current *C. elegans* genomic sequences averaging 30 kilobases per clone. tRNAscan 1.3 search speed decreases approximately linearly with length. Search speed for tRNAscan-SE is approximately constant, but varies based on tRNA density within the sequence.

Sensitivity

tRNAscan-SE was shown to be more sensitive than tRNAscan 1.3 by several measures, the first being a search of the Sprinzl and Genbank databases subsets (Table 2). In the Sprinzl test set, tRNAscan-SE detected 586 of 589 known tRNAs (99.5%), versus 560 of 589 (95.1%) for tRNAscan 1.3. Of all 1144 non-organellar tRNAs in the complete Sprinzl database, tRNAscan-SE fails to recognize seven. One was a eukaryotic sequence from *Trypanosoma brucei* (Sprinzl ID DT6050, Genbank TBTRNA3) which has been previously noted by Pavesi *et al.* (13) as being missed by both tRNAscan 1.3 and the Pavesi search algorithm. The other six tRNAs missed by tRNAscan-SE were from various eubacteria (Sprinzl ID's: DA1543, DE2180, DG1351, DG1482, DS1250, RG1380). Several of these undetected tRNAs appear to be irregular in source or function. DE2180 is derived from DNA from the cyanelle (a photosynthetic organelle) of the unicellular eukaryote *Cyanophora paradoxa* and is thus misclassified as eubacterial in the database.

DG1482 and RG1380 both contain substitutions of four highly conserved bases within the TΨC loop, an indication that the tRNAs are probably used in synthesis of the peptidoglycan instead of protein translation (29). All seven of these atypical tRNAs were detected using covariance model analysis. The tRNA covariance model search does miss two tRNAs within the 1144-member Sprinzl database subset, both selenocysteine tRNAs (Sprinzl ID DZ1430 & DZ7742) that pass below the 20.0 bit cutoff at 0.60 and 14.19 bits, respectively. EufindtRNA, designed to search eukaryotic sequences exclusively, shows improved sensitivity for eukaryotic tRNAs (98.6%) over tRNAscan 1.3 (95.0%), but is still slightly less sensitive than tRNAscan-SE (100%). Over the three phylogenetic domains, tRNA covariance model analysis appears to be the most sensitive detection method, yet tRNAscan-SE trails by as little as one third of one percentage point.

Searching the Genbank subset sequences which contain less reliable tRNA annotation, tRNAscan-SE detects 98.5% of the 1462 tRNAs verified by either covariance model analysis or visual inspection, whereas tRNAscan 1.3 has a 93.4 % detection rate (Table 2). All prediction discrepancies were visually inspected. Of the 18 tRNAs that covariance model analysis detected but were missed by all three other methods, all had scores over 36 bits, and were annotated in the Genbank entries. The two tRNAs detected by tRNAscan-SE but missed by covariance model analysis were a selenocysteine tRNA (CTTRSEL; same as previously noted Sprinzl DZ1430 tRNA), and a long tRNA from *Haloferax volcanii* (HALTGW) whose 104 bp intron caused the tRNA to exceed the maximum total length limit for normal tRNA covariance model analysis (150 bp). Of the 9 sequences annotated as tRNAs but missed by all four detection methods, four have large group I or group II introns of 241 bp or larger (ANATGL, SSU10482, PHU29955, SYOTRNLUAA), and five appear to have either sequencing errors or modified bases which appear in the Genbank annotation but not in the sequence (corresponding tRNAs within the Sprinzl database were identified correctly by all four detection methods). Because of sequence discrepancies between the Genbank sequences and corresponding Sprinzl entries, these five Genbank tRNAs were not included in the 1462-member test set.

Genome Analysis

Another measure of sensitivity was derived from searching complete or partial genomic sequence data from eubacterial, archaebacterial, yeast, and *C. elegans* sequencing projects (Table 3). For *Mycoplasma genitalium*, 33 tRNAs were noted in the published (30) and on-line gene identifications (http://www.tigr.org/tdb/mdb/mgdb/mgdb.html), whereas 36 tRNAs were detected by three tRNA detection methods (tRNAscan 1.3, tRNAscan-SE, covariance model analysis). The three tRNAs not appearing in the literature are for Arg (anticodon: CCT, bounds: 306615-306686, upper strand), Leu (anticodon: CAA, bounds: 448783-448861, upper strand), and Leu (anticodon: GAG, bounds: 446265-446181, reverse strand). For the completed *Haemophilus influenzae* genome, 56 tRNAs are noted in the literature (31) and on-line gene identifications (http://www.tigr.org/tdb/mdb/hidb/hidb.html). tRNAscan-SE and covariance model analysis both identify the tRNAs noted in the literature, plus two potentially novel tRNAs not noted in the literature: SelCys (anticodon: TCA, bounds: 753291-753201, reverse strand), and Leu (anticodon: GAG, tRNA bounds: 1576453-1576372, intron bounds: 1576419-1576408, reverse strand). The first is a selenocysteine tRNA and the other appears to be either a pseudogene or a true tRNA containing a short intron. The selenocysteine tRNA identification is not unexpected; BLAST searches identify two enzymes in the selenocysteine insertion pathway, as well formate dehydrogenase containing a 'UGA' selenocysteine-insertion codon. The evidence for the other potentially novel tRNA is less certain. The short 12 bp "intron" would presumably require protein-splicing to generate a functional tRNA, a feature that would be novel among eubacterial tRNAs. However, the covariance model score of 36.88 bits for the tRNA is well above the minimum cutoff of 20 bits, indicating that the sequence is likely to have evolutionary homology with tRNA. It is possible that it is a pseudogene. tRNAscan 1.3 identifies 55 of the 56 tRNAs noted in the literature (Gly-B, by TIGR nomenclature, is not detected), and does not detect either of the novel tRNAs detected by tRNAscan-SE and covariance model analysis.

The genomic sequence of the archaebacterium *Methanococcus jannaschii* was also analyzed. Both tRNAscan-SE and covariance model analysis identified all 37 tRNAs as given in the literature (32). tRNAscan 1.3 identified 36 of the 37 tRNAs, missing the single selenocysteine tRNA in the set. We also scanned the recently completed genomic sequence of the budding yeast *Saccharomyces cerevisiae* (12 Mbp). The covariance model search took 14 days to complete, and produced 275 tRNAs. Based either on inspection for ability to form correct tRNA secondary structure, or exact identity with previously characterized yeast tRNAs, we believe 274 predicted tRNAs are true tRNAs, and one is a pseudogene with an 7 bp 5' truncation. One of these 274 tRNAs was missing from the yeast genome project web site annotation (http://speedy.mips.biochem.mpg.de/mips/yeast), but this is probably an oversight since a tRNA of identical sequence is correctly annotated elsewhere in the genome (tRNA_i_S (GCT)LR2). tRNAscan-SE took 19 minutes and detected the same 275 tRNAs found by covariance model analysis. EufindtRNA found the same 275 tRNAs in just over one minute. tRNAscan 1.3 took about 10 hours to complete, and missed 4 (2 pairs identical in sequence) of the 274 true tRNAs found by the other three methods. The 4 Mbp of available genomic sequence from *Schizosaccharomyces pombe* (fission yeast) was also analyzed. tRNAscan-SE and covariance model analysis both predict 48 tRNAs. tRNAscan 1.3 identifies 45 of the 48 predicted by covariance model analysis (2 of 3 missed were identical in sequence), whereas EufindtRNA identifies 46 of the 48 total tRNAs.

Finally, we scanned the largest set of genomic sequence currently available, 58.4 Mbp from the *C. elegans* genome project. Since only a handful of the tRNAs detected have been previously published in the literature, we again relied on covariance model detection of tRNAs as our best measure for "true" tRNAs. Conflicts in tRNA predictions between tRNAscan 1.3, tRNAscan-SE

and covariance model analysis were all examined manually for highly conserved primary sequence motifs and proper secondary structure. As most tRNA species are multicopy in eukaryotes, BLAST similarity searches were used to help discern "false positives" from pseudogenes. We define false positives as predicted tRNAs which do not appear to be evolutionarily derived from true tRNAs. These false positives are assessed by failure to form recognizable tRNA secondary structure and the lack of related tRNAs elsewhere in the genome. Pseudogenes, on the other hand, usually have at least partial tRNA secondary structure, plus clear deletions or insertions relative to at least one related, intact tRNA elsewhere in the genome. tRNA-derived mobile elements also have recognizable primary sequence similarity to tRNAs, although most have poor tRNA secondary structure similarity. Of the 403 complete tRNAs detected by covariance model analysis, tRNAscan-SE detected all 403 tRNAs (100%), whereas tRNAscan 1.3 detected 389 (96.5%), and EufindtRNA found 400 (99.2%).

Taken together, the data analyzed from the *M. genitalium*, *H. influenzae*, *M. jannaschii*, *S. cerevisiae*, *S. pombe*, and *C. elegans* genomes, 100% of the 856 tRNAs detected by covariance model analysis were found by tRNAscan-SE. tRNAscan 1.3 detected 831, missing 25 tRNAs identified by covariance models, a 97.1 % detection rate. EufindtRNA detects 93.5% of the 856 tRNA set, but if only eukaryotic genomes are considered, the program finds 720 of 725 (99.3%).

Selectivity

While the "sensitivity" of an algorithm is measured by the proportion of true positives identified in reference sequences, a method's "selectivity" is measured by its ability to avoid misidentifying unrelated sequences as true tRNAs. Increased sensitivity is usually gained at the expense of an increased false positive rate. A rate of one false positive per five to ten million bases of sequence has, in the past, been acceptable since the total amount of uncharacterized or non-protein coding sequence in the databases has been relatively small. However, with the advent of whole-genome sequencing projects on the megabase scale, this false positive rate is of much greater concern.

Assessing the ability of an algorithm to discriminate between true and false positives using biological sequence data can be difficult. At false positive rates of less than one per million bases, there is not enough well annotated sequence in the public databases to give a reliable indication of an algorithm's true performance. Even for the data that is available, it is uncertain whether or not an accurate prediction has been made in the absence of biochemical experimental evidence. An alternative strategy is to generate random nucleotide sequence which is known to have no biologically-derived genes. An unlimited amount of random sequence can be generated based on a general or species-specific genomic nucleotide frequency. Each identification of a tRNA gene in this random sequence can then be confidently counted as a false positive. False positives due to biologically-derived repetitive elements or pseudogenes are not taken into account in these synthetic test sequences, and must be addressed separately.

We generated two types of random sequence sets to simulate the size and GC content of the *C. elegans* and human genomes (100 million and 3 billion bases of random sequence, respectively, as described in Methods). The number of false positives found with each algorithm appear in Table 4 along with false positive rates from actual genomic sequence (discussed below). Analysis of the simulated genomes gave consistent false positive rates between the various trials, at approximately 0.40 false positives per million bases for tRNAscan 1.3, a little more than half that for EufindtRNA, and zero for both tRNAscan-SE and covariance model analysis. In ten independent *C. elegans* genome simulations, an average of 42.5 tRNAs were identified by tRNAscan 1.4. The sequences for the false positive tRNAs were saved and analyzed with the original tRNAscan 1.3 program to confirm that false positives were due to the tRNAscan 1.3 algorithm, not the modifications introduced in tRNAscan 1.4. EufindtRNA misidentified an average of 26 false

positives per simulated *C. elegans* genome. Both tRNAscan-SE and the tRNA covariance model searches found zero positives for every trial (only one genome simulation was searched with the tRNA covariance model due to the extreme CPU demands). As seen in Table 5, minor differences among analysis times for the various methods for microbial genomes become substantial when analyzing larger eukaryotic genomes. Analysis of the single *C. elegans* genome simulation with covariance models required almost four CPU-months.

For the five human genome simulations, tRNAscan 1.4 produced an average of 1118 false positives per genome (had tRNAscan 1.3 been used, it would have taken almost half a CPU year per trial). EufindtRNA searched the simulated genomes in just over seven hours per trial, giving an average of 684 falsely predicted tRNAs for each. Had we searched the entire 3 billion nucleotide human genome simulation with tRNA covariance model analysis, it would have taken over nine CPU-years for each trial (Table 5). Based on the histogram of covariance model scores against 500 million bases of simulated human sequence data (not shown), we estimate that the tRNA covariance model search of the simulated human genome would have produced zero false positives. tRNAscan-SE required an average of a day and a half to scan each of the three billion nucleotide test sets, and produced no false positives in any of the five trials (the exact same sequences were used as in the trials described above for tRNAscan 1.4 and EufindtRNA).

A concern not addressed by the random sequence genome simulations is the “false positive” rate caused by certain classes of SINEs that are suspected to be derived from tRNA genes (4,5). These elements have similarity to known tRNA genes and contain well conserved RNA polymerase III internal A and B box promoters. To assess tRNAscan-SE’s ability to identify and exclude these types of pseudo-tRNAs, the repeat element database *Repbase* maintained by Jerzy Jurka (ftp://ncbi.nlm.nih.gov/repository/repbase) was scanned. Of the reference sequences searched, tRNAscan-SE did not produce any false positive tRNA identifications. Covariance model analysis, however, did misidentify 12 of 775 rodent B2 SINE sequences and two ALU-like sequences (bovine ALU-like repetitive element & rat ALU type III-like repetitive element), all with scores between 20 and 28 bits. Rat identifier (ID or R.dre.1) sequences, also known to have high similarity to alanine, proline, and other tRNAs, were searched within Genbank and dbEST (database of expressed sequence tags, 33). tRNAscan-SE misidentified four rat ID element sequences total, one from Genbank (RATRSIDH) and three from dbEST (R46943, R46943, R82886). The extreme sensitivity of covariance model analysis is also unable to distinguish between these SINEs and true tRNAs, giving bit scores between 24.5 and 33.1 bits. tRNAscan 1.3 requires strong adherence to secondary structure rules, thus does not call any of these pseudogenes as tRNAs. The rest of *Repbase*, including consensus and database collections of ALU, L1, THE, MIR, MIR2, THR, and B1 repetitive elements, were also searched with tRNAscan-SE, giving no other false positives.

The selectivity of tRNAscan has already affected genome sequence annotation detrimentally. In 58.4 Mbp of *C. elegans* genomic sequence, tRNAscan 1.3 produced 29 tRNAs which were judged to be false positives (0.50 FP /Mbp) based on searching with the tRNA covariance model, visual inspection of secondary structure, and lack of primary sequence similarity to any other tRNAs within the genome. Since both the Washington University Genome Sequencing Center (St. Louis) and the Sanger Center (Cambridge, UK) used tRNAscan 1.3 in semi-automated sequence annotation until very recently, 16 of these 29 false positives are annotated as tRNAs in finished, submitted Genbank entries. This false positive rate is very close to that seen in the random *C. elegans* genome simulation (0.42 FP/Mbp), giving additional confidence to the estimates based on simulated sequence data.

tRNAscan-SE produced no obvious false positives in the C. elegans genomic sequence, but did identify 8 tRNAs that were judged to be possible pseudogenes by manual inspection (Table 3). Eleven other tRNAs were automatically identified as pseudogenes via primary or secondary structure scores that fell below minimum values described in the methods. All 19 pseudogenes had

strong similarity to other tRNAs within the genome, and contained unusual features such as 3-16 bp truncations of the 5' end of the gene, or other large insertions or deletions within the sequence. One could consider detection of these possible pseudogenes a desirable feature of tRNAscan-SE's sensitivity. Further studies of these unusual tRNAs may help better elucidate aspects of genome dynamics, genetic element mobility, and evolution.

Selenocysteine tRNA Detection

There are not enough selenocysteine tRNA sequences to properly evaluate tRNAscan-SE's selenocysteine detection accuracy. Three selenocysteine tRNAs (one each from *H. influenzae*, *M. jannaschii*, and *C. elegans*) were detected in recent genome sequence data. The *H. influenzae* tRNA, previously unrecognized in the literature, was detected by the prokaryotic selenocysteine-specific routines and covariance model. The tRNA from the distantly related *M. jannaschii*, however, was detected by the standard EufindtRNA algorithm and general tRNA covariance model. The failure of the specialized routines may have been due in part to the fact that this is the first and only archaebacterial selenocysteine tRNA available to date. For the remaining non-archaeal selenocysteine tRNAs, use of the specialized models boosts covariance model scores from the 20-40 bit range to 45-72 bits. Since accurate tRNA secondary structure prediction relies on correct alignment of the tRNA sequence to the covariance model, use of selenocysteine-specific models for these tRNAs improves the accuracy of structure predictions. A search of the non-redundant database (nrdb) maintained at NCBI revealed no new selenocysteine tRNAs from species for which there was no previously noted sequence.

Intron Detection

tRNAscan-SE correctly predicted the introns for the 13 species of intron-containing tRNAs in the *S. cerevisiae* genome (34). tRNAscan 1.3 often gives multiple intron predictions for each tRNA, making correct placement uncertain. EufindtRNA does not attempt to predict intron boundaries at all (13).

Detection of tRNAs containing long introns, usually group I or group II, is problematic. The default maximum tRNA length for tRNAscan-SE is 192 bp, but this can be increased (option -L <max length>) to allow searches with no practical limit on tRNA length. In the first phase of tRNAscan-SE, EufindtRNA searches for A and B boxes of the specified maximum distance apart, and passes only the 5' and 3' tRNA ends to covariance model analysis for confirmation (removing the bulk of long intervening sequences). Using this option, tRNAscan-SE was able to detect three of the four long tRNAs initially missed by all four methods in the Genbank tRNA subset search (the fourth tRNA was undetectable with EufindtRNA even with the intron removed before analysis). Group I or II introns in tRNAs tend to occur in positions other than the canonical position of protein-spliced introns, so tRNAscan-SE mispredicts the intron bounds and anticodon sequence for these cases. 5' and 3' tRNA bounds were correct for all three unusual tRNAs.

Performance on Mitochondrial tRNAs

Although tRNAscan-SE was designed with non-organellar tRNA detection in mind, we also tested it on a complete mitochondrial genome, that of *Podospora anserina* (Genbank ID PANMTPACGA). tRNAscan-SE detected 22 of the 27 annotated tRNAs (81.5 %), tRNAscan 1.3 detected 18 of 27 (66.7%), and covariance model analysis detected all 27 tRNAs (Table 3). Since organellar genomes are usually small, the computational demand of covariance model analysis

alone (without the use of fast first-pass scanners) is not prohibitive. For this reason, tRNAscan-SE can be run in covariance model analysis-only mode (-C option) for maximum sensitivity, bypassing dependence on tRNAscan 1.4 and EufindtRNA. This mode gives the same results as would be obtained by running the covariance model search program alone, but in addition, produces annotated tRNA output identical in format to that found in the default tRNAscan-SE search mode.

DISCUSSION

Speed, Sensitivity, and Selectivity

The most sensitive and selective tRNA detection method that we are aware of utilizes probabilistic RNA covariance models (22), which are based on stochastic context-free grammar techniques. However, searching with covariance models has two drawbacks. First, it is extremely CPU-intensive, requiring days to weeks of processor time to scan megabase-size genomic data from higher eukaryotes. Second, the general nature of the approach hampers output of tRNA-specific feature information such as anticodon, isotype, and intron position. Our goal in the development of tRNAscan-SE was to produce a practical (i.e. *fast*) application of stochastic context-free grammar-based RNA analysis methods with sensitivity and selectivity as close as possible to using native covariance model searches. tRNAscan-SE achieves this goal.

tRNAscan-SE increases tRNA covariance model search speed by 1,000 to 3,000 fold while offering nearly equal sensitivity and slightly improved selectivity. Selenocysteine tRNA detection features are built into tRNAscan-SE, including modifications to EufindtRNA and the use of selenocysteine tRNA covariance models. With these additions, tRNAscan-SE correctly identifies both of the selenocysteine tRNAs in the Sprinzl database not detected by normal covariance model analysis. The Genbank version of one of these two selenocysteine tRNA sequences, CTTRSEL from *C. thermoaceticum*, was also detected within the Genbank tRNA subset (the other selenocysteine tRNA was not in the Genbank subset).

tRNAscan-SE also extends the maximum length of tRNAs detectable to almost any length. In covariance model analysis, search time increases as the square of the maximum tRNA length, so the search window has typically been limited to 150 bp. In tRNAscan-SE, the first-pass scanners define the approximate bounds of a tRNA, and for tRNAs with very long introns, intervening sequences can be cut out based on the first-pass analysis. This allows detection of rare, abnormally long tRNAs without greatly increasing the overall average search time. In the Genbank subset, tRNAscan-SE detected four tRNAs (HALTGW plus three detected with the -L option) whose introns, ranging from 104 to 850 bp, exceeded the normal length limit for covariance model detection.

tRNA False Positives & Pseudogenes

Of the 5,591 total false positives identified by tRNAscan 1.4 in 15 gigabases of simulated human sequence (Table 4), in only six instances did it agree with EufindtRNA (relaxed parameters) in falsely identifying a sequence as a tRNA. The majority of false positives found by tRNAscan 1.4 seem to have tRNA-like secondary structure but lack similarity to conserved tRNA primary sequence. EufindtRNA, on the other hand, identifies correctly spaced primary sequence promoter elements, yet tends to err because it does not check for proper tRNA secondary structure.

These observations hold up on examination of false positives from actual genomic sequence from *C. elegans*. Most of the 29 false positives identified by tRNAscan 1.3 were discarded by covariance model analysis because of the lack of primary sequence similarity to the general tRNA model. EufindtRNA, on the other hand, more commonly identifies pseudogene tRNA fragments, SINE-like repetitive elements, or other tRNA-like sequences containing A and B boxes (Table 3). Pseudogenes are recognizable since part of the sequence is very similar to other intact tRNAs, in spite of truncations or large insertions elsewhere in the pseudogene. However, tRNA secondary structure in pseudogenes and SINE-like elements tends to be lost more quickly than primary sequence promoter elements. This may not be surprising in light of the observation that portions of tRNA sequences are thought to help provide mobility for some tRNA-derived repetitive elements (35). Since EufindtRNA (relaxed parameters) only looks for canonical promoter regions, it is prone to finding these instances of pseudogenes and repetitive elements with tRNA promoters in the absence of structural tRNA features.

To some extent, covariance model analysis is also apt to identify truncated tRNAs and other tRNA-derived sequence elements. The minimum cutoff score of 20 bits has been set to include outlying tRNAs with low overall homology to the general tRNA model. However, if a part of a high-scoring tRNA is truncated, the score may be much lower, but still exceed the 20 bit threshold. The most extreme example of this occurs with a tRNA in the *C. elegans* cosmid W03A3. The tRNA has 100% identity with tRNAs on at least four other cosmids, except for a truncation of the first 16 bases that removes the 5' side of the aminoacyl acceptor stem and the first half of the A box promoter sequence (part of the D-loop). tRNAscan 1.3 did not detect this pseudogene because of the lost base pairings in the D-loop and aminoacyl stems, whereas EufindtRNA could not locate the A box promoter sequence. Covariance model analysis similarly identified three other pseudogenes that neither tRNAscan 1.3 nor EufindtRNA found: one appears to have a 13 bp truncation relative to tRNAs in two other cosmids; one has a peculiar 21 bp insertion in the middle of the A box promoter sequence that makes three near-perfect repeats of the 7-mer "GTCGCGA"; and one cosmid has a pseudo tRNA containing a 55 bp insert in the anticodon loop that does not appear to be a true intron. Since none of these were identified by either tRNAscan 1.3 or EufindtRNA, tRNAscan-SE necessarily does not detect them.

tRNAscan-SE does, however, detect 19 other tRNA-like sequences that are identified by EufindtRNA and "confirmed" by covariance model analysis (scores greater than 20 bits). These may or may not be pseudogenes. Nine of these involve 5' truncations of 3 to 15 nucleotides relative to other tRNAs in the nematode. It is impossible to determine by computational analysis alone if these are functional tRNAs or inactive pseudogenes. In either case, it is important to be aware of these possible tRNA pseudogenes for possible further experimental and/or computational study. Elucidating a common transpositional mechanism for preferential loss of the 5' end of these tRNAs is a question of interest.

Conclusion

tRNAscan-SE has been designed with the demands of human genome analysis in mind, but can be used for any DNA sequence. We estimate that tRNAscan-SE will detect about 99.5 % of the true tRNAs in the human genome, give zero false positives (except for tRNA-derived SINEs and tRNA pseudogenes), and take approximately 36 hours.

tRNAscan-SE demonstrates that general RNA structural profiles, covariance models, can be used as the basis for very sensitive RNA similarity searching. The primary limitation is speed. Although the strategy of using fast first-pass tRNA scanners in combination with second-stage covariance model analysis is effective here, this is not an attractive general strategy for searching for other RNA gene family members. Except for group I introns (36), there are no fast, specialized

algorithms for detection of other RNA gene families, and much effort is required for creating these highly specialized new programs. Further work will focus on algorithmic development of covariance model search methods that will reduce both time and memory requirements, allowing faster searches for larger RNA genes without the need for first-pass screens.

REFERENCES

1. Hatlen, L. and Attardi, G. (1971) *J. Mol. Biol*., **56**, 535-553.
2. Oba, T., Andachi, Y., Muto, A. and Osawa, S. (1991) *Proc. Nat. Acad. Sci., U.S.A.*, **88**, 921-925.
3. Kanna, A., Ohama, T., Abe, R. and Osawa, S. (1993) *J. Mol. Biol.*, **230**, 51-56.
4. Daniels, G.R. and Deininger, P.L. (1984) *Nature*, **317**, 819-822.
5. Deininger, P.L. (1989) In Berg, D.E. and Howe, M.M. (eds.), *Mobile DNA*. ASM, Washington, pp. 619-636.
6. Dandekar, T. and Hentze, M.W. (1995) *Trends In Genet.*, **11**, 45-50.
7. Staden, R. (1980) *Nucl. Acids Res.*, **8**, 817-825.
8. Paolella, G. and Russo, T. (1985) *Comput. Appl. Biosci.*, **1**, 149-151.
9. Shortridge, R.D., Pirtle, I.L. and Pirtle, R.M. (1986) *Comput. Appl. Biosci.*, **2**, 13-17
10. Marvel, C.C. (1986) *Nucl. Acids Res.*, **14**, 431-435.
11. Wozniak, P. and Makalowski, W. (1990) *Comput. Appl. Biosci.*, **6**, 49-50.
12. Fichant, G.A. and Burks, C. (1991) *J. Mol. Biol*., **220**, 659-671.
13. Pavesi, A., Conterio, F., Bolchi, A., Dieci, G. and Ottonello, S. (1994) *Nucl. Acids Res*, **22**, 1247-1256.
14. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol*., **215**, 403-410.
15. Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci.*, **85**, 2444-2448.
16. Saurin, W. and Marliere, P. (1987) *Comput. Appl. Biosci*., **3**, 115-120.
17. Staden, R. (1988) *Comput. Appl. Biosci*., **4**, 53-60.
18. Gautheret, D., Major, F. and Cedergren, R.J. (1990) *Comput. Appl. Biosci.*, **6**, 325-331.
19. Sibbald, P.R., Sommerfeldt, H. and Argos, P. (1992) *Comput. Appl. Biosci*., **8**, 45-48.
20. Laferrire, A., Gautheret, D. and Cedergren, R.J. (1994) *Comput. Appl. Biosci.*, **10**, 211-212.
21. Billoud, B., Kontic, M. and Viari, A. (1996) *Nucl. Acids Res.*, **24**, 1395-1403.
22. Eddy, S.R. and Durbin, R. (1994) *Nucl. Acids Res.*, **22**, 2079-2088.
23. Grate, L., Herbster, M., Hughey, R., Haussler, D., Mian, I.S. and Noller, H. (1994) *Proceedings, Second International Conference on Intelligent Systems for Molecular Biology*, **2**, 138-146.
24. Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C. and Haussler, D. (1994) *Nucl. Acids Res.*, **22**, 5112-5120.
25. Gribskov, M., Luthy, R. and Eisenberg, D. (1990) *Methods Enzymol.*, **183**, 146-159.
26. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) *J. Mol. Biol.*, **235**, 1501-1531.
27. Steinberg, S., Misch, A. and Sprinzl, M. (1993) *Nucl. Acids Res.*, **21**, 3011-3015.
28. Bernardi, G. (1993) *Gene*, **135**, 57-66.
29. Green, C.J. and Vold, B.S. (1993) *J. Bacteriol.*, **175**, 5091-5096.
30. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelly, J.M., *et al.*(1995) *Science*, **270**, 397-403.
31. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995) *Science*, **269**, 496-512.
32. Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., *et al.* (1996) *Science*, **273**, 1058-1073.
33. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) *Nat. Genet.*, **4**, 332-333.
34. Westaway, S.K. and Abelson, J. (1995) In Soll, D. and RajBhandary, U.L. (eds.), *tRNA: Structure, Biosynthesis, and Function*. ASM, Washington, pp.79-80.
35. Keeney, J.B., Chapman, K.B., Lauermann, V., Voytas, D.F., Astrom, S.U., von Pawel-Rammingen, U., Bystrom, A. and Boeke, J.D. (1995) *Mol. Cell. Biol.*, **15**, 217-226.

36. Lisacek, F, Diaz, Y. and Michel, F. (1994) *J. Mol. Biol.*, **235**, 1206-1217.

Table 1. Overall detection rates of tRNA search programs. True positives are based on detection rates within a non-organellar, non-viral subset of the Sprinzl tRNA database (see Table 2). False positive rates are estimates based on searches of randomly generated human sequence (see Table 4). Search speeds are from a search of 58.4 Mbp of *C. elegans* cosmid sequences on a Silicon Graphics Indigo2 R4400 200 Mhz workstation.

	True Positives (%)	False Positives (per Mbp)	Search Speed (bp/sec)
tRNAscan 1.3	95.1	0.37	400
EufindtRNA	88.8[1]	0.23	373,000
tRNA covariance model search	99.8	< 0.002	20
tRNAscan-SE	99.5	< 0.00007	30,000

[1]EufindtRNA is based on the Pavesi search algorithm which was designed to detect eukaryotic tRNAs only; searching only eukaryotic tRNAs, EufindtRNA has a 98.6% true positive detection rate (see Table 2)

Table 5. Analysis time in hours required for various complete genomes & tRNA search algorithms. Actual genome scan times are given for tRNAscan-SE and EufindtRNA (genome simulation times used for human). Estimated scan times are given for tRNAscan 1.3 (400 bp/sec) and tRNA covariance model analysis (tRNA CM; 20 bp/sec).

Complete Genome	Size (Mbp)	tRNAscan 1.3 (CPU hours)	EufindtRNA (CPU hours)	tRNA CM (CPU hours)	tRNAscan-SE (CPU hours)
P. anserina mitochondrion	0.1	0.14	< 0.001	2.8	0.019
H. influenzae	1.8	2.54	< 0.001	51	0.069
S. cerevisiae	12	16.7	0.02	333	0.33
C. elegans	100	139	0.15	2,780	1.8
Human	3,000	>4170	7.1	83,300	36.6

(NOT included)

Figure 1. Schematic diagram of tRNAscan-SE algorithm. Steps carried out by tRNAscan-SE are shown in ovals and rounded-edge boxes. tRNA selection and analysis performed by external programs are shown in rectangles.

Table 2. tRNA prediction within annotated database subsets. The detection rates for the Sprinzl tRNA database are broken down by phylogenetic domain. The Sprinzl subset tested contains only non-organellar, non-viral tRNAs which were not used in training of the tRNA covariance model. For the Sprinzl database subset, numbers in parentheses indicate percentage of correct tRNA identifications relative to total in the literature. The Genbank subset sequences were selected by retrieving non-organellar, non-viral, full-length tRNA sequences with 'tRNA' indicated in the feature field of the entry. Since Genbank tRNA annotation is less reliable, the numbers in parentheses for this row are the percentage of correct tRNA identifications relative to all tRNAs verified by either covariance model analysis or visual inspection.

Sequence Source	Literature	tRNAscan 1.3		EufindtRNA		tRNA CM		tRNAscan-SE	
	tRNAs	Tot	(%)	Tot	(%)	Tot	(%)	Tot	(%)
Sprinzl db (Archaea)	70	69	(98.6)	43	(61.4)[1]	70	(100)	70	(100)
Sprinzl db (Eubacteria)	240	226	(94.2)	205	(85.4)[1]	239	(99.6)	237	(98.7)
Sprinzl db (Eukarya)	279	265	(95.0)	275	(98.6)	279	(100)	279	(100)
Sprinzl db (total)	589	560	(95.1)	523	(88.8)	588	(99.8)	586	(99.5)
Genbank tRNA subset	1462	1366	(93.4)	760	(52.0)	1456	(99.6)	1440	(98.5)

[1] EufindtRNA is based on the Pavesi search algorithm (13) which was designed specifically to find only cytoplasmic eukaryotic tRNAs

Table 4. False positive rates for actual & simulated genomes. "Actual FP" rows contain false positives detected in actual genomic sequence. "Simulated FP" rows contain the false positives found in whole-genome scale random sequence simulations (10 trials for *C. elegans*, 5 for human). For tRNA covariance model searches (tRNA CM), only one random *C. elegans* and no human genome simulations were performed due to extreme CPU demands (ND=not done).

	Size (Mbp)	tRNAscan 1.3[1]		EufindtRNA		tRNA CM		tRNAscan-SE	
		FP	FP/Mbp	FP	FP/Mbp	FP	FP/Mbp	FP	FP/Mbp
S. cerevisiae									
Actual FP (completed genome)	12.0	4	0.33	10	0.83	0	< 0.08	0	< 0.08
C. elegans									
Actual FP (portion completed)	58.4	29	0.50	355	6.08	0	< 0.03	0	< 0.03
Simulated FP (total genome)	100	42.5	0.42	26	0.26	0	< 0.01	0	< 0.001
Human									
Actual FP (portion completed)	5.32	3	0.56	5	0.94	0	< 0.19	0	< 0.19
Simulated FP (total genome)	3000	1118	0.37	684	0.23	ND	--	0	< 0.00007

[1]Searches performed with tRNAscan 1.4, but all false positives verified with unaltered tRNAscan 1.3

Table 3. tRNAs identified in genomic databases by various search methods. "Literature" column represents the published number of tRNAs found within genomes. "Tot" columns indicate total number of tRNAs found in searches for each program. Numbers in parentheses in (%) columns indicate percentage of tRNAs detected relative to literature (*H. influenzae, M. jannaschii, P. anserina)*, or when published tRNA annotation is incomplete or uncertain (*M. genitalium, S. pombe, S. cerevisiae, C. elegans*), detection percentages are relative to total tRNAs found by tRNA covariance model analysis and supported by manual inspection. "FP" = false positives determined by covariance model analysis and manual inspection (these do not include pseudogenes that have strong similarity to known tRNAs). "pseudo" = tRNA identifications which appear to be pseudogenes containing 5' truncations of 3-16 bp, large insertions or deletions elsewhere, or other characteristics of tRNA-derived repetitive elements. "id pseudo" = tRNAs automatically identified by tRNAscan-SE as likely pseudogenes which have qualities similar to manually detected pseudogenes described above.

Sequence Source	Size	Literature	tRNAscan 1.3		EufindtRNA[1]		tRNA CM		tRNAscan-SE	
	(Kbp)	tRNAs	Tot	(%)	Tot	(%)	Tot	(%)	Tot	(%)
M. genitalium	580	33	36	(100)	19 +1 FP	(52.8)	36	(100)	36	(100)
H. influenzae	1,830	56	55	(98.2)	42 +2 FP	(73.7)	58	(103.6)	58	(103.6)
M. jannaschii	1,730	37	36	(97.3)	20 +1 FP	(54.0)	37	(100)	37	(100)
S. pombe (through 9/96)	4,176	--	45 +4 FP	(93.7)	46 +1 FP	(95.8)	48		48	(100)
S. cerevisiae	12,057	273	270 +4 FP	(98.5)	274 +10 FP +1 pseudo	(100)	274 +1 pseudo		274 +1 pseudo	(100)
C. elegans (through 11/13/96)	58,402	-- 16 FP	389 +29 FP	(96.5)	400 +355 FP +19 pseudo	(99.2)	403 +23 pseudo		403 +11 id pseudo +8 pseudo	(100)
P. anserina mitochondrion	100	27	18	(66.7)	11	(40.7)	27	(100)	22	(81.5)

[1] EufindtRNA is based on an algorithm (13) which was designed specifically to find only cytoplasmic eukaryotic tRNAs