W686–W689 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki366

The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs

Peter Schattner, Angela N. Brooks[1] and Todd M. Lowe*

Department of Biomolecular Engineering and the UCSC RNA Center, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA and [1]Division of Biological Sciences, Cell and Developmental Biology Section and Center for Molecular Genetics, University of California at San Diego, La Jolla, CA 92093, USA

Received January 12, 2005; Revised and Accepted February 28, 2005

ABSTRACT

Transfer RNAs (tRNAs) and small nucleolar RNAs (snoRNAs) are two of the largest classes of non-protein-coding RNAs. Conventional gene finders that detect protein-coding genes do not find tRNA and snoRNA genes because they lack the codon structure and statistical signatures of protein-coding genes. Previously, we developed tRNAscan-SE, snoscan and snoGPS for the detection of tRNAs, methylation-guide snoRNAs and pseudouridylation-guide snoRNAs, respectively. tRNAscan-SE is routinely applied to completed genomes, resulting in the identification of thousands of tRNA genes. Snoscan has successfully detected methylation-guide snoRNAs in a variety of eukaryotes and archaea, and snoGPS has identified novel pseudouridylation-guide snoRNAs in yeast and mammals. Although these programs have been quite successful at RNA gene detection, their use has been limited by the need to install and configure the software packages on UNIX workstations. Here, we describe online implementations of these RNA detection tools that make these programs accessible to a wider range of research biologists. The tRNAscan-SE, snoscan and snoGPS servers are available at http://lowelab.ucsc.edu/tRNAscan-SE/, http://lowelab.ucsc.edu/snoscan/ and http://lowelab.ucsc.edu/snoGPS/, respectively.

INTRODUCTION

Transfer RNA (tRNA) genes make up the single largest gene family. A typical eukaryotic genome contains hundreds of tRNA genes; the human genome contains an estimated 600 tRNA loci (1). Knowing the tRNA repertoire of an organism is important because it affects the codon bias seen in highly expressed protein-coding genes. The small nucleolar RNAs (snoRNAs) are involved at various stages of eukaryotic rRNA and small nuclear RNA (snRNA) biogenesis (2). In archaea, homologous classes of snoRNA-like small RNAs are involved in the biogenesis of rRNAs (3) and tRNAs (4,5). The two major families of snoRNAs are involved in guiding the two most common types of RNA modification: Box H/ACA snoRNAs are required for RNA pseudouridylation, while most of the C/D box snoRNAs guide RNA ribose methylation (2).

In a time when complete genomes are being rapidly sequenced, it is important to have an accurate means of tRNA and snoRNA gene identification. However, conventional gene finders that detect protein-coding genes do not detect genes for tRNAs and snoRNAs because RNA genes lack the sequence signals used by these programs (6). As a result, custom-designed gene finders have been required for the computational identification of non-protein-coding RNAs, such as tRNAs and snoRNAs (6).

Previously, we developed tRNAscan-SE (7), snoscan (8) and snoGPS (9) for the detection of tRNAs, methylation-guide snoRNAs and pseudouridylation-guide snoRNAs, respectively. tRNAscan-SE has been applied to all completed genomes, including the human genome (T. M. Lowe, manuscript in preparation; see http://lowelab.ucsc.edu/GtRNAdb/). Snoscan has detected scores of methylation-guide snoRNAs in eukaryotic (8,10,11) and archaeal genomes (12) and snoGPS has identified novel, experimentally verified pseudouridylation-guide snoRNAs in yeast (9) and mammals (P. Schattner, S. Barberan and T. M. Lowe, manuscript in preparation).

However, applying these programs to search for tRNAs and snoRNAs is not always straightforward. The programs need to be downloaded and installed on the user's UNIX-compatible computer system, and various options and data files must be configured or specified. Consequently, these programs have not been applied as widely or effectively as possible. In order to facilitate the application of these programs to a wider range of genomic searches, we have implemented a web server

*To whom correspondence should be addressed. Tel: +1 831 459 1511; Fax: +1 831 459 3139; Email: lowe@soe.ucsc.edu
Correspondence may also be addressed to Peter Schattner. Email: schattner@soe.ucsc.edu

interface for them. With these servers, research biologists will be able to apply tRNAscan-SE, snoscan and snoGPS to standard tRNA and snoRNA searches without having to install the programs on their local systems.

DESCRIPTION

The tRNAscan-SE, snoscan and snoGPS servers are accessed via the Lowe Lab Web server Interface at http://lowelab.ucsc.edu/tRNAscan-SE/, http://lowelab.ucsc.edu/snoscan/ and http://lowelab.ucsc.edu/snoGPS/, respectively. All three servers have similar user interfaces with differences limited to search-specific configuration options.

Each user interface consists of four major components:

(i) Search mode selection.
(ii) Query sequence selection.
(iii) Target sequence selection (not applicable for tRNAscan-SE).
(iv) Configuration of search-mode and options for displaying results.

The search mode determines which probabilistic model to be used in searches—each model is based on tRNA or snoRNA training data from selected species or phylogenetic groups (i.e. mammals, yeasts and archaea). If no explicit model for the species of interest is available in the user interface, specifying either a general model or a model from a related species generally yields good results. Different search modes can offer varying speed and sensitivity.

Query sequence selection is used to specify the sequences to be searched for tRNAs or snoRNAs. Raw or formatted sequence data can be pasted directly into the query sequence box or can be uploaded from a local file. Each server also has a set of program-specific search and output-display options. Examples include choice of genetic code for determining tRNA isotype for tRNAscan-SE, limits on the distances between some of the sequence motifs (e.g. C and D boxes) in snoscan and limits on the minimum number of base pairings in the guide region for snoGPS. In addition, each program has adjustable cutoff scores enabling tradeoffs between scan sensitivity and specificity. In most cases, the default parameter choices will be satisfactory and should be selected, especially by new users. However, more experienced users are able to exert some control over the program's results by manipulating these parameters.

The default output for tRNAscan-SE includes the location of each identified tRNA, the predicted anticodon, introns (if present) and the tRNA covariance model score. Also included are the overall statistics for the various component programs [trnascan (13), eufindtrna (14) and cove (15)] as well as summary data for the entire search. The summary data include counts of the total number of tRNAs found, the number of tRNA pseudogenes found, number of tRNAs with introns and the anticodons that were detected.

The snoscan and snoGPS outputs consist of a summary information line for each predicted snoRNA sequence, followed by the candidate in FASTA format. The summary listing for each hit includes:

(i) Query sequence name and snoRNA start and end positions within the query sequence.
(ii) Overall bit score.
(iii) Target sequence name and target position.
(iv) Total number of base pairings and mismatches in the guide region.
(v) The length of the candidate subsequence.

Also included in the display are graphical representations of the base pairing in the target-guide region and the secondary structure of the stem motif(s). Snoscan and snoGPS scores for known snoRNA sequences for various species are available on the website for comparison. Sample abbreviated output records from each of the three servers are shown in Figure 1. Further details on the use of the servers and the interpretation of their results are available in README files accessible from each of the servers.

DISCUSSION

Ideally, genome centers would annotate newly sequenced genomes with our software (in downloaded form), and individual users might not need to use our programs. However, to date, genome centers generally have not used snoscan or snoGPS. As a result, snoRNAs are missing from many annotated genomes. Other uses of the server implementations of our software include screening for tRNAs and snoRNAs in 'Rnomics' (16) experiments of genomes that are not completely sequenced, and checking potential target sites of known or putative snoRNAs. With the introduction of the tRNAscan-SE, snoscan and snoGPS servers, such computational detection and analysis of tRNA and snoRNA sequences become available to a much larger class of researchers.

Although the web server produces identical results to those generated by the original standalone programs, the web server implementation is less powerful for the scanning of large genomes and in terms of flexibility. The probabilistic algorithms used by the snoRNA programs can be relatively slow. Depending on the program, the search parameters used and the query and target sequence lengths, execution times can take a long time: hours or even days. Since the web server is intended as a shared resource, limitations on query and target sequence lengths are incorporated. Whole-genome snoRNA searches on the server are not currently possible; whole genome tRNA searches are limited to queries of five million base pairs or less.

The web server implementation is also limited in its ability to modify the underlying search model and training parameters. Generally, this is not an issue since the models and parameters have been optimized for their specific applications. In some cases, such as scanning genomes with unusual background base compositions, retraining the scan parameters can improve performance. In practice, we have found that search performance is only minimally impacted by inaccurate background base compositions. However, for optimal performance, creating a new model with correct background base compositions (i.e. using the downloaded software) is preferable.

Notwithstanding these limitations, in most cases the web server implementations of tRNAscan-SE, snoscan and snoGPS have essentially all of the capabilities of the standalone software with much easier user interfaces and gentler learning curves. We expect that access to these tools will enable more researchers to search for additional examples of these

(A) tRNAscan-SE Output

```
Sequence                        tRNA  Bounds   tRNA Anti  Intron Bounds Cove
Name                  tRNA #    Begin End      Type Codon Begin  End    Score
--------              ------    ---- ------    ---- ----- -----  ----   ------
chr6.trna18-AlaAGC      1         1    73       Ala  AGC    0     0      40.39

tRNAs decoding Standard 20 AA:            1
Selenocysteine tRNAs (TCA):               0
Possible suppressor tRNAs (CTA,TTA):      0
tRNAs with undetermined/unknown isotypes: 0
Predicted pseudogenes:                    0
                                         ---
Total tRNAs:                              1

Isotype / Anticodon Counts:
Ala   : 1     AGC: 1       GGC:         CGC:         TGC:
Gly   : 0     ACC:         GCC:         CCC:         TCC:

Predicted tRNA Secondary Structures:
chr6.trna18-AlaAGC.trna1 (1-73)   Length: 73 bp
Type: Ala   Anticodon: AGC at 34-36 (34-36)   Score: 40.39
         *    |    *    |    *    |    *    |    *    |    *    |    *    |
Seq: GGGGGATTAGCTCAAGCGGTAGGGTGCCTGCTTAGCATGCAAGAGGtAGCAGGATCGACGCCTGCATTCTCCA
Str: >>>>>>>..>>>>........<<<<.>.>>>.......<<<.<.....>>>>>.......<<<<<<<<<<<.
```

(B) snoscan Output

```
>> snR24  26.40  (1-87)  Cmpl: ySc-25S-Am1447 (U24)  12/0 bp  Gs-DpBox: 18 (18)  Len: 87  TS
Meth site found: 1447 (U24)     Guide Seq Sc: 11.88  (21.36 -1.12 -7.36 -1.00)
                      *
Db seq:  5'-      AGUAGCAAAUAU -3'   ySc-25S   (1444-1456)
                  ||||||||||||
Qry seq: 3'- AGACUUCAUCGUUUAUA -5'     snR24   (29-18)

Terminal stem:      +- [C Box]-N- ACU - 5'        Stem Sc: 0.84 (3 bp)
                    |             |||
                    +---[D Box] - UGAA - 3'       Stem Transit Sc: -1.11

>Summary      [ C Box ] --         -- [ Cmpl/ Mism ]  X [D'Bx] --       -- [D Bx]  Length
>Meth Am1447  [AUGAUGU] --   6 bp  -- [  12 / 0    ]  1 [CAGA] -- 47 bp -- [CUGA]   87 bp
>Sc    26.40  [  7.48 ] --  -1.59  -- [ 21.36 bits ]    [3.94]    -2.44    [8.05]

Candidate sequence:
>snR24  26.40  (1-87)  Cmpl: ySc-25S-Am1447  Len: 87
TCAAATGATGTAATAACATATTTGCTACTTCAGATGGAACTTTGAGTTCCGAATGAGACA
TACCAATTATCACCAAGATCTCTGATG
```

(C) snoGPS Output

```
snoRNA Hits
>YourSeq.3 33.73 (70-135) Cmpl: H_sapiens_LSU.U4417 U65 Pairs: 11/0/10/6/c 66 nt (W)
CCCCAGCTTAGGAAACAGGGTTGTTCTTCATGTGGATGACTCTGTGCCGAAAGCATGGGAACAGCT
X12  56XLLLLLLI12345678I           I87654321IRRRRR X65  21X AAAAAA
              ---
             /   \
             \   /
              G-C
              G.U
              G-C
              A-U
              C-G
              A-U
5'-(N7)UAGGAA    GCCGA(N11)ACA
       ||||||    |||||
       AUCCUU CY CGGCU
```

Figure 1. Sample web server output. Typical (**A**) tRNAscan-SE, (**B**) snoscan and (**C**) snoGPS outputs generated by the web server. For all three sample outputs, the result data have been abbreviated for clarity.

important RNA families in the ever-increasing number of sequenced genomes.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by a lab start up fund.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Decatur,W.A. and Fournier,M.J. (2003) RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.*, **278**, 695–698.
3. Dennis,P.P., Omer,A. and Lowe,T. (2001) A guided tour: small RNA function in Archaea. *Mol. Microbiol.*, **40**, 509–519.
4. Clouet d'Orval,B., Bortolin,M.L., Gaspin,C. and Bachellerie,J.P. (2001) Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. *Nucleic Acids Res.*, **29**, 4518–4529.
5. Ziesche,S.M., Omer,A.D. and Dennis,P.P. (2004) RNA-guided nucleotide modification of ribosomal and non-ribosomal RNAs in Archaea. *Mol. Microbiol.*, **54**, 980–993.
6. Schattner,P. (2003) Computational gene-finding for noncoding RNAs. In Barciszewski,J. (ed.), *NonCoding RNAs: Molecular Biology and Molecular Medicine*. Landes Bioscience, Georgetown, TX, pp. 33–49.
7. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
8. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
9. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M.,Jr, Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.*, **32**, 4281–4296.
10. Brown,J.W., Echeverria,M., Qu,L.H., Lowe,T.M., Bachellerie,J.P., Huttenhofer,A., Kastenmayer,J.P., Green,P.J., Shaw,P. and Marshall,D.F. (2003) Plant snoRNA database. *Nucleic Acids Res.*, **31**, 432–435.
11. Accardo,M.C., Giordano,E., Riccardo,S., Digilio,F.A., Iazzetti,G., Calogero,R.A. and Furia,M. (2004) A computational search for box C/D snoRNA genes in the *D.melanogaster* genome. *Bioinformatics,* 3293–3301.
12. Omer,A.D., Lowe,T.M., Russell,A.G., Ebhardt,H., Eddy,S.R. and Dennis,P.P. (2000) Homologs of small nucleolar RNAs in Archaea. *Science*, **288**, 517–522.
13. Fichant,G.A. and Burks,C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, **220**, 659–671.
14. Pavesi,A., Conterio,F., Bolchi,A., Dieci,G. and Ottonello,S. (1994) Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res.*, **22**, 1247–1256.
15. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
16. Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.