W350–W353 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm393

TFAM 1.0: an online tRNA function classifier

Helena Tåquist, Yuanyuan Cui and David H. Ardell*

Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, SE-751 24 Uppsala, Sweden

Received January 31, 2007; Revised April 3, 2007; Accepted May 1, 2007

ABSTRACT

We have earlier published an automated statistical classifier of tRNA function called TFAM. Unlike tRNA gene-finders, TFAM uses information from the total sequences of tRNAs and not just their anticodons to predict their function. Therefore TFAM has an advantage in predicting initiator tRNAs, the amino acid charging identity of nonstandard tRNAs such as suppressors, and the former identity of pseudo-tRNAs. In addition, TFAM predictions are robust to sequencing errors and useful for the statistical analysis of tRNA sequence, function and evolution. Earlier versions of TFAM required a complicated installation and running procedure, and only bacterial tRNA identity models were provided. Here we describe a new version of TFAM with both a Web Server interface and simplified standalone installation. New TFAM models are available including a proteobacterial model for the bacterial lysylated isoleucine tRNAs, making it now possible for TFAM to correctly classify all tRNA genes for some bacterial taxa. First-draft eukaryotic and archaeal models are also provided making initiator tRNA prediction easily accessible genes to any researcher or genome sequencing effort. The TFAM Web Server is available at http://tfam.lcb.uu.se

INTRODUCTION

The vast majority of new tRNA gene sequence data accumulates today from analysis of genome sequences. The major tRNA gene-finders in use today, tRNAscan-SE (1) and ARAGORN (2), classify the functions of predicted tRNA genes by structurally locating and decoding their inferred anticodons according to an assumed genetic code. As a result, genome projects regularly misclassify initiator tRNAs in genomes from all three phylogenetic domains and lysylated isoleucine tRNA (kIle) genes from bacteria (described further below). These two types of tRNAs carry the same genetically templated anticodons as methionine elongators and hence cannot be distinguished from them by anticodon-based tRNA classifiers. It may also happen that these genes are entirely missing in the annotation of a complete genome. Because genome projects regularly verify the completion of their assemblies by checking for the presence of a complete set of tRNA gene classes in their genome data, the ability to identify these two additional classes of tRNA genes provides additional power for this important task [although in very rare cases the lysylated isoacceptor may be missing along with corresponding metabolic pathways (3)]. Furthermore, a method that uses entire sequence information to classify tRNA gene function is more robust to sequencing error, can correctly predict tRNA charging specificity in organisms with altered genetic codes, can predict the identity of suppressors, and can predict the potential or former charging identity of tRNA-like molecules and pseudo-tRNAs (in this article we loosely use "tRNA" to mean both tRNA and tRNA gene sequences).

We present here a new version (1.0) of such a method—an update to the TFAM statistical classifier of tRNA function published earlier (4). TFAM 1.0 is now available online as a Web Server requiring no installation. Version 1.0 of TFAM also provides eukaryotic and archaeal tRNA functional models for the purpose of identifying initiator tRNAs, and an expanded bacterial model that can predict some kIle tRNAs. For certain bacterial species such as proteobacteria and gram-positive bacteria, TFAM 1.0 can be expected to correctly annotate the function of all tRNAs with good confidence.

In bacteria, the cytidine in the CAU anticodons of lysylated isoleucine tRNAs are post-transcriptionally modified to lysidine (5). This modification simultaneously changes the codon reading specificity of this usually minor isoacceptor from AUG to AUA and its amino acid charging specificity from methionine to isoleucine in keeping with the genetic code. Because the unmodified (CAU) tRNA is charged with methionine (6) it cannot be expected that TFAM would recognize this tRNA as an isoleucine tRNA. However, the determinants that target this tRNA for post-transcriptional modification are themselves genetically templated in the tRNA gene (7,8), and it is possible for TFAM to distinguish this class of tRNA with fairly good confidence when suitably trained (3). Nonetheless, both experimental (8) and bioinformatic (3) evidence suggest that the determinants for this class have diverged in bacteria. Rather than iteratively accreting all

*To whom correspondence should be addressed. Tel: +46 18 471 66 94; Fax: +46 18 471 66 98; Email: david.ardell@lcb.uu.se.

divergent sequences of the same apparent type into one model class (3), we have taken a different, perhaps more conservative approach of making smaller, more phylogenetically restricted models, in the hopes that this will allow us to study the evolution and diversification of tRNA identity determinants. Our ultimate aim is to model and understand the constellation of identity determinants that actually can or could function together in the same cellular context. In TFAM 1.0 we release a model of lysylated isoleucine tRNAs based only on proteobacterial data.

Earlier versions of TFAM were only available for standalone compilation and installation on UNIX-like platforms and required complicated installations of prerequisites. The new web-based interface, besides obviating the installation burden, provides additional functionality over the standalone interface including color visualization and sorting of TFAM scores. Installation of the standalone version has also been simplified by the removal of dependencies.

MODELS

Both the Web Server and standalone interfaces to TFAM version 1.0 provide new models for tRNA classification. The bacterial model of earlier versions (earlier called 'MSDB' for 'Modified Sprinzl DataBase' and provided with TFAM versions 0.2 and 0.3, henceforth 'bacterial TFAM model 0.1') has been complemented and corrected by genomic tDNA sequences from 46 proteobacterial species classified on the basis of clustering to known methionine elongator, unmodified isoleucine elongator, initiator and lysylated isoleucine (kIle) elongator tRNAs. This data, generously provided by Paul Higgs and available here as Supplementary Data to this article, is an extension to the dataset analyzed in (9). These sequences and their classifications are a consistent subset of those provided in an independently made TFAM model (3). A full description of how we used these proteobacterial sequences to modify bacterial TFAM model 0.1 from (4) is provided as Supplementary Data here and at the TFAM Web Server. The model presented here is called 'bacterial TFAM model 0.2'.

Bacterial TFAM model 0.2 also provides the five selenocysteine sequences from the year 2000 release of the Sprinzl Search Server http://www.uni-bayreuth.de/departments/biochemie/trna/, that were removed in bacterial TFAM model 0.1. A leave-one-out cross-validation of bacterial TFAM model 0.2 without selenocysteine tRNAs misclassified 16 of 759 sequences (2.1%) of which 2 of 50 kIle tRNAs were misclassified (96% sensitivity) and 2 of 50 sequences classified as kIle were not kIle (96% specificity). These results were identical when the Sel-Cys model sequences were included, but 2 of the 5 selenocysteine tRNAs were misclassified, underscoring the need to make this model larger and increase its generality.

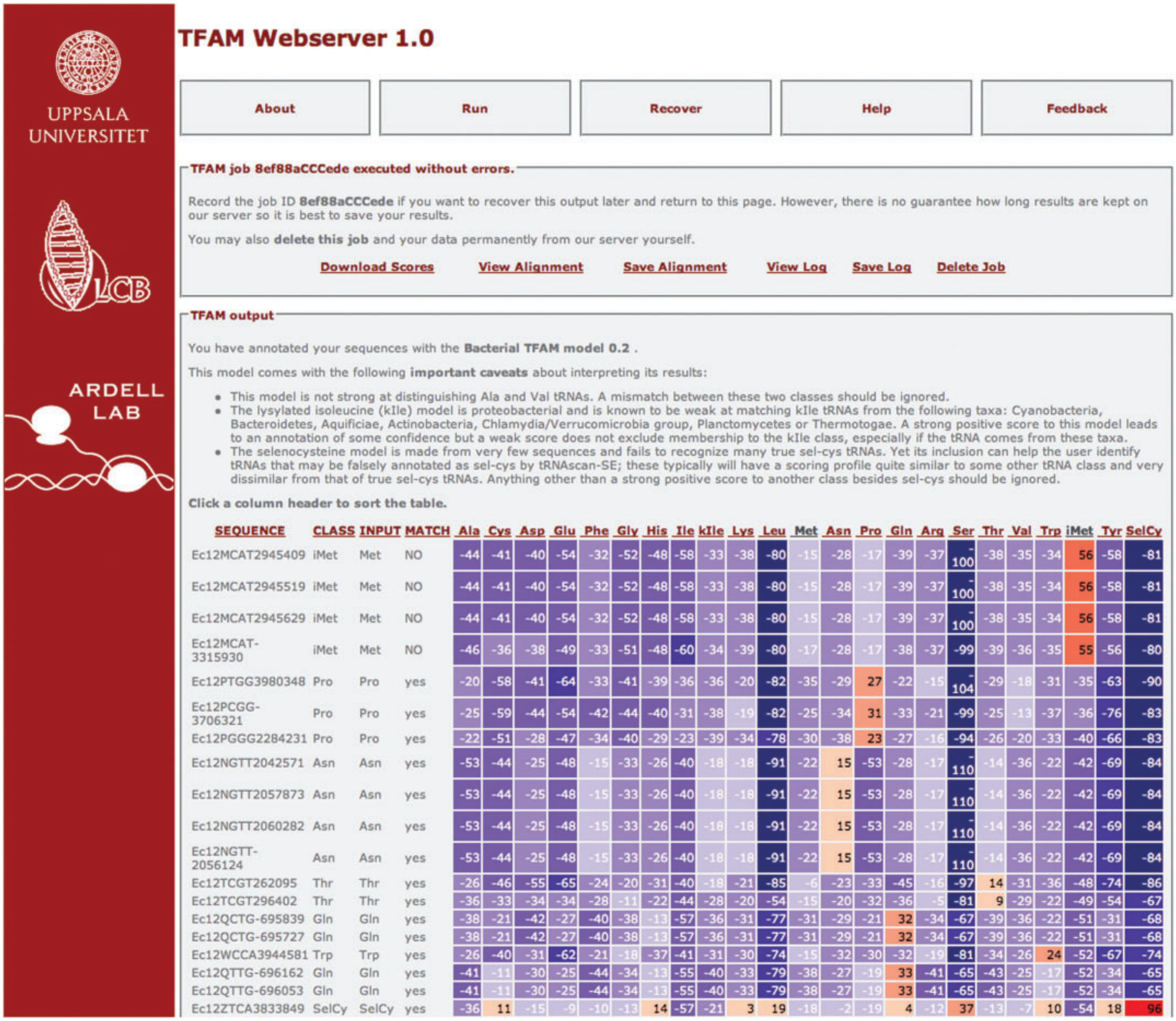
In TFAM 1.0 we introduce first-cut versions of eukaryotic and archaeal TFAM models, each given version numbers 0.1, with the main purpose of enabling convenient and automated initiator tRNA identification in all phylogenetic domains. Initiator tRNA determinants are highly conserved within phylogenetic domains (10). Both of these models were created from data downloaded on February 23, 2006 from the Sprinzl 2000 tRNA Search Server http://www.uni-bayreuth.de/departments/biochemie/trna/. For the eukaryotic TFAM model 0.1, redundant sequences were removed and no corrections were made to the identity annotations. The sensitivity of eukaryotic TFAM model 0.1 for detecting known initiator tRNAs in *Drosophila melanogaster* was verified to be perfect. Initiator tRNA predictions with eukaryotic TFAM model 0.1 in twelve *Drosophila* genomes have been released on the web at http://www.bioinf.manchester.ac.uk/bergman/data/ncRNA/tRNA/. Resubstitution analysis showed that all training data initiators from the Sprinzl 2000 dataset were correctly classified, including those from yeast, worm, flies, vertebrates and plants.

For the archaeal TFAM model 0.1, redundant sequences were removed and the identity annotations of three tRNAs were changed as described in Supplementary Data. The initiator tRNA sequences were verified to be consistent with other archaeal initiator identity determinants as described (11). We also verified the perfect performance of this model on two experimentally characterized archaeal initiator tRNA genes not included and differing in sequence from those in the model: those of *Pyrodictium occultum* (12) and *Pyrococcus abyssi* (13). The accuracy of the eukaryotic and archaeal TFAM models 0.1 for other tRNA types has not been verified or studied in any way.

RESULTS AND DISCUSSION

A user can input tRNA or tDNA sequences in multi-fasta format to the TFAM Web Server by cut-and-paste or file-upload, select a model with which to classify the input sequences, and then push a button to run the server. The browser window will wait while the computation takes place, after which the output is loaded in tabular format. A computation on 514 human tRNA sequences from the Genomic tRNA Database (GtRDB, http://lowelab.ucsc.edu/GtRNAdb/) took ~50 s. The maximum input size is currently set to 200 KB, which accommodates more than 1300 tRNA sequences at once.

A sample output on bacterial tRNA sequences is shown in Figure 1. Scores are visualized with a color scheme increasing from dark purple for negative numbers of large magnitude to bright red for large positive numbers. The table can be sorted by any column making it easy to find sequences with properties of interest. In addition, TFAM can indicate when its prediction conflicts with an annotated function. To get this functionality, the user may either name sequences with identifiers from which the annotated sequence can be parsed by TFAM, or TFAM can auto-detect anticodons and use these to annotate identity. A table of recognized sequence identifier formats is shown in Supplementary Data as well as in the Web-Server documentation. TFAM recognizes the format of identifiers from the GtRDB and the Sprinzl database (14) from which sequences may be pasted directly.



TFAM Webserver 1.0

TFAM job 8ef88aCCCede executed without errors.

Record the job ID **8ef88aCCCede** if you want to recover this output later and return to this page. However, there is no guarantee how long results are kept on our server so it is best to save your results.

You may also **delete this job** and your data permanently from our server yourself.

Download Scores   View Alignment   Save Alignment   View Log   Save Log   Delete Job

TFAM output

You have annotated your sequences with the **Bacterial TFAM model 0.2** .

This model comes with the following **important caveats** about interpreting its results:

- This model is not strong at distinguishing Ala and Val tRNAs. A mismatch between these two classes should be ignored.
- The lysylated isoleucine (kIle) model is proteobacterial and is known to be weak at matching kIle tRNAs from the following taxa: Cyanobacteria, Bacteroidetes, Aquificiae, Actinobacteria, Chlamydia/Verrucomicrobia group, Planctomycetes or Thermotogae. A strong positive score to this model leads to an annotation of some confidence but a weak score does not exclude membership to the kIle class, especially if the tRNA comes from these taxa.
- The selenocysteine model is made from very few sequences and fails to recognize many true sel-cys tRNAs. Yet its inclusion can help the user identify tRNAs that may be falsely annotated as sel-cys by tRNAscan-SE; these typically will have a scoring profile quite similar to some other tRNA class and very dissimilar from that of true sel-cys tRNAs. Anything other than a strong positive score to another class besides sel-cys should be ignored.

Click a column header to sort the table.

SEQUENCE	CLASS	INPUT	MATCH	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	kIle	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	iMet	Tyr	SelCy
Ec12MCAT2945409	iMet	Met	NO	-44	-41	-40	-54	-32	-52	-48	-58	-33	-38	-80	-15	-28	-17	-39	-37	-100	-38	-35	-34	56	-58	-81
Ec12MCAT2945519	iMet	Met	NO	-44	-41	-40	-54	-32	-52	-48	-58	-33	-38	-80	-15	-28	-17	-39	-37	-100	-38	-35	-34	56	-58	-81
Ec12MCAT2945629	iMet	Met	NO	-44	-41	-40	-54	-32	-52	-48	-58	-33	-38	-80	-15	-28	-17	-39	-37	-100	-38	-35	-34	56	-58	-81
Ec12MCAT-3315930	iMet	Met	NO	-46	-36	-38	-49	-33	-51	-48	-60	-34	-39	-80	-17	-28	-17	-38	-37	-99	-39	-36	-35	55	-56	-80
Ec12PTGG3980348	Pro	Pro	yes	-20	-58	-41	-64	-33	-41	-39	-36	-36	-20	-82	-35	-29	27	-22	-15	-104	-29	-18	-31	-35	-63	-90
Ec12PCGG-3706321	Pro	Pro	yes	-25	-59	-44	-54	-42	-44	-40	-31	-38	-19	-82	-25	-34	31	-33	-21	-99	-25	-13	-37	-36	-76	-83
Ec12PGGG2284231	Pro	Pro	yes	-22	-51	-28	-47	-34	-40	-29	-23	-39	-34	-78	-30	-38	23	-27	-16	-94	-26	-20	-33	-40	-66	-83
Ec12NGTT2042571	Asn	Asn	yes	-53	-44	-25	-48	-15	-33	-26	-40	-18	-18	-91	-22	15	-53	-28	-17	-110	-14	-36	-22	-42	-69	-84
Ec12NGTT2057873	Asn	Asn	yes	-53	-44	-25	-48	-15	-33	-26	-40	-18	-18	-91	-22	15	-53	-28	-17	-110	-14	-36	-22	-42	-69	-84
Ec12NGTT2060282	Asn	Asn	yes	-53	-44	-25	-48	-15	-33	-26	-40	-18	-18	-91	-22	15	-53	-28	-17	-110	-14	-36	-22	-42	-69	-84
Ec12NGTT-2056124	Asn	Asn	yes	-53	-44	-25	-48	-15	-33	-26	-40	-18	-18	-91	-22	15	-53	-28	-17	-110	-14	-36	-22	-42	-69	-84
Ec12TCGT262095	Thr	Thr	yes	-26	-46	-55	-65	-24	-20	-31	-40	-18	-21	-85	-6	-23	-33	-45	-16	-97	14	-31	-36	-48	-74	-86
Ec12TCGT296402	Thr	Thr	yes	-36	-33	-34	-34	-28	-11	-22	-44	-28	-20	-54	-15	-20	-32	-36	-5	-81	9	-29	-22	-49	-54	-67
Ec12QCTG-695839	Gln	Gln	yes	-38	-21	-42	-27	-40	-38	-13	-57	-36	-31	-77	-31	-29	-21	32	-34	-67	-39	-36	-22	-51	-31	-68
Ec12QCTG-695727	Gln	Gln	yes	-38	-21	-42	-27	-40	-38	-13	-57	-36	-31	-77	-31	-29	-21	32	-34	-67	-39	-36	-22	-51	-31	-68
Ec12WCCA3944581	Trp	Trp	yes	-26	-40	-31	-62	-21	-18	-37	-41	-31	-30	-74	-15	-32	-30	-32	-19	-81	-34	-26	24	-52	-67	-74
Ec12QTTG-696162	Gln	Gln	yes	-41	-11	-30	-25	-44	-34	-13	-55	-40	-33	-79	-38	-27	-19	33	-41	-65	-43	-25	-17	-52	-34	-65
Ec12QTTG-696053	Gln	Gln	yes	-41	-11	-30	-25	-44	-34	-13	-55	-40	-33	-79	-38	-27	-19	33	-41	-65	-43	-25	-17	-52	-34	-65
Ec12ZTCA3833849	SelCy	SelCy	yes	-36	11	-15	-9	-10	-13	14	-57	-21	3	19	-18	-2	-19	4	-12	37	-13	-7	10	-54	18	96

Figure 1. Sample output of the TFAM Web Server for tRNA genes from *Escherichia coli* K12. Colors indicate the magnitude and sign of the scores of each sequence to different tFAM models. Initiator tRNAs are sorted to the top of the table by clicking on the 'iMet' column header.

As described more fully in (4), TFAM works by structurally aligning the input sequences with the trusted model sequences using COVE software (15) and tRNA covariance models from (1). It then makes profiles of each model class and scores each input sequence according to the log-odds of belonging to a specific functional class versus belonging to any of the others. This log-odds computation is repeated for every functional class represented in the model. A positive score indicates a match to a particular functional class. The TFAM classification of a tRNA sequence is the functional class against which it has the highest score.

All results of TFAM computations are available for download from the Web Server including a structural alignment of their sequences (in aligned multi-fasta format) with TFAM classifications in the description line. The user is also provided a job ID with which she may return to the server for some unspecified time after visiting the Web Server and recover the results of their computation. Job IDs are not in any way public, so only users with a specific job ID in hand can recover results from that job. Jobs may be stored on the server for some unspecified time and then deleted, therefore it is safest to download results after computation. For more data privacy, the user may completely delete the data from our server after downloading their results. We log user-statistics but do not store any data input to the server for private use.

The tRNA structural alignment service of the TFAM Web Server may be useful in its own right. If the TFAM Web Server is used exclusively for this purpose users should additionally cite COVE software (15) and the tRNA covariance models from (1).

Currently the Web Server does not support uploading of custom tRNA classification models or the exclusion of certain positions from scoring such as the anticodon. However, these functionalities are available in the standalone version which is available free for download and obtainable through http://tfam.lcb.uu.se

TFAM results should not be used blindly. We emphasize that TFAM classifications are statistical and depend on the quality of trusted classifications that are used to make the models. Different models and even different 'tfams' for different functional classes vary in their statistical power and generality. We have provided a guide to the intended use of every component of every model available at the TFAM Web Server, and we also refer to earlier provided statistical results (4). To briefly summarize here, TFAM is not strong at distinguishing alanine and valine tRNAs in bacteria. The selenocysteine model is not very general and can fail to recognize true Sel-Cys tRNAs, nonetheless, the model is useful to distinguish falsely predicted Sel-Cys tRNAs such as called by tRNAscan-SE for firmicutes with an altered genetic code (data not shown). The kIle model is proteobacterial and does not score well for kIle tRNAs from Cyanobacteria, Bacteroidetes, Aquificiae, Actinobacteria, Chlamydia/Verrucomicrobia group, Planctomycetes or Thermotogae [data not shown, but see also (3)]. This is probably because the determinants for this class are divergent in bacteria, the subject of our current research. Finally, the eukaryotic and archaeal TFAM 0.1 models are provided only for the purpose of initiator tRNA identification.

We plan major refinements and expansions of the TFAM models, including specializing them to more narrow phyla, towards the goal of general models that nonetheless accurately represent the actual constellation of identity determinants and antideterminants that function together in the context of a specific translational system. In the meantime, we believe that the model quality and ease of use of the TFAM 1. 0 Web Server can promote practical improvements in the annotation of tRNA genes in genome projects and encourage computational biological research in tRNA structure, function and evolution already today.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank Paul Higgs and Philippe Boisvert for supplying an initial and unpublished version of a classification of the CAU-anticodon-containing tRNAs in proteobacteria, Casey Bergman for verifying the eukaryotic model for *Drosophila* initiator tRNAs, Emil Lundberg for system administration and support, Erik Bongcam-Rudloff and Apple, Inc. for hardware support and the Linnaeus Centre for Bioinformatics for financial support. Funding to pay the Open Access publication charges for this article was provided by the Linnaeus Centre for Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Laslett,D. and Canback,B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.
3. Silva,F.J., Belda,E. and Talens,S.E. (2006) Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. *Nucleic Acids Res.*, **34**, 6015–6022.
4. Ardell,D.H. and Andersson,S.G. (2006) TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res.*, **34**, 893–904.
5. Grosjean,H. and Bjork,G.R. (2004) Enzymatic conversion of cytidine to lysidine in anticodon of bacterial isoleucyl-tRNA–an alternative way of RNA editing. *Trends Biochem. Sci.*, **29**, 165–168.
6. Muramatsu,T., Nishikawa,K., Nemoto,F., Kuchino,Y., Nishimura,S., Miyazawa,T. and Yokoyama,S. (1988) Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. *Nature*, **336**, 179–181.
7. Ikeuchi,Y., Soma,A., Ote,T., Kato,J., Sekine,Y. and Suzuki,T. (2005) Molecular mechanism of lysidine synthesis that determines tRNA identity and codon recognition. *Mol. Cell*, **19**, 235–246.
8. Nakanishi,K., Fukai,S., Ikeuchi,Y., Soma,A., Sekine,Y., Suzuki,T. and Nureki,O. (2005) Structural basis for lysidine formation by ATP pyrophosphatase accompanied by a lysine-specific loop and a tRNA-recognition domain. *Proc. Natl Acad. Sci. USA*, **102**, 7487–7492.
9. Tang,B., Boisvert,P. and Higgs,P.G. (2004) Comparison of tRNA and rRNA phylogenies in proteobacteria: implications for the frequency of horizontal gene transfer. *arXiv:q-bio.PE/0404030*.
10. Marck,C. and Grosjean,H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189–1232.
11. Mallick,B., Chakrabarti,J., Sahoo,S., Ghosh,Z. and Das,S. (2005) Identity elements of archaeal tRNA. *DNA Res.*, **12**, 235–246.
12. Ushida,C., Muramatsu,T., Mizushima,H., Ueda,T., Watanabe,K., Stetter,K.O., Crain,P.F., McCloskey,J.A. and Kuchino,Y. (1996) Structural feature of the initiator tRNA gene from Pyrodictium occultum and the thermal stability of its gene product, tRNA(imet). *Biochimie*, **78**, 847–855.
13. Yatime,L., Schmitt,E., Blanquet,S. and Mechulam,Y. (2004) Functional molecular mapping of archaeal translation initiation factor 2. *J. Biol. Chem.*, **279**, 15984–15993.
14. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139–D140.
15. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.