Published online 27 October 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 20* ***6015–6022***
doi:10.1093/nar/gkl739

Differential annotation of tRNA genes with anticodon CAT in bacterial genomes

Francisco J. Silva*, Eugeni Belda and Santiago E. Talens

Departament de Genètica, Institut Cavanilles de Biodiversitat i Biologia Evolutiva,
Universitat de València, Apartat 22085, 46071 Valencia, Spain

Received June 26, 2006; Revised and Accepted September 25, 2006

ABSTRACT

We have developed three strategies to discriminate among the three types of tRNA genes with anticodon CAT (tRNA^{Ile}, elongator tRNA^{Met} and initiator tRNA^{fMet}) in bacterial genomes. With these strategies, we have classified the tRNA genes from 234 bacterial and several organellar genomes. These sequences, in an aligned or unaligned format, may be used for the identification and annotation of tRNA (CAT) genes in other genomes. The first strategy is based on the position of the problem sequences in a phenogram (a tree-like network), the second on the minimum average number of differences against the tRNA sequences of the three types and the third on the search for the highest score value against the profiles of the three types of tRNA genes. The species with the maximum number of tRNA^{fMet} and tRNA^{Met} was *Photobacterium profundum*, whereas the genome of one *Escherichia coli* strain presented the maximum number of tRNA^{Ile} (CAT) genes. This last tRNA gene and *tilS*, encoding an RNA-modifying enzyme, are not essential in bacteria. The acquisition of a tRNA^{Ile} (TAT) gene by *Mycoplasma mobile* has led to the loss of both the tRNA^{Ile} (CAT) and the *tilS* genes. The new tRNA has appropriated the function of decoding AUA codons.

INTRODUCTION

The prediction of non-coding RNA genes during the course of the annotation of a genome is a difficult task, which requires not only the search by sequence similarity but also the prediction of secondary structures in the transcribed RNAs. Because each RNA type presents its own structure, which includes essential and optional parts, the development of specific methods for each type is required. The cloverleaf secondary structure of tRNAs has served as an approach to identify putative tRNA specifying sequences in the DNA. This is one of the strategies of the program tRNAscan-SE (1) to identify and annotate tRNA genes. This program is probably the most widely used in genome annotation. After the identification of the anticodon loop, each tRNA gene is marked with the anticodon sequence and the associated amino acid, giving a score for this assignment. The accuracy of this program is high, although with limitations due to the post-transcriptional anticodon modifications, or difficulty in identifying pseudogenes. The initiator tRNA may not be distinguishable from the elongator tRNA^{Met}.

An alternative program, called TFAM (2), has recently been developed. It is based on the proximity to profiles that are mainly due to the presence of determinants in the tRNA sequences, which would putatively be associated with the binding by the aminoacyl-tRNA synthetases or the tRNA modification enzymes. By using this program, each tRNA sequence receives a score of proximity to the 21 tRNA profiles. They include the initiator formylmethionine tRNA and the 20 types of elongator tRNAs. This approach, combined with the sequence of the anticodon, may detect some special cases such as the Trp tRNAs from some *Mycoplasma* species that are incorrectly identified as Selenocysteine tRNAs by tRNAscan-SE. It also permits the detection of situations where the anticodon and the class against which TFAM has maximum score do not coincide, as a consequence of several situations such as, for example, post-transcriptional nucleotide substitutions compared with the anticodon DNA sequence.

The programs described previously, as well as other tRNA prediction programs, are unable to identify one special type of tRNA with anticodon CAU which, after modification to convert cytidine into lysidine, a lysine-containing cytidine, is recognized by isoleucine tRNA synthetase, charging isoleucine and changing codon recognition to AUA (3). The decoding of the codons AUN with the correct discrimination between AUG and the remaining codons to specifically translate Met or Ile is solved by several strategies in Archaea, Bacteria, Eukarya and Organelles (3,4). The strategy in bacterial genomes is exemplified in *Escherichia coli* or *Bacillus subtilis* by the presence of four tRNA species: two tRNAs with anticodon CAU for the decoding of the AUG codon as initiator or elongator, one tRNA (GAU) to decode AUY codons, and one tRNA (LAU) where L (lysidine) is a C_{34} modified with lysine to restrict decoding specifically to AUA. The enzyme responsible for this last modification is

*To whom correspondence should be addressed. Tel: +34 963543650; Fax: +34 963543670; Email: francisco.silva@uv.es

TilS (tRNA^{Ile}-lysidine synthase) and its gene was recently identified and given the name *tilS* (alternative names *mesJ* and *yacA*) (5). TilS is an RNA-modifying enzyme found in all the complete genomic sequences of bacteria and, for that reason, has been proposed as one of the 206 essential protein-coding genes required for maintaining bacterial cell life (6). Eukaryotes have solved this problem by producing a special tRNA where A_{34} is modified to Inosine (I). The anticodon IAU binds to the three codons (AUH) decoding as Ile. The two types of tRNA (CAU) decode only the AUG codon, either initiator or elongator as Met (3). In some eukaryotes, an additional tRNA (UAU) with U_{34} modified is used to decode AUA preferentially or restrictively (7). This type of tRNA gene with this anticodon may also be detected in a few bacterial genomes (8).

Bacterial tRNA types with anticodon CAU have to be recognized correctly by isoleucyl- and methionyl-tRNA synthetases in order to charge Ile or Met, respectively. The first step for tRNA^{Ile} (CAU) before being charged with isoleucine is the conversion of C_{34} to lysidine. TilS discriminates this tRNA from tRNA^{Met}. Once the tRNA (LAU) is produced, the isoleucyl-tRNA synthetase is able to charge it with Ile. It has been proposed based on the analysis of *E.coli* and *Aquifex aeolicus* that although tRNA^{Ile} (CAU) and tRNA^{Met} are very similar, their sequences are equipped with four sets of determinants that are positively or negatively recognized by the two aminoacyl-tRNA synthetases, by TilS and by a putative acetyltransferase which could modify C_{34} from the elongator tRNA^{Met} to acetylcytidine (9). The action of TilS is very important because in many bacterial species the modified tRNA^{Ile} (LAU) is the only tRNA able to read AUA codons. However, the presence of a few unmodified tRNA^{Ile} (CAU) molecules in the cell does not produce translating problems, because these molecules behave as elongator tRNA^{Met}, being recognized by the methionyl-tRNA synthetase and charged with Met. They decode AUG codons.

The positive or negative determinants of these tRNAs are not universal and *E.coli* TilS is unable to recognize *A.aeolicus* tRNA^{Ile} (CAU), probably because the two pairs of positive determinants C_4G_{69} and C_5G_{68}, at the aminoacyl stem are not conserved (9). Analysis of TilS and tRNA sequences in several bacteria indicates that this protein, tRNA^{Ile} (CAU) and tRNA^{Met} are coevolving with the aim of discriminating between both tRNA types (9,10).

In this study, we analysed the tRNA gene sequences with an anticodon CAT (the term anticodon is used by extension to describe at the DNA level the corresponding nucleotides to those present in the tRNA molecule) of 234 bacterial genomes and 10 organellar genomes. The aim of this work was to classify them into the three known types (Ile, initiator and elongator Met) and to develop methods to discriminate among them, especially to identify the tRNA^{Ile} (CAT) genes.

MATERIALS AND METHODS

Bacterial tRNA gene sequences

A total of 234 bacterial genomes were used in this study to identify the three types of tRNAs with anticodon CAT. They were all bacterial genomes present in the Genomic tRNA database (http://lowelab.ucsc.edu/GtRNAdb/) plus *Mycoplasma capricolum*. This database contains tRNA gene sequences identified and classified using the program tRNAscan-SE (1). Twelve tRNA genes that were not included in the genome lists of the database were extracted from the NCBI genome annotation. Two tRNAs, required to complete the three-type set, were identified by BLAST (see, in Supplementary Data, a table with the number of genes of each type in the analysed genomes). The sequences of the tRNA genes from 10 organellar genomes (6 chloroplast and 4 mitochondria) were also extracted from the NCBI database. Finally, we observed in the nucleotide alignments that some tRNA gene sequences of *Vibrio fischeri* did not contain the first and last nucleotides. Looking at the genome sequence, we realized that this was a mistake of the tRNA database. We included the complete sequences of these genes.

To test the annotation strategies, the tRNA genes with anticodon CAT of *Hahella chejuensis*, *Pelobacter carbinolicus* and *Salinibacter ruber* were extracted from the NCBI genomes database.

Computational sequence analyses

Sequences were aligned using the program CLUSTAL X (11). For the iterative incorporation (see Results) of the sequences of new taxonomic groups to the previous alignment, the option of profile alignment was used.

Alignment files were converted to MEGA in order to perform analyses using the program MEGA3 (12). Tree-like networks were obtained based on the number of pairwise differences and, may therefore be defined as phenograms. They were obtained with the neighbour-joining program (13) with the option of complete deletion, which removes any nucleotide site that does not contain a nucleotide in each one of the analysed genomes. After classification of the sequences into groups, distance matrices based on the number of differences were estimated with the option *Between Groups Means*.

The proximity to a specific tRNA profile was performed using the program TFAM (2). This program uses unaligned tRNA or tDNA sequences in the Fasta format as input. It first needs to produce position-specific scoring matrices for each tRNA gene type and later compares problem sequences with these profiles, producing a positive or negative score in front of them. In order to make comparisons, the program produces an alignment based on sequence similarity and secondary structural information. Sequences classified as the tRNA gene types Ile, Met and fMet were introduced into the program to create the three profiles. Later, the program TFAM was run with each one of these sequences (as well as with the sequences of *H.chejuensis*, *P.carbinolicus* and *S.ruber*) giving the scores against each of the three profiles.

RESULTS

Analysis of tRNA genes with anticodon CAT in Enterobactericeae and Clostridia/Mollicutes

Sequences of tRNAs with anticodon CAT in Enterobacteriaceae were aligned and a phenogram was obtained by using the neighbour-joining method and a pairwise distance matrix

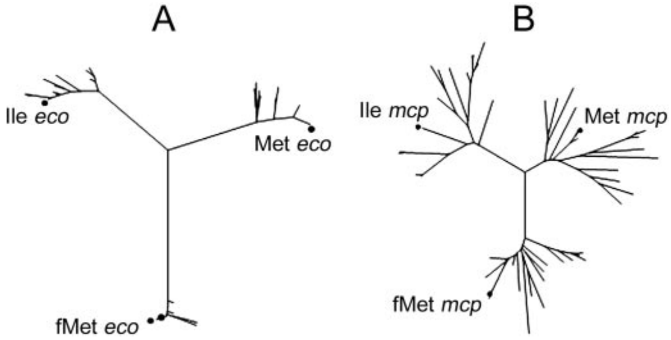

Figure 1. Phenograms of tRNA gene sequences with anticodon CAT. (**A**) Enterobacteriaceae. (**B**) Clostridia and Mollicutes. Filled circles show the location of the tRNA genes of a known type from *E.coli* (*eco*) and *M.capricolum* (*mcp*).

obtained with the number of differences (Figure 1A). Three well-defined clusters were obtained with an average number of 22–30 differences among them. The identity of each cluster was determined based on the known sequences of the three types in *E.coli* (14). The same strategy to identify the three groups was followed with the taxonomic groups of Clostridia and Mollicutes. The sequences of the tRNAs from *M.capricolum* (15) were used to identify each tRNA type. In spite of the use of a more divergent group of species, the three tRNA clusters were well-separated and identified based on the positions of the three *M.capricolum* tRNAs (Figure 1B). When the sequences of both taxonomic groups (Enterobacteriaceae and Clostridia/Mollicutes) were aligned and used to construct the phenogram, the tree obtained correctly clustered the tRNAs of each type (data not shown).

Identification of tRNA genes with anticodon CAT in other bacterial taxonomic groups

In order to determine for any taxonomic group which tRNA corresponded to each type and whether the three clusters could be easily identified, we continued as follows:

(i) The sequences of the tRNAs with anticodon CAT of a specific taxonomic group were obtained and aligned.
(ii) A phenogram was constructed using the neighbour-joining method and a pairwise distance matrix obtained with the number of differences.
(iii) The number of tRNAs for each species was checked and, in case that there was a species without a sequence of the three tRNA types, the genome annotation file was revised and/or a BLAST search against the genome sequence was carried out.
(iv) The previous alignment was then aligned to a total tRNA alignment which started with the Enterobacterial, Clostridia and Mollicutes sequences but was continuously increasing at each step with the incorporation of each new taxon.
(v) The number of nucleotide differences among the three tRNAs groups of the new taxon and the previously identified groups of the total tRNA alignment was estimated. Each group was identified based on the average number of differences against the $tRNA^{Ile}$ (CAT), $tRNA^{Met}$ (CAT) and $tRNA^{fMet}$ (CAT) of all taxonomic groups.
(vi) After identification, the new taxon tRNA group sequences were maintained in the total alignment and a new taxonomic group was analysed from step (i).

Once every taxonomic group had been incorporated into the alignment, including samples of mitochondrial and chloroplast tRNA gene sequences, the average number of differences between the sets at each taxonomic group and the whole sets was re-estimated (Table 1). The $tRNA^{fMet}$ groups were very similar, with a range of average differences of 6.2–10.5, except for the more divergent mitochondrial tRNAs (14.7). It permitted an easy identification of this type of tRNA. The discrimination between the two other tRNA types was more difficult. However, given a taxonomic group, we can compare the number of differences of the two unclassified tRNA clusters against the remaining taxonomic groups and estimate the quotient of the average number of differences against $tRNA^{Ile}$ (CAT) by those against $tRNA^{Met}$ (Table 1). This quotient produced values of 0.6–0.9 for the groups identified as $tRNA^{Ile}$ and 1.2–1.5 for those identified as $tRNA^{Met}$. The closest values were obtained for Cyanobacteria with 0.9 and 1.2 for the tRNA groups identified as $tRNA^{Ile}$ and $tRNA^{Met}$, respectively. In some taxonomic groups such as Actinobacteria, Cyanobacteria and Planctomyces, the $tRNA^{Ile}$ gene sequences were only slightly more similar to the Ile than to the Met type (Ile/Met ratio 0.9). However, their $tRNA^{Met}$ sequences were more dissimilar (1.3, 1.2 and 1.4, respectively), indicating that the identification of $tRNA^{Ile}$ genes is more related to their dissimilarity to the Met than to the similarity to the Ile-type sequence.

The identification of the three groups was also supported by the production of three clusters in the phenogram with the complete tRNA sets of the 234 analysed genomes plus the 10 organellar genomes (Figure 2). The previously known tRNAs types from *B.subtilis*, *A.aeolicus*, Chloroplasts and mitochondria clustered correctly with their corresponding tRNA types. Several nucleotide sites could be established as positive or negative discriminators between $tRNA^{Ile}$ (CAT) and $tRNA^{Met}$ (CAT) genes. Especially, the base pairs at the acceptor stem are remarkable: 3–70, 4–69 and 5–68, according to the Sprinzl position indexing (16). Nucleotides G_3, A_3, C_{70} and T_{70} could be a complete negative discriminator for either $tRNA^{Met}$ or $tRNA^{fMet}$ genes. At these positions both tRNAs usually have the pair C_3–G_{70}; therefore, G_3, A_3, C_{70} and T_{70} indicate a $tRNA^{Ile}$ gene. On the other hand, pairs C_4–G_{69} and C_5–G_{68} are very frequent in the $tRNA^{Ile}$ (CAT) gene, whereas the pairs A_{11}–T_{24} and G_{12}–C_{23} at the D stem seem to be complete discriminators of fMet unlike the two other tRNAs.

Finally, we created three tRNA profiles with 457 fMet, 293 Ile and 288 Met tRNA sequences that comprise the complete tRNA set with the exception of the organellar sequences and a few tRNAs of uncertain classification. These profiles were obtained using the program TFAM (2). Each of the 1070 tRNA sequences analysed in this study (including organellar and those difficult to classify) was compared with the 3 profiles and only in 10 cases there was a discrepancy from our previous classification. Six cases corresponded to the $tRNA^{Met}$ of *Bordetella* spp. and some Cyanobacteria which gave a positive value against the profiles of Ile and Met, slightly higher for the former. Our identification of

Table 1. Average pairwise number of differences between the three types of tRNAs from one taxonomic group and the whole set of sequences for fMet, Ile and Met tRNA types

	fMet	Ile	Met	Ile/Met
Actino_Fmet	7.0	23.7	20.3	
Actino_Ile	21.3	16.2	18.2	0.9
Actino_Met	19.4	19.5	15.1	1.3
Alpha_Fmet	7.9	24.2	21.9	
Alpha_Ile	23.5	15.2	20.3	0.7
Alpha_Met	18.3	20.5	14.1	1.5
Bacillales_Fmet	6.6	23.8	20.7	
Bacillales_Ile	23.3	13.5	20.5	0.7
Bacillales_Met	18.3	20.0	14.7	1.4
Bacteroid_Fmet	9.7	25.5	21.7	
Bacteroid_Ile	25.4	17.3	21.8	0.8
Bacteroid_Met	19.4	21.8	17.8	1.2
Beta_Fmet	6.2	22.8	20.2	
Beta_Ile	25.1	15.7	19.8	0.8
Beta_Met	21.1	20.3	16.2	1.3
Chlamydiae_Fmet	10.5	24.6	21.1	
Chlamydiae_Ile	18.1	15.9	18.7	0.8
Chlamydiae_Met	18.8	20.0	16.2	1.2
Chloroflexi_Fmet	8.5	25.2	21.7	
Chloroflexi_Ile	23.3	14.8	20.8	0.7
Chloroflexi_Met	22.1	22.5	19.2	1.2
Chlorop_Fmet	10.1	24.4	20.5	
Chlorop_Ile	32.1	21.9	27.3	0.8
Chlorop_Met	30.3	28.1	23.7	1.2
ClosMolli_Fmet	7.9	23.4	20.4	
ClosMolli_Ile	24.7	15.6	21.1	0.7
ClosMolli_Met	20.2	20.6	17.1	1.2
Cyano_Fmet	8.6	23.7	19.5	
Cyano_Ile	28.1	27.1	28.8	0.9
Cyano_Met	20.9	20.2	16.8	1.2
Deinoc_Fmet	7.1	23.5	19.8	
Deinoc_Ile	21.9	13.6	21.4	0.6
Deinoc_Met	20.2	18.3	13.9	1.3
Delta_Fmet	6.3	23.8	20.3	
Delta_Ile	22.8	13.4	20.6	0.7
Delta_Met	21.2	21.7	16.7	1.3
Entero_Fmet	6.4	24.2	20.3	
Entero_Ile	23.4	14.7	18.5	0.8
Entero_Met	24.2	20.8	17.1	1.2
Epsilon_Fmet	10.1	24.8	19.8	
Epsilon_Ile	22.5	14.3	20.3	0.7
Epsilon_Met	20.2	22.8	17.2	1.3
Fusobact_Fmet	7.4	22.4	19.5	
Fusobact_Ile	23.6	14.9	19.7	0.8
Fusobact_Met	23.2	22.0	18.3	1.2
Gamma_Fmet	6.8	23.9	20.6	
Gamma_Ile	25.1	15.5	20.4	0.8
Gamma_Met	24.5	20.9	17.0	1.2
GreenSulfur_Fmet	9.2	23.1	22.9	
GreenSulfur_Ile	24.8	13.0	19.7	0.7
GreenSulfur_Met	17.5	19.6	15.3	1.3
Hyperterm_Fmet	7.1	24.0	19.3	
Hyperterm_Ile	22.0	14.9	18.9	0.8
Hyperterm_Met	18.9	19.6	15.2	1.3
Lbacillales_Fmet	6.7	23.2	21.1	
Lbacillales_Ile	24.5	14.8	21.8	0.7
Lbacillales_Met	18.1	20.3	15.2	1.3
Mit_Fmet	14.7	25.9	23.7	
Mit_Ile	28.8	21.6	27.8	0.8
Mit_Met	22.2	24.6	19.4	1.3
Planctomy_Fmet	9.3	24.6	22.0	
Planctomy_Ile	20.9	17.0	18.9	0.9
Planctomy_Met	19.0	19.9	14.0	1.4
Spirochete_Fmet	9.7	23.9	21.2	
Spirochete_Ile	23.6	13.9	19.7	0.7
Spirochete_Met	20.4	20.7	16.5	1.3

Abbreviations: Fmet, tRNA^{fMet}; Ile, tRNA^{Ile}; Met, tRNA^{Met}; Actino, actinobacteria; Alpha, alpha-Proteobacteria; Bacteroid, Bacteroidetes; Beta, beta-Proteobacteria; Chlorop, chloroplasts; ClosMolli, Clostridia-Mollicutes; Cyano, cyanobacteria; Deinoc, Deinococcus-Thermus; Delta, delta-Proteobacteria; Entero, enterobacteria; Epsilon, epsilon-Proteobacteria; Fusobact, fusobacteria; Gamma, gamma-Proteobacteria excluding Enterobacteria; GreenSulfur, green sulphur bacteria; Hyperterm, hyperthermophilic bacteria; Lbacillales, Lactobacillales; Mit, Mitochondria; and Planctomy, Planctomyces.

The minimum number of differences among the three pairwise comparisons is highlighted. Ile/Met is the quotient of the number of differences against the whole Ile set by the number of differences against the whole Met set. Values higher than 1 indicate more similarity to tRNA^{Met} whereas values smaller than 1 to tRNA^{Ile}.

these tRNAs as Met is based on the fact that two of the three types of tRNA sequences in these species have a high proximity to the profiles of fMet and Ile (score values higher than 40 for their corresponding tRNA type and negative for the others), whereas the sequences of the third group have positive scores for both Met and Ile and negative for fMet. Classification was difficult for two divergent (against other delta-Proteobacteria) *Bdellovibrio bacteriovorus* tRNAs, for one *Bordetella bronchiseptica* tRNA and a mitochondrial tRNA (*Pseudendoclonium akinetum*). A 3D plot with the TFAM scores shows the isolation of the sequences belonging to the three groups, except for a few points (Figure 3).

There were two genomes in which at least one copy of each type of tRNA gene could not be identified. The first, *Mycoplasma mobile* has actually lost the tRNA^{Ile} (CAT), whereas the second was one of the strains of *Streptococcus pyogenes* (strain SF370 serotype M1), which has only two fMet tRNAs and no other tRNAs (CAT) could be detected in the genome sequence. A putative sequence assembly problem in a genome region with 28 tandem tRNAs in other strains and only 21 in this strain could be the reason.

The maximum number of genes in tRNAs with anticodon CAT corresponds to *Photobacterium profundum* with 15 copies. It is also the species with the maximum number of tRNA^{fMet} and tRNA^{Met} (8 and 5 genes, respectively). The genome of *E.coli* strain O157:H7 possesses 9 tRNA^{Ile} (CAT) genes, the maximum number among the genomes analysed (see a table with the number of genes of each type in the analysed genomes in Supplementary Data).

Annotation of tRNA genes with anticodon CAT in bacterial genomes

We propose three methods for the annotation of the three types of tRNA (anticodon CAT) in bacterial genomes based on the aligned sequence groups described previously.

(i) After identification with the tRNAscan program of the tRNA sequences with anticodon CAT in a genome, they will be aligned with the total tRNA set alignment (Supplementary Data) or with the tRNA alignment of the taxonomic group to which the species belongs. A phenogram, such as those shown in Figures 1 and 2,

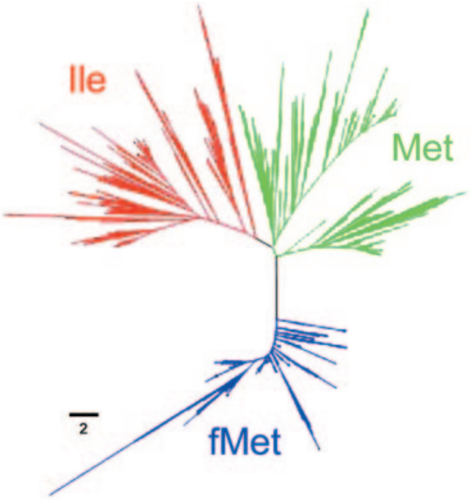

Figure 2. Phenogram of the complete set of tRNA genes with anticodon CAT. Red (tRNA^{Ile}), green (tRNA^{Met}) and blue (tRNA^{fMet}). The bar shows the branch length for 2 nt differences. The largest branch at the fMet group corresponds to the mitochondrial tRNA genes.

will be obtained and the position of the new sequences in the tree will indicate how the tRNA gene should be annotated.

(ii) A second strategy is the use of the average number of differences of each problem tRNA sequence compared with the three groups of sequences containing the Ile, Met and fMet tRNA types annotated in this paper. After alignment, the sequence file is converted to the MEGA format. The file is then opened with MEGA and groups are created for each one of the problem tRNA sequences and for the complete sets of fMet, Ile and Met sequences. Later, a matrix is obtained with the tool *Compute Between Groups Means*. The matrix is extracted and the average number of differences between each new sequence and the tRNA groups for fMet, Ile and Met, respectively, are estimated. Sequences corresponding to tRNA^{fMet} are easily identified based on the small number of differences with the fMet group. Sequences corresponding to the two other types of tRNAs are identified based on the smallest number of differences. The quotient of the value for tRNA^{Ile} by tRNA^{Met} may indicate the confidence of the identification. An example of this strategy for annotating tRNA genes is shown in Table 2 (A) for the genomes of *H.chejuensis*, *P.carbinolicus* and *S.ruber*. The restriction of the analyses to the use of the tRNA sequences of the species' taxonomic group produces more extreme Ile/Met values and a more precise identification (see Table 2, B). In a few cases the use of the total tRNA set produces an incorrect annotation (see tRNA gene sru2 in Table 2) and, for that reason, we recommend the use of the species' taxonomic group set.

(iii) Finally, we may use the profiles for three tRNA types and run TFAM with the problem sequences (Table 2, C) (see Supplementary Data for files containing the unaligned tRNA gene sequences in a TFAM input format, the *coveam* file and a table of name equivalences).

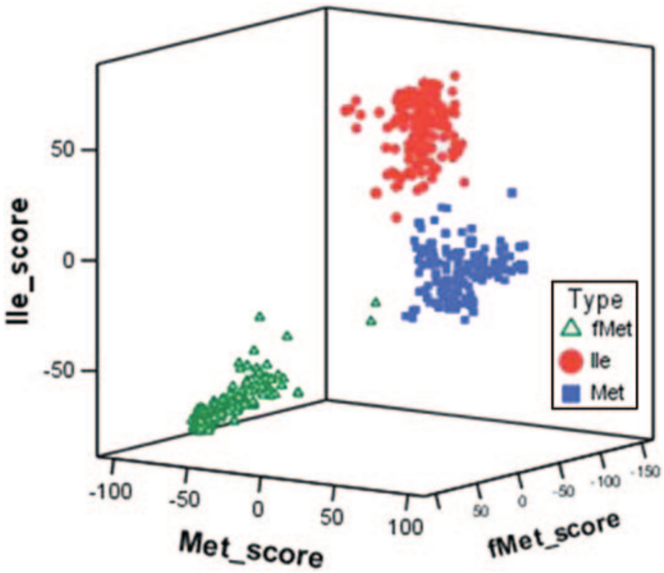

Figure 3. Three-dimensional plot of TFAM scores for tRNA gene sequences. Positive values show proximity to the profile of a specific tRNA type.

We have tested the first and third strategies with the complete datasets producing a discrepancy <1% due to a few atypical sequences and to the six genes producing an incorrect TFAM annotation due to the positive scores for both Ile and Met profiles (see above). We have not tested the complete dataset with the second strategy but, for any of the tested sequences, it produces the same results than those of the first strategy when the species' taxonomic group set is used.

TilS protein and tRNA^{Ile} (LAU) may be substituted by a tRNA^{Ile} (UAU) in bacteria

Classification of the tRNAs with anticodon CAT revealed that the genomes of *M.mobile* did not contain any tRNA^{Ile} (CAT) gene. However, thousands of AUA codons of the coding mRNAs of this species ought to be read. The explanation was that a new tRNA, very unusual in bacteria, was present. This tRNA contained an anticodon UAU (TAT in DNA). Its sequence was more similar to tRNA^{Ile} (GAT) than to the other types. It could be produced through the substitution of G_{34} by T at the tRNA gene. The U_{34} nucleotides at anticodons may read codons ending with any of the four nucleotides, A- or G-ending codons, only A, or even other cases depending of the type of tRNA, U_{34} modification and species (4). The genome of *M.mobile* contains the genes whose encoded proteins are required to produce some modified U_{34}. This will reduce the decoding on this tRNA to several possibilities including (i) the equivalent decoding capability of A and G codons, (ii) the preferential decoding of A- versus G-ending or, even (iii) the complete restriction to the AUA codon.

The *tilS* gene was not annotated in the *M.mobile* genome. We searched the genome using BLAST (tblastn with the *Mycoplasma pulmonis* TilS protein as a query, cut-off expected value = 1) without success and finally we compared the genome region where the gene was present in other

Table 2. Annotation of tRNA genes with anticodon CAT

(A) Average number of differences against the complete sequence sets of fMet, Ile and Met

	fMet	Ile	Met	Ile/Met
hch1	23.16	19.94	21.4	0.9
hch2	20.51	21.85	16.03	1.4
hch3	12.57	22.9	20.16	
hch4	3.65	23.16	20.7	
hch5	3.65	23.16	20.7	
hch6	22.81	13.73	20.21	0.7
pca1	5.53	23.9	20.25	
pca2	5.53	23.9	20.25	
pca3	22.21	20.77	16.21	1.3
pca4	23.34	12.93	20.32	0.6
sru1	23.73	12.81	18.44	0.7
sru2	14.34	19.43	15.0	1.3
sru3	8.79	24.79	20.01	

(B) Average number of differences against the taxon-specific gene sets

	Gamma fMet	Gamma Ile	Gamma Met	Ile/Met
hch1	28.79	22.04	28.24	0.8
hch2	24.28	23.76	13.7	1.7
hch3	15.32	28.55	27.06	
hch4	2.77	28.83	29.51	
hch5	2.77	28.83	29.51	
hch6	28.36	13.82	22.16	0.6

	Delta fMet	Delta Ile	Delta Met	Ile/Met
pca1	4.43	29	24.25	
pca2	4.86	30	24.75	
pca3	26.86	25.67	13	2
pca4	29	7.67	25.75	0.3

	Bacter fMet	Bacter Ile	Bacter Met	Ile/Met
sru1	31.25	16.75	26	0.6
sru2	24.5	24.25	18	1.3
sru3	15.75	31	26.5	

(C) TFAM scores against the fMet, Ile and Met profiles

	fMet	Ile	Met
hch1	−82.66	29.10	−4.67
hch2	−70.60	−24.52	34.30
hch3	17.24	−34.26	−19.69
hch4	78.36	−71.66	−63.71
hch5	78.36	−71.66	−63.71
hch6	−20.07	51.08	−103.41
pca1	68.01	−51.24	−70.65
pca2	69.16	−51.08	−71.27
pca3	−84.68	−6.02	23.33
pca4	−97.05	50.92	−26.58
sru1	−113.45	56.33	−15.78
sru2	−38.51	−18.65	11.14
sru3	44.07	−59.72	−36.65

hch, *H.chejuensis*; *pca*, *P.carbinolicus*; *sru*, *S.ruber*; Gamma, gamma-Proteobacteria; Delta, delta-Proteobacteria; Bacter, Bacteroidetes. Highlighted numbers are minimum numbers of differences (A and B) or the maximum TFAM scores (**C**).

related *Mycoplasma* such as *Mycoplasma synoviae*, *M.pulmonis* and *Mycoplasma hyopneumoniae* (Figure 4). The *tilS* gene was flanked by *pth* and *ftsH* in the other genomes whereas in *M.mobile* not only was the gene lost, but also the DNA was disintegrated indicating that *Mycoplasma* spp., in spite of their small genomes, are still able to lose the non-functional DNA as has been previously stated for bacterial endosymbiont genomes (17,18).

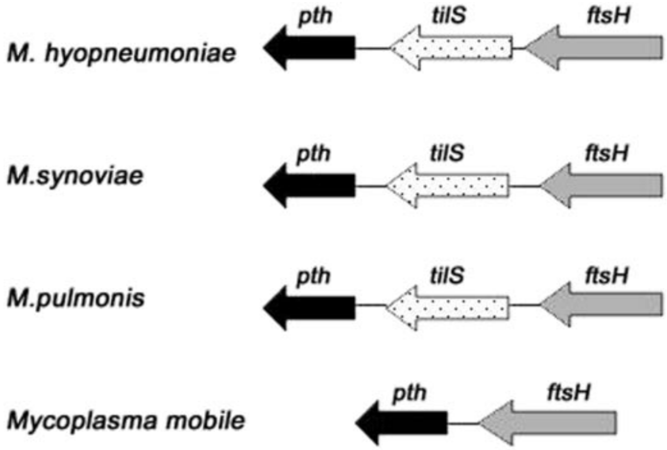

Figure 4. Comparative maps of the region including *tilS* gene.

DISCUSSION

Living organisms use different strategies for decoding AUN codons (3). Although eukaryotic nuclear genomes maintain the standard codon amino acid correspondence (AUH for Ile and AUG for Met), a few mitochondrial genomes have reassigned the AUA codon to Met. Bacteria and Archaea decode AUY codons with tRNAs with anticodon GAU, whereas the AUA codon requires a special tRNA type in which C_{34} is modified to specifically recognize the A-ending codons. In bacteria, the modified nucleotide is lysidine (14). Three types of tRNA genes with anticodon CAT are detected in bacterial genomes (tRNA^{Ile}, tRNA^{Met} and tRNA^{fMet}). Our results show that all three types of tRNA genes may be detected in spite of the small number of nucleotide sites that can be used for their classification. In this paper we have used a system of recruitment based on previous knowledge of the type of tRNA in *E.coli* and *M.capricolum*. Based on the proximity to the *E.coli* and *M.capricolum* sequences, we were able to completely classify the tRNAs from 234 genomes. Initiator tRNAs were more easily classified because of higher conservation. Discrimination between elongator tRNA^{Met} and tRNA^{Ile} was more difficult, especially in some taxonomic groups such as Cyanobacteria. Our results have shown that for any taxonomic group it is easy to produce a phenogram with three well-isolated groups. However, the maintenance of the three groups became more difficult as more distant taxonomic groups were included in the analysis. The reason is because tRNA genes and gene encoding enzymes involved in nucleotide modifications and aminoacylation are coevolving in such a way that positive or negative determinants in some species may not be important for others. Thus, the pairs C_4–G_{69} and C_5–G_{68} in the tRNA^{Ile} (CAU) are required for the *E.coli* TilS enzyme for use as a substrate. However, the loss of the CTD2 domain in the *A.aeolicus* TilS enzyme has meant that these 2 nt pairs were not conserved in the tRNA^{Ile} (CAU), with the consequence that the *E.coli* enzyme is now unable to use *in vitro* the *A.aeolicus* tRNA as a substrate (9).

We propose three methods for the annotation of the three types of genes in bacterial genomes, all of them based on our previous identification of more than 1000 tRNA genes. Identification may be performed based on tree topology, the number of differences compared with type-known sequences or the proximity to three profiles built on the frequencies of nucleotides at each nucleotide position. This is the first time that a method for distinguishing between elongator $tRNA^{Met}$ and $tRNA^{Ile}$ has been developed and it improves the annotation by other systems such as tRNAscan-SE (1) and TFAM (2).

Several nucleotide positions distinguish $tRNA^{fMet}$ from the other two with complete precision. A comparison of our results with those features of eubacterial initiator $tRNA^{fMet}$, previously described in the literature (19), revealed that the feature GGG..CCC (positions 29–31 and 39–41), important for targeting the tRNA to the ribosome P-site, is not conserved in the 15% of the tRNA genes that we have classified as fMet. Most of the differences are in alpha-Proteobacteria and in Chloroplast (they show the signature AGG..CCT), but also in some *Mycoplasma* spp. and mitochondria. Our alignment showed that the conserved feature is $R_{29}G_{30}G_{31}$.. $Y_{39}C_{40}Y_{41}$. Two other features proposed to be required for formylation are the A_{11}–U_{24} base pair at the D stem and the bases C_1, G_2, C_3, G_{70}, C_{71} and A_{72} at the acceptor stem (19). The A_{11} and T_{24} nucleotides are conserved in 100% of the $tRNA^{fMet}$ genes (except the divergent previously described *B.bronchiseptica* tRNA). The nucleotides at the acceptor stem were almost conserved in 100% of the sequences. In 1.5% of them, the important mismatch C_1–A_{72} was replaced by the weak base pair A_1–U_{72} or U_1–A_{72}. This change is probably not affecting the formylation of these tRNAs.

The distinction of $tRNA^{Met}$ and $tRNA^{Ile}$ is more difficult and requires more from the proximity to a profile than from the presence of a special positive or negative determinant.

The analysis of 234 genomes has shown one species which lacks the gene $tRNA^{Ile}$ (CAT). We have been able to establish a scenario in which a new tRNA gene with anticodon TAT has substituted not only the $tRNA^{Ile}$ (CAT) but also the *tilS* gene. This means that *tilS* is not an essential gene as proposed previously (6), based on several mutagenesis studies, in its presence (with the name *mesJ*) in the small genomes of many bacterial endosymbionts (18,20) and even in all gamma-Proteobacterial analysed genomes (21). However, the fact that *M.mobile* has up to now been the only bacterial genome without these two genes may point to a non-perfect discrimination of the AUA and AUG codons. Probably the tRNA with anticodon UAU is erroneously decoding some AUG codons, introducing Ile instead of Met in some polypeptides. Some tRNAs with unmodified U_{34} anticodons are able to decode the four codons of a codon family. This is the case of *Mycoplasma* spp. and mitochondria (22,23). This possibility is clearly stated for some codon families when the analysis of the tRNA gene content in some species reveals a single tRNA (UNN) for the Ala, Val, Pro and Thr codons (8). Several modifications of U_{34} restrict pairing to A- or G-ending codons, and may even enhance the decoding of the A-ending codons (4,22). The s^2 modification (2-thiouridine) could be predicted to enhance the reading of, for example, the GAA versus GAG codon, as was demonstrated in *in vivo* experiments (22). Modifications at position 5 of U_{34} are also involved in the restriction of tRNAs to the A and G-ending codons. *M.mobile* contains the genes encoding the enzymes required for several of these modifications (*mnmE*, *mnmG*, *nmmA* and *iscS*) (24). With these enzymes it would be able to produce 5-carboxymethylaminomethyl-2-thiouridine, a modified uridine detected in several *M.capricolum* tRNAs (15). The lack of the gene *mnmC* in *Mycoplasma* spp. is probably the reason why the common 5-methylaminomethyluridine and 5-methylaminomethyl-2-thiouridine are not detected in *M.capricolum* tRNAs (15).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank David Ardell and Pascual Asensi for their help during the installation of the TFAM program. We also want to thank to two anonymous reviewers for their helpful comments. This work was supported by *Fons d'Investigació de la Universitat de València* (FIU). EB was funded by a FPU fellowship (Ministerio de Educación y Ciencia). Funding to pay the Open Access publication charges for this article was provided by Fons d'Investigació de la Universitat de València (FIU).

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. Ardell,D.H. and Andersson,S.G.E. (2006) TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res.*, **34**, 893–904.
3. Grosjean,H. and Bjork,G.R. (2004) Enzymatic conversion of cytidine to lysidine in anticodon of bacterial tRNA(Ile)—an alternative way of RNA editing. *Trends Biochem. Sci.*, **29**, 165–168.
4. Agris,P.F. (2004) Decoding the genome: a modified view. *Nucleic Acids Res.*, **32**, 223–238.
5. Soma,A., Ikeuchi,Y., Kanemasa,S., Kobayashi,K., Ogasawara,N., Ote,T., Kato,J., Watanabe,K., Sekine,Y. and Suzuki,T. (2003) An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. *Mol. Cell*, **12**, 689–698.
6. Gil,R., Silva,F.J., Pereto,J. and Moya,A. (2004) Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.*, **68**, 518–537.
7. 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.
8. Rocha,E.P. (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.*, **14**, 2279–2286.
9. 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.
10. 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.
11. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible

strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.

12. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinformatics*, **5**, 150–163.
13. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
14. 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.
15. Andachi,Y., Yamao,F., Muto,A. and Osawa,S. (1989) Codon recognition patterns as deduced from sequences of the complete set of transfer-RNA species in *Mycoplasma capricolum*—resemblance to mitochondria. *J. Mol. Biol.*, **209**, 37–54.
16. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139–D140.
17. Silva,F.J., Latorre,A. and Moya,A. (2001) Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet.*, **17**, 615–618.
18. Gomez-Valero,L., Latorre,A. and Silva,F.J. (2004) The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola*. *Mol. Biol. Evol.*, **21**, 2172–2181.
19. RajBhandary,U.L. (1994) Initiator transfer RNAs. *J. Bacteriol.*, **176**, 547–552.
20. Gil,R., Silva,F.J., Zientz,E., Delmotte,F., Gonzalez-Candelas,F., Latorre,A., Rausell,C., Kamerbeek,J., Gadau,J., Holldobler,B. *et al.* (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA*, **100**, 9388–9393.
21. Belda,E., Moya,A. and Silva,F.J. (2005) Genome rearrangement distances and gene order phylogeny in gamma-proteobacteria. *Mol. Biol. Evol.*, **22**, 1456–1467.
22. Takai,K. and Yokoyama,S. (2003) Roles of 5-substituents of tRNA wobble uridines in the recognition of purine-ending codons. *Nucleic Acids Res.*, **31**, 6383–6391.
23. Santos,M.A.S., Moura,G., Massey,S.E. and Tuite,M.F. (2004) Driving change: the evolution of alternative genetic codes. *Trends Genet.*, **20**, 95–102.
24. Jaffe,J.D., Stange-Thomann,N., Smith,C., DeCaprio,D., Fisher,S., Butler,J., Calvo,S., Elkins,T., Fitzgerald,M.G., Hafez,N. *et al.* (2004) The complete genome and proteome of *Mycoplasma mobile*. *Genome Res.*, **14**, 1447–1461.