Evolution of *tuf* genes: ancient duplication, differential loss and gene conversion

Warren C. Lathe III[a,b], Peer Bork[a,b,*]

[a] *EMBL, Meyerhofstr. 1, D69012 Heidelberg, Germany*
[b] *Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany*

Received 21 May 2001; revised 12 June 2001; accepted 20 June 2001

First published online 18 July 2001

Edited by Gunnar von Heijne

Abstract The *tuf* gene of eubacteria, encoding the EF-tu elongation factor, was duplicated early in the evolution of the taxon. Phylogenetic and genomic location analysis of 20 complete eubacterial genomes suggests that this ancient duplication has been differentially lost and maintained in eubacteria. © 2001 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Elongation factor tu; Molecular phylogeny; Bacterial evolution; Gene duplication; Genome; Gene conversion

1. Introduction

The Ef-Tu protein, an elongation factor which loads the amino acyl tRNA molecule onto the ribosome during translation, is a monomeric GTPase similar to Ras proteins. The Ef-Tu protein is encoded by the *tuf* gene in eubacteria. Because of this very important function within the cell, both the nucleotide sequence identity and the genomic location of the *tuf* gene are well conserved between taxa. The amino acid sequence differs by no more than 27% between even the most divergent eubacterial species. In all genomes the *tuf* gene is found in only a small number of different transcriptional 'neighborhoods' or genomic locations [1]. There is only one exception to this conserved genomic location (*Mycoplasma genitalium*).

Early on in the study of the *tuf* gene it was discovered that in many proteobacterial species the *tuf* gene is duplicated. In both *Salmonella typhimurium* and *Escherichia coli*, these duplicate *tuf* genes are nearly identical in nucleotide sequence and experimental evidence shows that either of the *tuf* genes may be deleted without effect on the viability of the cell [2,3]. Because this duplication seems to be widespread in proteobacterial species, though the distribution of the duplication is not universal, it is assumed that the duplication preceded the divergence of the proteobacteria [4,5].

As the high similarity of *tuf* duplicates within species does not easily fit with an ancient duplication, we systematically surveyed 40 eubacterial genomes completed to date for the *tuf* gene. Genomic location information, sequence and phylogenetic analysis suggest that the *tuf* gene underwent a single ancient duplication before the divergence of eubacteria. Whereas most proteobacteria have maintained this duplication, with sparse occurrences outside this group, one or the other of the duplicated genes have been differentially lost in other taxa.

2. Materials and methods

2.1. Sequence and phylogenetic analysis

Amino acid and DNA alignments were done by ClustalX [6]. Phylogenetic and distance analysis was performed with the software package PAUP 4.0b [7]. Though a Maximum Parsimony method was used (heuristic search, closest addition), the neighbor joining distance method gave similar results.

2.2. Genomic analysis

Blast searches of the *tuf* gene were done using NCBI Psi blast [8] using default parameters. *Tuf* gene annotation searches, genomic location determinations and additional blast searches were performed using publicly available (as of November 2000) complete genomes as reported in TIGR [9].

3. Results

3.1. Presence of the tuf gene in eubacteria

We surveyed all complete and annotated eubacterial genomes for the presence of the *tuf* gene and copy number (Table 1). Every genome so far completed has at least one copy of the *tuf* gene, as is expected due to its functional importance in the cell. The distribution of the duplication, on the other hand, is spotty. Five of eight proteobacteria have duplicated *tuf* genes, whereas outside the proteobacteria, only *Deinococcus radiodurans* contains two *tuf* genes.

We performed blastn and Psi blast [8] searches to determine if duplicated *tuf* genes remain undiscovered in genomes where only one *tuf* gene is annotated. There was no evidence that a second *tuf* gene exists in these genomes.

3.2. Sequence analysis of duplications

Sequence analysis of all genomes with duplicated *tuf* genes shows duplicate *tuf* genes within a genome differ by less than 1.4% in nucleotide sequence (average difference is approximately 0.7%, see Table 2). This and phylogenetic analysis (Fig. 1) might indicate that the duplications within a genome are more closely related to each other (orthologous) than they are to *tuf* genes from without the genome. At first glance this suggests that each duplication is an independent duplication within that lineage. Several lines of evidence, on the other hand, point to a different explanation. For example, six independent duplications in seven genomes of proteobacteria

*Corresponding author. Fax: (49)-6221-387 517.
E-mail addresses: lathe@embl-heidelberg.de (W.C. Lathe III); bork@embl-heidelberg.de (P. Bork).

0014-5793/01/$20.00 © 2001 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.
PII: S0014-5793(01)02639-4

Table 1
List of bacterial genomes and number of *tuf* genes through annotated genomes and blast search

Species	Number of *tuf* genes
Proteobacteria	
Escherichia coli	2
Haemophilus influenzae	2
Buchnera sp.	1
Vibrio cholerae	2
Pseudomonas aeruginosa	2
Neisseria meningitidis	2
Campylobacter jejuni	1
Helicobacter pylori	1
Low-GC Gram-positives	
Bacillus subtilis	1
Ureaplasma urealyticum	1
Mycoplasma genitalium	1
High-GC Gram-positives	
Mycobacterium tuberculosis	1
Spirochetes	
Borrelia burgdorferi	1
Treponema pallidum	1
Chlamydia	
Chlamydia pneumoniae	1
Chlamydia trachomatis	1
Cyanobacteria	
Synechocystis sp.	1
Deinococcus	
Deinococcus radiodurans	2

would be an unlikely evolutionary scenario. A more parsimonious scenario would be a single duplication before the divergence of the proteobacteria, with the loss of one *tuf* gene in *Buchnera*. Also, here are several possible fates of duplicated genes. Presumably, the most likely fate of a duplicated gene is loss of function due to mutation. A duplicated gene might obtain a new function through gaining mutations that give the protein a new function selectively advantageous to the organism. In addition, a duplicated gene might be maintained within a genome through a process of gene conversion. Previous experimental research in *Salmonella* shows that the duplicated *tuf* genes in this genome maintain a high level of sequence identity due to gene conversion events [4,10].

3.3. Genomic location

The genomic neighborhood or location of the duplicated and single *tuf* genes in these genomes studied supports a possible scenario for a single early duplication with differential loss and maintenance of the duplicates. From phylogenetic and genomic data, it has been suggested that the ancestral bacterial genome possessed a single 'transcriptional unit' which contained many of the ribosomal proteins and related regulatory proteins [11]. The gene order of this 'transcriptional unit' was likely to be similar to that which can be seen today in several species including *B. subtilis*, *P. aeruginosa* and *N. meningitidis*. This gene order began with the genes *rpmG* (large subunit ribosomal protein) and *secE* (secretory protein) and included approximately 42 ribosomal and translation-related genes. The *tuf* gene was in the latter half of this unit, directly downstream of the *fusA* gene (translation elongation factor G) and upstream of the *rpsJ* gene (small unit ribosomal protein). In genomes where *tuf* is duplicated, the two *tuf* genes are found in one of two genomic locations originating from this ancient gene order, even if broken up and recombined into several new transcriptional units. The *tuf* gene is either upstream of the *rpmG* or *secE* genes, or directly downstream of *fusA* and/or upstream of *rpsJ*. Interestingly, in genomes where there is a single *tuf* gene, this gene is found in either one of the two neighborhoods mentioned.

Based on this and phylogenetic data, a reasonably confident picture of the evolution of the *tuf* gene in eubacteria can be drawn (Fig. 2). In this picture, the *tuf* gene was duplicated very early in the evolution of the eubacteria. We base this on the fact that the duplication is found not only in proteobacteria, but also in *Deinococcus*, suggesting the duplication occurred before this very deep branch separating these two divergent groups. The duplication placed the *tuf* gene in one of two locations (discussed above) in a long ancestral ribosomal array of genes. Though this array of genes was broken up and

Table 2
Amino acid sequence divergence in duplicated *tuf* genes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. *V.c.* A	–	**0.003**	0.003	0.119	0.127	0.127	0.190	0.190	0.193	0.190	0.259	0.259	0.168	0.245	0.631
2. *V.c.* B	**1**	–	0.122	0.122	0.129	0.129	0.193	0.193	0.195	0.193	0.261	0.261	0.170	0.247	0.631
3. *H.i.* A	47	48	–	**0.000**	0.074	0.074	0.170	0.170	0.178	0.175	0.251	0.251	0.124	0.237	0.631
4. *H.i.* B	47	48	**0**	–	0.074	0.074	0.170	0.170	0.178	0.175	0.251	0.251	0.124	0.237	0.631
5. *E.c.* A	50	51	29	29	–	**0.003**	0.150	0.150	0.162	0.160	0.241	0.241	0.102	0.232	0.647
6. *E.c.* B	50	51	29	29	**1**	–	0.150	0.150	0.162	0.160	0.241	0.241	0.102	0.232	0.647
7. *P.a.* A	75	76	67	67	59	59	–	**0.000**	0.185	0.188	0.259	0.259	0.155	0.269	0.655
8. *P.a.* B	75	76	67	67	59	59	**0**	–	0.185	0.188	0.259	0.259	0.155	0.269	0.655
9. *N.m.* A	76	77	70	70	64	64	73	73	–	**0.003**	0.223	0.223	0.180	0.206	0.628
10. *N.m.* B	75	76	69	69	63	63	74	74	**1**	–	0.221	0.221	0.183	0.207	0.629
11. *D.r.* A	102	103	99	99	95	95	103	103	88	87	–	**0.000**	0.267	0.243	0.641
12. *D.r.* B	102	103	99	99	95	95	103	103	88	87	**0**	–	0.267	0.243	0.641
13. *Buc.* A	76	76	49	49	40	40	75	75	71	72	105	105	–	0.250	0.636
14. *B.s.* A	106	106	93	93	91	91	96	96	81	81	96	96	98	–	0.617
15. *A.f.* A	243	243	243	243	249	249	254	254	242	242	250	250	245	238	–

Shown here are all six genomes with duplicated genomes and three others (including the archaebacterium *Archaeoglobus fulgidus*) without duplications for comparison. Above the diagonal is percent difference for a length of approximately 396 amino acids. Below the diagonal is the raw number of differences excluding missing data. Duplicated *tuf* genes are designated A and B depending on genomic location (A is upstream of *rpsJ* and/or downstream of *fusA*, B is upstream of *rpmG* and/or *secE*). Bold numbers are differences between duplicated *tuf* genes within a genome. The number of nucleotide differences in duplicated genes is 8/1182 in *Haemophilus influenzae*, 12 1185 *Escherichia coli*, 9/1182 in *Vibrio cholerae*, 9/1194 in *Pseudomonas aeruginosa*, 17/1182 in *Neisseria meningitidis* and 4/1215 in *Deinococcus radiodurans*. All but one nucleotide difference each in *V. cholerae, E. coli* and *N. meningitidis* are in synonymous sites. Sixteen of 17 nucleotide differences in *N. meningitidis* are in the first 252 bp of the 1182 bp sequence.

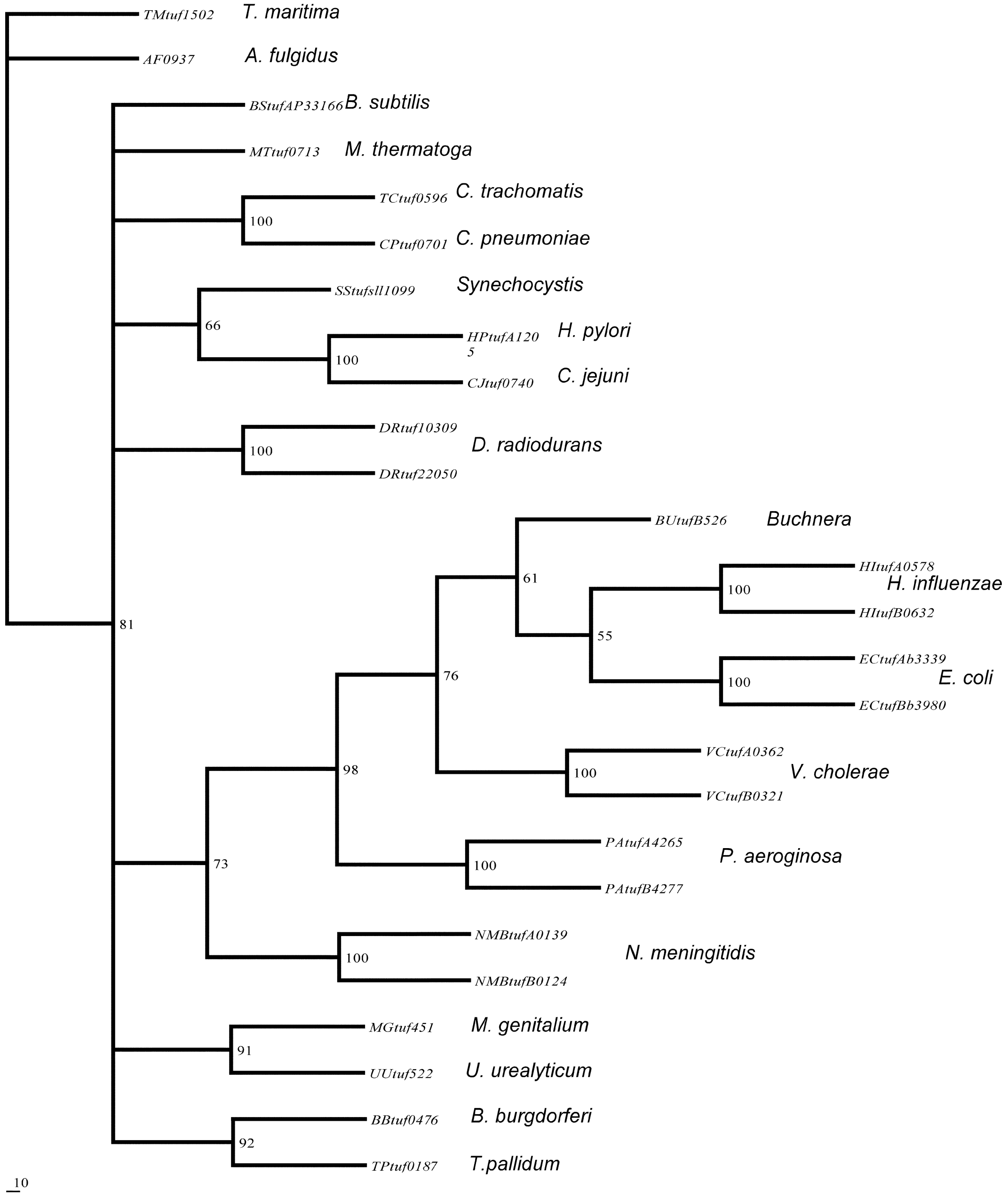

Fig. 1. Duplication phylogeny. 50% consensus bootstrap (1000 replications) tree using Maximum Parsimony algorithm, heuristic search, stepwise addition (closest). *Thermotoga maritima* and *Archaeoglobus fulgidus* were used as outgroup.

recombined over evolutionary time in eubacteria, the neighborhood context of the two *tuf* genes has remained relatively constant (one exception is *M. genitalium*). Though one or other of the *tuf* genes was lost in most genomes, many (mainly proteobacteria, but *D. radiodurans* also) have retained the duplication by way of gene conversion.

4. Discussion

Much recent work on bacterial genome evolution has suggested that genes and operons have been massively transferred horizontally across bacterial genomes [12]. Even the *rps14* gene, encoding a small subunit ribosomal protein, seems to have had several cases of introduction into a bacterial genome through horizontal transfer [13]. Though many of these cases are strongly supported by the evidence, the evolution of the *tuf* gene should provide a cautionary note when invoking horizontal transfer in cases of spotty distribution, incongruent gene and species phylogenies and seemingly confusing evolutionary scenarios. The *tuf* gene is one of the most conserved genes in the bacterial genome and and is well characterized in experimental studies. As such, it is a simple matter to show this gene's evolution since the divergence of eubacteria. Though its duplication has a spotty distribution [5], gene conversion and genomic location allow us to show that the duplication has been differentially lost and maintained in different genomes. A scenario for a less conserved and functionally important gene could be simply invoked for a greater number of duplications, genomic rearrangements, and differential maintenance by gene conversion and loss, that would over

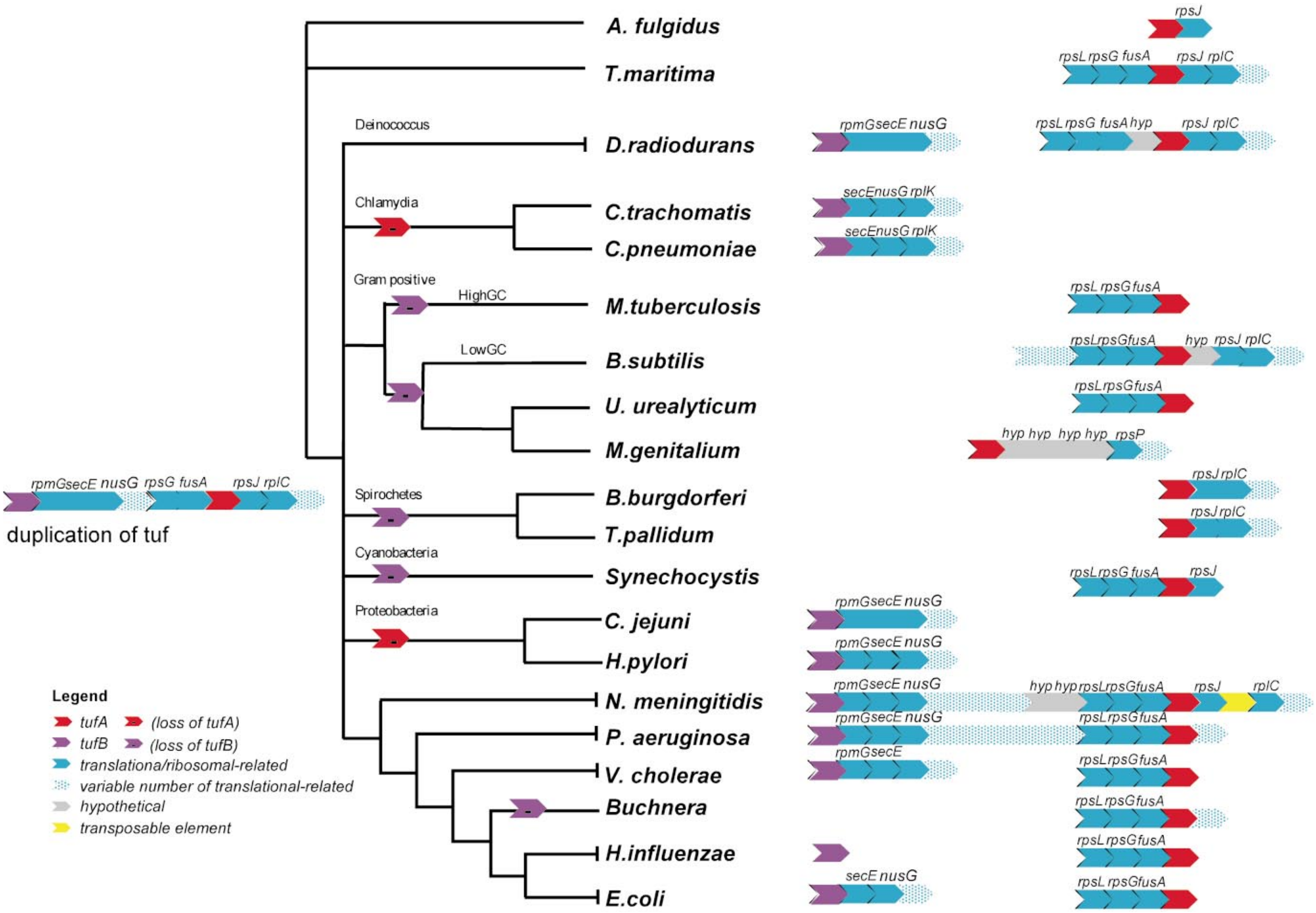

Fig. 2. A cladogram of the evolution of *tuf* in eubacteria based on *tuf* and *rpsJ* phylogeny and accepted taxonomy. An ancient duplication of the *tuf* gene (*tufA*) occurred early in the divergence of eubacteria with the addition of a second *tuf* gene (*tufB*) upstream of the *rpmG* gene, the first gene in the proposed ancient transcriptional unit (shown). The proposed original *tuf* gene (based on location of *tuf* in *T. maritima* and archaebacteria) is shown in red, duplicated *tufB* in purple. Other genes of the ribosomal/translational unit are shown in blue, with light blue designating several ribosomal and translation-related genes. Units separated by white space denote transcriptional units separated by three or more open reading frames (non-ribosomal/translation-related) in the genome. Based on genome location, *tufA* is subsequently lost in both the *Chlamydia* and ε subdivision proteobacteria whereas *tufB* is lost in *Buchnera*, *Synechocystis*, the spirochetes, and low- and high-GC Gram-positive bacteria. Early studies suggest clostridia, a Gram-positive low-GC taxon, has maintained the *tuf* duplication [4], suggesting the other low-GC Gram-positive taxa lost *tufB* in a separate event from the high-GC Gram-positives. The duplication is maintained in the remaining proteobacteria and *D. radiodurans*.

time cloud the picture of the evolution of the gene and a mistaken appeal to horizontal transfer to explain its evolution.

Why gene conversion has maintained the *tuf* gene duplication in most proteobacteria yet it is lost in all other complete genomes save *D. radiodurans* is a matter of conjecture, but it is interesting to note that the recBCD repair system that has been shown to result in gene conversion of the *tuf* gene in *Salmonella* is seemingly lacking in many if not most non-proteobacteria. Perhaps this specific mechanism of repair and conversion has the side effect of often maintaining gene duplications through gene conversion, whereas other mechanisms might be less likely to do so.

The completion and annotation of bacterial genomes is giving us a more complete picture of the evolution not only of the *tuf* gene as shown here, but of the genome as a whole. Further analysis of these individual genes and the genomes as a whole will give us a clearer picture of the roles of duplication, gene conversion, recombination and horizontal transfer in the evolution of the bacterial genome.

References

[1] Lathe, W.C. and Snel, B. (2000) Trends Biochem. Sci. 25, 474–479.

[2] Zuurmond, A.M. and Rundlof, A.K. (1999) Mol. Gen. Genet. 260, 603–607.

[3] Hughes, D. (1990) J. Mol. Biol. 215, 41–51.

[4] Abdulkarim, F. (1996) J. Mol. Biol. 260, 506–522.

[5] Sela, S., Yogev, D. and Razin, S. (1989) J. Bacteriol. 171, 581–584.

[6] Jeanmougin, F., Thompson, J.D., Gouy, M. and Higgins, D.G. (1998) Trends Biochem. Sci. 23, 405.

[7] Swofford, D.L. (1998) PAUP v4.0b, Sinauer, Sunderland, MA.

[8] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.

[9] TIGR: http://www.tigr.org/tdb/mdb/mdbcomplete.html.

[10] Hughes, D. (2000) J. Mol. Biol. 297, 355–364.

[11] Wachterhauser, G. (1998) Syst. Appl. Microbiol. 21, 473–477.

[12] Lawrence, J.G. (1996) Genetics 143, 1843–1860.

[13] Brochier, C. and Philippe, H. (2000) Trends Genet. 16, 529–533.