UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
Departament de Bioquímica i Biotecnologia

Codon usage adaptation in prokaryotic genomes

Memòria presentada per optar al grau de
Doctor per la Universitat Rovira i Virgili.
Amb menció de Doctorat Europeu.
Tarragona, 29 de Novembre de 2007.

Vist i plau del director de tesi: L'interessat:

Dr Santiago Garcia-Vallvé **Pere Puigbò Avalos**

Departament de Bioquímica i Biotecnologia

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

AGRAÏMENTS

En primer lloc agrair a les persones més importants i que més han fet per a què jo realitzi aquesta tesi, els meus pares. A ells els hi dedico aquesta tesi per tot el suport que m'han donat i que sempre em donen per a seguir endavant. Bé i també a la Laura, la Nina i el Rudy, clar! Ells sempre m'escolten pacientment quan assajo una presentació!

Agrair al meu director de tesi, el Santi Garcia-Vallvé (Vallvé), per tota l'ajuda que m'ha donat durant els moments bons i els dolents. Tot el que m'ha ensenyat m'ha fet créixer no tan sols com a científic, sinó com a persona. També a l'Anton Romeu, ell va ser qui em va permetre entrar a la URV i així he pogut gaudir d'aquests 4 anys de tesi.

Als meus companys de laboratori, sobretot a la Montserrat (és difícil escriure els agraïments sense la teva ajuda, vull dir, gairebé sense la teva ajuda!) i a l'Albert (pelut! com trobaré a faltar els moments "llançament de pilota"!) que van iniciar aquesta aventura amb mi i els he hagut ... ai!, m'han hagut d'aguantar durant aquests anys. Amb ells he passat els millors moments de la tesi. Ànims, que ja us queda menys! També a tota la gent que en un moment o altre han residit en laboratori de Bioinformàtica, a la Marina i al Pep (molt d'èxit en les vostres noves aventures, se us troba molt a faltar!), a l'Eduard (l'"Avi", que sempre està disposat a donar un cop de mà quan el necessites i no em refereixo a... quan diu obre la boca i tanca els ulls que no farà mal!), a l'Esther (en el laboratori va haver un abans i un després de l'Esther, gràcies per ser com ets!), a la Laura (molt d'èxit en la tesi, estàs al començament i segur que t'ho passaràs molt bé), a la Safae (trobo a faltar les converses polítiques del divendres per la tarda, molt d'èxit a tu també!) i al Gerard

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

("Colombo" ets collonut, jo de gran vull ser com tu!).

També agrair a tots els que no han residit en el laboratori, però que amb les seves visites han contribuït a fer millor la meva estada a la URV. A la Montse Pinent (ella em va posar en els seus agraïments i si no la poso protestarà!), a l'Helena (vigila que el Santi vagi pel bon camí quan jo no hi sigui!), a la Gemma (sempre ens quedarà Llangollen!), la Sabina (pel dia de la tesi espero un dels teus plats!) i el David Pajuelo (Pajuel! Aquest any repetim Schrödinger, no?). També al Josep Mª del Bas (Josepet, molt d'èxit a Liverpool i ens veiem a Anfield!) i a la gent d'orgànica, vull dir ... a la Lídia, clar! (on anirem a parar? Gonsales, molts ànims a tu també que ja queda menys!). També a aquells que estaven al començament de la tesi, al Cesc, la Vanesa, el Nino i la Montse Vadillo, i també a la resta de doctorands del departament. Al tots els professors del departament, especialment al Gerard Pujadas (ja que els agraïments són lo més important!) i al Miguel Angel Montero. També a la gent de secretaria, especialment, a la Cristina per la seva ajuda en els primers anys i a la Maribel que sense ella mai hagués pogut entregar la tesi a temps. Bé, per no allargar-me i no deixar-me a ningú, a tota la gent del departament i de la URV en general, gràcies a tots i totes! I també (last but not least...) a gent de fora de la URV, especialment a la Cris (la meva germaneta!), el Jaume (xurri!), la Dolors (noieta, busca una bona xocolateria a Alemanya!) i a lo Jordi Puxeu!

I would also like to thank James McInerney, people from the Bioinformatics lab at NUIM (not only Angela, also Davide, James Cotton, Fergal and Vicky) and Mary. Thank you very much and see you next year in Barcelona!

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

"...when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"

Sir Arthur Conan Doyle (Sherlock Holmes)

The Sign of Four (1890)

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

CONTENTS

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

INTRODUCTION

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Translational selection and highly expressed genes

Since the first nucleic acid sequences were sequenced and compared, several hypotheses about the evolution of genes and genomes have been proposed. Today, with the availability of a vast amount of sequences of proteins, genes and even genomes from all kinds of species, some of these hypotheses remain unchanged. From the limited nucleic acid sequences available, Grantham et al. (1) proposed the "genome hypothesis", postulating that genes in any given bacterial genome show a very similar pattern of choices among synonymous codons. In *E. coli*, *S. cerevisiae* and other model organisms (2), ribosomal-protein genes and other highly expressed genes were found to have a pronounced codon usage bias because they use a small subset of synonymous codons, i.e. codons that are recognized by the most abundant tRNA species (3). This bias is the result of "translational selection", i.e. using a codon that is translated by an abundant tRNA species will increase efficiency and accuracy (4). The fast accumulation of genes from the same species led to a series of multifactorial codon usage analyses to check whether translational selection was a general phenomenon between prokaryotic species. Theoretical studies showed that the occurrence of codon usage bias depends on factors such as the effective size of haploid population and reflects a balance between several forces like translational selection, mutational and positional bias and random genetic drift (4, 5). One of the currently accepted views is that genome-wide codon bias is determined primarily by mutational processes and only secondarily by translational selection (6). The available sequences of more than a hundred prokaryotic complete genomes have not changed this panorama. The genome hypothesis has enabled the search for compositionally anomalous genes or genome regions that are expected to be horizontally transferred (7, 8). Translational selection has helped to computationally predict a group of highly expressed genes in some genomes (9, 10). These studies have shown that genes that codify ribosomal proteins, translation and transcription processing factors and chaperone-degradation proteins are usually highly expressed. In addition, specific taxonomic groups also express genes involved in some of their metabolic characteristic pathways in high amounts. For example, among the highly expressed genes from *Deinococcus radiodurans*, some genes are involved in radiation resistance (11).

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Not all bacterial species seemed to be under this selection and species such as *M. tuberculosis* (12), *H. pylori* (13), *Mycoplasma genitalium* and *M. pneumoniae* (Kerr et al 1997) or Spirochaetes like *Borrelia burdorgferi* and *Treponema pallidum* (14) were found to be under a weak or no selection, respectively. Other species were there is a weak or no evidence of translational selection are *Rickettsia prowazekii, Chlamydia trachomatis, Chlamydophila pneumoniae, Blochmannia floridanus and Buchnera sp.* (15). However, Karlin and coworkers predicted a group of highly expressed genes in these species using the E(g) 'expression measure' (9, 16, 17, 18). In genomes with a high or low G+C content, it is difficult to evaluate translational selection because of the effect of the extreme (high or low) G+C content on the codon usage of genes. The genome from *Pseudomonas aeruginosa* is an example of this. Carbone and coworkers (10) used an iterative algorithm to suggest that translational selection bias does not dominate in this species. However, other researchers have shown that in this species the variation in codon usage among genes is associated with expression, although this is not the major trend (19).

Methods to predict highly expressed genes

In 1987, Sharp and Li (20) developed the Codon Adaptation Index (CAI) to measure the resemblance between the synonymous codon usage of a gene and the synonymous codon frequencies of a reference set. The CAI index ranges from zero to one: it is 1 if a gene always uses, for each encoded amino acid, the most frequently used synonymous codon in the reference set. Though it was developed to assess the extent to which selection has been effective at moulding the pattern of codon usage(21), it has other uses, e.g. for assessing the adaptation of viral genes to their hosts (21), for giving an approximate indication of the likely success of heterologous gene expression (22, 23), for making comparisons of codon usage in different organisms (21, 24)(21), for detecting dominating synonymous codon usage bias in genomes (10) and for studying cases of horizontally transferred genes (8). The CAI, developed by Sharp and Li (20) is the index that is most commonly used, by itself (22, 25, 26, 27) or in combination with an iterative algorithm (10), to predict highly expressed genes that use the degree of bias in their codon usage. However, recently an improved modification of the CAI has been proposed by Xia (28). The highly expressed genes predicted using the codon usage bias are expected to be

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

genes with a high expression in different situations, e.g. different media or growth phases. In such situations, translational selection is strong enough to modulate the codon usage of highly expressed genes. Independently of the method used to predict a group of highly expressed it must first be checked if a genome is under translational selection or not. In absence of translational selection, the expression levels of genes cannot be predicted from comparisons of codon usage (15). However, some authors predict a group of highly expressed genes without checking, by any method, whether a genome is under translational selection or not.

Defining a group of highly expressed genes is interesting not only for determining the metabolic capabilities of the genomes under translational selection but also for other reasons. Groups of highly expressed genes can be used to reduce the false positives of the predictions of acquired genes because they are compositionally different from the other genes in a genome (8, 29). The prediction of highly expressed genes can also be used to re-design synthetic genes to increase their expression level. If a gene contains codons that are rarely used by the host, its expression level will not be maximal.

Codon usage adaptation

Codon usage adaptation to a new genome (Amelioration)

The prediction of horizontally transferred genes using atypical nucleotide composition is based on the genome hypothesis (1) that assumes that codon usage and G+C content are distinct global features of each prokaryotic genome. With this method, a significant number of prokaryotic genes have been proposed as having been acquired by HGT (7, 8, 30). However, it cannot predict all acquired genes unambiguously (31) because genes may have adjusted to the base composition and codon usage of the host genome or because an unusual composition may be due to factors other than HGT (7). The term ‘amelioration’ is used to describe how genes acquired by horizontal gene transfer adapt their DNA composition to a new genome (32) . At the time of introduction, horizontally transferred genes have the base composition and codon usage pattern of the donor genome. But because transferred genes are subject to those mutational processes affecting the recipient genome, the acquired sequences will incur substitutions and eventually come to reflect the DNA

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

composition of the new genome. This process of ‘‘amelioration’’—whereby a sequence adjusts to the base composition and codon usage of the resident genome—is a function of the relative rate of G/C to A/T mutations (23). Models of amelioration can be used to estimate the time of introgression of foreign genes in a chromosome (32).

The genes originally encoded in the proto-mitochondria and now encoded in the nuclear genome could be a good example to assess the amelioration process. It is widely accepted that mitochondrion had a single origin, arising from a bacterial symbiont whose closest contemporary relatives are found within the a-proteobacteria (33, 34). Since its origin, the mitochondrial genome has undergone a streamlining process of genome reduction with intense periods of loss of genes. Currently, mitochondrial genomes exhibit a great variation in protein gene content among most major groups of eukaryotes, but only limited variation within large and ancient groups. This suggests a very episodic, punctuated pattern of mitochondrial gene loss over the broad sweep of eukaryotic evolution (35). Mitochondrial genomes have lost genes that lack a selective pressure for their conservation. This may include genes whose function may no longer be necessary, genes whose function has been superseded by some pre-existing nuclear genes or genes that have been transferred to the nucleus (36). The gene content of present mitochondrial genomes varies from 67 protein-coding genes in Reclinomonas americana, a flagellate protozoon, to 3 genes in other species. Mitochondria in humans and animals encode 13 respiratory-chain proteins and a minimal set of tRNAs that suffices to translate all codons. However, the vast majority of mitochondrial proteins are the products of nuclear genes. These genes are translated in the nucleus and the proteins are later transported to the mitochondria. Some of them i.e. those with a prokaryote homologue are thought to be the result of horizontal gene transfer events from the proto-mitochondrial to the nuclear genome. This hypothesis is reinforced by the fact that several of these genes are encoded in the mitochondrial genome in other eukaryotic species (37).

Codon usage adaptation to termophily

G+C content and optimal growth temperature are the two factors that most influence differences in amino acid composition and codon usage between organisms.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Analysis of the optimal temperatures of the enzymes extracted from hyperthermophilic organisms showed that thermal resistance was an intrinsic property of these enzymes (38). Comparative analysis of the amino acid composition of orthologous proteins from several mesophilic and thermophilic organisms indicated some amino acid substitutions that are preferred in thermophiles (38). However, the small number of sequences analyzed and the fact that factors other than temperature can affect the amino acid composition of proteins revealed the inconsistency of theses results (39). Comparison of the first completely sequenced genomes of several thermophiles and mesophiles showed that proteins from thermophiles contain higher levels of both charged and hydrophobic residues and lower levels of polar and uncharged ones (40). Once more complete genomes were sequenced, new analyses were performed using different methods and different datasets (41, 42, 43, 44, 45, 46, 47, 48). Although these studies show several discrepancies in the role of each amino acid, there is a consensus that glutamate (E) and, to a lesser extent, valine (V) are the amino acids that are more represented in thermophiles than in mesophiles.

There are greater discrepancies, however, over which amino acids are used with the lowest frequency in thermophiles or with the highest frequency in mesophiles. For example, Singer and Hickey [25] found that these amino acids were A, H, Q and T; Kreil and Ouzounis (41) found that they were Q and T; and Tekaia and coworkers (42) found only Q. These discrepancies indicate that hyperthermophilic and mesophilic enzymes may be very similar – their difference being that hyperthermophilic enzymes are more rigid than mesophilic enzymes (38). To increase their rigidity, hyperthermophilic enzymes may adopt several strategies but a common rule could be that more charged residues are found in hyperthermophilic proteins, mostly at the expense of uncharged polar residues (38). Computational, biochemical, and structural evidence now supports the hypothesis that ion pair formation, hydrogen bonds, and hydration, rather than hydrophobic interactions, play important roles in the stabilization of enzymes from extremophiles (49). Also, we cannot talk of a common amino acid usage in mesophiles because an adaptation to live at intermediate temperatures is unnecessary. When comparing the amino acid compositions of thermophilic and mesophilic proteins, therefore, different datasets and methods obtain different results. The relationship between genomic G+C content

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

and optimal growth temperature in prokaryotes has been debated recently in the literature (50, 51, 52, 53). Because G:C pairs in DNA are more thermally stable than A:T pairs, it has been suggested that a high G+C content may be a selective response to high temperature. In this sense, a significant correlation has been observed between optimal growth temperature and the G+C content of structural RNAs (51, 52). When open reading frames are analyzed, some studies have concluded that there is no correlation between G+C content and optimal growth temperature (50, 51, 52) and others have found a positive correlation among some families of prokaryotes (53). However, Pasamontes and Garcia-Vallve (54), using a multi-way method for comparing the amino acid composition of several groups of orthologous proteins from the same group of species, have showed that amino acid variations related to variations of G+C content and optimal growth temperature are independent and that the observed G+C-dependence is not a consequence of a thermophily dependence

It has been shown that thermophilic species have distinguishable patterns of synonymous codon usage (45, 48) and there are evidences that this difference is the result of selection related to thermophily (55). However, some authors argue that the difference in synonymous codon usage between (hyper)thermophilic and non-thermophilic species cannot be clearly attributed to a selective pressure linked to growth at high temperatures (56). Several authors have found that such a pattern was not simply due to the fact that most of the thermophiles studied were archaea rather than eubacteria, i.e. a distinguishable patterns of synonymous codon usage between thermophiles and mesophiles, and not between eubacteria and archaea (55, 57). Analyzing the synonymous codon usage of 16 genomes, Singer and Hickey observed differences in the frequencies of 23 codons (48). Among the thermophilies, they found increases in the relative frequencies of 11 codons (GGA, AGG, AGA, AAG, AAC, ATA, TAC, TTC, CAC, CTT and CTC) and decreases in 12 codons (AAT, ATT, ATC, TAT, TTG, TTT, CGG, CGA, CGT, CGC and CAT) (48). Most of these changes are due to: (i) a preference for C over T in the two-fold degenerate NNY codon groups and (ii) increase in ‘purine-rich’ codons (48). These patterns highlight an increase in AGR codons for arginine and ATA codons for isoleucine, and a decrease in CGN codons for arginine (45, 48, 55). Comparing the mRNA sequences of 72 fully sequenced prokaryotic proteomes (14 thermophilic and 58 mesophilic

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

species) Paz and coworkers (58) has showed that the thermophile purine-pyrimidine (R/Y) ratio within their mRNAs is significantly higher than that of the mesophiles, suggesting that mixed adenine·guanine and polyadenine tracts in mRNAs increase their thermostability (58).

Codon usage adaptation to increase gene expression (optimization)

Gene expression levels depend on many factors, such as promoter sequences and regulatory elements. One of the most important factors is the adaptation of the codon usage of the transcript gene to the typical codon usage of the host (59). Therefore, highly expressed genes in prokaryotic genomes under translational selection have a pronounced codon usage bias. This is because they use a small subset of codons that are recognized by the most abundant tRNA species (60). The force that modulates this codon adaptation is called translational selection and its strength is important in fast-growing bacteria (61, 62). If a gene contains codons that are rarely used by the host, its expression level will not be maximal. This may be one of the limitations of heterologous protein expression (63) and the development of DNA vaccines (64). A high number of synthetic genes have been re-designed to increase their expression level. The Synthetic Gene Database (SGDB) (65) contains information from more than 200 published experiments on synthetic genes. In the design process of a nucleic acid sequence that will be inserted into a new host to express a certain protein in large amounts, codon usage optimization is usually one of the first steps (63). Codon usage optimization basically involves altering the rare codons in the target gene so that they more closely reflect the codon usage of the host without modifying the amino acid sequence of the encoded protein (63). The information usually used for the optimization process is therefore the DNA or protein sequence to be optimized and a codon usage table (which we call the reference set) of the host.

There are several public web servers and stand-alone applications that allow some kind of codon optimization. ‘GeneDesign’ (66), ‘Synthetic Gene Designer’ (67) and ‘Gene Designer’ (68) are packages that provide a platform for synthetic gene design, including a codon optimization step. Other programs, such as DNAWorks (69) and GeMS (70), focus more on the process of oligonucleotide design for synthetic gene construction. The stand-alone application INCA provides an array of features,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

including now codon optimization, which are useful for analyzing synonymous codon usage in whole genomes (71). JCAT (72), 'Codon optimizer' (73), UpGene (74) and the server presented here focus on the codon optimization process. Although each server and application has its own features, all of them have several features in common. Most offer several options for the input of the codon usage reference set. One of these options is the possibility of using the tables from the Codon Usage database (75). Usually, a limited number of pre-computed tables of codon usage are available to be used as a reference set in the optimization process. In addition, not all of the available pre-computed reference sets correspond to a group of highly expressed genes (the proper reference set needed to optimize for increasing gene expression level). Though most of the programs and servers use a group of highly expressed genes from E. coli as a pre-computed reference set, only the 'Synthetic Gene Designer' and 'GeneDesign' servers provide a pre-computed group of highly expressed genes for 11 and 4 organisms, respectively. The exception is the JCAT web server, which offers pre-computed tables of predicted highly expressed genes from more than 200 bacterial species. However, this server uses the method of Carbone et al. (76) to predict a group of genes with a biased codon usage. These groups of genes do not always correspond to a group of highly expressed genes because not all bacterial species are under translational selection (76, 77). The high number of pre-computed codon usage tables from bacteria and archaea that are not under translational selection available in JCAT therefore creates some confusion.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

1. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.,* **8,** r49-r62.

2. Sharp,P.M., Cowe,E., Higgins,D.G., Shields,D.C., Wolfe,K.H. and Wright,F. (1988) Codon usage patterns in escherichia coli, bacillus subtilis, saccharomyces cerevisiae, schizosaccharomyces pombe, drosophila melanogaster and homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Res.,* **16,** 8207-8211.

3. Ikemura,T. (1981) Correlation between the abundance of escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.,* **146,** 1-21.

4. Sharp,P.M., Stenico,M., Peden,J.F. and Lloyd,A.T. (1993) Codon usage: Mutational bias, translational selection, or both? *Biochem. Soc. Trans.,* **21,** 835-841.

5. Carbone,A., Kepes,F. and Zinovyev,A. (2005) Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Mol. Biol. Evol.,* **22,** 547-561.

6. Chen,S.L., Lee,W., Hottes,A.K., Shapiro,L. and McAdams,H.H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. U. S. A.,* **101,** 3480-3485.

7. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.,* **10,** 1719-1725.

8. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: A database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.,* **31,** 187-189.

9. Karlin,S. and Mrazek,J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.,* **182,** 5238-5250.

10. Carbone,A., Zinovyev,A. and Kepes,F. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics,* **19,** 2005-2015.

11. Karlin,S. and Mrazek,J. (2001) Predicted highly expressed and putative alien genes of deinococcus radiodurans and implications for resistance to ionizing radiation damage. *Proc. Natl. Acad. Sci. U. S. A.,* **98,** 5240-5245.

12. Andersson,G.E. and Sharp,P.M. (1996) Codon usage in the mycobacterium

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

tuberculosis complex. *Microbiology,* **142 (Pt 4),** 915-925.

13. Lafay,B., Atherton,J.C. and Sharp,P.M. (2000) Absence of translationally selected synonymous codon usage bias in helicobacter pylori. *Microbiology,* **146 (Pt 4),** 851-860.

14. Lafay,B., Lloyd,A.T., McLean,M.J., Devine,K.M., Sharp,P.M. and Wolfe,K.H. (1999) Proteome composition and codon usage in spirochaetes: Species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.,* **27,** 1642-1649.

15. Henry,I. and Sharp,P.M. (2007) Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.,* **24,** 10-12.

16. Karlin,S., Mrazek,J., Campbell,A. and Kaiser,D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J. Bacteriol.,* **183,** 5025-5040.

17. Karlin,S., Barnett,M.J., Campbell,A.M., Fisher,R.F. and Mrazek,J. (2003) Predicting gene expression levels from codon biases in alpha-proteobacterial genomes. *Proc. Natl. Acad. Sci. U. S. A.,* **100,** 7313-7318.

18. Karlin,S., Theriot,J. and Mrazek,J. (2004) Comparative analysis of gene expression among low G+C gram-positive genomes. *Proc. Natl. Acad. Sci. U. S. A.,* **101,** 6182-6187.

19. Grocock,R.J. and Sharp,P.M. (2002) Synonymous codon usage in pseudomonas aeruginosa PA01. *Gene,* **289,** 131-139.

20. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.,* **15,** 1281-1295.

21. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.,* **15,** 1281-1295.

22. Wu,G., Culley,D.E. and Zhang,W. (2005) Predicted highly expressed genes in the genomes of streptomyces coelicolor and streptomyces avermitilis and the implications for their metabolism. *Microbiology,* **151,** 2175-2187.

23. Lawrence,J.G. and Ochman,H. (1998) Molecular archaeology of the escherichia coli genome. *Proc. Natl. Acad. Sci. U. S. A.,* **95,** 9413-9417.

24. Grote,A., Hiller,K., Scheer,M., Munch,R., Nortemann,B., Hempel,D.C. and Jahn,D. (2005) JCat: A novel tool to adapt codon usage of a target gene to its

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

potential expression host. *Nucleic Acids Res.,* **33,** W526-31.

25. Goetz,R.M. and Fuglsang,A. (2005) Correlation of codon bias measures with mRNA levels: Analysis of transcriptome data from escherichia coli. *Biochem. Biophys. Res. Commun.,* **327,** 4-7.

26. Wu,G., Nie,L. and Zhang,W. (2006) Predicted highly expressed genes in nocardia farcinica and the implication for its primary metabolism and nocardial virulence. *Antonie Van Leeuwenhoek,* **89,** 135-146.

27. Martin-Galiano,A.J., Wells,J.M. and de la Campa,A.G. (2004) Relationship between codon biased genes, microarray expression values and physiological characteristics of streptococcus pneumoniae. *Microbiology,* **150,** 2313-2325.

28. Xia,X. (2007) An improved implementation of codon adaptation index. *Evolutionary Bioinformatics,* **3,** 53-58.

29. Garcia-Vallve,S., Palau,J. and Romeu,A. (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in escherichia coli and bacillus subtilis. *Mol. Biol. Evol.,* **16,** 1125-1134.

30. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature,* **405,** 299-304.

31. Lawrence,J.G. and Ochman,H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.,* **10,** 1-4.

32. Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: Rates of change and exchange. *J. Mol. Evol.,* **44,** 383-397.

33. Burger,G., Gray,M.W. and Lang,B.F. (2003) Mitochondrial genomes: Anything goes. *Trends Genet.,* **19,** 709-716.

34. Gray,M.W., Burger,G. and Lang,B.F. (1999) Mitochondrial evolution. *Science,* **283,** 1476-1481.

35. Adams,K.L. and Palmer,J.D. (2003) Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Mol. Phylogenet. Evol.,* **29,** 380-395.

36. Gray,M.W., Burger,G. and Lang,B.F. (2001) The origin and early evolution of mitochondria. *Genome Biol.,* **2,** REVIEWS1018.

37. Gabaldon,T. and Huynen,M.A. (2004) Shaping the mitochondrial proteome. *Biochim. Biophys. Acta,* **1659,** 212-220.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

38. Vieille,C. and Zeikus,G.J. (2001) Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.,* **65,** 1-43.

39. Bohm,G. and Jaenicke,R. (1994) Relevance of sequence statistics for the properties of extremophilic proteins. *Int. J. Pept. Protein Res.,* **43,** 97-106.

40. Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M., et al. (1998) The complete genome of the hyperthermophilic bacterium aquifex aeolicus. *Nature,* **392,** 353-358.

41. Kreil,D.P. and Ouzounis,C.A. (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.,* **29,** 1608-1615.

42. Tekaia,F., Yeramian,E. and Dujon,B. (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. *Gene,* **297,** 51-60.

43. Lobry,J.R. and Chessel,D. (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J. Appl. Genet.,* **44,** 235-261.

44. Cambillau,C. and Claverie,J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.,* **275,** 32383-32386.

45. Farias,S.T. and Bonato,M.C. (2003) Preferred amino acids and thermostability. *Genet. Mol. Res.,* **2,** 383-393.

46. Nakashima,H., Fukuchi,S. and Nishikawa,K. (2003) Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J. Biochem. (Tokyo),* **133,** 507-513.

47. Saunders,N.F., Thomas,T., Curmi,P.M., Mattick,J.S., Kuczek,E., Slade,R., Davis,J., Franzmann,P.D., Boone,D., Rusterholtz,K., et al. (2003) Mechanisms of thermal adaptation revealed from the genomes of the antarctic archaea methanogenium frigidum and methanococcoides burtonii. *Genome Res.,* **13,** 1580-1588.

48. Singer,G.A. and Hickey,D.A. (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene,* **317,** 39-47.

49. Rossi,M., Ciaramella,M., Cannio,R., Pisani,F.M., Moracci,M. and Bartolucci,S. (2003) Extremophiles 2002. *J. Bacteriol.,* **185,** 3683-3689.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

50. Wang,H.C., Susko,E. and Roger,A.J. (2006) On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. *Biochem. Biophys. Res. Commun.,* **342,** 681-684.

51. Galtier,N. and Lobry,J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.,* **44,** 632-636.

52. Hurst,L.D. and Merchant,A.R. (2001) High guanine-cytosine content is not an adaptation to high temperature: A comparative analysis amongst prokaryotes. *Proc. Biol. Sci.,* **268,** 493-497.

53. Musto,H., Naya,H., Zavala,A., Romero,H., Alvarez-Valin,F. and Bernardi,G. (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.,* **573,** 73-77.

54. Pasamontes,A. and Garcia-Vallve,S. (2006) Use of a multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes. *BMC Bioinformatics,* **7,** 257.

55. Lynn,D.J., Singer,G.A. and Hickey,D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.,* **30,** 4272-4277.

56. Lobry,J.R. and Necsulea,A. (2006) Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene,* **385,** 128-136.

57. Montanucci,L., Martelli,P.L., Fariselli,P. and Casadio,R. (2007) Robust determinants of thermostability highlighted by a codon frequency index capable of discriminating thermophilic from mesophilic genomes. *J. Proteome Res.,* **6,** 2502-2508.

58. Paz,A., Mester,D., Baca,I., Nevo,E. and Korol,A. (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.,* **101,** 2951-2956.

59. Lithwick,G. and Margalit,H. (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res.,* **13,** 2665-2673.

60. Ikemura,T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. differences in synonymous codon choice patterns of yeast and escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.,* **158,** 573-597.

61. Rocha,E.P. (2004) Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

14, 2279-2286.

62. Sharp,P.M., Bailes,E., Grocock,R.J., Peden,J.F. and Sockett,R.E. (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.,* **33,** 1141-1153.

63. Gustafsson,C., Govindarajan,S. and Minshull,J. (2004) Codon bias and heterologous protein expression. *Trends Biotechnol.,* **22,** 346-353.

64. Ivory,C. and Chadee,K. (2004) DNA vaccines: Designing strategies against parasitic infections. *Genet. Vaccines Ther.,* **2,** 17.

65. Wu,G., Zheng,Y., Qureshi,I., Zin,H.T., Beck,T., Bulka,B. and Freeland,S.J. (2007) SGDB: A database of synthetic genes re-designed for optimizing protein over-expression. *Nucleic Acids Res.,* **35,** D76-9.

66. Richardson,S.M., Wheelan,S.J., Yarrington,R.M. and Boeke,J.D. (2006) GeneDesign: Rapid, automated design of multikilobase synthetic genes. *Genome Res.,* **16,** 550-556.

67. Wu,G., Bashir-Bello,N. and Freeland,S.J. (2006) The synthetic gene designer: A flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr. Purif.,* **47,** 441-445.

68. Villalobos,A., Ness,J.E., Gustafsson,C., Minshull,J. and Govindarajan,S. (2006) Gene designer: A synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics,* **7,** 285.

69. Hoover,D.M. and Lubkowski,J. (2002) DNAWorks: An automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.,* **30,** e43.

70. Jayaraj,S., Reid,R. and Santi,D.V. (2005) GeMS: An advanced software package for designing synthetic genes. *Nucleic Acids Res.,* **33,** 3011-3016.

71. Supek,F. and Vlahovicek,K. (2004) INCA: Synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics,* **20,** 2329-2330.

72. Grote,A., Hiller,K., Scheer,M., Munch,R., Nortemann,B., Hempel,D.C. and Jahn,D. (2005) JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.,* **33,** W526-31.

73. Fuglsang,A. (2003) Codon optimizer: A freeware tool for codon optimization. *Protein Expr. Purif.,* **31,** 247-249.

74. Gao,W., Rzewski,A., Sun,H., Robbins,P.D. and Gambotto,A. (2004) UpGene:

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Application of a web-based DNA codon optimization algorithm. *Biotechnol. Prog.,* **20,** 443-448.

75. Nakamura,Y., Gojobori,T. and Ikemura,T. (2000) Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Res.,* **28,** 292.

76. Carbone,A., Zinovyev,A. and Kepes,F. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics,* **19,** 2005-2015.

77. Willenbrock,H., Friis,C., Juncker,A.S. and Ussery,D.W. (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol.,* **7,** R114.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

OVERVIEW AND OBJECTIVES

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

This PhD thesis has been developed at the Evolutionary Genomics Group of the Biochemistry and Biotechnology Department of the 'Rovira i Virgili' University of Tarragona. The Evolutionary Genomics Group's research interest is in analyzing the molecular evolution of prokaryotes using the information extracted from completely sequenced genomes. One of the main research areas of this group has been predicting horizontally transferred genes in archaeal and bacterial genomes. The methodology developed by the group for predicting transferred genes is based on detecting compositionally anomalous genes, i.e. on genes with a G+C content and/or codon usage which is very different from the other genes in a given genome. In this context, the Evolutionary Genomics Group proposed at the beginning of this PhD thesis that we should develop a new method for predicting highly expressed genes in prokaryotic genomes. The purpose was to reduce false positives when predicting transferred genes and to filter the highly expressed genes. The new method for predicting highly expressed genes was a success and is an important part of this PhD thesis.

Our specific objectives were:

1. TO EVALUATE WHICH PROKARYOTIC SPECIES ARE UNDER TRANSLATIONAL SELECTION.

Evaluating translational selection is the first step in predicting a group of highly expressed genes, since only genomes affected by translational selection' have a group of genes with a codon usage adapted to the most abundant tRNA species. Traditionally, in order to detect whether a genome is under translational selection, researchers have analyzed the correspondence of the relative synonymous codon usage in all genes. In genomes under translational selection, the group of highly expressed genes forms a cluster in the correspondence analysis plot because they have a different codon usage

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

from the other genes in a genome. We wanted to analyze a large group of prokaryotic complete genomes, and so our first objective was to develop a new and automatic method, based on correspondence analysis, for evaluating which prokaryotic genomes are under a strong translational selection (chapter 1). As a result, we hoped to be able to predict a group of highly expressed genes within this particular set of genes.

2. TO PREDICT HIGHLY EXPRESSED GENES IN GENOMES UNDER TRANSLATIONAL SELECTION.

A group of highly expressed genes can be defined in a genome under translational selection by analyzing the codon usage bias of all the genes in the genome and by finding the differences between them. We needed to develop a new and automatic method in order to predict a group of highly expressed genes in all prokaryotic complete genomes. The method we developed is based on the Codon Adaptation Index (CAI). This uses the group of genes that codify for ribosomal protein genes as a seed and defines, through a series of iterations, a group of putative highly expressed genes (chapter 1). To further support our predictions, we analyzed the functions (chapter 2 and 3) and the essentiality (chapter 3) of the putative highly expressed genes.

3. TO ESTIMATE CODON USAGE ADAPTATION.

The CAI measures the similarity between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set. If this reference set is a group of highly expressed genes and if translational selection is present, the CAI values can be used to predict the expression level of a gene. However some of the adaptations detected with the CAI may merely be artefacts that arise from internal biases in the G+C composition and/or amino acid composition of the query sequences. We thought that an

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

expected CAI value (eCAI) may be useful for finding out whether the differences in the CAI are statistically significant or whether they are the product of biased nucleotide and/or amino acid composition. Therefore, we developed a new method for estimating an eCAI which generated random sequences with the same aminoacid and G+C composition as the query sequences (chapter 4). The eCAI therefore represents the upper limit of the CAI for those sequences whose a codon usage was solely due to mutational bias. To show how to use this eCAI, we analyzed the codon adaptation (or amelioration) of human mitochondrial genes encoded in the nuclear genome that were originally encoded in the proto-mitochondrial genome (chapter 4). Additionally, to assess the codon usage adaptation in various situations, we developed a new web server with a complete set of tools related to the CAI, such as the expected CAI value, the calculation and graphical representation of the CAI along a sequence and a protein multialignment translated to DNA (chapter 5).

4. TO DEVELOP A NEW WEB SERVER TO OPTIMIZE THE CODON USAGE OF A GENE IN ORDER TO INCREASE ITS GENE EXPRESSION.

Predicting highly expressed genes in genomes under translational selection allowed us to develop a new web server to optimize the codon usage of DNA or RNA sequences (chapter 1). Although several codon usage optimization programs exist, none of them specializes in optimizing the codon usage of a gene by using a group of highly expressed genes as a reference set in order to maximize its expression in a bacterial host. Therefore, we were interested in developing a new optimization program that uses the predictions of highly expressed genes. The server uses our prediction of the mean codon usage of

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

the group of highly expressed genes, but also introduces some new features such as using the tRNA gene copy numbers.

5. TO DISSEMINATE THE NEW ALGORITHMS DEVELOPED BY CREATING NEW DATABASES AND SERVERS.

To facilitate the use of the new algorithms and methods developed in this PhD thesis we created new programs, servers and databases freely available through Internet. The servers and databases that we originally planned to developed included: (i) a new genomic database that predicts which genes are highly expressed in prokaryotic complete genomes under strong translational selection (chapter 2); (ii) a new web-server with a complete set of tools related to calculating the CAI and the expected CAI value (chapter 4,5); (iii) a new web server for optimizing the codon usage of DNA or RNA sequences to increase their expression (chapter 1).

Although the main part of my thesis is based on analyzing codon usage and predicting highly expressed genes, I have also been working on other biological problems. Some of this work is related to the main topic of my thesis, while some of it has no relation at all. However, all of it has helped me to gain a better overall understanding of bacterial genomics. Among other things, I have been working on the evolution of codon usage and amino acid adaptation in thermophilic species. Initially our purpose was to analyze the differences in amino acid composition between highly expressed genes and the rest of the genes in a genome. While we were analyzing the compositional differences between species, we observed interesting differences between thermophile and non-thermophile species in terms of the putative evolution of thermophily in prokaryotes (chapter 6). During my thesis I have worked for four months in the Bioinformatics Laboratory of the Biology Department at the

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

National University of Ireland under the supervision of Dr James O. McInerney where I developed a new software program to compare phylogenetic trees (chapter 7).

Several parts of this PhD thesis have been published or submitted to international journals. These articles are:

- Puigbò P., Romeu A. and Garcia-Vallvé S. 2008. **HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection.** *Nucleic Acids Research*. doi:10.1093/nar/gkm831.

- Puigbò P., Guzmán E., Romeu A. and Garcia-Vallvé S. 2007. **OPTIMIZER: A web server for optimizing the codon usage of DNA sequences.** *Nucleic Acids Research* 35:W126-W131.

- Puigbò P., Garcia-Vallvé S. and McInerney J.O. 2007. **TOPD/FMTS: a new software to compare phylogenetic trees.** *Bioinformatics* 23(12):1556-1558.

- Puigbò P., Pasamontes A. and Garcia-Vallvé S. 2008. **Gaining and losing the capacity of thermophilic adaptation in prokaryotes.** *Trends in Genetics*. Accepted in press.

- Puigbò P., Bravo IG. and Garcia-Vallvé S. **E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI).** Submitted to *BMC Bioinformatics*.

- Puigbò P., Bravo IG. and Garcia-Vallvé S. **CAIcal: set of tools to assess codon usage adaptation.** In preparation.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

- Puigbò P., Guzmán E., Romeu A. and Garcia-Vallvé S. **Predicted highly expressed genes reveal common essential genes in prokaryotic genomes.** In preparation.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

CHAPTERS

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

1. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. **Pere Puigbò**, Eduard Guzmán, Antoni Romeu and Santiago Garcia-Vallvé. Nucleic Acids Research, 2007. 35:W126-W131.

ABSTRACT

OPTIMIZER is an on-line application that optimizes the codon usage of a gene to increase its expression level. Three methods of optimization are available: the 'one amino acid–one codon' method, a guided random method based on a Monte Carlo algorithm, and a new method designed to maximize the optimization with the fewest changes in the query sequence. One of the main features of *OPTIMIZER* is that it makes it possible to optimize a DNA sequence using pre-computed codon usage tables from a predicted group of highly expressed genes from more than 150 prokaryotic species under strong translational selection. These groups of highly expressed genes have been predicted using a new iterative algorithm. In addition, users can use, as a reference set, a pre-computed table containing the mean codon usage of ribosomal protein genes and, as a novelty, the tRNA gene-copy numbers. *OPTIMIZER* is accessible free of charge at http://genomes.urv.es/OPTIMIZER.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

INTRODUCTION

Gene expression levels depend on many factors, such as promoter sequences and regulatory elements. One of the most important factors is the adaptation of the codon usage of the transcript gene to the typical codon usage of the host (1). Therefore, highly expressed genes in prokaryotic genomes under translational selection have a pronounced codon usage bias. This is because they use a small subset of codons that are recognized by the most abundant tRNA species (2). The force that modulates this codon adaptation is called translational selection and its strength is important in fast-growing bacteria (3,4). If a gene contains codons that are rarely used by the host, its expression level will not be maximal. This may be one of the limitations of heterologous protein expression (5) and the development of DNA vaccines (6). A high number of synthetic genes have been re-designed to increase their expression level. The Synthetic Gene Database (SGDB) (7) contains information from more than 200 published experiments on synthetic genes. In the design process of a nucleic acid sequence that will be inserted into a new host to express a certain protein in large amounts, codon usage optimization is usually one of the first steps (5). Codon usage optimization basically involves altering the rare codons in the target gene so that they more closely reflect the codon usage of the host without modifying the amino acid sequence of the encoded protein (5). The information usually used for the optimization process is therefore the DNA or protein sequence to be optimized and a codon usage table (which we call the reference set) of the host.

Here we present a new web server, called OPTIMIZER, for codon usage optimization focused on the heterologous, or even homologous, gene expression in bacterial hosts. OPTIMIZER allows three optimization methods and uses several valuable, new reference sets. OPTIMIZER can therefore be used to optimize the expression level of a gene, to assess the adaptation of

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

alien genes inserted into a genome (8), or to design new genes from protein sequences. The server is freely available at http://genomes.urv.es/OPTIMIZER. It has been running since July 2005 and it is updated twice a year with new features and reference sets.

PROGRAM OVERVIEW

Implementation and input data

OPTIMIZER is an on-line application and its methods are implemented in PHP (hypertext pre-processor) programming language. The pre-calculated reference tables are stored into a MySQL database. The data input and the selection of the server options have been organized in four steps. These steps are: (1) Input the sequence to be optimized. DNA or protein sequences can be used, although further steps are slightly different depending on whether a DNA or protein sequence has been input. (2) Input the reference set. Users can insert a codon usage table in a variety of formats, including tables from the Codon Usage Database (9), or they can choose between 153 pre-computed codon usage tables for ribosomal protein genes or a group of highly expressed genes from prokaryotic genomes under translational selection. Users can also choose a reference set consisting of the tRNA gene-copy numbers. (3) Choose the genetic code. (4) Choose the method to be used in the optimization process. Depending on the type of sequence introduced (DNA or protein) and the reference set chosen, different optimization methods are available (see below for a description of the optimization methods).

Calculation of the reference sets

One of the main features of the OPTIMIZER server is that it contains a series of pre-computed reference sets that can be used in the optimization process. These reference sets can be a table containing the codon usage of the host

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

(or the codon usage of a group of genes, such as the group of highly expressed genes) or, as a novelty, the number of tRNA gene copies predicted with the tRNA-scan software (10). The pre-computed reference sets available in the server are from more than 150 prokaryotic genomes that are under a strong translational selection. The codon usage reference tables available for these genomes contain the mean codon usage of genes that encode ribosomal proteins or a group of highly expressed genes. Although the optimization process can be carried out using the mean codon usage of the host organism as a reference set, if the aim of the optimization process is to increase the expression level of a gene, it is preferable to use the codon usage of a group of highly expressed genes. The mean codon usage of bacteria is highly influenced by mutational bias (i.e. their G + C content). The optimal codons (those most frequently used in highly expressed genes) are usually those that agree with the mutational bias (i.e. G- or C-ending codons for G + C-rich organisms). However, the optimal codons are not always in agreement with mutational bias. For example, in the amino acids that are coded by only two synonymous codons ending in C or T, the C-ending codon is usually preferred, independently of the mutational bias (3). Therefore, using the mean codon usage of a genome may cause the wrong choice of optimal codons.

A new feature of the OPTIMIZER server is that it can use tRNA gene-copy numbers as a reference set for the optimization process. If the codon usage bias of highly expressed genes is caused by differences in tRNA gene-copy numbers, why not use this information for the optimization process? At present, information about tRNA gene-copy numbers is used in the OPTIMIZER server only with the ‘one amino acid–one codon’ optimization method (for a complete description of the methods available, see the ‘Optimization methods’ section below).

Evaluation of which bacterial genomes are under translational selection

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Not all prokaryotic species are under translational selection (4,11). It would be pointless to optimize the codon usage of a gene in order to increase its expression level in a species such as Helicobacter pylori, which is not under translational selection (i.e. in which the highly expressed genes do not have a different pattern of codon usage from the other genes of their genome) (12). Traditionally, correspondence analysis of the relative synonymous codon usage of all genes from a genome has been used to detect whether a genome is under translational selection (13). In genomes under translational selection, the ribosomal protein genes and other highly expressed genes form a cluster in the correspondence analysis plot, which confirms that highly expressed genes have a different codon usage from the other genes of a genome. This is the method we have used to detect which bacterial species are under translational selection. For each bacterial complete genome available, we made a correspondence analysis using the Relative Synonymous Codon Usage (RSCU) values of all the genes of a genome. To automate the analysis of the correspondence plots, we analyzed the position of the ribosomal protein genes (expected to be highly expressed genes) along the four principal axes obtained in the correspondence analysis. If a genome is under translational selection, ribosomal proteins and other highly expressed genes will show a codon usage bias and they will form a cluster in the correspondence plot. To make the prediction of translational selection, we checked whether the mean position of the ribosomal protein genes along any of the four principal axes was significantly different (evaluated with a t-test) from the mean position of the other genes of their genome. To check our predictions, we also visually inspected the correspondence plots (correspondence analysis plots are available from the homepage of the server) and analyzed the metabolic function of the predicted highly expressed genes obtained. Analysis of 334 prokaryotic genomes revealed that 153 genomes (the total number of different species and genera was 108 and 63, respectively) were under a strong translational selection. These genomes were then used to calculate the pre-

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

computed reference sets.

Prediction of highly expressed genes

The predicted highly expressed genes were obtained using an iterative algorithm that we have developed. This algorithm uses the group of genes that encode ribosomal proteins as a seed and, through a series of iterations, define a group of putative highly expressed genes. This algorithm works as follows:

Using the functional annotation, gene names or COG families, genes that encode ribosomal proteins are detected. Using the codon usage of these genes as a reference set, the Codon Adaptation Index (CAI), (14), at this stage namely CAIrp (15), is calculated for each gene of a genome.

Using now the group of genes with the highest CAI values as a reference set, the CAI for all genes is recalculated. This process is repeated until a homogeneous group is reached, i.e. when the group of genes with the highest CAI values in one iteration is the same as the group in the next iteration.

To provide further support for our predictions, we analyzed the metabolic functions of the putative highly expressed genes. As expected, ribosomal proteins and other expected highly expressed genes (16) were found in the final group of predicted highly expressed genes. To check our algorithm, we also analyzed species not under translational selection. With these species, either the algorithm never ended or the final group of genes had a high codon usage bias but was not related to their expression level. In this situation, neither ribosomal protein genes nor genes expected to have a high expression were included in the final group of genes with a codon usage bias. Our method is similar to the one developed by Carbone and co-workers (17). However, these authors used all the genes of an organism as the initial

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

reference set, whereas we used ribosomal protein genes.

Optimization methods

The OPTIMIZER server provides three methods for optimizing the codon usage of the query sequence. In the first method, the 'one amino acid–one codon' method, all the codons that encode the same amino acid are substituted by the most commonly used synonymous codon in the reference set. However, this approach has several drawbacks: for example, translational errors may be made due to an imbalanced tRNA pool and it is impossible to avoid repetitive elements or cleavage sites of restriction enzymes (5,18). To overcome these drawbacks, a second method, which we call the 'guided random' method, can be used. This method consists of a Monte Carlo algorithm that selects codons at random based on the frequencies of use of each codon in the reference set. The third method, which we call the 'customized one amino acid–one codon' method, is an intermediate method in which users choose how many of the 59 codons (if the standard genetic code has been selected) will be optimized with the 'one amino acid–one codon' approach. 'Rare codons' (i.e. the least used codons in the reference set) are the first codons changed with this approach. The aim of this third method is to maximize the optimization by making the fewest changes in the query sequence.

If the input sequence is a protein, it can be back-translated to DNA using the 'one amino acid–one codon' or the 'guided random' approach. If the 'one amino acid–one codon' approach is selected, the protein sequence can be back-translated to DNA using codons with the highest G + C or A + T content, or codons defined by Archetti (19) that minimize mutation errors.

Outputs

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Two indices, CAI and ENc (effective number of codons), are used to measure the optimization process. CAI measures the similarity between the codon usage of a gene and the codon usage of a reference group of genes (14). Its values range from 0 (when the codon usage of a sequence and that of the reference set are very different) to 1 (when both codon usages are the same). This index is the most effective of all codon bias measures for predicting gene expression levels (12,20). The second index is ENc, which is a measure of codon usage bias (21). Its values range from 20 (if only one codon per amino acid is used) to 61 (if all synonymous codons are used equally). Because highly expressed genes usually use the minimal subset of codons that are recognized by the most abundant tRNA species, their ENc values are expected to be low. Figure 1 shows some of the outputs provided by the optimization of a DNA sequence: for example, the query and optimized sequences and an alignment between them, a chart of the relative frequencies of each codon of the reference set and a codon usage table of the query and optimized sequences. In addition, the OPTIMIZER server has options for viewing or avoiding the cleavage sites of the selected restriction enzymes (22) and for splitting the optimized sequence into several overlapping oligonucleotides for the construction of a synthetic gene.

Comparison with other servers and programs

Table 1 shows a comparison of several public web servers and stand-alone applications that allow some kind of codon optimization. ‘GeneDesign’ (23), ‘Synthetic Gene Designer’ (24) and ‘Gene Designer’ (18) are packages that provide a platform for synthetic gene design, including a codon optimization step. Other programs, such as DNAWorks (25) and GeMS (26), focus more on the process of oligonucleotide design for synthetic gene construction. The stand-alone application INCA provides an array of features, including now codon optimization, which are useful for analyzing synonymous codon usage in whole genomes (27). JCAT (28), ‘Codon optimizer’ (29), UpGene (30) and

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

the server presented here focus on the codon optimization process. Although each server and application has its own features, all of them have several features in common. Most offer several options for the input of the codon usage reference set. One of these options is the possibility of using the tables from the Codon Usage database (9). Usually, a limited number of pre-computed tables of codon usage are available to be used as a reference set in the optimization process. In addition, not all of the available pre-computed reference sets correspond to a group of highly expressed genes (the proper reference set needed to optimize for increasing gene expression level). Though most of the programs and servers use a group of highly expressed genes from E. coli as a pre-computed reference set, only the 'Synthetic Gene Designer' and 'GeneDesign' servers provide a pre-computed group of highly expressed genes for 11 and 4 organisms, respectively. The exception is the JCAT web server, which offers pre-computed tables of predicted highly expressed genes from more than 200 bacterial species. However, this server uses the method of Carbone et al. (17) to predict a group of genes with a biased codon usage. These groups of genes do not always correspond to a group of highly expressed genes because not all bacterial species are under translational selection (11,17). The high number of pre-computed codon usage tables from bacteria and archaea that are not under translational selection available in JCAT therefore creates some confusion. The OPTIMIZER server presented here provides the most pre-computed codon usage tables for use as a reference set. The OPTIMIZER server provides pre-computed tables for more than 150 prokaryotic genomes that are under strong translational selection. In addition, two groups of genes are available in each reference set: a group of highly expressed genes predicted using a new prediction algorithm and the group of ribosomal protein genes. OPTIMIZER is the only server or stand-alone application that introduces a new kind of reference set such as information about the number of copies of tRNA genes for all the species included in the server. With regard to the methods for codon

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

usage optimization available in each server or program, the first programs developed used only the 'one amino acid–one codon' approach. More recent programs and servers now include further methods to create some codon usage variability. This variability reflects the codon usage variability of natural highly expressed genes and enables additional criteria to be introduced (such as the avoidance of restriction sites) in the optimization process. The OPTIMIZER server presented here provides three methods of codon optimization: a complete optimization of all codons, an optimization based on the relative codon usage frequencies of the reference set that uses a Monte Carlo approach (similar to methods from other programs and servers) and a novel approach designed to maximize the optimization with the minimum changes between the query and optimized sequences. Finally, note that only the 'Synthetic Gene Designer,' INCA and OPTIMIZER allow users to choose a non-standard genetic code.

CONCLUSIONS

OPTIMIZER is a new codon optimization web server focused on maximizing the gene expression level through the optimization of codon usage. It has unique features, such as a novel definition of a group of highly expressed genes from more than 150 prokaryotic species under translational selection, and the possibility of using information on tRNA gene-copy numbers in the optimization process. OPTIMIZER provides several pre-computed tables to specify a reference set and combines three different methods of codon optimization. The OPTIMIZER server can be used to optimize the expression level of a gene in heterologous gene expression or to design new genes that confer new metabolic capabilities in a given species.

ACKNOWLEDGEMENTS

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

This work has been financed by project BIO2003-07672 of the Spanish Ministry of Science and Technology. We thank Kevin Costello and John Bates of the Language Service of the Rovira i Virgili University for their help in writing the manuscript and two anonymous referees for their helpful comments. Funding to pay the Open Access publication charges for this article was provided by project BIO2003-07672 of the Spanish Ministry of Science and Technology.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

FIGURES

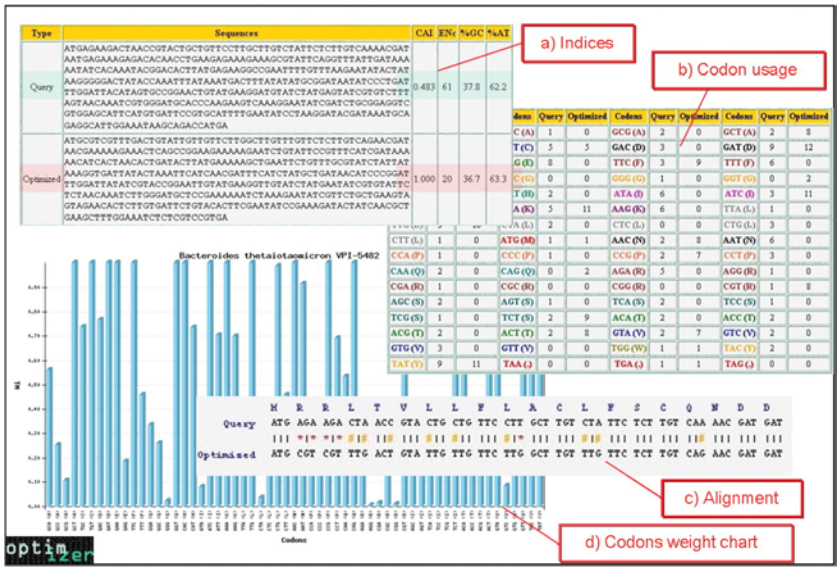


Figure 1. Outputs provided from the optimization of a DNA sequence: (a) the optimized and query sequences and the indices (CAI, ENc and %G + C) for evaluating the optimization process, (b) codon usage tables of the query and optimized sequences, (c) query and optimized sequence alignment to show changes in nucleotides (transitions or transversions) and (d) graphical view of the codon weight chart.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

TABLES

Table 1. Comparison of OPTIMIZER with other similar public web servers and softwares

Name	Methods	Genetic code	Reference set	Ref.
Web servers				
OPTIMIZER	- One AA - one codon - Guided Random (Monte Carlo algorithm) (2) - Customized one AA - one codon	Multiple	- HEG from >150 bacterial genomes under TS - RPG - tGCN - Codon usage database - Defined by users	This paper
JCAT	- One AA - one codon	Standard	- HEG from >200 bacterial genomes - Defined by users	28
Synthetic Gene Designer	- One AA - one codon (1) - Selective	Multiple	- HEG from 6 bacterial genomes - Codon usage	24

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

(SGD)	optimization[1] - Probabilistic optimization[1] [2]		database - Defined by users	
DNAWorks	- Use of the two highest frequency codons - Random	Standard	- HEG from *E. coli* - Codon usage tables for 10 species - Codon usage database - Defined by users	25
GeneDesign	- One AA - one codon - The next most optimal algorithm - The most different algorithm - Random	Standard	- HEG from 4 species - Defined by users	23
Stand-alone applications				
Gene Designer	- One AA - one codon - Monte Carlo algorithm [2]	Standard	- HEG from *E. coli* - Codon usage tables for 25	18

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			species - Codon usage database - Defined by users	
Codon optimizer	- One AA - one codon	Standard	- HEG for several bacterial species - Defined by users	29
INCA 2.1	- One AA - one codon	Multiple	- Mean codon usage of a whole genome or selection of any group of genes	27
UPGene	- One AA - one codon	Standard	- Eukaryotic, Bacteria, Yeast, Plant and Worm predefined codon usage frequency tables - Defined by users	30
GeMS	- Monte Carlo	Standard	- Codon usage	26

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	algorithm [2]		database - Defined by users	

Abbreviations used: HEG, codon usage of predicted highly expressed genes; RPG, codon usage of ribosomal protein genes; tGCN, tRNA Gene copy number; TS, translational selection.

[1] It uses an “optimality factor”, defined as a scaling factor, to control the optimality of codon usage. Higher values of this factor mean low CAI values and less optimized and more random codon usage.

[2] These methods are essentially the same. They use the relative codon usage frequencies of the reference set as the relative probability that each codon will be used in the optimization process.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

1. Lithwick, G. and Margalit, H. (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation *Genome Res*, . **13**, 2665–2673

2. Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system *J. Mol. Biol*, . **151**, 389–409

3. 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

4. Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., Sockett, R.E. (2005) Variation in the strength of selected codon usage bias among bacteria *Nucleic Acids Res*, . **33**, 1141–1153

5. Gustafsson, C., Govindarajan, S., Minshull, J. (2004) Codon bias and heterologous protein expression *Trends Biotechnol*, . **22**, 346–353

6. Ivory, C. and Chadee, K. (2004) DNA vaccines: designing strategies against parasitic infections *Genet. Vaccines Ther*, . **2**, 17

7. Wu, G., Zheng, Y., Qureshi, I., Zin, H.T., Beck, T., Bulka, B., Freeland, S.J. (2007) SGDB: a database of synthetic genes re-designed for optimizing protein over-expression *Nucleic Acids Res*, . **35**, D76–D79

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

8. Garcia-Vallve, S., Guzman, E., Montero, M.A., Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes *Nucleic Acids Res*, . **31**, 187–189

9. Nakamura, Y., Gojobori, T., Ikemura, T. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000 *Nucleic Acids Res*, . **28**, 292

10. 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

11. Willenbrock, H., Friis, C., Friis, A.S., Ussery, D.W. (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices *Genome Biol*, . **7**, R114

12. Henry, I. and Sharp, P.M. (2007) Predicting gene expression level from codon usage bias *Mol. Biol. Evol*, . **24**, 10–12

13. Perriere, G. and Thioulouse, J. (2002) Use and misuse of correspondence analysis in codon usage studies *Nucleic Acids Res*, . **30**, 4548–4555

14. Sharp, P.M. and Li, W.H. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications *Nucleic Acids Res*, . **15**, 1281–1295

15. Nakamura, Y. and Tabata, S. (1997) Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes *Microb. Comp. Genomics*, **2**, 299–312

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

16. Karlin, S. and Mrazek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes *J. Bacteriol*, . **182**, 5238–5250

17. Carbone, A., Zinovyev, A., Kepes, F. (2003) Codon adaptation index as a measure of dominating codon bias *Bioinformatics*, **19**, 2005–2015

18. Villalobos, A., Ness, J.E., Gustafsson, C., Minshull, J., Govindarajan, S. (2006) Gene designer: a synthetic biology tool for constructing artificial DNA segments *BMC Bioinformatics*, **7**, 285

19. Archetti, M. (2004) Selection on codon usage for error minimization at the protein level *J. Mol. Evol*, . **59**, 400–415

20. Goetz, R.M. and Fuglsang, A. (2005) Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli *Biochem. Biophys. Res. Commun*, . **327**, 4–7

21. Wright, F. (1990) The 'effective number of codons' used in a gene *Gene*, **87**, 23–29

22. Roberts, R.J., Vincze, T., Posfai, J., Macelis, D. (2005) REBASE – restriction enzymes and DNA methyltransferases *Nucleic Acids Res*, . **33**, D230–D232

23. Richardson, S.M., Wheelan, S.J., Yarrington, R.M., Boeke, J.D. (2006) GeneDesign: rapid, automated design of multikilobase synthetic genes *Genome Res*, . **16**, 550–556

24. Wu, G., Bashir-Bello, N., Freeland, S.J. (2006) The synthetic gene designer: a flexible web platform to explore sequence manipulation for

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

heterologous expression *Protein Expr. Purif*, . **47**, 441–445

25. Hoover, D.M. and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis *Nucleic Acids Res*, . **30**, e43

26. Jayaraj, S., Reid, R., Santi, D.V. (2005) GeMS: An advanced software package for designing synthetic genes *Nucleic Acids Res*, . **33**, 3011–3016

27. Supek, F. and Vlahovicek, K. (2004) INCA: Synonymous codon usage analysis and clustering by means of self-organizing map *Bioinformatics*, **20**, 2329–2330

28. Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D.C., Jahn, D. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host *Nucleic Acids Res*, . **33**, W526–W531

29. Fuglsang, A. (2003) Codon optimizer: a freeware tool for codon optimization *Protein Expr. Purif*, . **31**, 247–249

30. Gao, W., Rzewski, A., Sun, H., Robbins, P.D., Gambotto, A. (2004) UpGene: Application of a web-based DNA codon optimization algorithm *Biotechnol. Prog*, . **20**, 443–448

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

2. HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. **Pere Puigbò**, Antoni Romeu and Santiago Garcia-Vallvé. *Nucleic Acids Research*. doi:10.1093/nar/gkm831.

ABSTRACT

The highly expressed genes database (HEG-DB) is a genomic database that includes the prediction of which genes are highly expressed in prokaryotic complete genomes under strong translational selection. The current version of the database contains general features for almost 200 genomes under translational selection, including the correspondence analysis of the relative synonymous codon usage for all genes, and the analysis of their highly expressed genes. For each genome, the database contains functional and positional information about the predicted group of highly expressed genes. This information can also be accessed using a search engine. Among other statistical parameters, the database also provides the Codon Adaptation Index for all of the genes using the codon usage of the highly expressed genes as a reference set. The “Pathway Tools Omics Viewer” from the BioCyc database enables the metabolic capabilities of each genome to be explored, particularly those related to the group of highly expressed genes. The HEG-DB is freely available at http://genomes.urv.cat/HEG-DB.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

INTRODUCTION

Some genomes contain a group of genes, like ribosomal protein genes or other highly expressed genes, which have a pronounced codon usage bias because they use a small subset of synonymous codons: that is to say, codons that are recognized by the most abundant tRNA species (1).This bias is the result of "translational selection", i.e. codons that are translated by the most abundant tRNA species will increase efficiency and accuracy (2). Therefore, when a genome is under translational selection, genes with biased codon usage are usually considered to be a group of genes with high expression. The Codon Adaptation Index (CAI), developed by Sharp and Li (3) is the index that is most commonly used, by itself (4, 5) or in combination with an iterative algorithm (6, 7), to predict highly expressed genes that use the degree of bias in their codon usage. Karlin and co-workers use the “expression measure” of a gene, E(g), to evaluate the expression of genes through their codon usage bias (8, 9, 10, 11). However, this index has the problem that it is not always the gene with the strongest codon usage bias that has the highest predicted expression level (12). In any case, it must first be checked if a genome is under translational selection or not, independently of the method used to predict a group of highly expressed genes.

Here we present the highly expressed genes database (HEG-DB) which includes the evaluation of genomes under translational selection and the prediction of highly expressed genes in these genomes. The HEG-DB contains several statistical parameters of genes and genomes and data for the functional and metabolic analysis of the genomes under translational selection. With the HEG-DB, users can make genomic and functional analyses of highly expressed genes and assess their general functions and how they relate to the lifestyle and metabolism of the species. Defining a group of highly expressed genes is interesting not only for determining the

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

metabolic capabilities of the genomes under translational selection but also for other reasons. Groups of highly expressed genes can be used to reduce the false positives of the predictions of acquired genes because they are compositionally different from the other genes in a genome (13, 14).

SOURCE OF GENOMIC DATA AND METHODS

The methods for determining whether a genome is under translational selection and predicting highly expressed genes are described in an article by Puigbò and co-workers (7). Briefly, to evaluate whether a genome is under translational selection we made a correspondence analysis of the Relative Synonymous Codon Usage for all the genes in a genome. This analysis is traditionally used to detect whether a genome is under translational selection (15). Genomes are considered to be under translational selection when the group of ribosomal protein genes shows a codon usage bias and they form a cluster in the correspondence analysis plot (7). To predict the group of highly expressed genes in each genome we use an algorithm that uses the group of genes that codify for ribosomal protein genes as a seed and, through a series of iterations, define a group of putative highly expressed genes (7). However, in genomes with a high or low G+C content, it is difficult to predict highly expressed genes because of the effect of the extreme (high or low) G+C content on the codon usage of genes. The genome from *Pseudomonas aeruginosa* is an example of this. Carbone and coworkers (6) used an iterative algorithm to suggest that translational selection bias does not dominate in this species. However, other researchers have shown that in this species the variation in codon usage among genes is associated with expression, although this is not the major trend (16). To solve this situation and predict the group of highly expressed genes in those genomes we have made a slight modification to the previously described method (7). A gene is included in the list of biased genes for the following iteration only if its ENc (Effective Number of codons) is lower than its expected ENc estimated from the synonymous

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

G+C content at the third codon position (17). Because highly expressed genes usually use the minimal subset of codons that are recognized by the most abundant tRNA species, their ENc values are expected to be low (7). With this modification to our algorithm, genes whose CAI values are high because of extreme G+C bias and not because of high expression are removed from the list of biased genes. To provide further support for our predictions, we analyzed the metabolic functions of the putative highly expressed genes and, as expected, ribosomal proteins and other expected highly expressed genes were found in the final group of predicted highly expressed genes.

Gene expression is probably a continuous variable, and defining a group with the highest expression is relative and depends on the limits used (12). Experimental microarray experiments have shown that, even in species under translational selection, genes without a biased codon usage can be highly expressed (5, 18). The relationship between codon usage and gene expression is therefore only partial and can only be observed in species under translational selection. Because gene expression is closely related to promoter sequences and translational machinery, the highly expressed genes that are predicted through codon usage analyses are expected to be genes that are highly expressed in several situations (e.g. different media or growth phases). In these situations, translational selection is strong enough to modulate the codon usage of highly expressed genes.

IMPLEMENTATION AND ORGANIZATION OF THE DATABASE

The information about genes and genomes is stored in a MySQL database that can be accessed through a series of PHP web pages. The current version of the database contains information about almost 200 genomes under translational selection. The HTML interface is divided into four sections (see

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

figure 1): 1) The first section contains information about the genomes under translational selection, including links to some statistical parameters for these genomes, such as mean and standard deviations of total and positional G+C content, codon usage per thousand, relative synonymous codon usage and amino acid content. This section also includes the correspondence analysis plots of the relative synonymous codon usage for all the genes of the genomes used to predict translational selection. 2) The second section contains the list of the predicted highly expressed genes for all of the genomes under translational selection with their functional and positional information. 3) Since the definition of highly expressed genes is relative and depends on the limits, for each gene in the genomes under translational selection, we have included its CAI value. This information can also be accessed via a search engine that searches for gene names or keywords for a specific organism and taxa. 4) To see the metabolic capabilities of genomes under translational selection, the fourth section enables all the genes in a genome to be represented according to their CAI value on a metabolic map, using the Pathway Tools Omics Viewer from Biocyc (19, 20). The group of predicted highly expressed genes can be located separately on the metabolic pathway map of each genome. This last section makes a detailed functional analysis of the group of highly expressed genes and the preferred metabolic pathways in each genome under translational selection. For example, a schematic representation of the *Lactococcus lactis* metabolism from the BioCyc database (19, 20) is shown in figure 2. The figure shows that proteins encoded by highly expressed genes predicted with our methodology are involved in the main metabolic pathways of *L. lactis*.

DATABASE ACCESS

HEG-DB is freely accessible at http://genomes.urv.cat/HEG-DB.The database will be regularly updated with more genomes and new features.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

ACKNOWLEDGMENTS

This work has been financed by project BIO2003-07672 of the Spanish Ministry of Science and Technology. We thank John Bates of the Language Service of the Rovira i Virgili University for his help with writing the manuscript.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

FIGURES

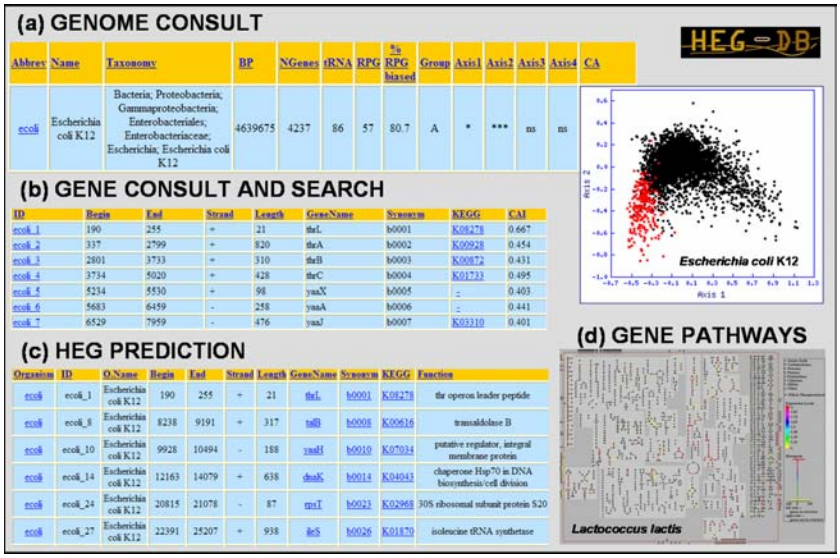


Figure 1. Outputs provided from the HEG-DB: (a) "Genomes consult" shows the list of all the genomes available in the database. In this section, users can select one or more genomes to see the statistical parameters (including the codon usage correspondence analysis plot used to predict translational selection) of the selected genomes. (b) The statistical and functional information available in each gene is accessible by a global consult of a specific genome or by a search engine. This section includes the CAI value of each gene. (c) List of predicted highly expressed genes in each genome. This section includes functional and positional information about each predicted gene. (d) The metabolic pathways which involve highly expressed genes can be viewed through the "pathway tools overview expression viewer" from the BioCyc database. In addition, this tool can be used to mark all genes according to their CAI on the pathway maps.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

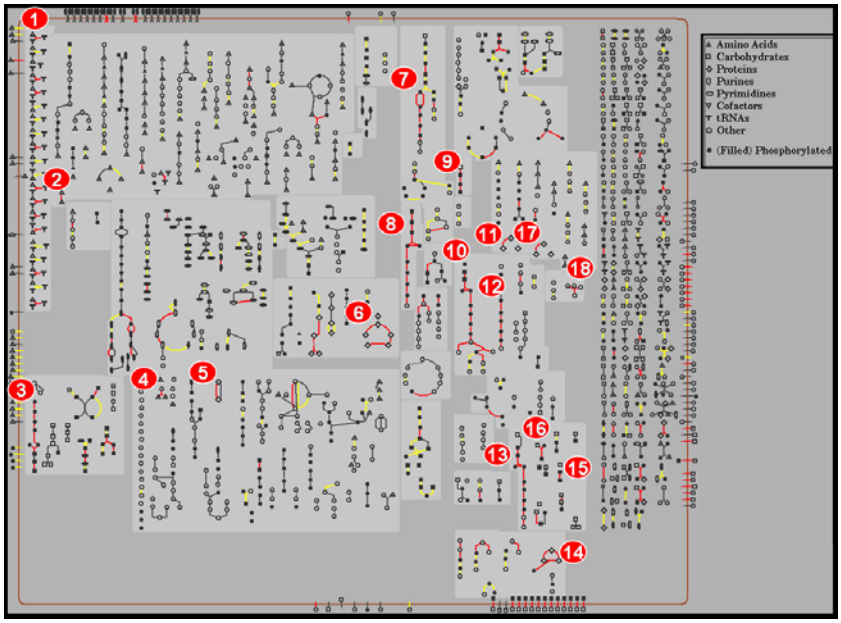


Figure 2. Schematic *Lactococcus lactis* metabolism. Reactions that are catalyzed with the product of a highly expressed gene are marked in red. Genes that have a slightly less biased codon usage than the predicted group of highly expressed genes are shown in yellow. Since gene expression is a continuous variable, this second group is also involved in several important pathways for this organism. This scheme was constructed using the "pathway tools overview expression viewer" from the BioCyc database (http://biocyc.org). The tool can be used directly through the HEG-DB. List of some pathways from *L. lactis* which involve highly expressed genes (shown in red): 1, tRNA charging pathway; 2, glutamine biosynthesis I; 3, gluconeogenesis; 4. γ-glutamyl cycle; 5, thioredoxin pathway; 6, fatty acid elongation – saturated; 7, glucose heterofermentation to lactate I; 8, glycolysis I; 9, N-acetyl-glucosamine degradation; 10, sorbitol fermentation to lactate, formate, ethanol and acetate; 11, branched-chain α-keto acid dehydrogenase

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

complex; 12, 2-dehydro-D-gluconate degradation; 13, fructose degradation to pyruvate and lactate (anaerobic); 14, pyruvate dehydrogenase complex; 15, mannose degradation; 16, sucrose degradation I; 17, 2-keto glutarate dehydrogenase complex; 18, removal of superoxide radicals. In addition, several highly expressed genes are involved in single reactions and in sucrose membrane transporters.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

1. Ikemura,T. (1981) Correlation between the abundance of escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.,* **146,** 1-21.

2. Sharp,P.M., Stenico,M., Peden,J.F. and Lloyd,A.T. (1993) Codon usage: Mutational bias, translational selection, or both? *Biochem. Soc. Trans.,* **21,** 835-841.

3. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.,* **15,** 1281-1295.

4. Wu,G., Nie,L. and Zhang,W. (2006) Predicted highly expressed genes in nocardia farcinica and the implication for its primary metabolism and nocardial virulence. *Antonie Van Leeuwenhoek,* **89,** 135-146.

5. Martin-Galiano,A.J., Wells,J.M. and de la Campa,A.G. (2004) Relationship between codon biased genes, microarray expression values and physiological characteristics of streptococcus pneumoniae. *Microbiology,* **150,** 2313-2325.

6. Carbone,A., Zinovyev,A. and Kepes,F. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics,* **19,** 2005-2015.

7. Puigbo,P., Guzman,E., Romeu,A. and Garcia-Vallve,S. (2007) OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. *Nucleic*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Acids Res., **35,** W126-W131.

8. Karlin,S. and Mrazek,J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.,* **182,** 5238-5250.

9. Karlin,S., Mrazek,J., Campbell,A. and Kaiser,D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J. Bacteriol.,* **183,** 5025-5040.

10. Karlin,S., Barnett,M.J., Campbell,A.M., Fisher,R.F. and Mrazek,J. (2003) Predicting gene expression levels from codon biases in alpha-proteobacterial genomes. *Proc. Natl. Acad. Sci. U. S. A.,* **100,** 7313-7318.

11. Karlin,S., Theriot,J. and Mrazek,J. (2004) Comparative analysis of gene expression among low G+C gram-positive genomes. *Proc. Natl. Acad. Sci. U. S. A.,* **101,** 6182-6187.

12. Henry,I. and Sharp,P.M. (2007) Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.,* **24,** 10-12.

13. Garcia-Vallve,S., Palau,J. and Romeu,A. (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in escherichia coli and bacillus subtilis. *Mol. Biol. Evol.,* **16,** 1125-1134.

14. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: A database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.,* **31,** 187-189.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

15. Perriere,G. and Thioulouse,J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.,* **30,** 4548-4555.

16. Grocock,R.J. and Sharp,P.M. (2002) Synonymous codon usage in pseudomonas aeruginosa PA01. *Gene,* **289,** 131-139.

17. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene,* **87,** 23-29.

18. dos Reis,M., Wernisch,L. and Savva,R. (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole escherichia coli K-12 genome. *Nucleic Acids Res.,* **31,** 6976-6985.

19. Paley,S.M. and Karp,P.D. (2006) The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Res.,* **34,** 3771-3778.

20. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.,* **33,** 6083-6089.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

3. Predicted highly expressed genes reveal common essential genes in prokaryotic genomes. **Pere Puigbò**, Eduard Guzmán, Antoni Romeu and Santiago Garcia-Vallvé. In preparation.

ABSTRACT

From the computational analysis of the codon usage bias in 173 genomes under translational selection we have defined a group of 184 highly expressed genes, common to all genomes analyzed. This group of genes may represent part of the essential group of genes for most of species. In addition, for several groups of taxonomically related species we have defined a group of taxon-specific highly expressed genes. We analyze our results by taking into account the metabolic categories, pathways or enzymes that are more represented in the group of highly expressed genes.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

INTRODUCTION

The methods to consider whether a genome is under translational selection and to predict the highly expressed genes are new methods that we have developed recently. These new methods are described in a recent article published in the Web-server 2007 special issue of Nucleic Acids Research 1. Our method developed to computationally predict a group of highly expressed genes follows the suggestions of Henry and Sharp 2, i.e. the use of the Codon Adaptation Index (CAI) and that it must be checked if the species analyzed are under translational selection, prior to the prediction of a group of highly expressed genes 2. We have introduced a slight modification in our algorithm to analyze genomes with a high or low G+C content, where the prediction of highly expressed genes is difficult because of the effect of the extreme (high or low) G+C content in the codon usage of genes. To provide further support for our predictions, we have analyzed the functions and leading or lagging chromosome position of the putative highly expressed genes (boxes 1 and 2). As these analyses reveals, the highly expressed genes that we predict are not a random group of genes but metabolic genes, with a putative function and located preferably at the leading strand. The analysis of the metabolic functions of the predicted highly expressed genes shows, as expected, that ribosomal proteins and other expected highly expressed genes (genes involved in translation, transcription, energy metabolism and the metabolism of biomolecules) were found in the final group of predicted highly expressed genes (figure 1).

Table 1 shows the 184 genes that are predicted to be highly expressed in most of the genomes under translational selection. They represent a core of metabolically important genes that are usually highly expressed in most

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

organisms. They include genes that codify translation and transcription factors, ribosomal proteins, aminoacyl-tRNA synthetases, replication complex proteins like ssb, GyrA and GyrB, chaperones like GroEL and GroES, and several genes involved in the metabolism of biomolecules (e.g. eight of the genes involved in glycolysis/gluconeogenesis). These 184 common highly expressed genes are almost universal in the bacterial world and could be considered essential genes for the survival of bacteria. However this group of genes neither defines the complete group of essential genes nor the minimal set of genes an organism needs for survival. Since this group of genes is essential in the maintenance of life in most of the species and they are detected as highly expressed genes in most of genomes under translational selection, they may be genes with a high expression in most of prokaryotic species, although in genomes under a weak or non-translational selection they do not show a different codon usage bias from the rest of genes of a genome. Table 1 only includes genes present in all the genomes analyzed. These genes do not form complete metabolic pathways, but they represent the universal steps in each pathway. In addition, gene expression is probably a continuous variable, and the definition of a group with the highest expression is relative and depends on the limits used. Thus, genes from table 1 must be used as a seed for the definition of universal and essential metabolic pathways, although some differences exist between species. For example, because of the differences in energy metabolism between prokaryotic species, only five of the eight subunits of the ATPase and the pyrophosphatase are the only genes of the “Energy Metabolism” section of table 1 predicted as highly expressed genes in the majority of the species analyzed. A group of genes that it would be interesting to study are the genes families with KEGG codes K09748, K09747 and K09710 (see table1). They are examples of genes with an unknown function, but predicted highly expressed genes in almost all the species analyzed (see also box 1).

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

A similar analysis for specific taxonomic groups of species shows which genes and which metabolic pathways are important in each group. Understanding the preferred metabolic pathways in each taxonomic group of species is crucial for a better knowledge of the life styles, habitats and main characteristics of microorganisms and may indicate possible drug targets to fight against pathogens. We have identified the common highly expressed genes for 12 taxonomic groups. Because these data is very large, we propose to present it as a supplementary table. This table shows that there is a good correlation between the definition of highly expressed genes in each particular group of species and their main metabolic capabilities.

Although highly expressed genes from a series of prokaryotes have been previously predicted from codon usage analyses (3, 4, and references therein, 5-8), and recent analyses have shown that codon bias signatures between microorganisms can be used to deduce environmental signatures related with the lifestyle of species 9, 10, we use our recent developed method of prediction of highly expressed genes to define a group of common essential genes in all the species analyzed and, as well, in several groups of taxonomically related species.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

FIGURES

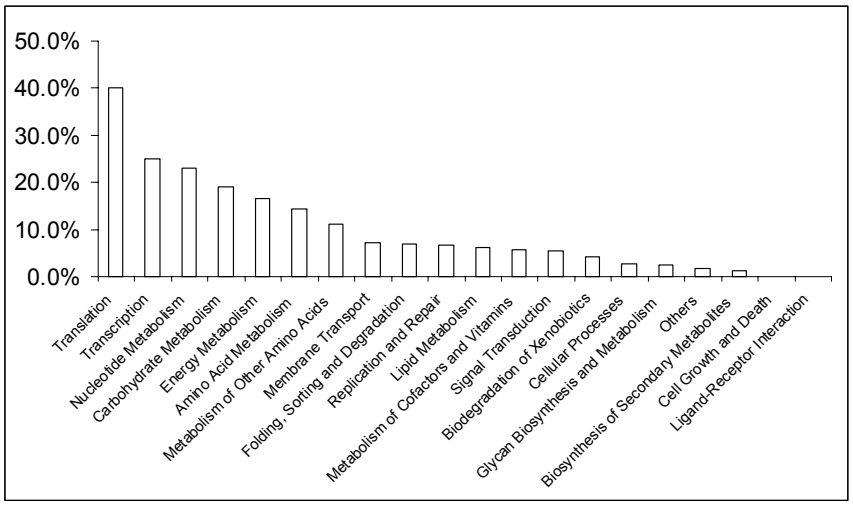


Figure 1. KEGG metabolic categories more represented in highly expressed genes from genomes under a strong translational selection. As expected, genes involved in translation, transcription, energy metabolism and the transport and metabolism of biomolecules are the metabolic categories with a greater number of highly expressed genes.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

TABLES

Table 1. Common highly expressed genes in the majority of species under translational selection.

KEGG Pathway		Genes	Comments
Amino Acid Metabolism	Alanine and aspartate metabolism	*purB*, *purA*	Catalyze the transformation of L-aspartate to fumarate in two steps. They form also part of the Purine metabolism pathway
	Glycine, serine and threonine metabolism	*asd*, *glyA*, *thrC*	Encode aspartate-semialdehyde dehydrogenase, glycine

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			hydroxymethyltransferase and threonine synthase
	Methionine metabolism	*metK*	Convert L-methionine into S-adenosyl-L-methionine
	Arginine and proline metabolism	*argF*, *argG*	ornithine carbamoyltransferase and argininosuccinate synthase
	Valine, leucine and isoleucine biosynthesis	*ilvC*, *ilvD*, *ilvE*	Encode the three last enzymes for the synthesis of valine and isoleucine
	Cysteine metabolism	*cysK*	Encodes cysteine synthase
	Glutamate metabolism	*glmS*, *glnA*, *guaA*, *purF*, *carB*	Related with the synthesis of glucosamine 6-phosphate and purine and pyrimidine metabolism

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Carbohydrate Metabolism	Aminosugars metabolism	*murA*	Related with the synthesis of peptidoglycan
	Fructose and mannose metabolism	*manB*	Catalyze the conversion of D-mannose 1-phosphate to D-mannose 6-phosphate
	Glycolysis / Gluconeogenesis	*pgi*, *gapA*, *pdhC*, *fbaA*, *pgk*, *tpiA*, *eno*, *pdhD*, *pyk*, *gpm*	Encode eight of the ten enzymes of glycolysis and two subunits of the pyruvate dehydrogenase multienzyme complex
	Citrate cycle (TCA cycle)	*gltA*, *icd*	Encode citrate synthase and isocitrate dehydrogenase
	Pentose phosphate pathway	*prsA*, *talB*, *rpe*, *tktB*	Form the non-oxidative portion of the pentose phosphate pathway

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Starch and sucrose metabolism	*galU*	UTP--glucose-1-phosphate uridylyltransferase
	Inositol phosphate metabolism	*suhB*	Encodes myo-inositol-1(or 4)-monophosphatase
	Pyruvate metabolism	*ackA*	Encodes acetate kinase
Cellular Processes	Cell division	*hflB*, *ftsZ*	Cell division proteins FtsH and FtsZ
Energy Metabolism	ATP synthesis	*atpD*, *atpF*, *atpC*, *atpA*, *atpE*	Subunits b, c, alpha, beta and epsilon of ATP synthase
	Oxidative phosphorylation	*ppa*	inorganic diphosphatase
Folding, Sorting and Degradation	Protein export	*lepB*, *secA*, *yajC*	signal peptidase I (catalyze the cleavage of N-terminal leader sequences) and two subunits of the Sec protein

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			secretion system
	Protein folding and associated processing	*clpP*, *cplX*, *ahpC*, *tig*, *trxA*, *groES*, *grpE*, *groEL*, *dnaJ*, *dnaK*	Protease and ATP-binding subunits of Clp protease, c-subunit of the alkyl hydroperoxide reductase, trigger factor, thioredoxin 1 and chaperonines GroES, GrpE, GroEL DnaJ and DnaK
Glycan Biosynthesis and Metabolism	Peptidoglycan biosynthesis	*murC*, *ddlA*, *glnA*	
Lipid Metabolism	Fatty acid biosynthesis	*acpP*, *fabF*, *fabG*	Encode acp (acyl carrier protein), 3-oxoacyl-[acyl-carrier-protein] synthase II and 3-oxoacyl-[acyl-carrier protein] reductase

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Membrane Transport	ABC transporters, prokaryotic	*pstS*	Encodes the substrate-binding subunit of the phosphate transport system
Nucleotide Metabolism	Purine metabolism	*purA*, *purB*, *purO*, *purL*, *hpt*, *guaB*, *purM*, *purC*, *apt*, *guaA*, *purF*, *purD*, *adk*, *ndk*, *pnp*, *nrdE*	Several genes of the purine metabolism pathway
	Pyrimidine metabolism	*pyrC*, *pyrE*, *carB*, *ndk*, *pnp*, *nrdE*, *trxB*, *thyA*, *upp*, *pyrG*	Several genes of the pyrimidine metabolism pathway
others	others	*K09748*, *sod2*, *K09747*, *K09710*, *E2.1.1.33*, *K06878*, *pyrH*, *bipA*, *K06942*, *nifS*	K09748, K09747, K09710: hypothetical proteins; sod2: Fe-Mn superoxide dismutase; E2.1.1.33: tRNA (guanine-N(7)-)-

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			methyltransferase¸K06942: Predicted GTPase, probable translation factor; nifS: cysteine desulfurase
Replication and Repair	Replication complex	*ssb*, *gyrA*, *gyrB*	Encode single-strand DNA-binding protein (ssb) and DNA gyrase subunits A and B
	Other replication, recombination and repair factors	*hupB*, *recA*	Encode DNA-binding protein HU-beta and recombination protein RecA
Transcription	Other transcription related proteins	*nusA*, *nusG*, *greA*	Encode transcription factors NusA, NusG and GreA
	Transcription factors	*cspA*	Encodes cold shock protein CspA
	RNA polymerase	*rpoA*, *rpoB*, *rpoC*, *rpoZ*, *rpoD*	Encode subunits alpha, ,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			beta, beta', omega and sigma of the DNA-directed RNA polymerase
Translation	Aminoacyl-tRNA biosynthesis	*thrS*, *serS*, *aspS*, *alaS*, *argS*, *proS*, *pheS*, *tyrS*, *trpS*, *leuS*, *ileS*, *valS*, *gltX*, *lysU*,	Several Aminoacyl-tRNA synthetases
	Other translation factors	*map*	Encodes a methionyl aminopeptidase

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Ribosome	*rpsJ, rpsU, rpsC, rplV, rpsR, rplE, rplL, rplS, rplB, rplX, rpsM, rplQ, rpsG, rpmB, rpsK, rpmI, rplR, rpsS, rplC, rplF, rpmE, rpsL, rplM, rpsN, rplK, rplW, rplI, rplN, rpsB, rplT, rpsO, rplD, rpsQ, rplO, rpmG, rpsI, rpmD, rpsD, rpmA, rpsF, rplA, rpsP, rpmF, rpsE, rpsT, rplJ, rpsH, rpsA, rplU, rplP, RBFA, rpmC, rpmH*	Encode 53 ribosomal proteins
Translation factors	*frr, prfA , infB, fusA, efp, tufA, tsf*	Encode ribosome recycling factor, peptide chain release factor RF-1, translation

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			initiation factor IF-2 and elongation factors EF-G, EF-P, EF-Tu and EF-Ts
	Other translation proteins	*trmU*, *tgt*	Encode a tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase and a queuine tRNA-ribosyltransferase

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

BOXES

Box 1. Genes with higher Codon Adaptation Index (CAI) are enriched in metabolic genes with a predicted function.

From an initial group of 173 genomes under translational selection, we checked those genomes that have at least a 50% of their genes annotated in a COG family[1]. 153 genomes followed this rule. We classified all genes from these 153 genomes in several CAI categories. Genes from categories A and B have CAIs higher than 1.5 and 0.5 standard deviations from their genomic CAI average, respectively. Category C corresponds to genes with CAI that do not deviate by more than 0.5 standard deviations from their mean species value. The CAI of genes from categories D and E are less than 0.5 and 1.5 standard deviations from the CAI average, respectively. Figure 1 shows the gene distribution in these five CAI categories. The group of predicted highly expressed genes is included in category A. Categories B-E include genes with progressively lower CAI values. The ratio between genes with a putative function (defined as genes that belong to some KEGG[2] or COG[1] family, excluding the R and S COG categories) and genes without a clear predicted function (defined as genes that belong to the R and S COG categories and genes that are not present in a COG or KEGG family) is 6.2 for genes from category A. This ratio decreases when CAI decreases and it is inverted when we look at category E. The genes predicted as highly expressed are therefore not a random group of genes but metabolic genes with a putative function. Among the group of genes without a clear predicted function we expect genes

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

that have not being well annotated, true genes with an unknown function and annotated ORFs that may be not true genes, also called ELFs[3]. The previous first and second groups of genes are expected to be in the categories with higher CAI values, and ELFs are expected to have the lowest CAI values (category E).

References

1 Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.

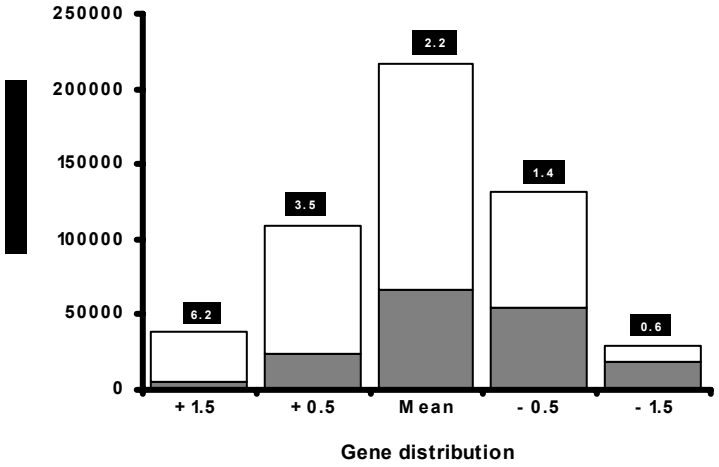
2 Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34:D354-357.

3 Ochman H. 2002. Distinguishing the ORFs from the ELFs: Short bacterial genes and the annotation of genomes. Trends in Genetics, 18 (7), pp. 335-337.

Figure 1. Distribution of genes from genomes under a strong translational selection classified in different CAI expression categories. The white bars represents the distribution of genes that belong to a COG (excluding the R and S categories) or KEGG family (i.e. genes with a putative known function) and the grey bars represents genes that are not present in a

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

COG or KEGG family (i.e. genes without a clear predicted function). The ratio between genes with a putative known function and genes with an unknown function is showed above each category.



UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Box 2. The common highly expressed genes in most prokaryotes are also part of the essential group of genes in prokaryotic species.

To test if the group of 184 common highly expressed genes in most prokaryotes under translational selection is also an essential group of genes, we have checked whether these genes are transcribed on the leading or lagging strand in 22 genomes under translational selection and in 21 genomes not under translational selection (Table 1). Due to "replicational selection" and "transcriptional selection" [1] it is expected that there is a higher number of genes, and especially essential genesl [2], on the leading strand. The collisions between RNA and DNA polymerases create interruptions in gene expression, and selection to minimize these interruptions can drive important genes to the leading strand [3]. Our results, see table 1, confirm this hypothesis. Since there are not differences between genomes that are under translational selection and genomes not under translational selection, this asymmetrical distribution may be universal and independent of translational selection. The 184 highly expressed genes conserved among prokaryotes are more frequent located on the leading strand than the average of all genes. This suggests that the predicted 184 highly expressed genes are essential in most of the genomes, even in genomes not under translational selection.

References

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

1 McInerney, J.O. (1998) Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. *Proc. Natl. Acad. Sci. U. S. A.* 95, 10698-10703

2 Rocha, E.P. and Danchin, A. (2003) Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* 31, 6570-6577

3 Price, M.N. *et al*. (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res.* 33, 3224-3234

Table 1. Relation of gene lying on the leading or lagging strand in 22 genomes under translational selection and in 21 genomes not under translational selection.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Translational selection	Genomes	Leading / Lagging (184 common heg)	Leading / Lagging (all genes)
Yes	*Bacillus halodurans*	**29.50**	**2.96**
	Staphylococcus aureus Mu50	**14.25**	**2.86**
	Staphylococcus aureus MW2	**14.25**	**3.16**
	Listeria monocytogenes	**14.17**	**3.57**
	Staphylococcus aureus N315	**13.08**	**2.96**
	Listeria innocua	**13.00**	**3.88**
	Lactococcus lactis	**9.65**	**4.03**
	Streptococcus pyogenes M1 GAS	**8.11**	**3.83**
	Streptococcus pneumoniae R6	**7.95**	**3.73**
	Streptococcus pyogenes MGAS8232	**7.70**	**3.99**
	Corynebacterium glutamicum ATCC 13032 Kitasato	**4.97**	**1.31**
	Escherichia coli O157H7 EDL933	**4.26**	**1.40**
	Escherichia coli K12	**3.97**	**1.25**
	Escherichia coli O157H7	**3.97**	**1.43**
	Salmonella typhimurium LT2	**3.84**	**1.43**
	Pseudomonas aeruginosa	**3.00**	**1.26**
	Haemophilus influenzae	**2.51**	**1.22**
	Sinorhizobium meliloti	**2.46**	**1.26**
	Bacteroides thetaiotaomicron VPI-5482	**2.16**	**1.39**
	Salmonella typhi	**2.12**	**1.15**
	Pasteurella multocida	**1.77**	**1.40**
	Neisseria meningitidis MC58	**1.51**	**1.18**

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Translational selection	Genomes	Leading / Lagging (184 common heg)	Leading / Lagging (all genes)
No	*Thermoanaerobacter tengcongensis*	**16.60**	**6.48**
	Clostridium perfringens	**12.38**	**4.87**
	Clostridium acetobutylicum	**11.43**	**3.71**
	Mycobacterium tuberculosis CDC1551	**7.42**	**1.43**
	Mycobacterium leprae	**7.24**	**1.91**
	Mycoplasma pneumoniae	**5.95**	**3.92**
	Mycobacterium tuberculosis H37Rv	**5.88**	**1.44**
	Mycoplasma genitalium	**5.74**	**4.55**
	Borrelia burgdorferi	**4.77**	**1.98**
	Xanthomonas citri	**4.48**	**1.25**
	Xanthomonas campestris	**4.32**	**1.23**
	Treponema pallidum	**4.00**	**1.94**
	Ureaplasma urealyticum	**4.00**	**2.36**
	Chlamydia muridarum	**2.51**	**1.22**
	Chlamydia trachomatis	**2.49**	**1.23**
	Campylobacter jejuni	**2.45**	**1.57**
	Mycoplasma pulmonis	**2.42**	**1.57**
	Caulobacter crescentus	**1.75**	**1.21**
	Chlorobium tepidum TLS	**1.68**	**1.30**
	Buchnera aphidicola	**1.63**	**1.28**
	Fusobacterium nucleatum	**1.47**	**1.42**

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

TABLES

Supplementary Table. Taxon-specific highly expressed genes for each taxonomic group of species under translational selection. Genes tha are common highly expressed genes in the majority of species are market in red.

Actinomycetales	Amino Acid Metabolism	Alanine and aspartate metabolism	*aspA, alr,* ***alaS****,* ***purB****,* ***purA****,* ***aspS***
		Arginine and proline metabolism	***proS****,* ***argS***
		Glycine, serine and threonine metabolism	*serA,* ***asd****, sdaA, lysC, thrA*
		Histidine metabolism	*hisG, hisC, hisB, hisD*
		Lysine biosynthesis	*lysA, dapA, dapB, dapD*
		Methionine metabolism	*metE, metY,* ***metK****, ahcY*
		Phenylalanine, tyrosine and tryptophan biosynthesis	*aroC, ARO1, aroE, aroH,* ***tyrS****, aroA, trpD, trpC, aroK*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	*proB, argC,* ***argF****, argB, argJ, proC, proA, argH,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			argG, *argD*
		Valine, leucine and isoleucine biosynthesis	*LEUD*, ***ilvD***, ***leuS***, ***ilvC***, ***valS***, *LEUC*, ***ileS***, ***ilvE***, *leuA*, *leuB*
	Carbohydrate Metabolism	Aminosugars metabolism	*murB*, ***glmS***, *pgm*, *glmU*
		Butanoate metabolism	*gabD*
		Citrate cycle (TCA cycle)	*PEPCK*, *sucA*, *SDHB*, *fumC*, *sucB*
		Fructose and mannose metabolism	***manB***
		Glycolysis / Gluconeogenesis	*GLPX*, ***tpiA***, *aceE*, *E1.2.1.3*, ***eno***, ***pdhD***, *ppgk*, *pfk*, ***fbaA***, ***pyk***, *ldh*, ***pgk***, ***gpm***, ***gapA***
		Glyoxylate and dicarBoxylate metabolism	*mdh*, *folD*, *acnA*, *glcB*, ***gltA***
		Nucleotide sugars metabolism	*galE*
		Pentose phosphate pathway	***talB***, ***rpe***, *gnd*, ***prsA***, *rpiB*, ***tktB***, *deoC*
		Propanoate metabolism	*pccB*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Pyruvate metabolism	*mqo, lldD*
		Starch and sucrose metabolism	*glgB, E3.2.1.21, ugd, otsA, PYG, glgC,* ***pgi***
	Cellular Processes	Cell division	***FTSZ***, ***hflB***
	Energy Metabolism	ATP synthesis	***atpA***, ***atpE***, ***atpD***
		Methane metabolism	*E1.11.1.6,* ***glyA***
		Nitrogen metabolism	*gdhA*
		Oxidative phosphorylation	*QCRB, COXA,* ***ppa****, NDH*
		Sulfur metabolism	*cysE*
	Folding, Sorting and Degradation	Protein export	*TATA,* ***SECA***
		Protein folding and associated processing	*CLPB,* ***clpP***, ***HSPE1***, ***GRPE***, ***HSPD1****, CLPC*
	Glycan Biosynthesis and Metabolism	Peptidoglycan biosynthesis	***ddlA***, ***glnA***, *mraY*
	Lipid Metabolism	Biosynthesis of steroids	*ispH, ispG, E2.5.1.30*
		Fatty acid biosynthesis (path 1)	***fabG****, accC*
		Glycerolipid metabolism	*gpsA, glpK*
	Membrane Transport	ABC transporters, prokaryotic	*metQ, ABC.FEV.S,* ***pstS****, metI,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			ABC.PE.S
		Other ion-coupled transporters	*TC.SSS*
		Pores ion channels	*TC.MSCL*, ***DNAK***
	Metabolism of Cofactors and Vitamins	Biotin metabolism	*bioB*
		Folate biosynthesis	*folK*, *folB*, *folP*, *folE*
		Nicotinate and nicotinamide metabolism	*iunH*, *nadC*
		One carBon pool by folate	*folA*
		Pantothenate and CoA biosynthesis	*COAA*
		Porphyrin and chlorophyll metabolism	*hemB*, *hemC*, *hemE*, ***gltX***, *hemL*
		Thiamine metabolism	*thiE*, *thiD*
		Ubiquinone biosynthesis	*menB*
		Vitamin B6 metabolism	***thrC***, *serC*
	Metabolism of Other Amino Acids	Glutathione metabolism	***icd***, *pepN*, *zwf*
		Selenoamino acid metabolism	***cysK***
	Nucleotide Metabolism	Purine metabolism	***purO***, ***purL***, ***purC***, ***guaB***, ***guaA***, ***purF***, ***adk***, ***purM***, ***purD***, ***hpt***
		Pyrimidine metabolism	***ndk***, ***pnp***, *pyrD*, ***nrdE***, ***upp***, ***trxB***, *nrdF*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			thyA, ***pyrE***, ***pyrC***, ***pyrG***, ***carB***
	others	others	***KARS***, ***E2.1.1.33***, ***K06878***, *E3.5.1.88, ABC-2.AB.A, tpx,* ***TIG***, ***bipA***, ***K06942***, *msrA, glf, K09772, SUFB, pdx1, PPIA, LIPA, E5.2.1.8, ligA,* ***pyrH***, ***tgt***, *sufC,* ***upp****S, pepA, YJFH, sseA,* ***K09710***, *K09761,* ***trmU***, *msrB*
	Replication and Repair	Replication complex	***GYRA***, *TOPA,* ***GYRB***
	Transcription	Basal transcription factors	***GREA***, ***NUSG***
		Other and unclassified family transcriptional regulators	***CSPA***
		RNA polymerase	***RPOC***, ***RPOZ***, ***RPOA***, ***RPOB***
	Translation	Aminoacyl-tRNA biosynthesis	***serS***, ***thrS***, *GRS1*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Other translation factors	*GATB*, *rph*, ***map***, *GATA*
Ribosome	***rpsK***, ***rpmI***, ***rpsP***, ***rplR***, ***rplI***, ***rplN***, ***rpsS***, ***rpsC***, ***rpsB***, ***rplT***, ***rplC***, ***rpsE***, ***rplV***, ***rpsR***, ***rplF***, ***rpsO***, ***rplD***, ***rpsQ***, ***rplE***, ***rplL***, ***rplS***, ***rplO***, ***rpmG***, *rplY*, ***rpsT***, ***rplJ***, ***rpsH***, ***rplB***, ***rpsA***, ***rplU***, ***rpmE***, ***rpsL***, ***rpsI***, ***rplX***, ***rpmD***, ***rplP***, ***rpsD***, ***rpsM***, ***rplQ***, ***rplM***, ***rpsN***, ***rplK***, ***rpmC***, ***rpsG***, ***rpmA***, ***rpmH***, ***rplW***, ***rpsF***, ***rplA***
Translation factors	***infB***, ***efp***, ***tufA***, ***fusA***, ***tsf***, ***frr***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Alteromonadales	Amino Acid Metabolism	Alanine and aspartate metabolism	***purB***, ***purA***, ***alaS***, *nadB*
		Arginine and proline metabolism	***argS***
		Glycine, serine and threonine metabolism	*kbl*, ***asd***
		Histidine metabolism	*hisG*, *hutU*
		Lysine biosynthesis	*dapD*
		Methionine metabolism	***metK***
		Phenylalanine metabolism	*E1.13.11.27*
		Phenylalanine, tyrosine and tryptophan biosynthesis	*aroK*, *aroH*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	*argC*, *proC*, ***argG***
		Valine, leucine and isoleucine biosynthesis	***ilvD***, ***ilvC***, ***ilvE***
	Carbohydrate Metabolism	Aminosugars metabolism	***glmS***
		Butanoate metabolism	*gabD*
		Citrate cycle (TCA cycle)	*sucA*, *sucB*, *idh*, *pckA*
		Glycolysis / Gluconeogenesis	***pgk***, ***gapA***, *E1.2.1.3*, ***eno***, ***pyk***, ***gpm***
		Glyoxylate and dicarBoxylate metabolism	***gltA***
		Pentose phosphate pathway	***prsA***, ***tktB***, *deoC*, ***talB***, ***rpe***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Propanoate metabolism	*prpC*
		Pyruvate metabolism	*gloA, maeB*
		Starch and sucrose metabolism	***pgi, galU***
	Energy Metabolism	Methane metabolism	***glyA***
		Oxidative phosphorylation	*etf,* ***ppa***
		Reductive carBoxylate cycle (CO2 fixation)	*ppsA*
	Folding, Sorting and Degradation	Protein export	***lepB****, lspA*
		Protein folding and associated processing	***clpP****, lon*
	Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	*lpxA, kdsB*
		Peptidoglycan biosynthesis	***glnA***
	Lipid Metabolism	Biosynthesis of steroids	*ispH, ispG*
		Fatty acid biosynthesis (path 1)	*fabA*
		Fatty acid biosynthesis (path 2)	*fadA*
		Glycerolipid metabolism	*psd, phoA*
		Synthesis and degradation of ketone bodies	*atoB*
	Metabolism of Cofactors and Vitamins	Biotin metabolism	*bioB*
		Pantothenate and CoA biosynthesis	*panB*
		Porphyrin and chlorophyll metabolism	*btuR, hemC, hemE*
		Vitamin B6 metabolism	*serC*
	Metabolism of	beta-Alanine metabolism	*panC*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Other Amino Acids	Glutathione metabolism	***icd***, *zwf, gshB*
	Nucleotide Metabolism	Purine metabolism	***purL***, ***purC***, ***apt***, ***purD***, *gmk*, ***guaB***, ***adk***, ***purM***
		Pyrimidine metabolism	***ndk***, ***pnp***, ***pyrC***, ***pyrE***, ***pyrG***
	others	others	*PREP*, ***K09747***, *msrA, K09780, ctpA*, ***tgt***, *pepA*, ***K09710***, ***E2.1.1.33***, *pepB, E3.5.1.88, dcp*, ***trmU***, *prlC*
	Signal Transduction	Phosphatidylinositol signaling system	***adk***
	Translation	Aminoacyl-tRNA biosynthesis	***serS***, *glnS*
		Other translation factors	*rph*
Bacillales	Amino Acid Metabolism	Alanine and aspartate metabolism	*asnS*, ***purB***, ***purA***, *pycB*, ***alaS***, ***aspS***
		Arginine and proline metabolism	***proS***, *pfs, putA*, ***argS***
		Glycine, serine and threonine metabolism	*gcvPA, gcvPB*, ***asd***, *tdcB*
		Methionine metabolism	***metK***, *metE*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Phenylalanine, tyrosine and tryptophan biosynthesis	***tyrS***, ***pheS***, *aroE*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	***argG***
		Valine, leucine and isoleucine biosynthesis	***leuS***, ***ileS***, ***ilvC***, *ilvB*, ***valS***, ***ilvE***
		Valine, leucine and isoleucine degradation	*bkdA2*
	Biodegradation of Xenobiotics	Benzoate degradation via CoA ligation	*E2.3.1.-*
	Carbohydrate Metabolism	Aminosugars metabolism	***glmS***, ***murA***
		Butanoate metabolism	*pflD*
		C5-Branched dibasic acid metabolism	*alsD*
		Citrate cycle (TCA cycle)	*sucD*, *SDHB*, *sucC*, *SDHA*, *sucB*, *pckA*
		Fructose and mannose metabolism	***manB***
		Glycolysis / Gluconeogenesis	*pfk*, ***fbaA***, *ldh*, *pdhB*, ***pgk***, ***gapA***, ***tpiA***, ***eno***, *adh*, ***pdhD***, ***pyk***, ***pdhC***, *acs*, ***gpm***, *pdhA*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Glyoxylate and dicarBoxylate metabolism	*fhs, folD, acnA,* ***gltA***
		Pentose phosphate pathway	***prsA***, ***tktB***, *deoC,* ***rpe***, *gnd, deoB*
		Pyruvate metabolism	*sfcA*
		Starch and sucrose metabolism	***pgi***, ***galU***, *glk*
	Cellular Processes	Cell division	***FTSZ***, ***hflB***, *DIVIVA*
	Energy Metabolism	ATP synthesis	***atpD***, ***atpA***, ***atpE***
		Methane metabolism	***glyA***, *CAT*
		Oxidative phosphorylation	***ppa***, *QOXD, QOXB*
		Reductive carBoxylate cycle (CO2 fixation)	*ald*
	Folding, Sorting and Degradation	Protein export	*ftsY,* ***lepB***, ***SECA***, ***YAJC***
		Protein folding and associated processing	***clpP***, ***HSPE1***, ***GRPE***, ***HSPD1***, *CLPC,* ***DNAJ***, ***CLPX***
	Glycan Biosynthesis and Metabolism	Peptidoglycan biosynthesis	*E3.5.1.28,* ***ddlA***, ***glnA***, ***murC***
	Lipid Metabolism	Fatty acid biosynthesis (path 1)	*fabH,* ***fabG***, ***ACPP***, *FABZ,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			fabD, fabI, accB
		Glycerolipid metabolism	*glpD, glpK*
	Membrane Transport	ABC transporters, ABC-2 and other types	*HIT*
		ABC transporters, prokaryotic	*ABC.FEV.S, zur,* ***pstS****, metQ, pstB, ABC.PA.S, ABC.PE.S*
		Phosphotransferase system (PTS)	*PTS-Glc-EIIA, PTS-EI, PTS-HPR*
		Pores ion channels	*GLPF,* ***DNAK***
	Metabolism of Cofactors and Vitamins	Nicotinate and nicotinamide metabolism	*nadE*
		One carBon pool by folate	*gcvT*
		Pantothenate and CoA biosynthesis	*ACPD*
		Porphyrin and chlorophyll metabolism	*hemB, hemH, hemL*
		Riboflavin metabolism	*RIBH*
		Ubiquinone biosynthesis	*menB*
	Metabolism of Other Amino Acids	D-Alanine metabolism	*dltC*
		Glutathione metabolism	***icd****, E1.11.1.9*
		Selenoamino acid metabolism	***cysK***
		Taurine and hypotaurine metabolism	*pta,* ***ackA***
	Nucleotide Metabolism	Purine metabolism	***purO****,* ***apt****,* ***guaA****,* ***purD****,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			guaB, ***purM***
		Pyrimidine metabolism	***ndk***, ***pnp***, *cmk*, ***nrdE***, ***trxB***, ***pyrC***, *DEOD*, ***upp***, *nrdF*, ***pyrE***, ***pyrG***
	others	others	*CSPR*, *K09162*, ***K09748***, ***SOD2***, ***K09747***, *E1.-.-.-*, *E3.4.11.-*, *SUFB*, *pdx1*, ***tgt***, *FER*, *GCVH*, *E4.4.1.21*, *E2.3.1.89*, ***K09710***, *pepQ*, ***ahpC***, ***KARS***, ***E2.1.1.33***, ***K06878***, *NIFU*, *OBG*, *tpx*, ***TIG***, ***fabF***, ***bipA***, ***K06942***, *aroA*, *LEPA*, *E5.2.1.8*, ***TRXA***, ***pyrH***, *queF*, *ABC.MN.S*, *sufC*, *pepB*, *PBUG*, *znuA*, ***trmU***, *E3.4.21.-*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Replication and Repair	Other replication, recombination and repair factors	***HUPB***, *DPS, nfo,* ***RECA***
	Transcription	Basal transcription factors	***NUSA***, ***NUSG***, ***GREA***
		HTH family transcriptional regulators	*LACI*
		Other and unclassified family transcriptional regulators	*CODY,* ***CSPA***
		RNA polymerase	*RPOE,* ***RPOA***, ***RPOC***, ***rpoD***, ***RPOZ***, ***RPOB***
	Translation	Aminoacyl-tRNA biosynthesis	***serS***, ***thrS***
		Other translation factors	***map***, *GATA, GATB, GATC*
		Ribosome	***rpsJ***, ***rplI***, ***rpsU***, ***rplN***, ***rpsC***, ***rpsB***, ***rplT***, ***rplV***, ***rpsR***, ***rpsO***, ***rplD***, ***rpsQ***, ***rplE***, ***rplL***, ***rplS***, ***rplO***, ***rplB***, ***rpsI***, ***rplX***, ***rpmD***, ***rpsD***, ***rpsM***, ***rplQ***, ***rpsG***, ***rpmA***, ***rpmB***, ***rpsF***, ***rplA***, ***rpsK***, ***rpmI***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			rpsP, ***rplR***, ***rpmF***, ***rpsS***, ***rplC***, ***rpsE***, ***rplF***, ***rpsT***, ***rplJ***, ***rpsH***, ***rpsA***, *rpmJ*, ***rplU***, ***rpmE***, ***rpsL***, ***rplP***, ***rplM***, ***RBFA***, ***rpsN***, ***rplK***, ***rpmH***, ***rplW***
		Translation factors	***fusA***, *infC*, ***frr***, ***infB***, ***efp***, ***tufA***, ***prfA***, *infA*, ***tsf***
Bacteroidales	Amino Acid Metabolism	Alanine and aspartate metabolism	*aspA*, *aspB*, *asnS*, *pepD*, ***alaS***, ***purB***, ***purA***, *nadB*, *panD*, ***aspS***, *gadB*
		Arginine and proline metabolism	***proS***, ***argS***
		Glycine, serine and threonine metabolism	*serA*, *kbl*, ***asd***, *lysC*, *gcvPB*
		Histidine metabolism	*hisC*, *hutH*, *hisS*
		Lysine biosynthesis	*E1.4.1.16*, *dapA*, *dapB*
		Methionine metabolism	*fmt*, ***metK***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Phenylalanine, tyrosine and tryptophan biosynthesis	***tyrS***, *qutE*, ***pheS***
		Tryptophan metabolism	***trpS***
		Valine, leucine and isoleucine biosynthesis	***leuS***, ***ilvC***, ***valS***, ***ileS***, ***ilvE***
	Biodegradation of Xenobiotics	1,4-Dichlorobenzene degradation	*nqrA*, *nqrC*, *nqrE*, *nqrB*, *nqrF*, *nqrD*
	Biosynthesis of Secondary Metabolites	Terpenoid biosynthesis	*ispA*
	Carbohydrate Metabolism	Aminosugars metabolism	*nagZ*, *NAGB*
		Butanoate metabolism	*nifJ*
		Citrate cycle (TCA cycle)	*korB*, *fumB*, *korA*, *SDHB*, *SDHC*, *pckA*
		Fructose and mannose metabolism	*gmd*, *pfk*, *fucI*, *manC*, *manA*, ***manB***
		Galactose metabolism	*galK*
		Glycolysis / Gluconeogenesis	***tpiA***, *fbaB*, *fbp*, ***eno***, ***fbaA***, *galM*, ***pgk***, ***gpm***, ***gapA***
		Glyoxylate and dicarBoxylate metabolism	*fhs*, *mdh*, *folD*, *E1.1.1.29*
		Nucleotide sugars metabolism	*galE*, *rfbD*, *rfbA*, *rfbB*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Pentose phosphate pathway	***talB***, ***rpe***, ***prsA***, *rpiB*, ***tktB***, *deoC*
		Propanoate metabolism	*E5.4.99.2*, *mcmA2*, *pccB*, *mmdC*
		Pyruvate metabolism	*gloA*, *maeB*, *ppdK*
		Starch and sucrose metabolism	*glgB*, *PYG*, *glk*, ***pgi***
	Cellular Processes	Cell division	*FTSQ*, *soj*, *MRP*, *MREB*, ***FTSZ***, ***hflB***
	Energy Metabolism	ATP synthesis	*ntpK*, *ntpB*, *ntpE*, *ntpI*, *ntpA*
		Methane metabolism	***glyA***
		Nitrogen metabolism	*gltD*, *gdhA*
		Oxidative phosphorylation	*FLDA*, *CYDA*, *etfA*
	Folding, Sorting and Degradation	Protein export	***YAJC***, *OXA1*, *ftsY*, ***lepB***, *SECG*, *SECY*, *SECF*, ***SECA***, *ffh*
		Protein folding and associated processing	*CLPB*, ***DNAJ***, ***clpP***, ***HSPE1***, *HSP90A*, ***GRPE***, ***HSPD1***, *lon*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			CLPC
	Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	*LPXC, lpxA, kdsA*
		N-Glycan degradation	*FUCA1,* ***manB***
		Peptidoglycan biosynthesis	*murD*
		Sphingoglycolipid metabolism	*E1.3.99.-*
	Lipid Metabolism	Biosynthesis of steroids	*dxs, ispF*
		Fatty acid biosynthesis (path 1)	***fabG****, fabH,* ***ACPP****, fabD, fabI*
		Fatty acid metabolism	*fadD*
		Glycerolipid metabolism	*phoA*
	Membrane Transport	ABC transporters, ABC-2 and other types	*ABC.CD.TX, ABC-2.A, HIT, ABC.CD.A*
		Other ion-coupled transporters	*TC.SULP, TC.POT, TC.HAE1*
		Other transporters	*spoIIIE*
		Pores ion channels	*ABC.FEV.OM, TC.MSCL,* ***DNAK***
	Metabolism of Cofactors and Vitamins	Biotin metabolism	*bioF*
		Folate biosynthesis	*folE*
		One carBon pool by folate	*gcvT*
		Pantothenate and CoA biosynthesis	*kdtB, panB*
		Porphyrin and chlorophyll	***gltX***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		metabolism	
		Riboflavin metabolism	*ribE, RIBH*
		Thiamine metabolism	*THIJ*
		Vitamin B6 metabolism	*pdxJ, serC*
	Metabolism of Other Amino Acids	Aminophosphonate metabolism	*E2.1.1.-*
		beta-Alanine metabolism	*panC*
		Selenoamino acid metabolism	*cysD*
		Taurine and hypotaurine metabolism	*pta,* ***ackA***
	Nucleotide Metabolism	Purine metabolism	***purO***, ***purL***, *gmk,* ***purC***, ***guaB***, ***guaA***, ***adk***, ***purM***, *purE, relA,* ***hpt***
		Pyrimidine metabolism	***pnp***, ***nrdE***, *pyrF,* ***trxB***, *pyrI, nrdD,* ***pyrE***, *PUNA,* ***pyrG***, ***carB***
	others	others	***KARS***, ***E2.1.1.33***, *MOXR, K07164,* ***K06878***, *RIBB, E3.5.1.88, K07166, E1.7.-.-, tpx, SPPA,* ***TIG***, *pepT, FKLB, BFR,* ***fabF***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			COML, ***bipA****,* ***K06942****, PQQL,* ***SOD2****, aroA, K07107, RNFC, E3.4.24.-, LEPA, K06969, SLYD, SUFB, K09117, BCP, K06861, E5.2.1.8,* ***TRXA****, RNFE, ctpA,* ***tgt****,* ***pyrH****, MLTD, ENGA, queF, SUFD, sufC, DPP3, RNFD, MRCA, RNFA, GCVH, K07011, PEPO, EXBB, RNFG, SURA, ERA, LEMA, DPP4, RLUB, E3.4.21.-,* ***ahpC****, gcp*
		Unclassified; Enzyme Complex; Pyruvate/Oxoglutarate oxidoreductases	*iorA, iorB*
	Replication and Repair	DNA polymerase	*dnaN, dnaQ*
		Other replication, recombination and repair factors	*xthA,* ***HUPB****, DPS, MUTS,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			RECA
		Replication complex	***GYRA***, *TOPA*, ***GYRB***, *DNAB*, *DNAA*
	Signal Transduction	Two-component system	*WECC*
		Wnt signaling pathway	*MAZG*
	Transcription	Basal transcription factors	***NUSA***, ***GREA***, ***NUSG***
		RNA polymerase	***RPOC***, ***rpoD***, ***RPOA***, *rpoE*, ***RPOB***
		Aminoacyl-tRNA biosynthesis	***serS***, *glnS*, *cysS*, ***thrS***, *GRS1*
		Other translation factors	***map***
		Ribosome	***rpsK***, ***rpmI***, ***rpsP***, ***rplR***, ***rpsJ***, ***rplI***, ***rplN***, ***rpsU***, ***rpsS***, ***rpsC***, ***rpsB***, ***rplT***, ***rplC***, ***rpsE***, ***rplV***, ***rpsR***, ***rplF***, ***rpsO***, ***rplD***, ***rpsQ***, ***rplE***, ***rplL***, ***rplS***, ***rplO***, ***rpmG***, *rplY*, ***rpsT***, ***rplJ***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			rpsH, ***rplB***, ***rpsA***, ***rplU***, ***rpmE***, ***rpsL***, ***rpsI***, ***rplX***, ***rpmD***, ***rplP***, ***rpsD***, ***rpsM***, ***rplQ***, ***rplM***, ***RBFA***, ***rpsN***, ***rplK***, ***rpmC***, ***rpsG***, ***rpmA***, ***rpmH***, ***rplW***, ***rpmB***, ***rpsF***, ***rplA***
		Translation factors	***infB***, ***efp***, ***tufA***, ***prfA***, *prfC*, ***fusA***, ***tsf***, *infC*, ***frr***
Burkholderiales	Amino Acid Metabolism	Alanine and aspartate metabolism	***purB***, *ansB*, ***purA***, ***alaS***, *E2.6.1.18*, *nadB*, ***aspS***
		Arginine and proline metabolism	***proS***, *putA*, ***argS***
		Glycine, serine and threonine metabolism	*thrA*, *serB*, ***asd***, *lysC*, *tdcB*
		Histidine metabolism	*hisG*, *hisA*, *hisC*, *hisB*, *hisI*, *hisI*, *hisD*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Lysine biosynthesis	*dapE, lysA, dapD, dapF, dapA, dapB*
		Methionine metabolism	*metX, fmt, metY,* ***metK****, ahcY, metE, metH*
		Phenylalanine metabolism	*E1.13.11.27*
		Phenylalanine, tyrosine and tryptophan biosynthesis	***trpe****,* ***tyrS****, trpD,* ***pheS****, aroC, aroE, aroH, trpD, aroA, trpC*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	*proB, argC,* ***argF****, argB, proC, proA, argH,* ***argG****, argD*
		Valine, leucine and isoleucine biosynthesis	***ilvD****,* ***leuS****,* ***ileS****, leuA, ilvH, LEUD,* ***ilvC****,* ***valS****,* ***ilvE****, leuB*
		Valine, leucine and isoleucine degradation	*ivd*
	Biodegradation of Xenobiotics	1,4-Dichlorobenzene degradation	*E3.1.1.45, catA*
		Benzoate degradation via CoA	*fadB, gcdH*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	ligation	
	Benzoate degradation via hydroxylation	*pcaI, hpaF*
	Tetrachloroethene degradation	*E1.1.1.-*
Biosynthesis of Secondary Metabolites	Alkaloid biosynthesis I	*tyrB*
Carbohydrate Metabolism	Aminosugars metabolism	***glmS***, *nagZ, E3.1.3.-,* ***murA***
	Ascorbate and aldarate metabolism	*E4.2.1.41*
	Butanoate metabolism	*gabD, phbB*
	Citrate cycle (TCA cycle)	*sucD, SDHB, sucA, sucC, SDHA, sucB*
	Glycolysis / Gluconeogenesis	***fbaA***, ***pgk***, ***gapA***, *aceE,* ***eno***, *adh,* ***pdhD***, ***pdhC***, *acs,* ***gpm***
	Glyoxylate and dicarBoxylate metabolism	*acnB, garR, mdh, folD, gcl, glcB, E1.2.1.2A, E1.2.1.2, acnA, gip, aceA, purU,* ***gltA***
	Inositol metabolism	*iolA*
	Pentose phosphate pathway	***prsA***, ***tktB***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		*edd, **talB**, **rpe***
	Propanoate metabolism	*prpe, prpC*
	Pyruvate metabolism	*gloA, dld*
	Starch and sucrose metabolism	*E2.7.1.-, **galU***
Energy Metabolism	ATP synthesis	***atpD**, **atpA***
	Methane metabolism	***glyA**, E1.1.1.284*
	Nitrogen metabolism	*nirB, gltD*
	Oxidative phosphorylation	*E1.9.3.1, etf, ppk, **ppa**, CYTB*
	Reductive carBoxylate cycle (CO2 fixation)	*ppc, ppsA*
	Sulfur metabolism	*cysH*
Folding, Sorting and Degradation	Protein export	*SECB, lspA*
	Protein folding and associated processing	***clpP**, DSBA, clpQ, HSLU, **HSPD1**, lon, CLPA*
Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	*lpxA, kdsA, lpxK*
	Peptidoglycan biosynthesis	*murE, mraY, murD, **ddlA**, **glnA**, **murC**, murF*
	Sphingoglycolipid metabolism	*E1.3.99.-*
Lipid Metabolism	Biosynthesis of steroids	*dxs, ispH, ispG*
	Fatty acid biosynthesis (path 1)	***fabG**, fabD*
	Fatty acid biosynthesis (path 2)	*fadA*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Fatty acid metabolism	*fadD*
	Glycerolipid metabolism	*plsC, glpK*
	Synthesis and degradation of ketone bodies	*hmgL, atoB, scoA*
Membrane Transport	ABC transporters, prokaryotic	*PHOU, ABC.PE.S*
Metabolism of Cofactors and Vitamins	Biotin metabolism	*bioF, bioA, bioD, bioB*
	Folate biosynthesis	*folE, folK, ptpS*
	Nicotinate and nicotinamide metabolism	*pncB, nadE, nadD, nadC, E2.7.1.23*
	One carBon pool by folate	*metF*
	Pantothenate and CoA biosynthesis	*ACPD, panB*
	Porphyrin and chlorophyll metabolism	*hemB, hemF, btuR,* ***gltX****, hemH, hemE, hemL, cobT*
	Riboflavin metabolism	*ribA, ribF, RIBH*
	Thiamine metabolism	*thiE, thiD, thiL*
	Vitamin B6 metabolism	*serC, pdxA,* ***thrC***
Metabolism of Other Amino Acids	Aminophosphonate metabolism	*E2.7.8.-, E2.6.1.-*
	beta-Alanine metabolism	*panC, paaG, acd*
	Glutathione metabolism	***icd****, pepN, ggt, E1.11.1.9, zwf,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			gst, gshA, gshB
		Selenoamino acid metabolism	*cysD, cysN*
		Taurine and hypotaurine metabolism	*pta,* ***ackA***
	Nucleotide Metabolism	Purine metabolism	***purO***, ***purL***, ***purC***, ***guaA***, ***purF***, *guaD,* ***purD***, *xdhB, gmk,* ***guaB***, ***adk***, ***purM***, *relA*
		Pyrimidine metabolism	***ndk***, ***pnp***, *pyrD,* ***nrdE***, *dcd,* ***trxB***, ***thyA***, ***pyrC***, *dut, carA,* ***upp***, *pyrF, tmk,* ***pyrE***, ***pyrG***, ***carB***
	others	others	*E1.14.12.17, pcnB, glnE, qor,* ***SOD2***, *E2.7.8.23, E1.-.-.-, miaA, glnD, K09767, ligA, ctpA,* ***tgt***, *pepA, tam, lexA,* ***K09710***, ***ahpC***, *PPIB,* ***KARS***, ***E2.1.1.33***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			K06878, *PMBA, E3.1.-.-, RIBB, E3.5.1.88, pepP, tpx, dacD, pcm, E3.8.1.2,* ***pyrH***, ***nifS***, *TLDD,* ***trmU***, *prlC*
	Replication and Repair	DNA polymerase	*dnaN*
		Other replication, recombination and repair factors	*xthA*
		Replication complex	*PARE, TOPB*
	Signal Transduction	Phosphatidylinositol signaling system	***adk***
	Transcription	Aminoacyl-tRNA biosynthesis	***serS***, *glnS, glyS,* ***thrS***
		Other translation factors	*rph,* ***map***, *GATA, GATB*
		RNA polymerase	***RPOC***, ***RPOB***
		Translation factors	***fusA***, ***tufA***
Enterobacteriales	Amino Acid Metabolism	Alanine and aspartate metabolism	*aspA, asnS, aspC,* ***purB***, *ansB,* ***purA***, *panD, pepD, asnB,* ***alaS***, *asnA,* ***aspS***
		Arginine and proline metabolism	***proS***, *speA, pfs, speD,* ***argS***
		Glycine, serine and threonine	*serA, kbl, tdh,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		metabolism	*sdaA, gcvPB, tdcB*
		Histidine metabolism	*hisG, HISF, HISH, hisS, hisA, hisI*
		Lysine biosynthesis	*dapD*
		Methionine metabolism	***metK**, metE, metH*
		Phenylalanine, tyrosine and tryptophan biosynthesis	***tyrS**, **pheS**, aroC, trpB, aroH*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	*speE, proA, argH, **argG***
		Valine, leucine and isoleucine biosynthesis	***ilvD**, **leuS**, **ileS**, leuA, LEUD, **ilvC**, **valS**, LEUC, **ilvE**, leuB*
	Biodegradation of Xenobiotics	Benzoate degradation via CoA ligation	*YHBS*
	Carbohydrate Metabolism	Aminosugars metabolism	*nanA, **glmS**, NAGB, pgm, **murA**, nagA*
		Butanoate metabolism	*pflD, adhE*
		Citrate cycle (TCA cycle)	*FRDD, FRDB, sucD, sucA, FRDA, sucC, fumB, SDHA,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			sucB, pckA
		Fructose and mannose metabolism	*mtlD, fruK*
		Glycolysis / Gluconeogenesis	*GLPX, bglA, fbp, pfk,* ***fbaA****,* ***pgk****,* ***gapA****,* ***tpiA****, aceE,* ***eno****,* ***pdhD****,* ***pyk****,* ***pdhC****,* ***gpm***
		Glyoxylate and dicarBoxylate metabolism	*acnB, mdh, folD, eda, aceA,* ***gltA***
		Nucleotide sugars metabolism	*galE*
		Pentose and glucuronate interconversions	*uxuB, araA, uxaC, uxuA*
		Pentose phosphate pathway	***prsA****, idnK,* ***tktB****, deoC,* ***talB****,* ***rpe****, gnd, deoB, rpiA*
		Pyruvate metabolism	*gloA, maeB, LDHA*
		Starch and sucrose metabolism	***pgi***
	Cellular Processes	Cell division	*MRP,* ***FTSZ****,* ***hflB****, MREB*
		Flagellar assembly	*FLIC, FLGE*
	Energy	ATP synthesis	***atpD****,* ***atpA****,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Metabolism		***atpF***, *atpG*, ***atpE***, ***atpC***
		Methane metabolism	***glyA***
		Nitrogen metabolism	*nirB*, *NAPA*, *nirD*, *gltD*, *gltB*
		Oxidative phosphorylation	*NUOL*, *NUOH*, *CYDA*, *nouD*, ***ppa***, *NUOG*, *NUOM*, *NUOJ*, *FLDA*, *NUOF*, *NUOB*, *NUOI*, *NUON*, *NUOE*, *NDH*, *CYOC*, *CYDB*
		Reductive carBoxylate cycle (CO2 fixation)	*ppsA*
		Sulfur metabolism	*cysE*
	Folding, Sorting and Degradation	Protein export	*OXA1*, *ftsY*, ***lepB***, *SECB*, *SECF*, ***SECA***, *ffh*, ***YAJC***, *SECD*
		Protein folding and associated processing	***clpP***, *DSBA*, ***HSPE1***, *HFLC*, *clpQ*, *HFLK*, *HSLU*, *HSP90A*, ***GRPE***, ***HSPD1***, *DSBC*, *lon*, ***DNAJ***, *IBPA*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			HTPX, HSLO, CLPA, ***CLPX***
	Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	*LPXC, RFAD, lpxA, RFAE, GMHA, kdsA*
		Peptidoglycan biosynthesis	*murE, mraY, murD,* ***glnA****,* ***murC***
	Lipid Metabolism	Biosynthesis of steroids	*ispH, ispG*
		Fatty acid biosynthesis (path 1)	*accD, fabH, fabA, fabB,* ***fabG****,* ***ACPP****, FABZ, fabD, E6.4.1.2, fabI, accA, accB*
		Glycerolipid metabolism	*glpK*
	Membrane Transport	ABC transporters, prokaryotic	*sbp, ABC.PA.A, livH,* ***pstS****, malE, metQ, proX, livK, pstB, ABC.PA.S, ABC.SS.S, PHOU, ABC.PE.S*
		Other ion-coupled transporters	*TC.CNT, TC.DAACS, SDAC, TC.AAT, TC.POT, ACRA*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Phosphotransferase system (PTS)	*PTS-Glc-EIIA, PTS-Nag-EIIC, PTS-Man-EIIC, PTS-Unk-EIIA, PTS-Mtl-EIIC, PTS-Man-EIID, PTS-Unk-EIIC, PTS-Glc-EIIC, PTS-Fru-EIIC, PTS-HPR*
		Pores ion channels	*TOLA, TC.OMF, lamB, TC.MIT,* ***DNAK****, TC.OOP*
	Metabolism of Cofactors and Vitamins	Folate biosynthesis	*MOAC*
		Nicotinate and nicotinamide metabolism	*pntB, udhA*
		One carBon pool by folate	*metF, folA, gcvT*
		Pantothenate and CoA biosynthesis	*ACPD, panB*
		Porphyrin and chlorophyll metabolism	*hemB,* ***gltX****, hemE, hemL*
		Riboflavin metabolism	*RIBH*
		Thiamine metabolism	*THIC, THII*
		Ubiquinone biosynthesis	*UBIE, menB*
		Vitamin B6 metabolism	*pdxJ, serC,* ***thrC***
	Metabolism of	beta-Alanine metabolism	*panC*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Other Amino Acids	Glutathione metabolism	***icd***, *pepN, gshB, gor*
		Selenoamino acid metabolism	*cysD, cysJ, cysI, cysN*
		Taurine and hypotaurine metabolism	***ackA***
	Nucleotide Metabolism	Purine metabolism	*spoT,* ***purO***, ***purL***, *purK, gpt,* ***purC***, ***apt***, ***guaA***, ***purF***, ***purD***, *purE,* ***hpt***, *guaC,* ***guaB***, ***adk***, ***purM***
		Pyrimidine metabolism	***ndk***, ***pnp***, ***nrdE***, ***trxB***, ***thyA***, *udp, pyrI, deoA, dut, carA, DEOD,* ***upp***, *cpdB, nrdF, nrdD,* ***pyrE***, ***pyrG***, ***carB***
	others	others	*MRDA, htrA, CSTA, pcnB, HFQ, K06866, K09158,* ***K09748***, *ENGC,* ***SOD2***, *E3.1.1.31YBHE,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			K09747, *E1.-.-.-*, *TOLB*, *LPP*, *cafA*, *SLYD*, *SUFB*, *TC.MSCS*, *MIAB*, *FKPA*, *BCP*, *K06861*, *K07115*, *PPIA*, *DKSA*, *PHNA*, *K09767*, *ctpA*, ***tgt***, *OSMY*, *rsmC*, *NLPB*, *pepA*, *GCVH*, *phoL*, *GNTT*, *lexA*, *PAL*, *GRXA*, *YFIF*, *FNR*, *K07223*, *K06959*, *K09136*, *K06941*, ***K09710***, *DCUA*, *SRMB*, *YAJG*, *pepQ*, *ompR*, ***ahpC***, *PPIB*, *YCBY*, *ABC.X1*, ***KARS***, *K06873*, *CYAY*, *SSPA*, *pepB*, ***K06878***, *NIFU*, *PMBA*, *PRMA*, *YHGI*, *K07034*, *K09774*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			arcA, FDX, RIBB, E1.7.-.-, pepP, tpx, dacD, ***TIG****, FKLB,* ***fabF****, YFCB, COML,* ***bipA****,* ***K06942****, CORC, PPID, glmM, K09802, LEPA, ptsI, GLNB, PSPA, cyaA, LIPA, GLPE, MRCB,* ***TRXA****, YJGF,* ***pyrH****, USPA, ENGA,* ***nifS****, K09807, TLDD, MRCA, PFLE, HUPA, YJFH, PPIC, FADL, SLYB, RNE, ompH, GLRX5, YAET, HNS, SURA, K08303, ERA, K07274, QUEA, K06911,* ***trmU****, RLUB, imp, SLPA, prlC, DEAD, OMPC, HAM1*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Replication and Repair	DNA polymerase	*dnaN*
		Other replication, recombination and repair factors	*pcrA, RDGC,* ***HUPB****, DPS, UVRA, HEPA, RECQ, RHLB, xthA,* ***RECA***
		Replication complex	***GYRA****, PARE, TOPA,* ***GYRB****, DNAB,* ***SSB****, PARC*
	Signal Transduction	Phosphatidylinositol signaling system	***adk***
		Two-component system	*WECC*
	Transcription	Basal transcription factors	***NUSA****, NUSB,* ***NUSG****, RHO,* ***GREA***
		HTH family transcriptional regulators	*FUR, PURR, FIS, GLPR*
		Other and unclassified family transcriptional regulators	*rnb, METJ,* ***CSPA***
		RNA polymerase	***RPOA****,* ***RPOC****,* ***rpoD****, rpoH,* ***RPOZ****,* ***RPOB***
	Translation	Aminoacyl-tRNA biosynthesis	***serS****, glnS, cysS, glyS, glyQ,* ***thrS***
		Other translation factors	***map***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Ribosome	***rpsJ***, ***rplI***, ***rpsU***, ***rplN***, ***rpsC***, ***rpsB***, ***rplT***, ***rplV***, ***rpsR***, ***rpsO***, ***rplD***, ***rpsQ***, ***rplE***, ***rplL***, ***rplS***, ***rplO***, ***rpmG***, ***rplB***, ***rpsI***, ***rplX***, ***rpmD***, ***rpsD***, ***rpsM***, ***rplQ***, ***rpsG***, ***rpmA***, ***rpmB***, ***rpsF***, ***rplA***, ***rpsK***, ***rpmI***, ***rpsP***, ***rplR***, ***rpmF***, ***rpsS***, ***rplC***, ***rpsE***, ***rplF***, *rplY*, ***rpsT***, ***rplJ***, ***rpsH***, ***rpsA***, ***rplU***, ***rpmE***, ***rpsL***, ***rplP***, ***rplM***, ***RBFA***, ***rpsN***, ***rplK***, ***rpmC***, ***rpmH***, ***rplW***
		Translation factors	*prfC*, ***fusA***, ***frr***, ***infB***, ***efp***, ***tufA***, ***prfA***, *infA*, ***tsf***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			prfB
Lactobacillales	Amino Acid Metabolism	Alanine and aspartate metabolism	*asnS*, ***alaS***
	Carbohydrate Metabolism	Glycolysis / Gluconeogenesis	***tpiA***, ***eno***, *pfk*, ***pyk***, *ldh*, ***pgk***, ***gpm***, ***gapA***
	Carbohydrate Metabolism	Starch and sucrose metabolism	***pgi***
	Energy Metabolism	Oxidative phosphorylation	***ppa***
	Folding, Sorting and Degradation	Protein folding and associated processing	***HSPD1***
	Metabolism of Other Amino Acids	D-Alanine metabolism	*dltC*
	Metabolism of Other Amino Acids	Taurine and hypotaurine metabolism	*pta*
	Nucleotide Metabolism	Purine metabolism	***guaB***, ***guaA***
	others	others	***K09747***
	Translation	Aminoacyl-tRNA biosynthesis	***serS***, ***thrS***
	Translation	Other translation factors	*GATB, GATA*
Methanococcales	Amino Acid Metabolism	Methionine metabolism	***metK***
	Carbohydrate	Glycolysis / Gluconeogenesis	***eno***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Metabolism		
	Carbohydrate Metabolism	Pentose phosphate pathway	***talB***
	Carbohydrate Metabolism	Starch and sucrose metabolism	*E3.2.1.4*
	Energy Metabolism	Oxidative phosphorylation	***ppa***
	Metabolism of Cofactors and Vitamins	Folate biosynthesis	*mtd, E1.12.98.1, hmd*
	Metabolism of Cofactors and Vitamins	Porphyrin and chlorophyll metabolism	***gltX***
	Nucleotide Metabolism	Purine metabolism	***adk***, ***purO***, ***purD***
	others	others	*gap, E1.1.1.272, arsA*
Neisseriales	Amino Acid Metabolism	Alanine and aspartate metabolism	*aspA,* ***alaS***, ***purB***, ***purA***, *nadB, panD,* ***aspS***
		Arginine and proline metabolism	***proS***, *putA,* ***argS***
		Glycine, serine and threonine metabolism	***asd***, *lysC, thrA*
		Histidine metabolism	*hisC, hisB, hisD, hisA*
		Lysine biosynthesis	*lysA, dapA,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			dapB, dapD
		Methionine metabolism	*metE, fmt,* ***metK***
		Phenylalanine, tyrosine and tryptophan biosynthesis	*aroC,* ***trpe****, ARO1,* ***tyrS****, aroA, trpF, trpC, aroK*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	*argC,* ***argF****, argJ, argH,* ***argG***
		Valine, leucine and isoleucine biosynthesis	*LEUD,* ***ilvD****,* ***leuS****,* ***ilvC****,* ***valS****,* ***ileS****, leuA, leuB*
	Biosynthesis of Secondary Metabolites	Alkaloid biosynthesis I	*tyrB*
	Carbohydrate Metabolism	Aminosugars metabolism	*nagZ,* ***glmS***
		Citrate cycle (TCA cycle)	*sucA, sucC, sucD, sucB*
		Glycolysis / Gluconeogenesis	***tpiA****, aceE,* ***eno****, adh,* ***pdhD****, fbp,* ***fbaA****,* ***pyk****,* ***pdhC****,* ***pgk****,* ***gpm****,* ***gapA***
		Glyoxylate and dicarBoxylate	*acnB, eda, fhs*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	metabolism	
	Nucleotide sugars metabolism	*rfbA, rfbB*
	Pentose phosphate pathway	***talB**, idnK, **tktB**, edd*
	Pyruvate metabolism	*gloB, sfcA*
	Starch and sucrose metabolism	***pgi***
Cellular Processes	Cell division	***hflB***
Energy Metabolism	Methane metabolism	***glyA**, CAT*
	Nitrogen metabolism	*E1.7.2.1, gdhA*
	Oxidative phosphorylation	*ppk, **ppa***
	Reductive carBoxylate cycle (CO2 fixation)	*ppc, ppsA*
	Sulfur metabolism	*cysE*
Folding, Sorting and Degradation	Protein export	*OXA1*
	Protein folding and associated processing	***HSPE1**, **HSPD1**, DSBC, lon*
Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	*LPXD, RFAD, lpxA, kdsA*
	Peptidoglycan biosynthesis	*murE, **ddlA**, **glnA**, **murC**, mraY*
Lipid Metabolism	Fatty acid biosynthesis (path 1)	***fabG**, **ACPP**, fabD, E6.4.1.2, fabI*
	Fatty acid metabolism	*fadD*
Membrane	ABC transporters, prokaryotic	*metQ,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	Transport		*ABC.FEV.A*
		Pores ion channels	***DNAK***
	Metabolism of Cofactors and Vitamins	Folate biosynthesis	*folC*
		Nicotinate and nicotinamide metabolism	*pncB, nadC*
		One carBon pool by folate	*metF, gcvT*
		Porphyrin and chlorophyll metabolism	*hemC, hemE,* ***gltX****, hemL*
		Riboflavin metabolism	*ribD*
		Thiamine metabolism	*thiE*
		Vitamin B6 metabolism	***thrC***
	Metabolism of Other Amino Acids	Aminophosphonate metabolism	*E2.1.1.-*
		Glutathione metabolism	***icd****, pepN, gshA, gshB*
		Selenoamino acid metabolism	***cysK***
		Taurine and hypotaurine metabolism	***ackA***
	Nucleotide Metabolism	Purine metabolism	***purO****,* ***purL****,* ***guaB****,* ***apt****,* ***guaA****,* ***purF****,* ***adk***
		Pyrimidine metabolism	***ndk****,* ***pnp****, carA, pyrD, cmk, dcd, pyrF,* ***trxB****,* ***pyrE****,* ***pyrC****,* ***pyrG***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	others	others	***KARS***, ***K06878***, *NIFU, fpr, glnE, RIBB,* ***K09748***, ***TIG***, ***fabF***, *K09794,* ***bipA***, ***K06942***, ***K09747***, *E1.-.-.-, miaA, E3.4.24.-, TDCF, K09801, MIAB, E5.2.1.8, glnD,* ***TRXA***, ***tgt***, ***pyrH***, ***nifS***, *RNE, K08303,* ***K09710***, *K09761,* ***trmU***, *RLUB, prlC, E3.4.21.-, PPIB*
	Replication and Repair	Other replication, recombination and repair factors	*xthA, pcrA,* ***HUPB***
		Replication complex	*PARE,* ***GYRB***
	Signal Transduction	Phosphatidylinositol signaling system	*dgkA*
	Transcription	Aminoacyl-tRNA biosynthesis	***serS***, *glnS, glyS*
		Basal transcription factors	***NUSA***
		Other translation factors	***map***
		Ribosome	***rpsP***, ***rpmF***, ***rpsU***, ***rpsB***,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			rpsO, ***rplS***, ***rpmG***, ***rpsT***, ***rpsA***, ***rpmE***, ***rplM***, ***rpmA***
		Translation factors	***efp***, ***tufA***, ***tsf***, ***frr***
Pasteurellales	Amino Acid Metabolism	Alanine and aspartate metabolism	*aspA*, *asnS*, ***alaS***, ***purB***, ***purA***, *asnA*, ***aspS***
		Arginine and proline metabolism	***proS***, ***argS***
		Glycine, serine and threonine metabolism	***asd***
		Lysine biosynthesis	*dapD*
		Methionine metabolism	***metK***
		Phenylalanine, tyrosine and tryptophan biosynthesis	***pheS***, *aroK*
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	***argG***
		Valine, leucine and isoleucine biosynthesis	***leuS***, ***ilvC***, ***valS***, ***ileS***
	Biodegradation of Xenobiotics	1,4-Dichlorobenzene degradation	*nqrA*, *nqrC*, *nqrB*, *nqrF*
	Carbohydrate Metabolism	Aminosugars metabolism	*nanA*, ***glmS***, ***murA***, *glmU*
		Butanoate metabolism	*pflD*
		Citrate cycle (TCA cycle)	*FRDA*, *sucB*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			pckA
		Fructose and mannose metabolism	***manB***
		Glycolysis / Gluconeogenesis	***tpiA***, *aceE*, ***eno***, ***pdhD***, *pfk*, ***fbaA***, ***pyk***, ***pdhC***, ***pgk***, ***gpm***, ***gapA***
		Glyoxylate and dicarBoxylate metabolism	*mdh*
		Nucleotide sugars metabolism	*galE*
		Pentose phosphate pathway	***talB***, *gnd*, ***prsA***, ***tktB***, *deoC*
		Pyruvate metabolism	*gloA*, *maeB*
		Starch and sucrose metabolism	***galU***, ***pgi***
	Energy Metabolism	ATP synthesis	***atpA***, ***atpF***, ***atpE***, ***atpD***
		Methane metabolism	***glyA***
		Nitrogen metabolism	*NAPC*, *NAPB*, *gdhA*
		Oxidative phosphorylation	*FLDA*, *CYDA*, ***ppa***, *CYDB*
	Folding, Sorting and Degradation	Protein export	***YAJC***, *SECB*
		Protein folding and associated processing	***HSPE1***, ***HSPD1***
	Glycan Biosynthesis	Lipopolysaccharide biosynthesis	*RFAD*, *GMHA*, *lpxA*, *kdsA*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	and Metabolism	Peptidoglycan biosynthesis	***glnA***
	Lipid Metabolism	Fatty acid biosynthesis (path 1)	***fabG***, ***ACPP***, *FABZ, fabD, E6.4.1.2, fabI, fabB*
		Glycerolipid metabolism	*psd*
	Membrane Transport	ABC transporters, prokaryotic	*potA, metQ, potD, ABC.PA.S, modA*
		Major facilitator superfamily (MFS)	*GLPT*
		Phosphotransferase system (PTS)	*PTS-Glc-EIIA*
		Pores ion channels	***DNAK***, *TC.OOP*
	Metabolism of Cofactors and Vitamins	Porphyrin and chlorophyll metabolism	***gltX***
		Riboflavin metabolism	*RIBH*
		Ubiquinone biosynthesis	*menB*
		Vitamin B6 metabolism	*serC*
	Metabolism of Other Amino Acids	Glutathione metabolism	*pepN, zwf, gor*
		Selenoamino acid metabolism	***cysK***
		Taurine and hypotaurine metabolism	***ackA***
	Nucleotide Metabolism	Purine metabolism	*gpt,* ***guaB***, ***apt***, ***guaA***, ***adk***, ***purM***, ***hpt***
		Pyrimidine metabolism	***pnp***, *DEOD,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			upp, *nrdF*, ***pyrG***
	others	others	***KARS***, *K06873*, *SSPA*, *HFQ*, *NIFU*, *K06866*, ***TIG***, *COML*, ***SOD2***, ***K09747***, *K09802*, *SLYD*, ***TRXA***, *YJGF*, ***pyrH***, ***tgt***, ***nifS***, *SLYB*, *GRXA*, *E4.4.1.21*, *znuA*, *prlC*, ***ahpC***, *PPIB*
	Replication and Repair	Replication complex	***SSB***
	Signal Transduction	Phosphatidylinositol signaling system	***adk***
		Wnt signaling pathway	*ftn*
	Transcription	Basal transcription factors	***NUSA***, *NUSB*, ***GREA***, ***NUSG***
		HTH family transcriptional regulators	*FIS*
		Other and unclassified family transcriptional regulators	***CSPA***
		RNA polymerase	***RPOC***, ***RPOZ***, ***RPOB***
	Translation	Aminoacyl-tRNA biosynthesis	***serS***, *glnS*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			thrS
		Ribosome	***rpsK***, ***rpmI***, ***rplR***, ***rpsJ***, ***rplI***, ***rpmF***, ***rpsU***, ***rpsC***, ***rpsB***, ***rplT***, ***rplC***, ***rpsE***, ***rplV***, ***rpsR***, ***rplF***, ***rpsO***, ***rplD***, ***rpsQ***, ***rplL***, ***rplS***, ***rplO***, ***rpmG***, *rplY*, ***rpsT***, ***rplJ***, ***rpsH***, ***rplB***, ***rpsA***, ***rpmE***, ***rpsL***, ***rpsI***, ***rplX***, ***rplP***, ***rpsD***, ***rplQ***, ***rplM***, ***rplK***, ***rpsG***, ***rpmA***, ***rpmH***, ***rplW***, ***rpmB***, ***rpsF***, ***rplA***
		Translation factors	***infB***, ***efp***, ***tufA***, *infA*, ***fusA***, ***tsf***, ***frr***
Pseudomonad	Amino Acid Metabolism	Alanine and aspartate metabolism	***purB***, ***purA***, ***alaS***, ***aspS***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Arginine and proline metabolism	***proS***, *putA*, ***argS***
		Glycine, serine and threonine metabolism	*serA*, *betB*
		Histidine metabolism	*hisA*
		Lysine biosynthesis	*dapD*
		Methionine metabolism	***metK***, *ahcY*, *metH*
		Urea cycle and metabolism of amino groups	*speF*, *argH*, ***argG***
		Valine, leucine and isoleucine biosynthesis	***ilvD***, ***leuS***, ***ileS***, *leuA*, ***ilvC***, ***valS***, ***ilvE***
	Carbohydrate Metabolism	Butanoate metabolism	*gabD*
		Citrate cycle (TCA cycle)	*sucD*, *sucB*
		Fructose and mannose metabolism	***manB***
		Glycolysis / Gluconeogenesis	*fbp*, ***fbaA***, ***pgk***, *E1.2.1.3*, ***eno***, ***pdhD***
		Glyoxylate and dicarBoxylate metabolism	*acnB*, *aceA*, ***gltA***
		Pentose phosphate pathway	***tktB***
		Propanoate metabolism	*prpC*
		Pyruvate metabolism	*gloA*
		Starch and sucrose metabolism	***galU***
	Energy Metabolism	Oxidative phosphorylation	*etf*, ***ppa***
		Reductive carBoxylate cycle	*ppsA*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		(CO2 fixation)	
	Folding, Sorting and Degradation	Protein export	*SECB*
		Protein folding and associated processing	*DSBA*, ***HSPE1***, ***HSPD1***, *DSBC*
	Glycan Biosynthesis and Metabolism	Peptidoglycan biosynthesis	***glnA***, ***murC***
	Lipid Metabolism	Fatty acid biosynthesis (path 1)	*fabH*, *fabB*
	Membrane Transport	Pores ion channels	***DNAK***
	Metabolism of Cofactors and Vitamins	One carBon pool by folate	*metF*
		Thiamine metabolism	*THIG*
		Ubiquinone biosynthesis	*ubiG*
	Metabolism of Other Amino Acids	Glutathione metabolism	***icd***
	Nucleotide Metabolism	Purine metabolism	***purO***, ***purL***, ***purC***, ***guaA***, ***purF***, ***purD***, ***hpt***, ***guaB***, ***adk***, ***purM***
		Pyrimidine metabolism	***ndk***, ***pnp***, ***nrdE***, ***trxB***, ***pyrC***, *dut*, ***pyrG***, ***carB***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	others	others	***SOD2***, ***K09747***, *K09780, SLYD, K09767, ctpA, pepA, ZUR,* ***K09710***, ***ahpC***, *PPIB,* ***KARS***, *NIFU,* ***bipA***, ***K06942***, *SOHB, LIPA, ABC-2.AB.P, znuA, gabT, prlC*
	Signal Transduction	Phosphatidylinositol signaling system	***adk***
	Translation	Aminoacyl-tRNA biosynthesis	***serS***, *glnS, glyS*
		Other translation factors	*GATB*
		Translation factors	***tsf***
Rhizobiales	Amino Acid Metabolism	Alanine and aspartate metabolism	*pycB, aspB,* ***alaS***, *E2.6.1.18,* ***purB***, ***purA***, ***aspS***
		Arginine and proline metabolism	***proS***, *pip,* ***argS***
		Glycine, serine and threonine metabolism	***ALAS***, *betB,* ***asd***, *lysC, thrA, soxD, gcvPB, ltaA*
		Histidine metabolism	*hisG, hisC, HISF, hisB, hisI,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			HISH, hutH, hisD, hisS
		Lysine biosynthesis	*dapE, lysA, dapA, dapB, dapD*
		Methionine metabolism	*metY,* ***metK****, metH, ahcY*
		Phenylalanine metabolism	*dadA*
		Phenylalanine, tyrosine and tryptophan biosynthesis	*aroC, ARO1, trpB, aroH,* ***tyrS****, trpD, trpA, aroA, tyrC, trpD,* ***pheS***
		Tryptophan metabolism	***trpS***
		Urea cycle and metabolism of amino groups	*argC, speF,* ***argF****, argB, argJ, proA, argH,* ***argG****, argD*
		Valine, leucine and isoleucine biosynthesis	*LEUD,* ***ilvD****,* ***leuS****,* ***ilvC****, ilvB,* ***valS****, LEUC,* ***ileS****, leuA, leuB, ilvH*
		Valine, leucine and isoleucine degradation	*bkdA1, bkdA2*
	Biodegradation of Xenobiotics	Benzoate degradation via CoA ligation	*paaH, E3.1.2.-*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Nitrobenzene degradation	*E1.2.1.-*
	Biosynthesis of Secondary Metabolites	Terpenoid biosynthesis	*ispA*
	Carbohydrate Metabolism	Aminosugars metabolism	*E3.1.4.-*, ***glmS***, *pgm*, ***murA***, *glmU*
		Butanoate metabolism	*gabD, E5.1.2.3*
		Citrate cycle (TCA cycle)	*sucA*, *sucC*, *SDHA*, *sucD*, *SDHB*, *fumC*, *sucB*, *SDHD*
		Glycolysis / Gluconeogenesis	*E1.2.1.3*, ***eno***, *adh*, ***pyk***, ***pdhC***, *pdhB*, *acs*, ***pgk***, ***gpm***, *pdhA*, ***gapA***
		Glyoxylate and dicarBoxylate metabolism	*glcD*, *mdh*, *folD*, *acnA*, *E1.1.1.37A*, *glcB*, *aceA*, *purU*, ***gltA***
		Inositol metabolism	*IOLD*, *iolA*
		Pentose and glucuronate interconversions	*xylA*
		Pentose phosphate pathway	***talB***, ***rpe***, *gnd*, ***prsA***, ***tktB***, *devB*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Pyruvate metabolism	*gloA, gloB, ppdK, E1.1.1.39*
		Starch and sucrose metabolism	*E3.6.1.-, ugd,* ***galU****, glk,* ***pgi***
	Cellular Processes	Cell division	*FTSQ, GID, MRP, FTSI,* ***FTSZ****,* ***hflB****, FTSJ*
		Flagellar assembly	*FLIE, FLIC, FLIQ, FLGL, FLIF, FLGB, FLHB, FLGG, FLGH*
	Energy Metabolism	ATP synthesis	***atpA****,* ***atpF****, atpG, atpB, atpH,* ***atpE****,* ***atpD****,* ***atpC***
	Energy Metabolism	Methane metabolism	***glyA***
		Nitrogen metabolism	*CCMF, gltD, gltB, CCME*
		Oxidative phosphorylation	*CYT1, NUOL, NUOH, etf, NUOF, COXB, etfA, nouD, COXA, NUOB, NUOI,* ***ppa****, NUON, NUOK, NUOG, NUOE,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			COXC, NUOM, NUOJ, etfB, CYTB, NUOC
		Photosynthesis	*SUFE*
		Sulfur metabolism	*cysE*
	Folding, Sorting and Degradation	Protein export	***YAJC***, *OXA1, ftsY,* ***lepB***, *SECB, TATC, TATA, SECY, SECF,* ***SECA***, *ffh*
		Protein folding and associated processing	*CLPB,* ***DNAJ***, *IBPA,* ***clpP***, ***HSPE1***, *HFLC, clpQ, HFLK, HSLU, HSLO,* ***GRPE***, *CLPA,* ***HSPD1***, ***CLPX***, *lon*
	Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	*HTRB, lpxA, kdsA, kdsB*
		Peptidoglycan biosynthesis	***glnA***, ***murC***, *dat*
		Sphingoglycolipid metabolism	*E1.3.99.-*
	Lipid Metabolism	Biosynthesis of steroids	*dxs, ispH, ispG*
		Fatty acid biosynthesis (path 1)	***fabG***, *accD,* ***ACPP***, *fabA, FABZ, fabD,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			E6.4.1.2, accA, fabB, accB
		Fatty acid biosynthesis (path 2)	*fadA*
		Fatty acid metabolism	*fadD*
		Glycerolipid metabolism	*psd, glpD, pssA*
	Membrane Transport	ABC transporters, ABC-2 and other types	*ABC.CD.A*
		ABC transporters, prokaryotic	*malG, ABC.PA.A, ABC.SS.P, livG, pstA, metQ, metN, malK, livK, potD, pstB, livH, ABC.FE.S, ABC.PA.S,* ***pstS****, ABC.SS.S, metI, tauB, tauA, PHOU, malF, livM, ABC.PE.S, livF, malE, pstC, ABC.SS.A*
		Other ion-coupled transporters	*TC.HAE1, ACRA, TC.AMT*
		Other transporters	*spoIIIE*
		Phosphotransferase system (PTS)	*PTS-HPR*
		Pores ion channels	*TC.MSCL,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			YEGD, MOTA, MOTB, ***DNAK***
	Metabolism of Cofactors and Vitamins	Folate biosynthesis	*MOAA, folC, folE*
		Nicotinate and nicotinamide metabolism	*pncB, iunH, pntB, pntA*
		One carBon pool by folate	*gcvT*
		Pantothenate and CoA biosynthesis	*COAA*
		Porphyrin and chlorophyll metabolism	*hemB, COBW, COBH, hemF, COBI,* ***gltX****, cobT*
		Ubiquinone biosynthesis	*UBIE*
		Vitamin B6 metabolism	*serC*
	Metabolism of Other Amino Acids	Aminophosphonate metabolism	*E2.6.1.-, E1.1.99.-*
		beta-Alanine metabolism	*acd, paaG*
		Glutathione metabolism	***icd****, pepN, gst, gshA, zwf, gshB*
		Selenoamino acid metabolism	*cysD,* ***cysK****, cysN, cysI*
	Nucleotide Metabolism	Purine metabolism	*spoT,* ***purO****,* ***purL****, adeC, purK,* ***purC****,* ***guaB****,* ***guaA****,* ***purF****, amn,* ***purM****,* ***purD****,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			hpt
		Pyrimidine metabolism	***ndk***, ***pnp***, *carA*, ***nrdE***, ***upp***, *polA*, *dcd*, ***pyrE***, ***pyrG***, ***carB***
	others	others	*MOXR*, ***K06878***, *PRMA*, *fpr*, *K07145*, *BIOY*, *K09774*, *E2.5.1.44*, *lolD*, *RIBB*, *E4.2.99.-*, *pepP*, ***K09748***, *S**PPA***, *RLUC*, ***TIG***, *FSR*, *BFR*, ***fabF***, *COML*, ***bipA***, ***K06942***, ***SOD2***, *PPID*, *MGTE*, *glmM*, ***K09747***, *K06890*, *TOLB*, *K09780*, *PDHR*, *IRR*, *LEPA*, *SUFB*, *MIAB*, *K09117*, *K06861*, *dcp*, *GLNB*, *LIPA*, *K06915*, *glnD*, *K07021*, *lolE*,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			pheA, ctpA, ABCF3, ***tgt****, YGCA, K07146, sufC, SPOU, RMUC, MVIN, ugpB, pepA, TLDD, ABCB-BAC, MRCA, pepB, ABC.X4.A, K07018, GCVH, phoL, YAET, ECM4, SURA, PHAC, K07736, YGIH, ERA, ptsP, K06941, K07457, ENGB, LEMA, RND,* ***trmU****, msrB, imp, YFIH, himD, FDXA, DEAD, E3.4.21.-, ptrB, HAM1, SMPB*
	Replication and Repair	DNA polymerase	*dnaN, dnaX, dnaQ*
		Other replication, recombination and repair factors	*UVRB, pcrA, XSEB, E3.1.11.5,* ***HUPB****, MFD,*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

			DPS, ***RECA****,* *RADC, UVRA, RUVA, radA*
		Replication complex	***GYRA****, PARE, TOPA,* ***GYRB****,* ***SSB****, DNAB, DNAA, PARC*
	Signal Transduction	Phosphatidylinositol signaling system	***adk***
	Transcription	Basal transcription factors	***NUSA****, NUSB, RHO,* ***GREA****,* ***NUSG***
		HTH family transcriptional regulators	*GLPR*
		Other and unclassified family transcriptional regulators	***CSPA***
		RNA polymerase	***RPOC****,* ***rpoD****, rpoH,* ***RPOZ****,* ***RPOA****, rpoE,* ***RPOB***
	Translation	Aminoacyl-tRNA biosynthesis	***serS****, cysS, glyS, glyQ,* ***thrS****, spoVC*
		Other translation factors	*GATB, GATC, rph,* ***map****, GATA*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

		Ribosome	***rpsK***, ***rpmI***, ***rpsP***, ***rplR***, ***rplI***, ***rplN***, ***rpsU***, ***rpsS***, ***rpsC***, ***rpsB***, ***rplT***, ***rplC***, ***rpsE***, ***rplV***, ***rpsR***, ***rplF***, ***rpsO***, ***rplD***, ***rpsQ***, ***rplE***, ***rplL***, ***rplS***, ***rplO***, ***rpmG***, *rplY*, ***rpsT***, ***rplJ***, ***rpsH***, ***rplB***, ***rpsA***, ***rplU***, ***rpmE***, *TRUB*, ***rpsL***, ***rpsI***, ***rplX***, *RIMM*, ***rpsD***, ***rpsM***, ***rplQ***, ***rplM***, ***RBFA***, ***rpsN***, ***rplK***, ***rpmC***, ***rpsG***, ***rpmA***, ***rplW***, ***rpmB***, ***rpsF***, ***rplA***
		Translation factors	***infB***, ***efp***, ***tufA***, ***prfA***, *prfC*, *infA*, ***fusA***, ***tsf***, *prfB*, *infC*, ***frr***

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

1 Puigbo, P. et al. (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. 35, W126-W131

2 Henry, I. and Sharp, P.M. (2007) Predicting gene expression level from codon usage bias. Mol. Biol. Evol. 24, 10-12

3 Karlin, S. et al. (2004) Comparative analysis of gene expression among low G+C gram-positive genomes. Proc. Natl. Acad. Sci. U. S. A. 101, 6182-6187

4 Karlin, S. et al. (2005) Predicted highly expressed genes in archaeal genomes. Proc. Natl. Acad. Sci. U. S. A. 102, 7303-7308

5 Carbone, A. et al. (2003) Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19, 2005-2015

6 Carbone, A. (2006) Computational prediction of genomic functional cores specific to different microbes. J. Mol. Evol. 63, 733-746

7 Martin-Galiano, A.J. et al. (2004) Relationship between codon biased genes, microarray expression values and physiological characteristics of Streptococcus pneumoniae. Microbiology 150, 2313-2325

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

8 Wu, G. et al. (2005) Predicted highly expressed genes in the genomes of Streptomyces coelicolor and Streptomyces avermitilis and the implications for their metabolism. Microbiology 151, 2175-2187

9 Carbone, A. et al. (2005) Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol. Biol. Evol. 22, 547-561

10 Willenbrock, H. and Ussery, D.W. (2007) Prediction of highly expressed genes in microbes based on chromatin accessibility. BMC Mol. Biol. 8, 11

ISBN: 978-84-691-0653-2 /DL: T.2224-2007

4. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). **Pere Puigbò**, Ignacio G. Bravo, Santiago Garcia-Vallvé. Submited to BMC Bioinformatics.

ABSTRACT

Background: The Codon Adaptation Index (CAI) is a measure of the synonymous codon usage bias for a DNA or RNA sequence. It quantifies the similarity between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set. Since extreme values in the nucleotide or in the amino acid composition have an impact on differential preference for synonymous codons, it is essential to define an expected value of CAI in order to properly interpret the CAI and provide statistical support to CAI analyses. Though several freely available programs calculate the CAI for a given DNA sequence, none of them corrects for compositional biases or provides confidence intervals for CAI values.

Results: The E-CAI server, available at http://genomes.urv.es/CAIcal/E-CAI, is a new web-application that calculates a novel expected value of CAI for a set of query sequences by generating random sequences with the same G+C and amino acid content to those of the input. An executable file, a tutorial, a Frequently Asked Questions (FAQ) section and several examples are also available. To exemplify the use of the E-CAI server, we have analysed the codon adaptation of human mitochondrial genes that codify a subunit of the mitochondrial respiratory chain (excluding those genes that lack a prokaryotic orthologue) and are encoded in the nuclear genome. It is assumed that these genes were transferred from the proto-mitochondrial to the nuclear genome

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

and that its codon usage was then *ameliorated*.

Conclusions: The E-CAI server provides a direct threshold value for discerning whether the differences in CAI are statistically significant or whether they are merely artifacts that arise from internal biases in the G+C composition and/or amino acid composition of the query sequences.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

BACKGROUND

The Codon Adaptation Index (CAI), introduced by Sharp and Li [1], is a measure of the synonymous codon usage bias for a DNA or RNA sequence and measures the resemblance between the synonymous codon usage of a gene and the synonymous codon frequencies of a reference set. The CAI index ranges from zero to one: it is 1 if a gene always uses, for each encoded amino acid, the most frequently used synonymous codon in the reference set. Though it was originally developed to assess how effective selection has been at moulding the pattern of codon usage [1], it has since been applied to problems such as predicting the expression level of a gene [2], predicting a group of highly expressed genes [3,4], assessing the adaptation of viral genes to their hosts [1], giving an approximate indication of the likely success of heterologous gene expression [5], making comparisons of codon usage preferences in different organisms [1], identifying horizontally transferred genes [6-8], detecting dominating synonymous genomic codon usage bias in genomes [9], acquiring new knowledge about species lifestyle [10], and identifying the causes of protein rate variation [11,12].

Since the absolute value of the CAI depends on the query sequence and the reference set, both of these parameters are important for correctly interpreting CAI values. On the one hand, if the reference set has a random synonymous codon usage with few differences in the use of synonymous codons, the CAI values will be high, i.e. close to one. On the other hand, extreme G+C and/or amino acid compositions on the query sequence may lead to extreme CAI values that are not directly linked to codon usage preferences. It is therefore essential to define a threshold level for the expected CAI value (eCAI) in order to interpret the significance of codon usage biases and to provide statistical support to CAI analyses. The eCAI estimated by our server makes it possible to discern whether differences in the CAI are statistically significant or whether

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

they cannot be distinguished from biases due to nucleotide or amino acid composition. Although several authors have used some kind of expected codon usage [13,14], there is no server or program available to estimate it.

IMPLEMENTATION

The E-CAI server uses a novel algorithm that calculates an expected CAI for a set of query sequences by generating random sequences with similar G+C content and amino acid composition to the query sequences. The server, implemented in PHP, is integrated with several tools for the calculation and graphical representation of CAI. An executable version has been written entirely in Perl and precompiled for use in Linux, Windows and Macintosh operating systems. The Perl source code is available on request. A tutorial, a Frequently Asked Questions (FAQ) section and several examples are available from the home page of the server.

Inputs of the server

The basic inputs for calculating the expected CAI value are the query sequences, the codon usage of the reference set and the genetic code used. The query sequences must be DNA or RNA sequences in fasta format. The codon usage of the reference set can be introduced in a variety of formats, including the format of the Codon Usage Database [15]. Optionally, the user can introduce a G+C percentage to generate the random sequences. If this G+C percentage is not introduced, the server uses the G+C percentage from the query sequences.

Generation of the random sequences and estimation of the expected CAI

The method for estimating an expected CAI is based on generating 500

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

random sequences with the same amino acid composition as the query but with codon usage assigned randomly, either on the basis of the average G+C content of the input, or on the basis of the G+C percentage introduced by the user. Once all random sequences are generated, their CAI values are calculated. The normality of the CAI values of the random generated sequences is assessed with a Kolmogorov-Smirnov Test. An expected CAI value is then estimated using an upper one-sided tolerance interval for a normal distribution and a confidence limit and a percentage of the population (also called coverage) chosen by the user [16]. A tolerance interval is a way to determine a range within which, with some confidence, a specified proportion of a population falls. The eCAI therefore represents the upper limit of the CAI for sequences with a codon usage caused solely by mutational bias. This means that if the CAI value of a gene is greater than the expected value estimated on composition bias alone, it may be considered evidence of codon usage adaptation or selection. An effective and intuitive way to compare the CAI value of a gene with its expected CAI value is to use that we call the normalised CAI value. This normalised CAI is defined as the quotient between the CAI of a gene and its expected value.

The E-CAI server allows two methods for generating the random sequences. The first one, called *Markov*, is a Markov Model of order 0. This means that the probability of finding an amino acid at a specific position is independent of the other amino acid positions. The Markov method generates the random sequences by adding one amino acid each time, using the frequencies of each amino acid in the query sequences and a random number. It chooses a random number in the interval (0,1), sums the fractions of the amino acid composition of the query and assigns as the next amino acid the one that causes the sum to exceed the random number [17]. This process is repeated until the desired length of the sequence is reached. The random sequences are then back-translated to DNA sequences, assigning randomly one of the synonymous codon to each amino acid, either on the basis of the average

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

G+C content of the input or on the basis of the G+C percentage introduced by the user. The second method for generating the random sequences, called *Poisson*, is based on the assumption that the number of occurrences for each amino acid in a sequence follows a Poisson distribution. The normalised amino acid frequencies in the query sequences multiplied by the length (*n*) of the generated random sequences are used as the expected numbers of occurrences of each amino acid in the random sequences. These values are used to calculate the probabilities that there were exactly *k* occurrences of each amino acid in a sequence of length *n*. From the sum of these probabilities and a random number, the expected number of occurrences for each amino acid in a random sequence is calculated in a similar way to the Markov method. This process is repeated until the desired number of sequences has been generated. Again, the random sequences are then back-translated to DNA sequences by the same method described above. The results generated by the Markov and Poisson methods are comparable, but the Markov method is more precise and the Poisson method is faster. In addition, similar values of eCAI are obtained when the GenRGenS software is used to generate the random sequences [18].

Interpretation of the results

The reference set used to calculate the CAI is important for the correct interpretation of its meaning. The CAI measures the similarity between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set. If this reference set is a group of highly expressed genes and in the presence of selected codon usage bias, the CAI values can be used to predict the expression level of genes [19]. If the average codon usage of a genome is used as a reference set, the CAI can be interpreted as a measure of the codon adaptation of a gene in the context of a genome. This information can be used to improve the expression of a gene in a heterologous expression system [5]. The values of eCAI calculated by the E-CAI server are

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

expected to be over-estimations because the synonymous codon usage of genes is highly influenced by the G+C content at the third codon position and because amino acid usage is also species-specific [20]. The query sequences define both nucleotide and amino acid composition and are therefore important factors in the calculation of eCAI. The expected CAI value would be meaningless if the composition of the query sequences were very heterogeneous. To assess the homogeneity of the sequences in the query set, a Chi-Square test is calculated to test the goodness-of-fit between the amino acid composition or G+C content of each of the query sequences and the average values used to generate the random sequences. The percentage of query sequences that fit the amino acid and/or G+C mean distributions are then shown. If the query sequences are compositionally very heterogeneous, these percentages will be small. In this case we suggest splitting the query sequences into smaller and homogeneous subsets and estimating the eCAI values for each of the subsets separately.

Executable version

To calculate CAI values for hundreds or thousands of sequences on a whole-genome scale and generate an eCAI, users can download an executable program that automatically performs these calculations. The inputs, methods and outputs of this executable version are the same as those of the web version. However, it enables one to choose the length and number of randomly generated sequences. More details about this script and how to use it are found in the tutorial.

RESULTS

EXAMPLE: The Amelioration of mitochondrial genes encoded in the human nuclear genome

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

It is widely accepted that mitochondria have their origin in a single event, arising from a bacterial symbiont whose closest contemporary relatives are found within the alfa-proteobacteria [21, 22]. Since its origin, the mitochondrial genome has undergone a streamlining process of genome reduction with intense periods of loss of genes [23]. Nowadays, mitochondrial genomes exhibit a great variation in protein gene content among most major groups of eukaryotes, but only limited variation within large and ancient groups. This suggests a very episodic, punctuated pattern of mitochondrial gene loss over the broad sweep of eukaryotic evolution [24]. Mitochondrial genomes have lost genes that lack a selective pressure for their conservation. This could include genes whose function may no longer be necessary, genes whose function has been superseded by some pre-existing nuclear genes or genes that have been transferred to the nucleus [23]. The gene content of present mitochondrial genomes varies from 63 protein-coding genes in *Reclinomonas americana*, a flagellate protozoon, to three genes in other species (see the GOBASE database [25], which contains information for more than 1500 complete mitochondrial genomes). Mitochondria in Vertebrates encode for 13 respiratory-chain proteins and for a minimal set of tRNAs that suffice to translate all codons. However, the vast majority of proteins located in the mitochondria are the product of nuclear genes. These genes are transcribed in the nucleus, translated in the cytoplasm and the proteins are subsequently vehiculated to the mitochondria. Some, those which show homology to present prokaryote genes, are thought to be the result of horizontal gene transfer events from the proto-mitochondrial to the nuclear genome. This hypothesis is reinforced by the fact that several of these genes are encoded in the mitochondrial genome in other eukaryotic species [26].

To exemplify the use of the CAI server and the significance of expected CAI values, we have analyzed the differential codon adaptation of human mitochondrial genes to both the human codon usage and the mitochondrial codon usage. We used the human codon usage table from Lander et al. [27]

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

and the mean codon usage of all genes from human mitochondrial genome (GenBank accession number AF347015) as human and mitochondrial reference sets, respectively. We have focused on genes that encode for a subunit of the mitochondrial respiratory chain complexes I to V, excluding those that lack a prokaryotic orthologue. Finally, we have divided the genes into two categories according to whether they are encoded in the nuclear or in the mitochondrial genome. More than half of the analyzed nuclear-encoded mitochondrial genes from human are present in the mitochondrial genome in other organisms, which reflects their proto-mitochondrial origin. Our results are summarised in Table 1, which shows the CAI values with respect to human codon usage (CAIhm) and to the average codon usage of genes encoded in the human mitochondrial genome (CAImt). Because of the heterogeneity in G+C content of the mitochondrial genes encoded in the nucleus, an expected value (eCAI) was estimated individually for each gene using the Poisson method, a 95% level of confidence and 99% coverage. These expected values are also shown in Table 1, as is the normalised CAI value, which is defined as the quotient between the CAI for each gene and its expected value. A value greater than one in this normalised expected CAI value means that the observed CAI is greater than its expected value, which could be interpreted as the result of an adaptation process in the codon usage. Table1 shows that most nuclear-encoded mitochondrial genes are better adapted to the nuclear codon usage than would be expected by chance, while mitochondrial-encoded mitochondrial genes are better adapted to the mitochondrial codon usage than would be expected by chance. The CAIhm values of all thirteen mitochondrial-encoded mitochondrial genes are below their expected upper limit, estimated using a sample of random genes with the same G+C content and amino acid composition (Table 1b). At the same time, twelve out of these thirteen genes have a CAImt above their expected upper limit at a 99% confidence level and 95% coverage. The obvious interpretation, therefore, is that mitochondrial-encoded mitochondrial

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

genes are better adapted to mitochondrial codon usage than to nuclear codon usage. Conversely, nuclear-encoded mitochondrial genes are better adapted to nuclear codon usage than to mitochondrial codon usage. Thirty-four of the 37 nuclear-encoded mitochondrial genes show a CAIhm above the expected upper limit at a 95% confidence level and 99% coverage, whereas only two genes have a CAImt above the expected upper limit at a 95% confidence level and 99% of coverage (Table 1a). This means that the codon usage of the genes originally encoded in the proto-mitochondria and that are now encoded in the human nuclear genome has been ameliorated and adapted to the human codon usage after their transfer to the nucleus. The E-CAI server provides individual CAI values for each gene with respect to both the nuclear and mitochondrial codon usages, as well as independent eCAI threshold values for differentiating true codon usage optimization from spurious random matches that may arise from compositional biases.

Several nuclear-encoded mitochondrial genes have a higher G+C content than mitochondrial-encoded mitochondrial ones. It could therefore be argued that the differences between CAI values of mitochondrial genes of different origin probably reflect differences in G+C content rather than differences in codon usage adaptation. To address this issue, in Figure 1 we have represented the normalised CAIhm of human mitochondrial genes against their G+C content at third codon position. Although some mitochondrial genes encoded in the nuclear genome have a higher G+C content than mitochondrial encoded ones, there are several mitochondrial genes, encoded in the nuclear and mitochondrial genome, with similar G+C contents. However, the normalised CAIhm is very different in both populations (figure 1), as is also demonstrated if a Kolmogorov-Smirnoff test ($D=1.0$, $P<0.0001$) is used. This clearly shows that the codon usage of the nuclear encoded genes is not only due to mutational pressure or G+C content, and that a certain degree of codon usage adaptation exists. In this sense, it has recently been reported that a weak positive correlation between gene expression

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

levels and the frequency of optimal codons exists in humans [28, 29].

CONCLUSIONS

The E-CAI server described here provides an expected value of CAI for discerning whether the differences in CAI are statistically significant and arise from the codon preferences or whether they are merely artifacts that arise from internal biases in the G+C composition and/or amino acid composition of the query sequences. Using a normalised CAI value, defined as the quotient between the CAI of a gene and its expected value, is an effective and intuitive way to analyze the codon usage bias of genes and codon usage adaptation.

AVAILABILITY AND REQUIREMENTS

Project name: E-CAI.

Project home page: http://genomes.urv.cat/CAIcal/E-CAI.

Operating system(s): Platform independent.

Programming language: PHP.

Other requirements: None.

Any restrictions to use by non-academics: license needed.

Authors' contributions: PP designed the server, made the programming task, helped to draft the manuscript and prepared the example. IGB participated in design of the server, developed the Poisson-based method, and helped to draft the manuscript. SG-V conceived and designed the server, coordinated the project and drafted the manuscript. All authors read and approved the final manuscript.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

ACKNOWLEDGEMENTS

This work has been financed by project BIO2003-07672 of the Spanish Ministry of Science and Technology. We thank John Bates and Kevin Costello of the Language Service of the Rovira i Virgili University for their help with writing the manuscript and Hervé Philippe for his comments on the manuscript.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

FIGURES

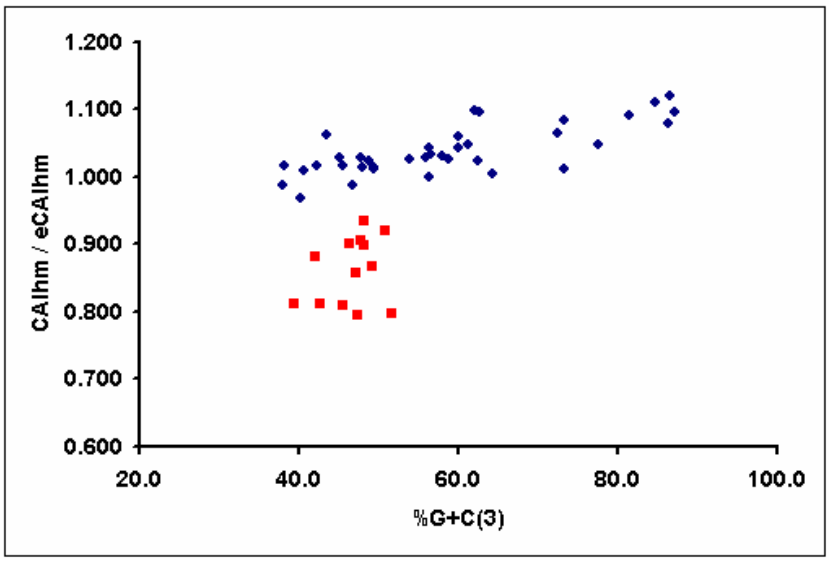


Figure 1. Graphical representation of the normalised CAIhm, defined as the quotient between the CAI of a gene and its expected value, versus G+C content at the third codon positions for the human genes that encode a subunit of a complex of the mitochondrial respiratory chain. Red squares represent mitochondrial genes encoded in the human mitochondrial genome and blue dots represent mitochondrial genes encoded in the human nuclear genome. An expected value of CAI was estimated for each gene with the E-CAI server, using the Poisson method and a 95% interval confidence and a 99% population coverage.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

TABLES

Table 1. Analysis of human mitochondrial genes that encode a subunit of complexes I-V of the mitochondrial respiratory chain encoded in the nuclear (a) or mitochondrial (b) genome.

a) Nuclear encoded genes										
Complex	**Gene name**	**Length**	**CAI_{hm}**		**$eCAI_{hm}$**	**CAI_{hm} / $eCAI_{hm}$**	**CAI_{mt}**		**$eCAI_{mt}$**	**CAI_{mt} / $eCAI_{mt}$**
					p=0.05	**p=0.05**			**p=0.05**	**p=0.05**
I	NDUFS1	2184	0.695	*	0.683	1.018	0.434		0.519	0.836
	NDUFS2	1392	0.765	**	0.734	1.042	0.391		0.500	0.782
	NDUFS3	795	0.754	*	0.750	1.005	0.402		0.488	0.824
	NDUFS7	642	0.867	**	0.780	1.112	0.442		0.446	0.991
	NDUFS8	633	0.868	**	0.796	1.090	0.439		0.465	0.944
	NDUFV1	1395	0.825	**	0.774	1.066	0.417		0.482	0.865
	NDUFV2	750	0.695		0.703	0.989	0.449		0.519	0.865
II	SDHC	510	0.699	*	0.679	1.029	0.377		0.457	0.825
	SDHD	480	0.663	*	0.654	1.014	0.387		0.464	0.834
	SDHA	1995	0.768	*	0.750	1.024	0.423		0.496	0.853
	SDHB	843	0.778	**	0.754	1.032	0.454		0.481	0.944
III	UQCRFS1	825	0.711	*	0.711	1.000	0.391		0.483	0.810
	CYC1	978	0.759	*	0.750	1.012	0.379		0.449	0.844
IV	COX10	1332	0.744	**	0.713	1.043	0.454		0.462	0.983
	COX11	831	0.738	*	0.725	1.018	0.407		0.513	0.793
	COX15	1140	0.707	*	0.688	1.028	0.411		0.472	0.871
V	ATP5B	1590	0.714	*	0.698	1.023	0.412		0.507	0.813

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	ATP5A1	1512	0.695	*	0.684	1.016	0.409		0.519	0.788
	ATP5C1	897	0.726	*	0.705	1.030	0.463		0.509	0.910
	ATP5O	642	0.700	**	0.681	1.028	0.429		0.486	0.883
	ATP5D	507	0.807	**	0.748	1.079	0.410		0.426	0.962
	ATP5G1	411	0.776	**	0.707	1.098	0.456		0.482	0.946
	ATP5G2	474	0.752	**	0.686	1.096	0.472	*	0.451	1.047
	ATP5G3	429	0.720	**	0.678	1.062	0.430		0.510	0.843
	ATP6V1A	1854	0.709	*	0.702	1.010	0.451		0.525	0.859
	ATP6V1B1	1536	0.703		0.711	0.989	0.439		0.514	0.854
	ATP6V1D	744	0.676		0.697	0.970	0.430		0.522	0.824
	ATP6V1E1	681	0.721	*	0.713	1.011	0.431		0.500	0.862
	ATP6V1E2	681	0.777	**	0.733	1.060	0.410		0.466	0.880
	TCIRG1	2493	0.857	**	0.781	1.097	0.421		0.434	0.970
	ATP6V0D2	1053	0.732	*	0.722	1.014	0.456		0.518	0.880
	ATP6V0C	468	0.838	**	0.748	1.120	0.511	**	0.461	1.108
	ATP6F	618	0.803	**	0.741	1.084	0.510		0.514	0.992
	ATP6V0D1	1056	0.831	**	0.793	1.048	0.457		0.495	0.923
	ATP6V0A1	2496	0.758	*	0.734	1.033	0.424		0.507	0.836
	ATP6V0A4	2523	0.770	**	0.735	1.048	0.458		0.494	0.927
	ATP6V0A2	2571	0.748	*	0.728	1.027	0.450		0.491	0.916

b) Mitochondrial encoded genes

Complex	**Gene Name**	**Length**	**CAI_{hm}**		**$eCAI_{hm}$**	**CAI_{hm} / $eCAI_{hm}$**	**CAI_{mt}**		**$eCAI_{mt}$**	**CAI_{mt} / $eCAI_{mt}$**
					p=0.05	**p=0.05**			**p=0.05**	**p=0.05**

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

I	ND1	957	0.635		0.796	0.798	0.760	**	0.456	1.667
	ND2	1044	0.616		0.774	0.796	0.677	**	0.457	1.481
	ND3	345	0.571		0.703	0.812	0.701	**	0.461	1.521
	ND4L	297	0.550		0.679	0.810	0.738	**	0.472	1.564
	ND4	1377	0.612		0.654	0.936	0.722	**	0.455	1.587
	ND5	1812	0.651		0.750	0.868	0.723	**	0.471	1.535
	ND6	525	0.612		0.754	0.812	0.361		0.551	0.655
III	CYTB	1134	0.655		0.711	0.921	0.758	**	0.481	1.576
IV	COX1	1542	0.644		0.750	0.859	0.715	**	0.509	1.405
	COX2	684	0.641		0.713	0.899	0.664	**	0.503	1.320
	COX3	780	0.656		0.725	0.905	0.704	**	0.497	1.416
V	ATP8	207	0.606		0.688	0.881	0.633	**	0.452	1.400
	ATP6	681	0.629		0.698	0.901	0.701	**	0.472	1.485

Expected CAIs (eCAIs) at 95% (p=0.05) and 99% (p=0.01) confidence and 99% coverage were calculated using the Poisson method of the E-CAI server. For the sake of clarity, only the eCAIs at p=0.05 are shown. CAIhm and CAImt mean CAI calculated using the mean nuclear and mitochondrial codon usage as a reference set, respectively. * and ** mean that the CAI is higher than the eCAI at $p<0.05$ and $p<0.01$, respectively. Normalised CAI values (defined as the quotient between the CAI and its expected value) greater than one are shaded and must be interpreted as evidence of adaptation to the reference codon usage beyond mere compositional biases.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

1. Sharp PM, Li WH: **The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.

2. Goetz RM, Fuglsang A: **Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli.** *Biochem Biophys Res Commun* 2005, **327:**4-7.

3. Wu G, Culley DE, Zhang W: **Predicted highly expressed genes in the genomes of Streptomyces coelicolor and Streptomyces avermitilis and the implications for their metabolism.** *Microbiology* 2005, **151:**2175-2187.

4. Wu G, Nie L, Zhang W: **Predicted highly expressed genes in Nocardia farcinica and the implication for its primary metabolism and nocardial virulence.** *Antonie Van Leeuwenhoek* 2006, **89:**135-146.

5. Puigbo P, Guzman, E, Romeu A, Garcia-Vallve S: **OPTIMIZER: A web server for optimizing the codon usage of DNA sequences.** *Nucleic Acids Res* 2007 *in press*.

6. Lawrence JG, Ochman H: **Molecular archaeology of the Escherichia coli genome.** *Proc Natl Acad Sci U S A* 1998, **95:**9413-9417.

7. Garcia-Vallve S, Palau J, Romeu A: **Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in Escherichia coli and Bacillus subtilis.** *Mol Biol Evol* 1999, **16:**1125-1134.

8. Garcia-Vallve S, Guzman E, Montero MA, Romeu A: **HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete**

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

genomes. *Nucleic Acids Res* 2003, **31:**187-189.

9. Carbone A, Zinovyev A, Kepes F: **Codon adaptation index as a measure of dominating codon bias.** *Bioinformatics* 2003, **19:**2005-2015.

10. Willenbrock H, Friis C, Juncker AS, Ussery DW: **An environmental signature for 323 microbial genomes based on codon adaptation indices.** *Genome Biol* 2006, **7:**R114.

11. Drummond DA, Raval A, Wilke CO: **A single determinant dominates the rate of yeast protein evolution.** *Mol Biol Evol* 2006, **23:**327-337.

12. McInerney JO: **The causes of protein evolutionary rate variation.** *Trends Ecol Evol* 2006, **21:**230-232.

13. Morton BR: **Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages.** *J Mol Evol* 1998, **46:**449-459.

14. Supek F, Vlahovicek K: **Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity.** *BMC Bioinformatics* 2005, **6:**182.

15. Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from international DNA sequence databases: status for the year 2000.** *Nucleic Acids Res* 2000, **28:**292.

16. Hahn GJ, Meeker WQ: Statistical intervals: a guide for practitioners. New York: Wiley, 1991.

17. Fitch WM: **Random sequences.** *J Mol Biol* 1983, **163:**171-176.

18. Ponty Y, Termier M, Denise A: **GenRGenS: software for generating**

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

random genomic sequences and structures. *Bioinformatics* 2006, **22:**1534-1535.

19. Henry I, Sharp PM: **Predicting gene expression level from codon usage bias.** *Mol Biol Evol* 2007, **24:**10-12.

20. Pasamontes A, Garcia-Vallve S: **Use of a multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes.** *BMC Bioinformatics* 2006, **7:**257.

21. Burger G, Gray MW, Lang BF: **Mitochondrial genomes: anything goes.** *Trends Genet* 2003, **19:**709-716.

22. Gray MW, Burger G, Lang BF: **Mitochondrial evolution.** *Science* 1999, **283:**1476-1481.

23. Gray MW, Burger G, Lang BF: **The origin and early evolution of mitochondria.** *Genome Biol* 2001, **2:**REVIEWS1018. .

24. Adams KL, Palmer JD: **Evolution of mitochondrial gene content: gene loss and transfer to the nucleus.** *Mol Phylogenet Evol* 2003, **29:**380-395.

25. O'Brien EA, Zhang Y, Yang L, et al: **GOBASE--a database of organelle and bacterial genome information.** *Nucleic Acids Res* 2006, **34:**D697-9.

26. Gabaldon T, Huynen MA: **Shaping the mitochondrial proteome.** *Biochim Biophys Acta* 2004, **1659:**212-220.

27. Lander ES, Linton LM, Birren B, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

28. Lavner Y, Kotlar D: **Codon bias as a factor in regulating expression via translation rate in the human genome.** *Gene* 2005, **345:**127-138.

29. Kotlar D, Lavner Y: **The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids.** *BMC Genomics* 2006, **7:**67.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

5. CAIcal: set of tools to assess codon usage adaptation. **Pere Puigbò**, Ignacio G. Bravo, Santiago Garcia-Vallvé. In preparation.

ABSTRACT

The Codon Adaptation Index (CAI) has been developed to measure the synonymous codon usage bias for a DNA or RNA sequence. The CAI quantifies the similarity between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set. CAIcal is a web-server available at http://genomes.urv.cat/CAIcal that includes a complete set of utilities related with the CAI. The server contains several important features, such as the calculation and graphical representation of the CAI along a sequence or a protein multialignment translated to DNA. The calculation of CAI and expected value of CAI (eCAI) is also included as one of the CAIcal tools.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

INTRODUCTION

Ever since a relatively high number of DNA sequences were publicly available in databases, several statistical analyses have been performed. One of the parameters that first interested the scientist was codon usage (Grantham and others 1980). It was soon discovered that there is a considerable heterogeneity in the codon usage between genes within species and that the degree of codon bias is positively correlated with gene expression (Gouy and Gautier 1982). (Carbone, Zinovyev, Kepes 2003)To quantify the degree of bias in the codon usage of genes, several parameters or indices have been developed. The Codon Adaptation Index (CAI), developed by Sharp and Li (Sharp and Li 1987), rapidly became one of the most used indices. The CAI is a measure of the synonymous codon usage bias for a DNA or RNA sequence and measures the similarity between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set. The index ranges from 0 to 1: it is 1 if a gene always uses the most frequently used synonymous codons in the reference set. Though it was developed to assess the extent to which selection has been effective at moulding the pattern of codon usage(Sharp and Li 1987), it has other uses, e.g. for predicting the level of expression of a gene (Goetz and Fuglsang 2005; Puigbo and others 2007; Wu, Culley, Zhang 2005), for assessing the adaptation of viral genes to their hosts (Sharp and Li 1987), for giving an approximate indication of the likely success of heterologous gene expression [3,6], for making comparisons of codon usage in different organisms (Grote and others 2005; Sharp and Li 1987)(Sharp and Li 1987), for detecting dominating synonymous codon usage bias in genomes (Carbone, Zinovyev, Kepes 2003) and for studying cases of horizontally transferred genes (Garcia-Vallve and others 2003).

The CAIcal web-server includes a complete set of tools related with codon usage adaptation. CAI is calculated as described in (Puigbo, Bravo, Garcia-

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Vallvé, 2007), i.e. following the original method proposed by Sharp and Li (Sharp and Li 1987) but using the recent computer implementation proposed by Xia (Xia 2007). The web-server calculates the CAI for a group of sequences using different reference sets and has other features, e.g. the representation of the CAI along a sequence or multialignment and the estimation of an expected CAI value and its confidence interval

DESCRIPTION OF THE CAIcal SERVER

The web-server created with PHP is available at http://genomes.urv.cat/CAIcal. The graphical user TCL/TK interface executes a Perl program to easily calculate the CAI and eCAI locally. The web-server that it has been running since 2005, it has been improved periodically with new features and it has been extensively proofed. In the following subsections we describe the inputs of the server and its main features.

Inputs of the server

The inputs for the server depend on the calculation to be performed. The basic inputs for calculating CAI are the query sequences, the reference set and the genetic code. The query sequences must be DNA or RNA sequences in fasta format. The server first checks whether the query sequences are a DNA or RNA region that codifies a protein. The reference set needed to calculate the CAI can be introduced in a variety of formats, including that of the Codon Usage Database (Nakamura, Gojobori, Ikemura 2000). A direct link to this database is provided in the CAIcal interface. This database contains codon usage tables extracted from GenBank and organized by species. Several of the calculations available in CAIcal, such as the CAI calculation and its representation in a sequence, can be used with two reference sets simultaneously. Therefore, it is easier to compare the codon usage of a gene with the codon usage of two different organisms and check whether it is more

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

adapted to one of them. See the tutorial available from the server home page for a complete description of errors and warnings and for more information about input requirements.

Set of tools

The server first provides several basic calculations that are also available elsewhere:

(i) The absolute and synonymous codon usage of a group of DNA sequences and other useful parameters such as length, total G+C content and G+C content at the three codon positions, and the effective number of codons (Wright 1990).

(ii) The CAI of a DNA sequence or group of sequences. This index measures the adaptation of the synonymous codon usage of a gene to the synonymous codon usage of up to two reference sets that can be chosen by the user.

(iii) An expected value of CAI (Puigbo, Bravo, Garcia-Vallvé, 2007) is determined by randomly generating 500 sequences from the G+C content and the amino acid composition of the query sequences. This expected CAI therefore provides a direct threshold value for discerning whether the differences in the CAI value are statistically significant and arise from the codon preferences or whether they are merely artefacts that arise from internal biases in the G+C composition and/or amino acid composition of the query sequences (Puigbo, Bravo, Garcia-Vallvé, 2007). Additionally, one of the tools included in CAIcal is a graphical local user interface that can be downloaded and allows the calculation of the CAI and eCAI of hundreds or thousands of sequences on a whole-genome scale easily.

There are other programs, such as CodonW and EMBOSS (Rice, Longden,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Bleasby 2000), and servers, such as JCAT (Grote and others 2005) and the CAI Calculator (Wu, Culley, Zhang 2005), that calculate the CAI for a gene or a group of genes. The differences between these programs mainly involve the way the reference set is introduced. As well as these basic calculations, the CAIcal server also compares features of codon usage and codon adaptation that have hitherto not been implemented online.

(iv) The weight of each codon (i.e. the frequency of codon use compared to the frequency of use of the optimal codon for that amino acid in the reference set) along a DNA sequence can be graphically represented using a window the length of which is defined by the user. This result provides an intuitive visualisation of the changes in the CAI throughout the input and identifies discontinuities that might correlate with informational and/or operational features of the DNA sequence.

(v) A graphical representation can be made of the weight of each codon along a multiple protein alignment that has been translated to a DNA alignment using a unique reference set for all the sequences of the alignment or using a reference set for each sequence. The inputs for this option are a protein multialignment in clustal format, the DNA sequence of each of the sequences of the multialignment (with the same identification field between the DNA and protein sequences) and one or more codon usage tables to use as reference sets. This result provides a graphical display that enables the protein sequence alignment to be correlated with the informational/compositional content of the DNA sequence that encodes them.

The options available in the server are summarized in figure 1. All these options are accessible from the main page of the server and several links have been created between them. For instance, after the CAI value of a group of sequences has been calculated, an expected CAI value can be estimated or the graphical representation of the CAI value along each sequence can be

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

visualized. Several parameters used in the calculations, such as the window length in the graphical representation of the CAI along a sequence or the upper confidence limit to estimate an expected CAI, are defined by the user. The results are therefore flexible and fit the needs of the user. For the results, the server provides several tables and graphs. Also, several text boxes containing the results in a tab-delimited format have been created, which makes it easy to copy and paste them into spreadsheet programs. Finally, a tutorial, a Frequently Asked Questions (FAQ) section and several examples are available from the home page of the server.

HOW TO USE THE CAIcal SERVER

The CAIcal helps to annotate genomic discontinuities such as the donor splicing site of the E4 ORF of papilomaviruses. Papillomaviruses (PVs) are a family of small dsDNA viruses that cause a variety of diseases including cervical cancer. The genome of PVs is modular with three different regions, each of which has a different evolutionary rate (Garcia-Vallve, Alonso, Bravo 2005; Garcia-Vallve and others 2006). These regions are: an upstream regulatory region (URR), an early region that codes for proteins (e.g. E1, E2, E4, E5, E6 and E7) involved in viral transcription, replication, cell proliferation and other steps of the viral life cycle, and a structural region that contains two genes that code for the capsid proteins L1 and L2. A general characteristic of genes encoded in human PVs is their peculiar codon usage preference compared to the preferred codon usage in human genes (Bravo and Muller 2005; Zhao, Liu, Frazer 2003), although the exact reason for this poor adaptation to the genome of their host is still unknown. Like other viral genomes, some of the PV genes overlap partially or completely. This is the case of the E4 gene, which completely overlaps the E2 gene in a different reading frame (Hughes and Hughes 2005). The function of E4 is not completely understood and its annotation is not very rigorous (Nakamura, Gojobori, Ikemura 2000). The mature E4 protein appears after splicing, with

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

the donor site situated some codons downstream from the start codon of the E1 gene, and the acceptor site situated close to the middle of the E2 gene (Doorbar 2005; Peh and others 2004). The fact that most of E4 overlaps with E2, that the mature E1^E4 protein contains a few amino acids from E1 and that the splice sites are not strictly conserved, makes it difficult to in silico determine the true E4 sequence. The E4 PVs genes available in the databases are therefore very different in length and similarity. Although the genomes of many PVs have been sequenced, information about the expression of their genes or cDNA sequences is only available for a few of them. One of these is HPV1. In this case, the annotation of the HPV1 E4 gene is confirmed by mRNA data (Palermo-Dilts, Broker, Chow 1990). However, the E4 gene from HPV63, a PV that is phylogenetically related to HPV1 (Garcia-Vallve, Alonso, Bravo 2005), is longer than the E4 gene from HPV1. The difference is 96 nucleotides that are located at the beginning of HPV63 E4. We can use the CAIcal server to show that the codon usage of these 96 nucleotides at the beginning of HPV63 E4 is very different from that of the rest of the E4 sequence, measured as the CAI value calculated with the human codon usage as reference (figure 2). This suggests that the acceptor splice site of HPV63 E4 is not well annotated and that the true E4 that overlaps with E2 probably starts downstream from the annotated position.

ACKNOWLEDGMENTS

We thank Kevin Costello of the Language Service of the Rovira i Virgili University for his help with writing the manuscript. We also thank Agnes Hotz-Wagenblatt from the HUSAR Bioinformatics Laboratory at Deutsches Krebsforschungszentrum and Obdulia Rabal from the “Institut Químic de Sarrià” for testing the server.

FIGURES

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

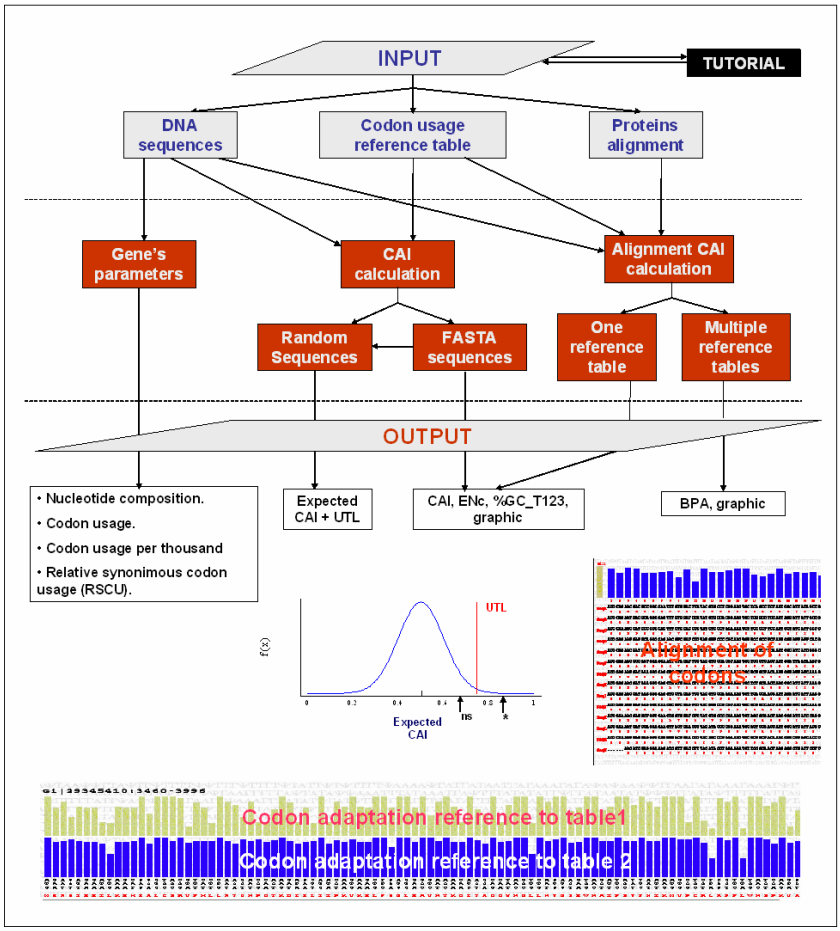


Figure 1. Schematic representation of the options available in the caical server. using a combination of three inputs (dna or rna sequences, a codon usage reference table and/or a protein alignment), the server calculates gene parameters such as %g+c, rscu and nc, the cai for one or more dna or rna sequences, an expected cai and upper tolerance limit and represents the cai along a dna sequence or in a protein multialignment translated to dna.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

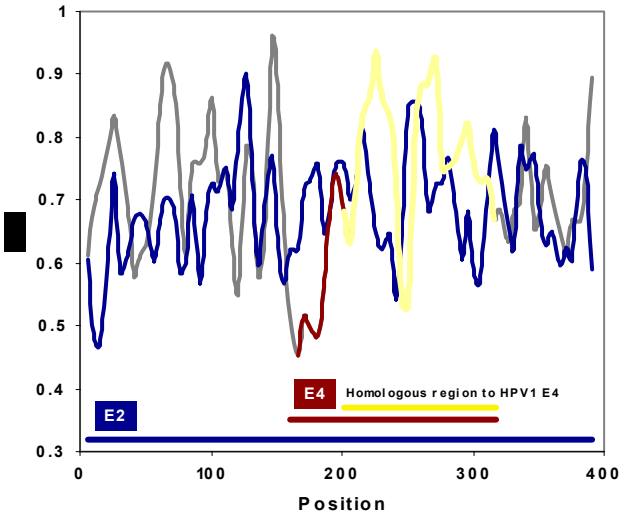


Figure 2. Representation of the cai, calculated using the human mean codon usage as a reference set, in the dna sequence that encodes hpv63 e2. The blue line represents the reading frame that encodes e2. The grey-red-yellow line represents the reading frame +1, which contains e4. The yellow line represents the fragment of hpv63 e4 homologous to the closely related hpv1 e4. The red line represents the stretch also annotated as hpv63 e4, but which lacks homology with hpv1 e4. Note that the initial e4 region from hpv63, which is not homologous to the hpv1 gene, has an extremely low cai, which suggests a wrong annotation for the e4 gene in hpv63. This figure was obtained using the output of the calculation of cai along a sequence of the caical server, with a window length of 11 and a window step of 5.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

Bravo IG and Muller M. 2005. Codon usage in papillomavirus genes: Practical and functional aspects. Papillomavirus Report 16:63-72.

Carbone A, Zinovyev A, Kepes F. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19(16):2005-15.

Doorbar J. 2005. The papillomavirus life cycle. J Clin Virol 32 Suppl 1:S7-15.
Garcia-Vallve S, Alonso A, Bravo IG. 2005. Papillomaviruses: Different genes have different histories. Trends Microbiol 13(11):514-21.

Garcia-Vallve S, Iglesias-Rozas JR, Alonso A, Bravo IG. 2006. Different papillomaviruses have different repertoires of transcription factor binding sites: Convergence and divergence in the upstream regulatory region. BMC Evol Biol 6:20.

Garcia-Vallve S, Guzman E, Montero MA, Romeu A. 2003. HGT-DB: A database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res 31(1):187-9.

Goetz RM and Fuglsang A. 2005. Correlation of codon bias measures with mRNA levels: Analysis of transcriptome data from escherichia coli. Biochem Biophys Res Commun 327(1):4-7.

Gouy M and Gautier C. 1982. Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res 10(22):7055-74.

Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8(1):r49-62.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC, Jahn D. 2005. JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res 33(Web Server issue):W526-31.

Hughes AL and Hughes MAK. 2005. Patterns of nucleotide difference in overlapping andnon-overlapping reading frames of papillomavirus genomes. Virus Research 113:81-8.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. Nucleic Acids Res 28(1):292.

Palermo-Dilts DA, Broker TR, Chow LT. 1990. Human papillomavirus type 1 produces redundant as well as polycistronic mRNAs in plantar warts. J Virol 64(6):3144-9.

Peh WL, Brandsma JL, Christensen ND, Cladel NM, Wu X, Doorbar J. 2004. The viral E4 protein is required for the completion of the cottontail rabbit papillomavirus productive cycle in vivo. J Virol 78(4):2142-51.

Puigbò P, Guzman E, Romeu A, Garcia-Vallve S. 2007. OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 35:W126-31.

Puigbò P, Bravo IG, Garcia-Vallve S. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). Submited.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The european molecular biology open software suite. Trends Genet 16(6):276-7.

Sharp PM and Li WH. 1987. The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Nucleic Acids Res 15(3):1281-95.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene 87(1):23-9.

Wu G, Culley DE, Zhang W. 2005. Predicted highly expressed genes in the genomes of streptomyces coelicolor and streptomyces avermitilis and the implications for their metabolism. Microbiology 151(Pt 7):2175-87.

Xia X. 2007. An improved implementation of codon adaptation index. Evolutionary Bioinformatics 3:53-8.

Zhao KN, Liu WJ, Frazer IH. 2003. Codon usage bias and A+T content variation in human papillomavirus genomes. Virus Res 98(2):95-104.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

6. Gaining and losing the thermophilic adaptation in prokaryotes. **Pere Puigbò**, Alberto Pasamontes, Santiago Garcia-Vallvé. *Trends in Genetics*. Accepted in press.

ABSTRACT

We have studied the evolution of thermophily in prokaryotes using the phylogenetic relationships between 279 bacteria and archaea and their thermophilic amino acid composition signature. Our findings suggest several examples in which the capacity of thermophilic adaptation has been gained or lost over relatively short evolutionary periods throughout the evolution of prokaryotes.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

AMINO ACID COMPOSITION SIGNATURE FOR THERMOPHILES

Since the first genome sequence of a prokaryotic organism was published in 1995, the number of completely sequenced genomes has grown exponentially. This fast accumulation of complete genomes enables genomes and proteomes to be compared. Differences in the amino acid composition between species were soon identified. Nucleotide bias and optimal growth temperature were shown to be the factors that most influence the differences in amino acid composition between organisms [1-4]. Recently we compared the amino acid composition of several groups of orthologous proteins from different species and showed that the differences in amino acid composition between thermophiles and mesophiles affect virtually all proteins within a proteome [5]. Because the first thermophilic and hyperthermophilic organisms to be sequenced were mainly archaea, the bias in amino acid composition observed in thermophiles and hyperthermophiles might have been related to their evolutionary relationships, and not an indication of their adaptation to high temperatures [4,6]. Recent analyses with more genomes have confirmed the initial finding that there is a relationship between amino acid composition and optimal growth temperature [7-10]. Together these findings enables us to define a thermophilic amino acid composition signature [7,10,11]. However, the basis for thermostability remains elusive and few general rules have been derived [10,12]. Comparisons between proteins from thermophiles and mesophiles using different datasets and methods usually show several discrepancies. This is probably because thermal stability is determined by a fine balance between several contributing factors (i.e. changes in surface charge distribution; helix dipole stabilization; packing and reduction in solvent-accessible hydrophobic surface; increased occurrences of hydrogen bonds, ion pairs, disulfide bridges or hydrophobic and aromatic interactions; the

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

contribution of specific chaperones; an increase in protein compactness or the decrease of polar and uncharged residues) and different strategies (structure-based or sequence-based [13]) might have been exploited by evolutionary distant organisms [12-14].

To determine whether the amino acid composition signature of thermophilic organisms is a general phenomenon, the result (or cause) of their thermal adaptation, and to study the evolution of thermophilic adaptation in prokaryotes, we used Correspondence analysis (CA) to analyse the mean amino acid composition of 279 prokaryotes (Figure 1). CA is used to reduce the dimensionality of an initial dataset by finding new variables or axes and project this dataset into a two-dimensional space with a minimum loss of information and maximum scattering. The greatest variation is shown on the first axis (CA1), and the other axes account for progressively less variation. The positions of the species on CA1 correlates (r=0.95) with their G+C content, which shows that the trend represented by CA1 is nucleotide bias. Variation along CA2 is due to the optimal growth temperature (r=-0.81). All hyperthermophiles and some thermophiles appear at the bottom of Figure 1, and show a thermophilic signature (i.e. they can be distinguished from the other species by their amino acid composition). The position of amino acids in this figure shows that hyperthermophiles use with a higher frequency glutamate (E) and with less frequency glutamine (Q). Some mesophiles, however, clearly cluster with hyperthermophiles and thermophiles: for example, *Methanococcus maripaludis* (mmp), several species of Clostridium (cac, ctc and cpe), *Fuseobacterium nucleatum* (fnu), and several species of Methanosarcina (mac, mba and mma). However, such bacterial thermophiles as *Chlorobium tepidum* (cte), *Geobacillus kaustophilus* (gka), *Methylococcus capsulatus (mca), Thermosynechococcus elongatus* (tel), *Thermobifida fusca* (tfu) and *Streptococcus thermophilus* (stc and stl) do not show the thermophilic signature and cluster with bacterial mesophiles. It could be argued that some thermophilic bacteria cluster with mesophilic bacteria in

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Figure 1 because thermophilic eubacteria and archaea have different mechanisms for the adaptation of proteins at high temperatures [12] and CA2 mainly reflects the thermophilic adaptation of archaea. However, the plot is similar if only bacteria are analyzed (see Figure S1 in the Online Supplementary Material). Our alternative interpretation of the amino acid differences between thermophilic species is that the thermophilic amino acid composition signature is an adaptation to living at high temperatures and reflects the time that has passed since the acquisition of thermophily.

POSITION OF THERMOPHILES IN THE TREE OF LIFE

Figure 2 shows the phylogenetic position of the thermophiles and hyperthermophiles analyzed in Figure 1. Hyperthermophiles and thermophiles are basically concentrated in three phylogenetic clusters: the Archaea, Clostridia, and the cluster containing Fusebacteria, Aquificae and Thermotogae. The cluster of *F. nucleatum* with *Aquifex aeolicus* and *Thermotoga maritima* is an undecided question [15], although it has been previously suggested [15,16]. Interestingly, mesophiles from these three clades have the thermophilic amino acid composition signature. These might suggest that the ancestors of each of these groups of organisms were thermophiles and adapted to living at high temperatures, although inferring characters, from sequence data, for ancient ancestors that existed long before is speculative and needs to be supported by stronger evidences. Many species from these three clades, including several mesophiles, which have the thermophilic signature, have a DNA-repair system specific for thermophilic archaea and bacteria (see Table S1 in the Online Supplementary Material), lending support to this hypothesis [17]. Thermophiles that do not have the thermophilic signature are scattered over the tree of life and belong to different taxonomic groups (Figure 2). Therefore, the ancestors of each of these taxonomic groups do not seem to have been thermophiles.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

THERMOPHILY HAS BEEN ACQUIRED AND LOST SEVERAL TIMES DURING PROKARYOTIC EVOLUTION

The combination of the analyses of the thermophilic amino acid signature and the position of thermophiles in the tree of life give important clues about the evolution of thermophily. Some of the possible evolutionary scenarios that include a case of acquisition or loss of the capability of living at high temperatures are summarized below. Mesophiles that are taxonomically related to a group of hyperthermophilic or thermophilic organisms generally have the thermophilic amino acid composition signature. This suggests a transversion to mesophily in these species [5]. The causes of this loss of thermophily are not known and might involve the loss of some of the elements (not fully understood) needed to live at high temperatures. These species retain the thermophilic amino acid signature, suggesting that the loss of thermophily is recent. Thermophiles that do not have the thermophilic signature (e.g. *C. tepidum* or *G. kaustophilus*) and are taxonomically related to mesophiles might be examples of recent transversions to thermophily. However, thermophiles (such as *Thermus thermophilus*), which have the thermophilic signature but are taxonomically related to mesophiles, might be examples of acquisition of thermophily [18], although in this case the acquisition might be ancient. To provide further evidence for our evolutionary hypotheses about the gain and lost of thermophily in relatively short evolutionary periods, we analyzed the characteristic patterns of synonymous codon usage in all the species of Figure 1. It has been shown that thermophilic prokaryotes have distinguishable patterns of synonymous codon usage [3,19]. These patterns include an increase in AGR codons for arginine and ATA codons for isoleucine, and a decrease in CGN codons for arginine [3,19]. The CA of the relative synonymous codon usage (RSCU) also differentiates hyperthermophiles and thermophiles from mesophiles on the CA2 axis (see Figure S2 in the online supplementary material). There are,

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

however, few mesophilic species with a synonymous codon-based thermophilic signature, perhaps because synonymous codon usage changes faster than amino acid composition does. Therefore, species that have recently lost their thermophilic capability still retain the amino acid composition signature but not the synonymous codon usage. More interestingly, the same thermophilic bacteria that cluster within mesophiles in Figure 1, cluster within mesophiles in the CA of the synonymous codon usage (see Figure S2 in the Online Supplementary Material). This evidence supports the hypothesis that these thermophiles have gained the thermophilic capability recently. There could be differential rates of gain and loss of thermophily. This rate difference probably reflects differences in selection intensity, i.e. a thermophile probably could survive at lower temperatures but a mesophile cannot survive at very high temperature.

Horizontal gene transfer (HGT) is an efficient way of acquiring new functionalities and capabilities [20]. There is a variety of evidence to show that HGT has had an important role in the adaptation of species to living at high temperatures [18,21-23]. Modifications at the proteomic level (adapting the amino acid composition) and nucleotide level (adapting the synonymous codon usage) might be required, especially for hyperthermophily. Comparative genomics and phyletic pattern analyses will be useful for identifying the genomic determinants of thermophily [17,24,25] and the role of HGT in the evolution of thermophily.

CONCLUDING REMARKS

The evolutionary scenario of thermophilic adaptation suggests that the amino acid composition signature in thermophilic organisms is a consequence of or an adaptation to living at high temperatures, not its cause. Our findings suggest that thermophilic adaptation at the level of protein composition is a

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

relatively labile character because it can be gained and lost several times over relatively short evolutionary periods.

ACKNOWLEDGEMENTS

We thank John Bates of the Language Service of the Rovira i Virgili University for his help with writing the manuscript.

FIGURES

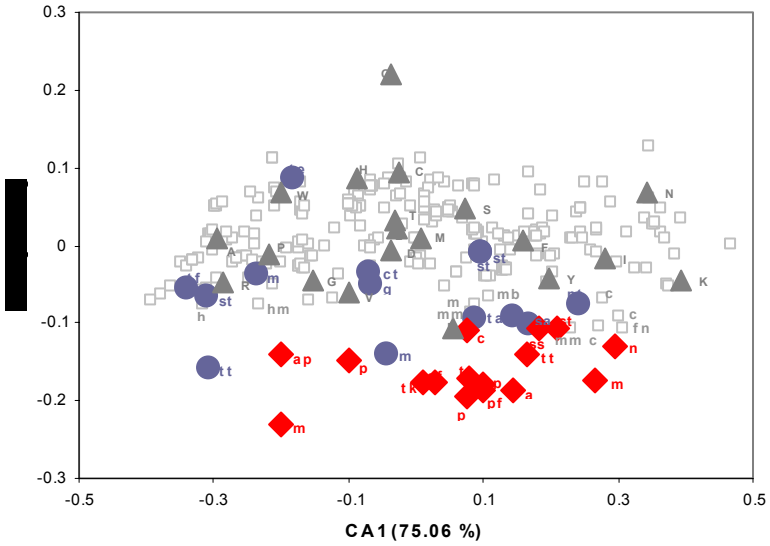


Figure 1. Correspondence analysis of the amino acid composition of the proteomes of the 279 species analyzed. Complete genomes were downloaded from the NCBI ftp site. Mean values of amino acid composition for each organism were analyzed using Correspondence analysis (CA). Mesophiles, thermophiles and hyperthermophiles are represented as white squares, blue dots and red diamonds, respectively. Grey triangles show the

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

loading scores of the amino acids. The CA1 and CA2 axes explain, respectively, 75.06% and 7.58% of the variability of the dataset. The positions of the organisms on these axes correlate with their G+C content (r= 0.95) and the optimal growth temperature (r= -0.81), respectively. All the hyperthermophiles and some thermophiles can be distinguished from mesophiles by their amino acid composition. However, some mesophiles (e.g. cac, ctc, cpe, fnu, mac, mba, mma, mmp) have the thermophilic amino acid signature and some thermophiles (e.g. cte, gka, mca, tel, tfu, stc, stl) cluster with mesophiles. The abbreviations used are (see also supplementary table 1): aae, *Aquifex aeolicus*; afu, *Archaeoglobus fulgidus*; ape, *Aeropyrum pernix*; cac, *Clostridium acetobutylicum*; chy, *Carboxydothermus hydrogenoformans*; cpe, *Clostridium perfringens*; ctc, *Clostridium tetani*; cte, *Chlorobium tepidum*; fnu, *Fusobacterium nucleatum*; gka, *Geobacillus kaustophilus*; hal, *Halobacterium sp*; hma, *Haloarcula marismortui*; mac, *Methanosarcina acetivorans*; mba, *Methanosarcina barkeri*; mca, *Methylococcus capsulatus*; mja, *Methanocaldococcus jannaschii*; mka, *Methanopyrus kandleri*; mma, *Methanosarcina mazei*; mmp, *Methanococcus maripaludis*; mth, *Methanothermobacter thermautotrophicus*; neq, *Nanoarchaeum equitans*; pab, *Pyrococcus abyssi*; pai, *Pyrobaculum aerophilum*; pfu, *Pyrococcus furiosus*; pho, *Pyrococcus horikoshii*; pto, *Picrophilus torridus*; sai, *Sulfolobus acidocaldarius*; sso, *Sulfolobus solfataricus*; stc, *Streptococcus thermophilus CNRZ1066*; sth, *Symbiobacterium thermophilum*; stl, *Streptococcus thermophilus LMG 18311*; sto, *Sulfolobus tokodaii*; tac, *Thermoplasma acidophilum*; tel, *Thermosynechococcus elongatus*; tfu, *Thermobifida fusca*; tko, *Thermococcus kodakarensis*; tma, *Thermotoga maritima*; tte, *Thermoanaerobacter tengcongensis*; tth, *Thermus thermophilus* HB27; tvo, *Thermoplasma volcanium*.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

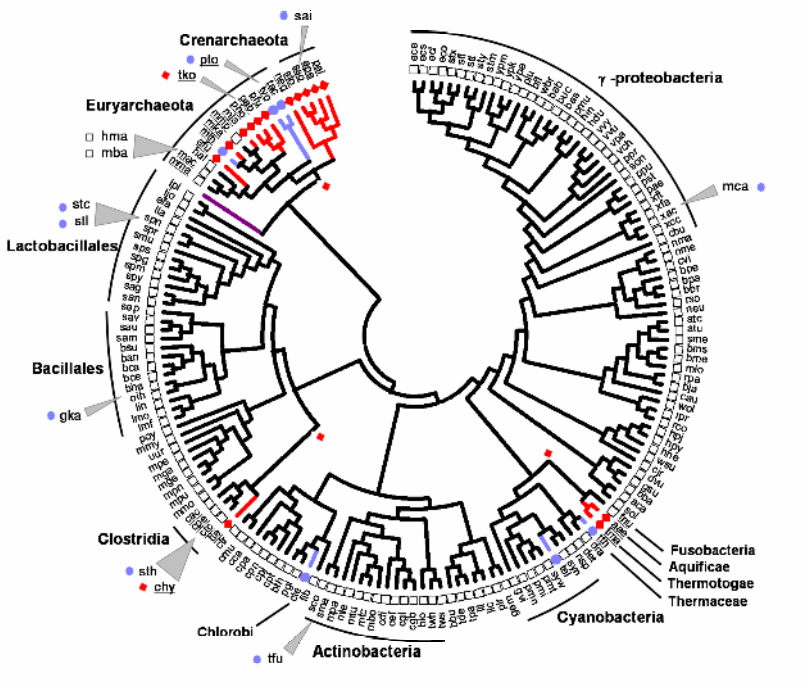


Figure 2. Phylogenetic position of the thermophiles and hyperthermophiles analyzed in figure 1. Based on the tree of life from Ciccarelli and coworkers [15]. Species not analyzed by Ciccarelli and coworkers [15] were included (i.e. the 31 proteins used by Ciccarelli and coworkers [15] were identified in these species, concatenated, aligned and analyzed) and their positions are shown in this figure. Branches containing eukaryotes were collapsed and are showed as a purple branch. The mesophiles that have the thermophilic signature in figure 1 (e.g. cac, ctc, cpe, fnu, mac, mba, mma, mmp) are archaea, clostridium species and *F. nucleatum* and are probably cases of loss of thermophily. The thermophiles that cluster with mesophiles in figure 1 (e.g. cte, gka, mca, tel, tfu, stc, stl) are the thermophiles that are phylogenetically related to a large group of mesophiles and are probably cases of recent gain of thermophily. *Thermus*

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

thermophilus (tth), which has the amino acid thermophilic signature but is phylogenetically related to the mesophile *D. radiodurans* (dra), is probably a case of an ancient gain of thermophily. See the legend to figure 1 for the abbreviations and symbols used.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

SUPPLEMENTARY DATA

Supplementary Table 1. List of species analyzed including their name, taxonomy, optimal growth temperature, position in the correspondence analysis and code used in this study. Hyperthermophilic (defined as organisms with an optimal growth temperature > 80ºC), Thermophilic (defined as organisms with an optimal growth temperature between 50 and 80ºC), Mesophilic (defined as organisms with an optimal growth temperature between 20 and 50ºC) and Psychrophilic (defined as organisms with an optimal growth temperature < 20ºC) organisms are represented as H, T, M and P respectively. Optimal growth temperature and range were extracted from the Genome Project Database at NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=genomepri).

Code	Organism	Range	Opt. Temp. (ºC)	Taxonomy		CA1	CA2
aae	*Aquifex aeolicus VF5*	H	96	Eubacteria	Aquificae	0.1430	-0.1874
aci	*Acinetobacter sp. ADP1*	M	37	Eubacteria	Proteobacteria;Gamm	0.0088	0.1108

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

					aproteobacteria		
afu	*Archaeoglobus fulgidus DSM 4304*	H	83	Archaeabacteria	Euryarchaeota;Archaeoglobi	0.0268	-0.1767
ama	*Anaplasma marginale str. St. Maries*			Eubacteria	Proteobacteria;Alphaproteobacteria	-0.0672	-0.0270
ape	*Aeropyrum pernix K1*	H	90-95	Archaeabacteria	Crenarchaeota	-0.2008	-0.1416
atc	*Agrobacterium tumefaciens str. C58*	M	25-28	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.1722	-0.0280
atu	*Agrobacterium tumefaciens str. C58*	M	25-28	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.1736	-0.0266
ava	*Anabaena variabilis ATCC 29413*			Eubacteria	Cyanobacteria	-0.0247	0.0868
bab	*Buchnera aphidicola str. Bp (Baizongia pistaciae)*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.3445	0.0608
ban	*Bacillus anthracis str. Ames*	M		Eubacteria	Firmicutes;Bacillales	0.1298	-0.0328
bar	*Bacillus anthracis str. 'Ames Ancestor'*	M		Eubacteria	Firmicutes;Bacillales	0.1298	-0.0328

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

bas	*Buchnera aphidicola str. Sg (Schizaphis graminum)*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.3627	0.0167
bat	*Bacillus anthracis str. Sterne*	M		Eubacteria	Firmicutes;Bacillales	0.1312	-0.0307
bba	*Bdellovibrio bacteriovorus HD100*	M	28-30	Eubacteria	Proteobacteria;Deltaproteobacteria	0.0024	0.0097
bbr	*Bordetella bronchiseptica RB50*	M	35-37	Eubacteria	Proteobacteria;Betaproteobacteria	-0.3064	0.0133
bbu	*Borrelia burgdorferi B31*	M		Eubacteria	Spirochaetes	0.3716	-0.0499
bca	*Bacillus cereus ATCC 10987*	M	25-35	Eubacteria	Firmicutes;Bacillales	0.1301	-0.0280
bce	*Bacillus cereus ATCC 14579*	M	25-35	Eubacteria	Firmicutes;Bacillales	0.1326	-0.0290
bcl	*Bacillus clausii KSM-K16*			Eubacteria	Firmicutes;Bacillales	0.0069	-0.0144
bcz	*Bacillus cereus E33L*	M	25-35	Eubacteria	Firmicutes;Bacillales	0.1336	-0.0309
bfl	*Candidatus Blochmannia floridanus*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.2508	0.0807
bfr	*Bacteroides fragilis YCH46*	M	37	Eubacteria	Bacteroidetes	0.0906	-0.0117
bfs	*Bacteroides fragilis NCTC 9343*	M	37	Eubacteria	Bacteroidetes	0.0886	-0.0115

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

bga	*Borrelia garinii PBi*	M		Eubacteria	Spirochaetes	0.3761	-0.0516
bha	*Bacillus halodurans C-125*	M		Eubacteria	Firmicutes;Bacillales	0.0227	-0.0235
bhe	*Bartonella henselae str. Houston-1*	M	37	Eubacteria	Proteobacteria;Alphaproteobacteria	0.0256	0.0305
bja	*Bradyrhizobium japonicum USDA 110*	M	25-30	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2278	-0.0222
bld	*Bacillus licheniformis ATCC 14580*	M		Eubacteria	Firmicutes;Bacillales	0.0598	-0.0343
bli	*Bacillus licheniformis ATCC 14580*	M		Eubacteria	Firmicutes;Bacillales	0.0598	-0.0341
blo	*Bifidobacterium longum NCC2705*	M	37-41	Eubacteria	Actinobacteria	-0.1602	0.0001
bma	*Burkholderia mallei ATCC 23344*	M		Eubacteria	Proteobacteria;Betaproteobacteria	-0.3206	-0.0160
bmb	*Brucella abortus biovar 1 str. 9-941*	M	37	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.1680	-0.0285
bme	*Brucella melitensis 16M*	M	37	Eubacteria	Proteobacteria;Alphapr	-0.1699	-0.0256

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

					oteobacteria		
bms	*Brucella suis 1330*	M	37	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.1651	-0.0288
bpa	*Bordetella parapertussis 12822*	M	35-37	Eubacteria	Proteobacteria;Betaproteobacteria	-0.3054	0.0147
bpe	*Bordetella pertussis Tohama I*	M	35-37	Eubacteria	Proteobacteria;Betaproteobacteria	-0.3009	0.0202
bpm	*Burkholderia pseudomallei 1710b*	M		Eubacteria	Proteobacteria;Betaproteobacteria	-0.3499	-0.0211
bpn	*Candidatus Blochmannia pennsylvanicus str. BPEN*			Eubacteria	Proteobacteria;Gammaproteobacteria	0.1866	0.0705
bps	*Burkholderia pseudomallei K96243*	M		Eubacteria	Proteobacteria;Betaproteobacteria	-0.3018	-0.0077
bqu	*Bartonella quintana str. Toulouse*	M	37	Eubacteria	Proteobacteria;Alphaproteobacteria	0.0147	0.0327
bsu	*Bacillus subtilis subsp. subtilis str. 168*	M	25-35	Eubacteria	Firmicutes;Bacillales	0.0759	-0.0166

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

bth	*Bacteroides thetaiotaomicron VPI-5482*	M		Eubacteria	Bacteroidetes	0.0964	-0.0054
btk	*Bacillus thuringiensis serovar konkukian str. 97-27*	M		Eubacteria	Firmicutes;Bacillales	0.1339	-0.0323
buc	*Buchnera aphidicola str. APS (Acyrthosiphon pisum)*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.3391	0.0344
bur	*Burkholderia sp. 383*			Eubacteria	Proteobacteria;Betaproteobacteria	-0.2801	0.0059
cab	*Chlamydophila abortus S26/3*	M	37	Eubacteria	Chlamydiae	0.0529	0.0488
cac	*Clostridium acetobutylicum ATCC 824*	M	10-65	Eubacteria	Firmicutes;Clostridia	0.2768	-0.0677
cbu	*Coxiella burnetii RSA 493*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0326	0.0361
cca	*Chlamydophila caviae GPIC*	M	37	Eubacteria	Chlamydiae	0.0557	0.0437
cch	*Chlorobium chlorochromatii CaD3*			Eubacteria	Chlorobi	-0.0197	0.0278
ccr	*Caulobacter crescentus CB15*	M	35	Eubacteria	Proteobacteria;Alphapr	-0.2943	-0.0418

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

					oteobacteria		
cdi	*Corynebacterium diphtheriae NCTC 13129*	M	37	Eubacteria	Actinobacteria	-0.1740	-0.0042
cef	*Corynebacterium efficiens YS-314*	M	30-45	Eubacteria	Actinobacteria	-0.2338	-0.0287
cgb	*Corynebacterium glutamicum ATCC 13032*	M	30-40	Eubacteria	Actinobacteria	-0.1563	-0.0154
cgl	*Corynebacterium glutamicum ATCC 13032*	M	30-40	Eubacteria	Actinobacteria	-0.1531	-0.0136
chy	*Carboxydothermus hydrogenoformans Z-2901*	H	78	Eubacteria	Firmicutes;Clostridia	0.0758	-0.1088
cje	*Campylobacter jejuni subsp. jejuni NCTC 11168*	M		Eubacteria	Proteobacteria;Epsilon proteobacteria	0.2608	-0.0061
cjk	*Corynebacterium jeikeium K411*	M		Eubacteria	Actinobacteria	-0.1999	-0.0216
cjr	*Campylobacter jejuni RM1221*	M		Eubacteria	Proteobacteria;Epsilon proteobacteria	0.2696	-0.0084

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

cmu	*Chlamydia muridarum Nigg*	M	37	Eubacteria	Chlamydiae	0.0290	0.0455
cpa	*Chlamydophila pneumoniae AR39*	M	37	Eubacteria	Chlamydiae	0.0501	0.0437
cpe	*Clostridium perfringens str. 13*	M	37	Eubacteria	Firmicutes;Clostridia	0.2706	-0.1054
cpj	*Chlamydophila pneumoniae J138*	M	37	Eubacteria	Chlamydiae	0.0480	0.0473
cpn	*Chlamydophila pneumoniae CWL029*	M	37	Eubacteria	Chlamydiae	0.0478	0.0477
cps	*Colwellia psychrerythraea 34H*	P	8	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0649	0.0748
cpt	*Chlamydophila pneumoniae TW-183*	M	37	Eubacteria	Chlamydiae	0.0481	0.0473
cta	*Chlamydia trachomatis A/HAR-13*	M		Eubacteria	Chlamydiae	0.0128	0.0480
ctc	*Clostridium tetani E88*	M	37	Eubacteria	Firmicutes;Clostridia	0.3024	-0.0903
cte	*Chlorobium tepidum TLS*	T	48	Eubacteria	Chlorobi	-0.0689	-0.0354

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

ctr	*Chlamydia trachomatis D/UW-3/CX*	M	37	Eubacteria	Chlamydiae	0.0138	0.0469
cvi	*Chromobacterium violaceum ATCC 12472*	M	25	Eubacteria	Proteobacteria;Betaproteobacteria	-0.2346	0.0367
dar	*Dechloromonas aromatica RCB*			Eubacteria	Proteobacteria;Betaproteobacteria	-0.1735	0.0043
dde	*Desulfovibrio desulfuricans G20*	M	25-40	Eubacteria	Proteobacteria;Deltaproteobacteria	-0.1900	-0.0037
deh	*Dehalococcoides sp. CBDB1*	M		Eubacteria	Chloroflexi	0.0006	-0.0165
det	*Dehalococcoides ethenogenes 195*	M	35	Eubacteria	Chloroflexi	-0.0076	-0.0208
dps	*Desulfotalea psychrophila LSv54*	P	7	Eubacteria	Proteobacteria;Deltaproteobacteria	-0.0024	0.0093
dra	*Deinococcus radiodurans R1*	M	30-37	Eubacteria	Deinococcus-Thermus	-0.3115	0.0030
dvu	*Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough*	M	25-40	Eubacteria	Proteobacteria;Deltaproteobacteria	-0.2486	-0.0423
eba	*Azoarcus sp. EbN1*	M	26	Eubacteria	Proteobacteria;Betapro	-0.2729	-0.0222

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

					teobacteria		
eca	*Erwinia carotovora subsp. atroseptica SCRI1043*	M	27-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0915	0.0681
ecc	*Escherichia coli CFT073*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0836	0.0627
ece	*Escherichia coli O157:H7 EDL933*	M	25-35	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0805	0.0514
ecn	*Ehrlichia canis str. Jake*			Eubacteria	Proteobacteria;Alphaproteobacteria	0.2379	0.0416
eco	*Escherichia coli K12*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0852	0.0528
ecs	*Escherichia coli O157:H7*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0806	0.0511
efa	*Enterococcus faecalis V583*	M	37	Eubacteria	Firmicutes;Lactobacillales	0.0984	0.0041
erg	*Ehrlichia ruminantium str. Gardel*			Eubacteria	Proteobacteria;Alphaproteobacteria	0.2443	0.0477

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

eru	*Ehrlichia ruminantium str. Welgevonden*			Eubacteria	Proteobacteria;Alphaproteobacteria	0.2449	0.0473
erw	*Ehrlichia ruminantium str. Welgevonden*			Eubacteria	Proteobacteria;Alphaproteobacteria	0.2436	0.0464
fnu	*Fusobacterium nucleatum subsp. nucleatum ATCC 25586*	M	37	Eubacteria	Fusobacteria	0.3071	-0.1079
ftu	*Francisella tularensis subsp. tularensis Schu 4*			Eubacteria	Proteobacteria;Gammaproteobacteria	0.2043	0.0376
gka	*Geobacillus kaustophilus HTA426*	T	60	Eubacteria	Firmicutes;Bacillales	-0.0659	-0.0508
gme	*Geobacter metallireducens GS-15*	M	30	Eubacteria	Proteobacteria;Deltaproteobacteria	-0.1227	-0.0674
gox	*Gluconobacter oxydans 621H*	M	25-30	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2397	0.0138
gsu	*Geobacter sulfurreducens PCA*	M	30	Eubacteria	Proteobacteria;Deltaproteobacteria	-0.1592	-0.0698
gvi	*Gloeobacter violaceus PCC*	M		Eubacteria	Cyanobacteria	-0.2389	0.0021

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	7421						
hal	*Halobacterium sp. NRC-1*	M	42	Archaeabacteria	Euryarchaeota;Halobacteria	-0.3158	-0.0756
hdu	*Haemophilus ducreyi 35000HP*	M	35-37	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0421	0.0818
hhe	*Helicobacter hepaticus ATCC 51449*	M	37	Eubacteria	Proteobacteria;Epsilonproteobacteria	0.1665	0.0388
hin	*Haemophilus influenzae Rd KW20*	M	35-37	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0482	0.0485
hit	*Haemophilus influenzae 86-028NP*	M	35-37	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0531	0.0495
hma	*Haloarcula marismortui ATCC 43049*	M	40-50	Archaeabacteria	Euryarchaeota;Halobacteria	-0.2336	-0.0748
hpj	*Helicobacter pylori J99*	M	37	Eubacteria	Proteobacteria;Epsilonproteobacteria	0.1841	0.0134
hpy	*Helicobacter pylori 26695*	M	37	Eubacteria	Proteobacteria;Epsilonproteobacteria	0.1865	0.0176

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

ilo	*Idiomarina loihiensis L2TR*	M	4-46	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0458	0.0470
lac	*Lactobacillus acidophilus NCFM*	M	25-35	Eubacteria	Firmicutes;Lactobacillales	0.1494	0.0076
lic	*Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130*	M	28-30	Eubacteria	Spirochaetes	0.1814	-0.0127
lil	*Leptospira interrogans serovar Lai str. 56601*	M	28-30	Eubacteria	Spirochaetes	0.1795	-0.0146
lin	*Listeria innocua Clip11262*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1098	-0.0388
ljo	*Lactobacillus johnsonii NCC 533*	M	25-35	Eubacteria	Firmicutes;Lactobacillales	0.1537	0.0115
lla	*Lactococcus lactis subsp. lactis Il1403*	M	40	Eubacteria	Firmicutes;Lactobacillales	0.1409	-0.0067
lmf	*Listeria monocytogenes str. 4b F2365*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1034	-0.0392
lmo	*Listeria monocytogenes EGD-e*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1047	-0.0392

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

lpf	*Legionella pneumophila str. Lens*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.0854	0.0785
lpl	*Lactobacillus plantarum WCFS1*	M	25-35	Eubacteria	Firmicutes;Lactobacillales	-0.0250	0.1054
lpn	*Legionella pneumophila subsp. pneumophila str. Philadelphia 1*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.0889	0.0777
lpp	*Legionella pneumophila str. Paris*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.0870	0.0771
lsa	*Lactobacillus sakei subsp. sakei 23K*	M		Eubacteria	Firmicutes;Lactobacillales	0.0217	0.0848
lxx	*Leifsonia xyli subsp. xyli str. CTCB07*	M	20-25	Eubacteria	Actinobacteria	-0.3112	-0.0549
mac	*Methanosarcina acetivorans C2A*	M	35-40	Archaeabacteria	Euryarchaeota;Methanomicrobia	0.0881	-0.0753
mba	*Methanosarcina barkeri str. fusaro*	M	35-40	Archaeabacteria	Euryarchaeota;Methanomicrobia	0.1089	-0.0659
mbo	*Mycobacterium bovis*	M	37	Eubacteria	Actinobacteria	-0.3303	-0.0193

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	AF2122/97						
mca	*Methylococcus capsulatus str. Bath*	T	45	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2352	-0.0374
mfl	*Mesoplasma florum L1*	M	20-40	Eubacteria	Firmicutes;Mollicutes	0.3253	-0.0306
mga	*Mycoplasma gallisepticum R*	M	37	Eubacteria	Firmicutes;Mollicutes	0.2882	0.0548
mge	*Mycoplasma genitalium G-37*	M	37	Eubacteria	Firmicutes;Mollicutes	0.2785	0.0808
mhj	*Mycoplasma hyopneumoniae J*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3532	0.0259
mhp	*Mycoplasma hyopneumoniae 7448*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3531	0.0277
mhy	*Mycoplasma hyopneumoniae 232*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3542	0.0305
mja	*Methanocaldococcus jannaschii DSM 2661*	H	85	Archaeabacteria	Euryarchaeota;Methanococci	0.2660	-0.1742
mka	*Methanopyrus kandleri AV19*	H	98	Archaeabacteria	Euryarchaeota;Methanopyri	-0.1996	-0.2306
mle	*Mycobacterium leprae TN*	M	37	Eubacteria	Actinobacteria	-0.2577	-0.0140

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

mlo	*Mesorhizobium loti MAFF303099*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2203	-0.0255
mma	*Methanosarcina mazei Go1*	M	30-40	Archaeabacteria	Euryarchaeota;Methanomicrobia	0.0827	-0.0863
mmo	*Mycoplasma mobile 163K*	M	20	Eubacteria	Firmicutes;Mollicutes	0.3839	-0.0117
mmp	*Methanococcus maripaludis S2*	M	35-40	Archaeabacteria	Euryarchaeota;Methanococci	0.2302	-0.1079
mmy	*Mycoplasma mycoides subsp. mycoides SC str. PG1*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3933	0.0342
mpa	*Mycobacterium avium subsp. paratuberculosis K-10*	M	37	Eubacteria	Actinobacteria	-0.3386	-0.0270
mpe	*Mycoplasma penetrans HF-2*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3690	0.0096
mpn	*Mycoplasma pneumoniae M129*	M	37	Eubacteria	Firmicutes;Mollicutes	0.1690	0.0940
mpu	*Mycoplasma pulmonis UAB CTIP*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3441	-0.0304
msu	*Mannheimia*	M	37	Eubacteria	Proteobacteria;Gamm	0.0310	0.0417

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	succiniciproducens MBEL55E				aproteobacteria		
msy	*Mycoplasma synoviae 53*	M	37	Eubacteria	Firmicutes;Mollicutes	0.3136	0.0175
mtc	*Mycobacterium tuberculosis CDC1551*	M	37	Eubacteria	Actinobacteria	-0.3322	-0.0188
mth	*Methanothermobacter thermautotrophicus str. Delta H*	T	65-70	Archaeabacteria	Euryarchaeota;Methanobacteria	-0.0417	-0.1399
mtu	*Mycobacterium tuberculosis H37Rv*	M	37	Eubacteria	Actinobacteria	-0.3307	-0.0206
neq	*Nanoarchaeum equitans Kin4-M*	H		Archaeabacteria	Nanoarchaeota	0.2939	-0.1303
neu	*Nitrosomonas europaea ATCC 19718*	M		Eubacteria	Proteobacteria;Betaproteobacteria	-0.0962	0.0380
nfa	*Nocardia farcinica IFM 10152*	M	37	Eubacteria	Actinobacteria	-0.3784	-0.0629
ngo	*Neisseria gonorrhoeae FA 1090*	M	35-37	Eubacteria	Proteobacteria;Betaproteobacteria	-0.0760	-0.0005
nma	*Neisseria meningitidis Z2491*	M	35-37	Eubacteria	Proteobacteria;Betaproteobacteria	-0.0615	0.0068

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

nme	*Neisseria meningitidis MC58*	M	35-37	Eubacteria	Proteobacteria;Betaproteobacteria	-0.0557	0.0086
noc	*Nitrosococcus oceani ATCC 19707*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1193	0.0219
nsp	*Nostoc sp. PCC 7120*	M		Eubacteria	Cyanobacteria	-0.0209	0.0865
nwi	*Nitrobacter winogradskyi Nb-255*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2322	-0.0304
oih	*Oceanobacillus iheyensis HTE831*	M	30	Eubacteria	Firmicutes;Bacillales	0.1270	-0.0070
pab	*Pyrococcus abyssi GE5*	H	103	Archaeabacteria	Euryarchaeota;Thermococci	0.0771	-0.1948
pac	*Propionibacterium acnes KPA171202*	M	37	Eubacteria	Actinobacteria	-0.2496	-0.0148
pae	*Pseudomonas aeruginosa PAO1*	M	25-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2551	0.0160
pai	*Pyrobaculum aerophilum str. IM2*	H	100	Archaeabacteria	Crenarchaeota	-0.0987	-0.1477

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

par	*Psychrobacter arcticus 273-4*	P	-2.5-20	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0123	0.0800
pca	*Pelobacter carbinolicus DSM 2380*	M		Eubacteria	Proteobacteria;Deltaproteobacteria	-0.1336	-0.0009
pcu	*Candidatus Protochlamydia amoebophila UWE25*	M		Eubacteria	Chlamydiae	0.1166	0.0841
pfl	*Pseudomonas fluorescens Pf-5*	M	25-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2094	0.0722
pfo	*Pseudomonas fluorescens PfO-1*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1708	0.0458
pfu	*Pyrococcus furiosus DSM 3638*	H	100	Archaeabacteria	Euryarchaeota;Thermococci	0.0990	-0.1874
pgi	*Porphyromonas gingivalis W83*	M	37	Eubacteria	Bacteroidetes	0.0036	-0.0238
pha	*Pseudoalteromonas haloplanktis TAC125*	P		Eubacteria	Proteobacteria;Gammaproteobacteria	0.0330	0.0871
pho	*Pyrococcus horikoshii OT3*	H	98	Archaeabacteria	Euryarchaeota;Thermococci	0.0986	-0.1780

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

plt	*Pelodictyon luteolum DSM 273*	M	25	Eubacteria	Chlorobi	-0.1128	-0.0378
plu	*Photorhabdus luminescens subsp. laumondii TTO1*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.0136	0.0652
pma	*Prochlorococcus marinus subsp. marinus str. CCMP1375*	M		Eubacteria	Cyanobacteria	0.0712	0.0339
pmi	*Prochlorococcus marinus str. MIT 9312*	M		Eubacteria	Cyanobacteria	0.2304	0.0088
pmm	*Prochlorococcus marinus subsp. pastoris str. CCMP1986*	M		Eubacteria	Cyanobacteria	0.2325	0.0087
pmn	*Prochlorococcus marinus str. NATL2A*	M		Eubacteria	Cyanobacteria	0.1136	0.0230
pmt	*Prochlorococcus marinus str. MIT 9313*	M		Eubacteria	Cyanobacteria	-0.1719	0.0812
pmu	*Pasteurella multocida subsp. multocida str. Pm70*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0180	0.0737
poy	*Onion yellows phytoplasma OY-M*	M		Eubacteria	Firmicutes;Mollicutes	0.3467	0.1278

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

ppr	*Photobacterium profundum SS9*	P	15	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0123	0.0626
ppu	*Pseudomonas putida KT2440*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2083	0.0495
psb	*Pseudomonas syringae pv. syringae B728a*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1694	0.0439
psp	*Pseudomonas syringae pv. phaseolicola 1448A*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1666	0.0451
pst	*Pseudomonas syringae pv. tomato str. DC3000*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1673	0.0438
pto	*Picrophilus torridus DSM 9790*	T	60	Archaeabacteria	Euryarchaeota;Thermoplasmata	0.2413	-0.0761
pub	*Candidatus Pelagibacter ubique HTCC1062*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	0.2985	-0.0227
rba	*Rhodopirellula baltica SH 1*	M	28	Eubacteria	Planctomycetes	-0.1677	0.0410
rco	*Rickettsia conorii str. Malish 7*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	0.2434	0.0164

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

reu	*Ralstonia eutropha JMP134*	M	30	Eubacteria	Proteobacteria;Betaproteobacteria	-0.2703	0.0087
rfe	*Rickettsia felis URRWXCal2*			Eubacteria	Proteobacteria;Alphaproteobacteria	0.2475	0.0094
rpa	*Rhodopseudomonas palustris CGA009*	M	25-35	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2404	-0.0196
rpr	*Rickettsia prowazekii str. Madrid E*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	0.2685	0.0196
rso	*Ralstonia solanacearum GMI1000*	M		Eubacteria	Proteobacteria;Betaproteobacteria	-0.2796	0.0160
rsp	*Rhodobacter sphaeroides 2.4.1*	M	25-35	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.3187	-0.0578
rty	*Rickettsia typhi str. Wilmington*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	0.2672	0.0209
sac	*Staphylococcus aureus subsp. aureus COL*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1766	0.0122
sag	*Streptococcus agalactiae*	M	37	Eubacteria	Firmicutes;Lactobacilla	0.1330	0.0110

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	2603V/R				les		
sai	*Sulfolobus acidocaldarius DSM 639*	T	70-75	Archaeabacteria	Crenarchaeota	0.1666	-0.1026
sak	*Streptococcus agalactiae A909*	M	37	Eubacteria	Firmicutes;Lactobacillales	0.1323	0.0127
sam	*Staphylococcus aureus subsp. aureus MW2*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1794	0.0151
san	*Streptococcus agalactiae NEM316*	M	37	Eubacteria	Firmicutes;Lactobacillales	0.1371	0.0095
sar	*Staphylococcus aureus subsp. aureus MRSA252*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1782	0.0187
sas	*Staphylococcus aureus subsp. aureus MSSA476*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1782	0.0113
sau	*Staphylococcus aureus subsp. aureus N315*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1795	0.0142
sav	*Staphylococcus aureus subsp. aureus Mu50*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1777	0.0178

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

sco	*Streptomyces coelicolor A3(2)*	M	25-35	Eubacteria	Actinobacteria	-0.3926	-0.0709
sec	*Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0979	0.0559
sep	*Staphylococcus epidermidis ATCC 12228*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1890	0.0218
ser	*Staphylococcus epidermidis RP62A*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1907	0.0162
sfl	*Shigella flexneri 2a str. 301*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0887	0.0485
sfx	*Shigella flexneri 2a str. 2457T*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0941	0.0515
sha	*Staphylococcus haemolyticus JCSC1435*	M	30-37	Eubacteria	Firmicutes;Bacillales	0.1804	0.0124
sil	*Silicibacter pomeroyi DSS-3*			Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2655	-0.0176
sma	*Streptomyces avermitilis MA-*	M	25-35	Eubacteria	Actinobacteria	-0.3622	-0.0525

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	4680						
sme	*Sinorhizobium meliloti 1021*	M	25-30	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.2131	-0.0502
smu	*Streptococcus mutans UA159*	M	37	Eubacteria	Firmicutes;Lactobacillales	0.1227	0.0137
son	*Shewanella oneidensis MR-1*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0415	0.0854
spa	*Streptococcus pyogenes MGAS10394*	M	35	Eubacteria	Firmicutes;Lactobacillales	0.0887	0.0135
spb	*Streptococcus pyogenes MGAS6180*	M	35	Eubacteria	Firmicutes;Lactobacillales	0.0935	0.0170
spg	*Streptococcus pyogenes MGAS315*	M	30-35	Eubacteria	Firmicutes;Lactobacillales	0.0903	0.0132
spm	*Streptococcus pyogenes MGAS8232*	M	30-35	Eubacteria	Firmicutes;Lactobacillales	0.0919	0.0196
spn	*Streptococcus pneumoniae TIGR4*	M	30-35	Eubacteria	Firmicutes;Lactobacillales	0.0899	-0.0067

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

spr	*Streptococcus pneumoniae R6*	M	30-35	Eubacteria	Firmicutes;Lactobacillales	0.0881	-0.0073
sps	*Streptococcus pyogenes SSI-1*	M	30-35	Eubacteria	Firmicutes;Lactobacillales	0.0912	0.0149
spt	*Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0991	0.0569
spy	*Streptococcus pyogenes M1 GAS*	M	30-35	Eubacteria	Firmicutes;Lactobacillales	0.0833	0.0206
spz	*Streptococcus pyogenes MGAS5005*	M	35	Eubacteria	Firmicutes;Lactobacillales	0.0871	0.0210
ssn	*Shigella sonnei Ss046*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0941	0.0499
sso	*Sulfolobus solfataricus P2*	H	85	Archaeabacteria	Crenarchaeota	0.1824	-0.1065
ssp	*Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305*	M		Eubacteria	Firmicutes;Bacillales	0.1617	0.0164

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

stc	*Streptococcus thermophilus CNRZ1066*	T	45	Eubacteria	Firmicutes;Lactobacillales	0.0977	-0.0090
sth	*Symbiobacterium thermophilum IAM 14863*	T	60	Eubacteria	Actinobacteria	-0.3094	-0.0666
stl	*Streptococcus thermophilus LMG 18311*	T	45	Eubacteria	Firmicutes;Lactobacillales	0.0962	-0.0088
stm	*Salmonella typhimurium LT2*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0986	0.0540
sto	*Sulfolobus tokodaii str. 7*	H	80	Archaeabacteria	Crenarchaeota	0.2102	-0.1057
stt	*Salmonella enterica subsp. enterica serovar Typhi Ty2*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0989	0.0536
sty	*Salmonella enterica subsp. enterica serovar Typhi str. CT18*	M	37	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0990	0.0526
syc	*Synechococcus elongatus PCC 6301*	M		Eubacteria	Cyanobacteria	-0.2130	0.1117
syd	*Synechococcus sp. CC9605*	M		Eubacteria	Cyanobacteria	-0.2325	0.0613

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

sye	*Synechococcus sp. CC9902*	M		Eubacteria	Cyanobacteria	-0.2069	0.0672
syf	*Synechococcus elongatus PCC 7942*	M		Eubacteria	Cyanobacteria	-0.2112	0.1115
syn	*Synechocystis sp. PCC 6803*	M		Eubacteria	Cyanobacteria	-0.0678	0.0843
syw	*Synechococcus sp. WH 8102*	M		Eubacteria	Cyanobacteria	-0.2449	0.0682
tac	*Thermoplasma acidophilum DSM 1728*	T	59	Archaeabacteria	Euryarchaeota;Thermoplasmata	0.0861	-0.0901
tbd	*Thiobacillus denitrificans ATCC 25259*	M	28-32	Eubacteria	Proteobacteria;Betaproteobacteria	-0.2646	-0.0164
tcx	*Thiomicrospira crunogena XCL-2*	M	28-32	Eubacteria	Proteobacteria;Gammaproteobacteria	0.0135	0.0550
tde	*Treponema denticola ATCC 35405*	M	30-42	Eubacteria	Spirochaetes	0.1929	-0.0472
tdn	*Thiomicrospira denitrificans ATCC 33889*	M	20-25	Eubacteria	Proteobacteria;Epsilonproteobacteria	0.2028	-0.0351
tel	*Thermosynechococcus elongatus BP-1*	T	55	Eubacteria	Cyanobacteria	-0.1822	0.0869

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

tfu	*Thermobifida fusca*	T	50-55	Eubacteria	Actinobacteria	-0.3400	-0.0542
tko	*Thermococcus kodakarensis KOD1*	H	85	Archaeabacteria	Euryarchaeota;Thermococci	0.0104	-0.1763
tma	*Thermotoga maritima MSB8*	H	80	Eubacteria	Thermotogae	0.0799	-0.1704
tpa	*Treponema pallidum subsp. pallidum str. Nichols*	M		Eubacteria	Spirochaetes	-0.1675	0.0058
tte	*Thermoanaerobacter tengcongensis MB4*	H	75	Eubacteria	Firmicutes;Clostridia	0.1633	-0.1411
tth	*Thermus thermophilus HB27*	T	68	Eubacteria	Deinococcus-Thermus	-0.3074	-0.1580
tvo	*Thermoplasma volcanium GSS1*	T	60	Archaeabacteria	Euryarchaeota;Thermoplasmata	0.1439	-0.0925
twh	*Tropheryma whipplei str. Twist*	M	37	Eubacteria	Actinobacteria	-0.0676	-0.0017
tws	*Tropheryma whipplei TW08/27*	M	37	Eubacteria	Actinobacteria	-0.0657	-0.0043
uur	*Ureaplasma parvum serovar 3 str. ATCC 700970*	M		Eubacteria	Firmicutes;Mollicutes	0.3634	0.0471
vch	*Vibrio cholerae O1 biovar eltor*	M	20-30	Eubacteria	Proteobacteria;Gamm	-0.0473	0.0790

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	str. N16961				aproteobacteria		
vfi	*Vibrio fischeri ES114*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.0486	0.0441
vpa	*Vibrio parahaemolyticus RIMD 2210633*	M	20-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0074	0.0512
vvu	*Vibrio vulnificus CMCP6*	M	20-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0293	0.0620
vvy	*Vibrio vulnificus YJ016*	M	20-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0227	0.0661
wbm	*Wolbachia endosymbiont strain TRS of Brugia malayi*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	0.2077	-0.0109
wbr	*Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	0.4670	0.0021
wol	*Wolbachia endosymbiont of Drosophila melanogaster*	M		Eubacteria	Proteobacteria;Alphaproteobacteria	0.1996	-0.0153
wsu	*Wolinella succinogenes DSM*	M		Eubacteria	Proteobacteria;Epsilon	0.0422	-0.0522

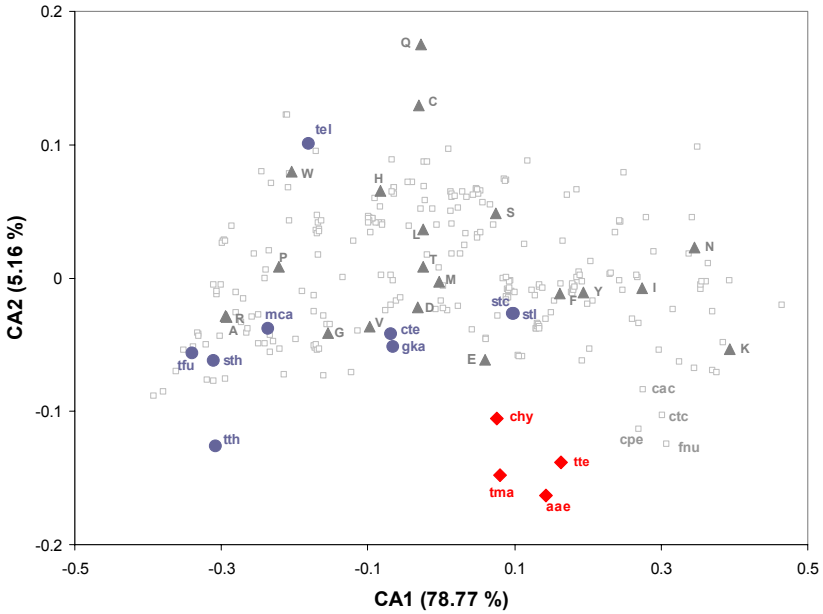
UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

	1740				proteobacteria		
xac	*Xanthomonas axonopodis pv. citri str. 306*	M	25-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2967	0.0500
xcb	*Xanthomonas campestris pv. campestris str. 8004*	M	25-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2977	0.0512
xcc	*Xanthomonas campestris pv. campestris str. ATCC 33913*	M	25-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2992	0.0499
xcv	*Xanthomonas campestris pv. vesicatoria str. 85-10*	M	25-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2951	0.0516
xfa	*Xylella fastidiosa 9a5c*	M	26-28	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1701	0.0546
xft	*Xylella fastidiosa Temecula1*	M	26-28	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.1643	0.0509
xoo	*Xanthomonas oryzae pv. oryzae KACC10331*	M		Eubacteria	Proteobacteria;Gammaproteobacteria	-0.2860	0.0547
ype	*Yersinia pestis CO92*	M	28-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0655	0.0714

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

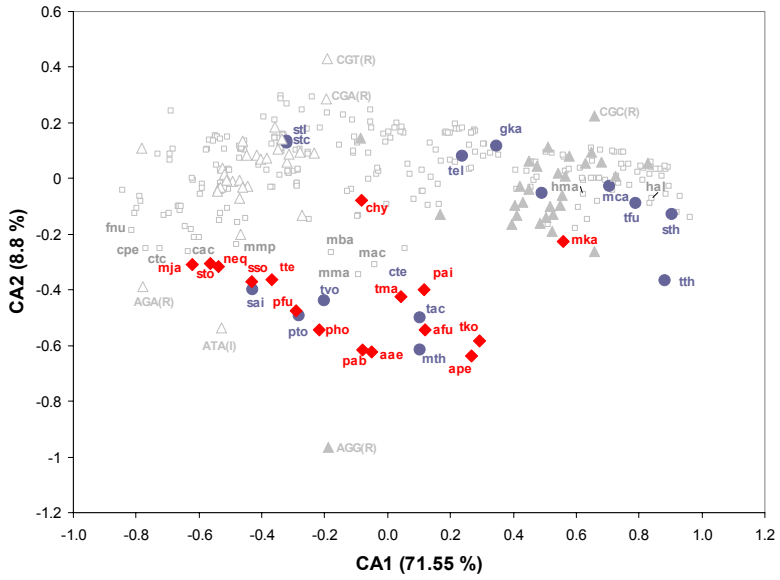
ypk	*Yersinia pestis KIM*	M	28-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0652	0.0756
ypm	*Yersinia pestis biovar Medievalis str. 91001*	M	28-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0641	0.0758
yps	*Yersinia pseudotuberculosis IP 32953*	M	28-30	Eubacteria	Proteobacteria;Gammaproteobacteria	-0.0677	0.0753
zmo	*Zymomonas mobilis subsp. mobilis ZM4*	M	25-30	Eubacteria	Proteobacteria;Alphaproteobacteria	-0.0824	0.0244

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007



Supplementary figure 1. Correspondence analysis of the amino acid composition of the proteomes analyzed in figure 1, excluding archaeal species. See the legend to figure 1 for the abbreviations and symbols used.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007



Supplementary figure 2. Correspondence analysis of the synonymous codon usage, measured with the RSCU values, of all species analyzed in figure 1. Grey and white triangles show the loading scores of G or C and A or T-ending codons, respectively. See the legend to figure 1 for the abbreviations and other symbols used.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

1 Kreil, D.P. and Ouzounis, C.A. (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* 29, 1608-1615

2 Tekaia, F. *et al*. (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 297, 51-60

3 Singer, G.A. and Hickey, D.A. (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317, 39-47

4 Hickey, D.A. and Singer, G.A. (2004) Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.* 5, 117

5 Pasamontes, A. and Garcia-Vallve, S. (2006) Use of a multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes. *BMC Bioinformatics* 7, 257

6 Vieille, C. and Zeikus, G.J. (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1-43

7 Suhre, K. and Claverie, J.M. (2003) Genomic correlates of hyperthermostability, an update. *J. Biol. Chem.* 278, 17198-17202

8 Takami, H. *et al*. (2004) Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. *Nucleic Acids Res.* 32, 6292-6303

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

9 Tekaia, F. and Yeramian, E. (2006) Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genomics* 7, 307

10 Zeldovich, K.B. *et al*. (2007) Protein and DNA Sequence Determinants of Thermophilic Adaptation. *PLoS Comput. Biol.* 3, e5

11 Pe'er, I. *et al*. (2004) Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins* 54, 20-40

12 Mizuguchi, K. *et al*. (2007) Environment specific substitution tables for thermophilic proteins. *BMC Bioinformatics* 8 Suppl 1, S15

13 Berezovsky, I.N. and Shakhnovich, E.I. (2005) Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 102, 12742-12747

14 Sadeghi, M. *et al*. (2006) Effective factors in thermostability of thermophilic proteins. *Biophys. Chem.* 119, 256-270

15 Ciccarelli, F.D. *et al*. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283-1287

16 Brochier, C. and Philippe, H. (2002) Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* 417, 244

17 Makarova, K.S. *et al*. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.*

30, 482-496 18 Omelchenko, M.V. *et al*. (2005) Comparative genomics of *Thermus thermophilus* and *Deinococcus radiodurans*: divergent routes of adaptation to thermophily and

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

radiation resistance. *BMC Evol. Biol.* 5, 57

19 De Farias, S.T. and Bonato, M.C. (2003) Preferred amino acids and thermostability. *Genet. Mol. Res.* 2, 383-393

20 Garcia-Vallve, S. *et al*. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* 10, 1719-1725

21 Boucher, Y. *et al*. (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu. Rev. Genet.* 37, 283-328

22 Schwarzenlander, C. and Averhoff, B. (2006) Characterization of DNA transport in the thermophilic bacterium *Thermus thermophilus* HB27. *FEBS J.* 273, 4210- 4218

23 Brochier-Armanet, C. and Forterre, P. (2006) Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea* 2, 83-93

24 Makarova, K.S. *et al*. (2003) Potential genomic determinants of hyperthermophily. *Trends Genet.* 19, 172-176

25 Bruggemann, H. and Chen, C. (2006) Comparative genomics of *Thermus thermophilus*: Plasticity of the megaplasmid and its contribution to a thermophilic lifestyle. *J. Biotechnol.* 124, 654-661

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

ISBN: 978-84-691-0653-2 /DL: T.2224-2007

7. TOPD/FMTS: a new software to compare phylogenetic trees. **Pere Puigbò**, Santiago Garcia-Vallvé and James O. McInerney. *Bioinformatics*, 2007. 23(12):1556-1558

ABSTRACT

Summary: TOPD/FMTS has been developed to evaluate similarities and differences between phylogenetic trees. The software implements several new algorithms (including the Disagree method that returns the taxa that disagree between two trees and the Nodal method that compares two trees using nodal information) and several previously described methods (such as the Partition method, Triplets or Quartets) to compare phylogenetic trees. One of the novelties of this software is that the FMTS program allows the comparison of trees that contain both orthologs and paralogues. Each option is also complemented with a randomisation analysis to test the null hypothesis that the similarity between two trees is not better than chance expectation.

Availability: The Perl source code of TOPD/FMTS is available at http://genomes.urv.es/topd.

Supplementary Information: A complete tutorial and several examples of how to use the software have been included on the home page of the application

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Chapter 7

INTRODUCTION

Phylogenetic trees have often been compared in molecular evolution studies because different sets of putatively orthologous genes often yield strongly supported but incompatible tree topologies (Beiko and Hamilton, 2006). Incongruence in tree topologies can be explained by such processes as horizontal gene transfer events (Garcia-Vallve et al., 2003; Creevey et al., 2004), hidden paralogy (Creevey et al., 2004) and model misspecification (Rokas et al., 2003). Most archaeal and bacterial genomes contain genes from multiple sources (Doolittle, 1999) and each phylogenetic tree constructed from a protein family reflects the evolutionary history of its sequences. There are also many methods of constructing phylogenetic trees (e.g. Distance, Parsimony or Likelihood), which can produce different trees. Given this situation, it is desirable to compare phylogenetic trees from a set of sequences constructed by different methods and/or to compare phylogenetic trees from different sets of homologs.

Although many methods for comparing phylogenetic trees have been described, for example, Nearest-neighbor interchanging (Waterman and Smith, 1978), Subtree transfer distance (Allen and Steel, 2001), Quartets (Estabrook et al., 1985), Partition or Symmetric difference metrics (Robinson and Foulds, 1981) and Path length metrics (Steel and Penny, 1993), very few have been implemented for their use in a program and there is no program with a comprehensive set of implemented methods. For this reason, we have developed the TOPD/FMTS software. TOPD/FMTS compares phylogenetic trees using some of the above methods, but also implements new algorithms.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

This means a sensitivity analysis can be carried out on any set of results to evaluate methodological properties and biases. TOPD/FMTS combines two programs: 1) the TOPD (TOPological Distance) program, which compares two trees with the same taxa or two pruned trees and 2) the FMTS (From Multiple To Single) program, which converts multigene family trees to singlegene family trees. The FMTS program is activated automatically only if one or both trees to be compared are multigene family trees, so both programs can work together depending on input data structure. Additionally, each option of this software is complemented with a randomisation analysis to test the null hypothesis that the similarity between two trees is not better than chance.

PROGRAM OVERVIEW

Inputs and outputs

The software minimally requires a file containing two trees in PHYLIP format to calculate a distance between them. Alternatively, a file containing a list of trees can be provided in order to calculate the differences between all of them or to compare them with a reference tree. The parameter '-f' followed by the name of the input file is the only mandatory parameter required to run the program. Other parameters can be modified according to the user's requirements (use '–h' to see the complete list of parameters). This software can compare trees with leafsets that either completely or partially overlap. If trees only partially overlap they are pruned to their common leafset in order to compare their topologies. The input trees can be rooted or unrooted. If a rooted tree is input, it will be automatically unrooted. Some results are printed in the standard output, by default, but can be easily redirected into an output

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

file using terminal commands. The final results (i.e. the values of the comparison and the percentage of overlapping taxa) are printed in an output file.

TOPD

The TOPD program compares trees using several methods, which are called Nodal, Split, Quartets, Triplets, and Disagree. The Split or Partition metrics (Robinson and Foulds, 1981) and quartets and triplets (Estabrook et al., 1985) have been described and implemented previously but this software offers additional possibilities such as the comparison of multigene family trees, the comparison of partially overlapping trees and randomisation tests. The Nodal method uses the path length metrics described by Steel and Penny (1993). The Disagree method uses a novel algorithm described and implemented in this software and is the opposite of the methods that find the Most Agreement Subtree. The Agreement method described by Kubicka et al. (1995) finds the single greatest agreement subtree when two trees are compared, while our disagree method finds the taxa that produce disagreement between two trees.

The Nodal method constructs pairwise distance matrices from the two input trees using only the leaves that are common to both trees. This is done by comparing the number of nodes that separate each taxon from the other taxa in the tree. If the two trees do not have the same taxa, but have overlapping leafsets, the trees are appropriately pruned so they can be compared. Then the differences between the two matrices are calculated to obtain the distance between the two trees. The nodal distance score is calculated using the rootmeansquared distance (RMSD) of these two matrices. The RMSD is 0 for

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

identical trees, and increases as the two trees become more dissimilar. In those cases where two leafsets are overlapping but not identical, we have added another score that considers the percentage of taxa that the two trees have in common. This second score is equal to the RMSD if both taxonsets are the same and becomes proportionally greater when this percentage is reduced, i.e. this score is 0 if two trees are equal and increase depending of two factors: the dissimilarity between the trees and the number of overlapping leaves (see the equation in http://genomes.urv.es/topd/nodal_e.html).

The Disagree method compares two trees and returns the taxa whose phylogenetic position disagrees in these trees. Penny and Hendy (1985) used the term “gain” to describe the reduction in the difference when two trees are compared after any taxon is removed. Our Disagree method uses an iterative algorithm and can work at four levels of comparison. The computational time needed at each level increases. The method works at level 1 by removing one taxon every time and calculating the gain (reduction in the split distance) between the two trees. The taxon that produces most gain is removed for the following iterations. This procedure is repeated until the split distance is zero (see the algorithm in http://genomes.urv.cat/topd/disagree.html). We have used this algorithm in a thousand comparisons of trees obtained from known protein families. At level 1, approximately 80% of the comparisons can be solved (i.e. the split distance becomes 0 after removing one taxon or set of taxa). The second, third and fourth levels remove 2, 3 and 4 taxa every time, respectively, and then calculate the gain. When a solution exists, every level solves the comparisons between trees that cannot be solved in the previous level.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Chapter 7

FMTS

The FMTS program can be used to compare two trees, one or both of which are multigene family trees. Until now, trees that contain more than one gene copy per genome could not be compared automatically using any software. The TOPD/FMTS program makes it possible by evaluating each gene copy independently. The FMTS program systematically prunes each gene copy from the multifamily tree to obtain all possible singlegene trees. The result is a set of singlegene family trees. Each tree can be then compared with TOPD, using any of the previously described methods and the result is the mean and standard deviation of all comparisons. In its standard output, the program provides the result of all comparisons and a text file of all of the pruned singlegene family trees. The use of the FMTS program may be computationally expensive when the number of singlegene family trees obtained from a multigene family tree is enormous. To overcome this limitation, the FMTS program allows the option of randomly pruning the multigene tree by default 100 times. Users, however, can modify this number.

The set of single gene trees obtained with FMTS would contain a mixture of orthologs and paralogues. Those trees can be checked individually, using the TOPD program and a reference species tree, to help to define orthologs and paralogues, or identify horizontally transferred genes. The identification of true orthologs is essential for studying the speciation process. On the other hand, the analysis of paralogues helps to understand the evolution by gene duplication, which is a major force in creating new functionalities (Jordan et al., 2001; Lynch and Conery 2000). Another method capable of dealing with paralogy is the reconciled trees method (Cotton and Page, 2002). But this

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

method tries to infer gene duplication events and estimate species phylogenies, while the FMTS algorithm is helpful to study phylogenies of protein families that contain orthologs and paralogues through the tree comparison with the program TOPD.

Randomisation analysis

This software implements two randomisation methods that evaluate whether the similarity between two trees is better than random. In the first method (Guided), all taxa are removed from the tree and randomly reassigned while maintaining the topology of the original tree. This means that the positions of the taxa have been randomly changed. The second method (Random), generates random trees, by a markovian method, with the same taxa as the original tree but randomly changes the topology of the tree and consequently, the relationships of the taxa. A similar method is used in the Clann program (Creevey and McInerney, 2005). Then a comparison between random trees is calculated using any of the methods allowed by the software. This is repeated as many times as the user requires. By default the program carries out this random analysis 100 times and the result is the mean and standard deviation of the different repetitions. A criticalpoint can be used to evaluate whether the similarity between two trees is better than random.

ACKNOWLEDGEMENTS

We thank John Bates of the Language Service of the Rovira i Virgili University for his help with writing the manuscript and members of the Bioinformatics Laboratory for discussions. This work has been financed by the project BIO200307672 from the MCT of the Spanish Government.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

REFERENCES

Allen,L. and Steel,M. (2001) Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics, 5, 1-15.

Beiko,R.G. and Hamilton,N. (2006) Phylogenetic identification of lateral gene transfer events. BMC Evolutionary Biology, 6, 15.

Cotton,J.A., Page,R.D (2002) Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. Lond. B, 269, 1555-1561.

Creevey,C.J., Fitzpatrick,D.A., Philip,G.K., Kinsella,R.J., O'Connell,M.J., Pentony,M.M., Travers,S.A., Wilkinson,M. and McInerney,J.O. (2004) Does a treelike phylogeny only exist at the tips in the prokaryotes? Proc. R. Soc. Lond. B. Biol. Sci., 271, 2551–2558.

Creevey,C.J. and McInerney, J.O. (2005) Clann: Investigating phylogenetic information through supertree analyses. Bioinformatic, 21: 390-392.

Doolittle,W.F. (1999) Phylogenetic Classifcation and the Universal Tree. Sience, 284, 2124-2128.

Estabrook,G.F., McMorris,F.R. and Meachan,A. (1985) Comparison of Undirected Phylogenetic Trees Based on Subtree of Four Evolutionary Units. Syst. Zool, 34, 193-200.

GarciaVallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGTDB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res., 31, 187-189.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. and Koonin, E.V. (2001) Lineagespecific gene expansions in bacterial and archaeal genomes. Genome Research, 11,555-565.

Kubicka,E., Kubicki,G. and McMorris,F.R. (1995) An algorithm to find agreement subtrees. Journal of Classification, 12, 91-99.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. Science, 290,1151-1155.

Penny,D. and Hendy,M.D. (1985) The Use of Tree Metrics. Syst. Zool., 34:75-82.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. Math. Biosci., 53, 131-147.

Rokas, A., Williams B. L., King N., and Carroll S.B. (2003) Genomescale approaches to resolving incongruence in molecular phylogenies. Nature, 425,798-803

Steel,M.A. and Penny,D. (1993) Distribution of tree comparison metrics some new results. Systematic Biology, 42, 126-141.

Waterman,M.S. and Steel,M. (1978) On the similarity of dendrograms. Journal of Theoretical Biology, 73, 789-800.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

CONCLUSIONS

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

Conclusions

1. Using the correspondence analysis of the relative synonymous codon usage of all genes we have developed a new and automatic method for detecting whether a genome is under translational selection. When applied to a set of prokaryotic complete genomes, approximately 45% of the species analyzed were predicted to be under translational selection. In these genomes, the group of highly expressed genes forms a cluster in the correspondence analysis plot because they have a different codon usage from the other genes of a genome.

2. We have developed a new iterative algorithm which predicts a group of highly expressed genes in genomes under translational selection by using the Codon Adaptation Index and the group of ribosomal protein genes as a seed. The highly expressed genes that we predict are not a random group of genes. They are metabolic genes with a putative function, which are located preferably in the leading strand. The functional analysis of these genes shows, as expected, that ribosomal protein genes and genes involved in translation, transcription, energy metabolism and the metabolism of biomolecules are found in the final group of predicted highly expressed genes. The predicted highly expressed genes are used as an initial filter to reduce the number of false positives of the Horizontal Gene Transfer Database (HGT-DB, http://genomes.urv.es/HGT-DB/) maintained in our group.

3. We have identified a group of 184 highly expressed genes, with a characteristic codon usage, conserved among all species under a strong translational selection. These genes define the universal steps of metabolic pathways essential for the life of bacteria in a competitive medium. We have also identified the common highly expressed genes for

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

12 taxonomic groups. There is a good correlation between the definition of highly expressed genes in each particular group of species and their main metabolic capabilities.

4. We have developed a new genomic database, called HEG-DB, which can predict which genes are highly expressed in prokaryotic complete genomes under strong translational selection. The database is freely available at http://genomes.urv.cat/HEG-DB.

5. We have developed a new web sever, called OPTIMIZER, to optimize the codon usage of DNA or RNA sequences. This new web server can be used to predict and optimize the level expression of a gene in heterologous gene expression or to express new genes that confer new metabolic capabilities in a given species. It has unique features, such as a novel definition of a group of highly expressed genes from over 150 prokaryotic species under translational selection, a new method designed to maximize the optimization with the fewest changes in the query sequence, and the possibility of using information on tRNA gene copy numbers in the optimization process. This web server is freely available at http://genomes.urv.cat/OPTIMIZER.

6. We have developed an expected value of CAI (eCAI) to find out whether the differences in the CAI are statistically significant or whether they are the product of biased nucleotide and/or amino acid composition.

7. The use of the eCAI has shown that nuclear-encoded mitochondrial genes from humans are better adapted to nuclear codon usage than to mitochondrial codon usage. These genes were originally encoded in the

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

proto-mitochondria but are now encoded in the human nuclear genome. This means that the codon usage of these genes has been ameliorated and adapted to the human codon usage.

8. We have developed a new web server, called CAIcal, with a complete set of tools related to the CAI that contain important features, such as showing the expected CAI value and calculating and representing graphically the CAI along a sequence or a protein multialignment translated to DNA. This web server is freely available at http://genomes.urv.cat/CAIcal.

9. We have developed a new software program, called TOPD/FMTS, to evaluate the similarities and differences between phylogenetic trees. The software uses several new algorithms and previously described methods to compare phylogenetic trees. One of the new features of this software is that the FMTS program can compare trees that contain both orthologs and paralogues. This program is freely available at http://genomes.urv.cat/topd.

10. The evolution of thermophilic adaptation suggests that the amino acid composition signature in thermophilic organisms is a consequence of or an adaptation to living at high temperatures, not its cause. Our findings suggest that there have been several cases where the capacity for thermophilic adaptation has been gained or lost throughout the evolution of prokaryotes.

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007

UNIVERSITAT ROVIRA I VIRGILI
CODON USAGE ADAPTATION IN PROKARYOTIC GENOMES
Pere Puigbò Avalós
ISBN: 978-84-691-0653-2 /DL: T.2224-2007