*D524–D527* *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
*doi:10.1093/nar/gkm831*

# HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection

**Pere Puigbò*, Antoni Romeu and Santiago Garcia-Vallvé**

Evolutionary Genomics Group, Biochemistry and Biotechnology Department, Faculty of Chemistry, Rovira i Virgili University (URV), c/Marcel-li Domingo, s/n. Campus Sescelades, 43007 Tarragona, Spain

Received July 24, 2007; Revised and Accepted September 21, 2007

## 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 (CAI) 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.**

## INTRODUCTION

Translational selection is the force that modulates the codon usage bias of a group of highly expressed genes in some genomes. This modulation consists in the adaptation of the synonymous codon usage to the most abundant tRNA species to increase traductional efficiency and accuracy (1,2). 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. However, not all the species are under translational selection (3). 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 (2,4). One of the currently accepted views is that genome-wide codon bias is determined primarily by mutational processes and only secondarily by translational selection (5). The Codon Adaptation Index (CAI), developed by Sharp and Li (6) is the index that is most commonly used, by itself (7,8) or in combination with an iterative algorithm (9,10), to predict highly expressed genes using 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 (11–14). 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 (15). 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 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 (16,17).

*To whom correspondence should be addressed. Tel: +34 977558778; Fax: +34 977558232; Email: ppuigbo@urv.cat

## 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 (10). 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 (18). 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 (10). The ribosomal proteins were obtained by keyword searching using the annotation information provided by the NCBI. 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 (10). The number of iterations is not constant and depends on the genome analyzed. 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 co-workers (9) 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). 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 (10). 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 G + C content at the third codon position (20). 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 (10). 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 (15). Experimental microarray experiments have shown that, even in species under translational selection, genes without a biased codon usage can be highly expressed (8,21). 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

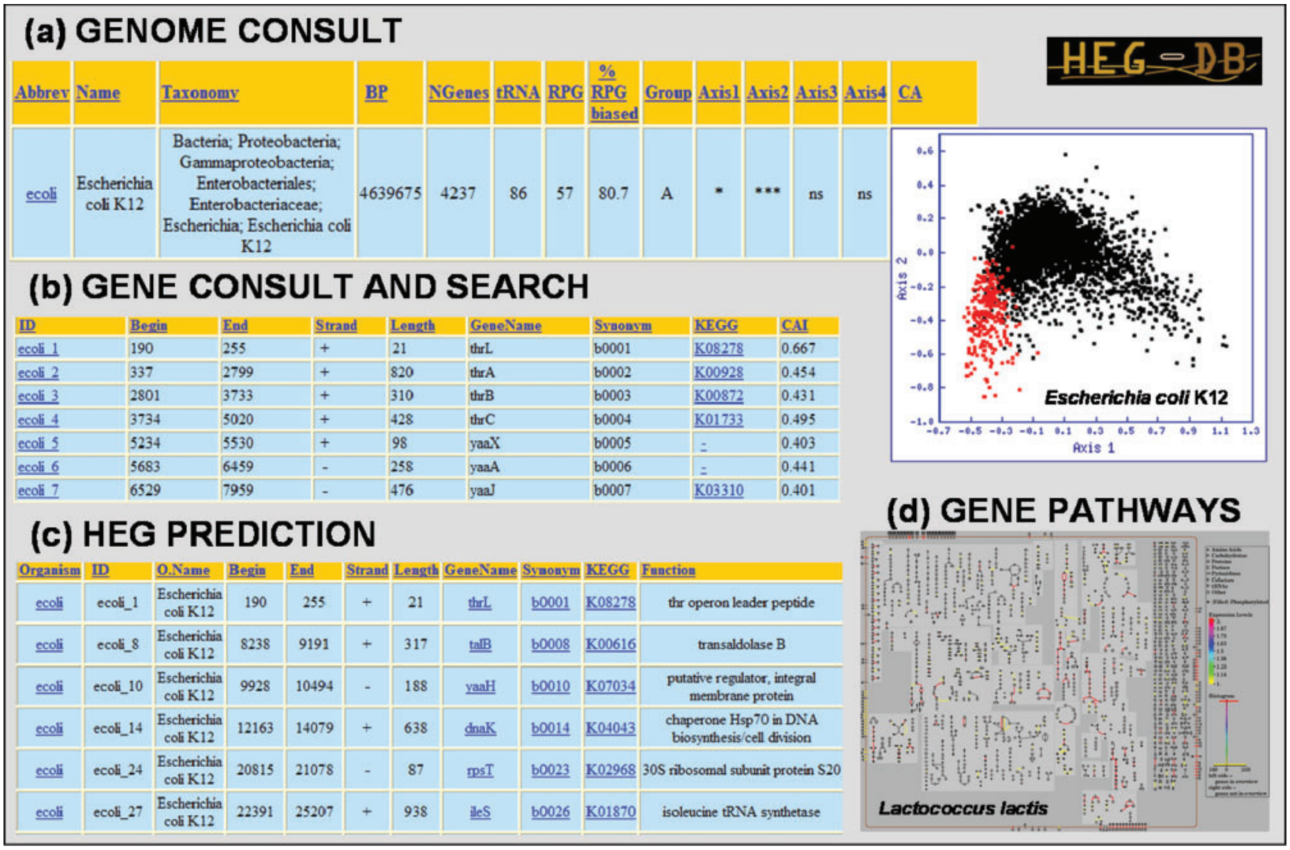
The information about genes and genomes is stored in a MySQL database that can be accessed through a series of PHP web pages. The current version of the database contains information about almost 200 genomes under translational selection. The HTML interface is divided into four sections (Figure 1): (i) 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. (ii) 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. (iii) 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. (iv) 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 (22,23). 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.

## 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.

## ACKNOWLEDGEMENTS

We thank John Bates of the Language Service of the Rovira i Virgili University for his help with writing the manuscript, Suzanne M. Paley from SRI International for her help with the Pathway Tools Omics Viewer from BioCyc and two anonymous referees for their helpful comments which contributed to improve the HEG-DB. This work has been financed by project BIO2003-07672 of the Spanish Ministry of Science and Technology. Funding to pay the Open Access publication charges for



(a) GENOME CONSULT

| Abbrev | Name | Taxonomy | BP | NGenes | tRNA | RPG | % RPG biased | Group | Axis1 | Axis2 | Axis3 | Axis4 | CA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ecoli | Escherichia coli K12 | Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia; Escherichia coli K12 | 4639675 | 4237 | 86 | 57 | 80.7 | A | * | *** | ns | ns | |

(b) GENE CONSULT AND SEARCH

| ID | Begin | End | Strand | Length | GeneName | Synonym | KEGG | CAI |
|---|---|---|---|---|---|---|---|---|
| ecoli 1 | 190 | 255 | + | 21 | thrL | b0001 | K08278 | 0.667 |
| ecoli 2 | 337 | 2799 | + | 820 | thrA | b0002 | K00928 | 0.454 |
| ecoli 3 | 2801 | 3733 | + | 310 | thrB | b0003 | K00872 | 0.431 |
| ecoli 4 | 3734 | 5020 | + | 428 | thrC | b0004 | K01733 | 0.495 |
| ecoli 5 | 5234 | 5530 | + | 98 | yaaX | b0005 | - | 0.403 |
| ecoli 6 | 5683 | 6459 | - | 258 | yaaA | b0006 | - | 0.441 |
| ecoli 7 | 6529 | 7959 | - | 476 | yaaJ | b0007 | K03310 | 0.401 |

(c) HEG PREDICTION

| Organism | ID | O.Name | Begin | End | Strand | Length | GeneName | Synonym | KEGG | Function |
|---|---|---|---|---|---|---|---|---|---|---|
| ecoli | ecoli_1 | Escherichia coli K12 | 190 | 255 | + | 21 | thrL | b0001 | K08278 | thr operon leader peptide |
| ecoli | ecoli_8 | Escherichia coli K12 | 8238 | 9191 | + | 317 | talB | b0008 | K00616 | transaldolase B |
| ecoli | ecoli_10 | Escherichia coli K12 | 9928 | 10494 | - | 188 | yaaH | b0010 | K07034 | putative regulator, integral membrane protein |
| ecoli | ecoli_14 | Escherichia coli K12 | 12163 | 14079 | + | 638 | dnaK | b0014 | K04043 | chaperone Hsp70 in DNA biosynthesis/cell division |
| ecoli | ecoli_24 | Escherichia coli K12 | 20815 | 21078 | - | 87 | rpsT | b0023 | K02968 | 30S ribosomal subunit protein S20 |
| ecoli | ecoli_27 | Escherichia coli K12 | 22391 | 25207 | + | 938 | ileS | b0026 | K01870 | isoleucine tRNA synthetase |

(d) GENE PATHWAYS

**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.

this article was provided by BIO2003-07672 of the Spanish Ministry of Science and Technology.

*Conflict of interest statement*. None declared.

## 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., 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.
4. 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.
5. 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. USA*, **101**, 3480–3485.
6. 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.
7. 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.
8. 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.
9. Carbone,A., Zinovyev,A. and Kepes,F. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, **19**, 2005–2015.
10. Puigbò,P., Guzman,E., Romeu,A. and Garcia-Vallve,S. (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.*, **35**, W126–W131.
11. Karlin,S. and Mrazek,J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.*, **182**, 5238–5250.
12. 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.
13. 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. USA*, **100**, 7313–7318.
14. Karlin,S., Theriot,J. and Mrazek,J. (2004) Comparative analysis of gene expression among low G + C gram-positive genomes. *Proc. Natl Acad. Sci. USA*, **101**, 6182–6187.
15. Henry,I. and Sharp,P.M. (2007) Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.*, **24**, 10–12.

16. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.
17. 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.
18. Perriere,G. and Thioulouse,J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.*, **30**, 4548–4555.
19. Grocock,R.J. and Sharp,P.M. (2002) Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene*, **289**, 131–139.
20. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
21. 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.
22. Paley,S.M. and Karp,P.D. (2006) The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Res.*, **34**, 3771–3778.
23. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.