Nucleic Acids Research, 2005, Vol. 33, No. 3 **1051–1057**
doi:10.1093/nar/gki261

Relative predicted protein levels of functionally associated proteins are conserved across organisms

Gila Lithwick and Hanah Margalit*

Department of Molecular Genetics and Biotechnology, The Faculty of Medicine, The Hebrew University,
Jerusalem 91120, Israel

Received January 3, 2005; Revised and Accepted February 1, 2005

ABSTRACT

We show that the predicted protein levels of functionally related proteins change in a coordinated fashion over many unicellular organisms. For each protein, we created a profile containing a protein abundance measure in each of a set of organisms. We show that for functionally related proteins these profiles tend to be correlated. Using the Codon Adaptation Index as a predictor of protein abundance in 48 unicellular organisms, we demonstrated this phenomenon for two types of functional relations: for proteins that physically interact and for proteins involved in consecutive steps within a metabolic pathway. Our results suggest that the protein abundance levels of functionally related proteins co-evolve.

INTRODUCTION

Over the past few years, the characterization and prediction of functional relations between proteins have been at the forefront of genomic research. Several types of features, based on experimental results and computational analyses, were shown to be associated with functionally related and interacting proteins. It was shown that functionally associated proteins exhibit correlated mRNA expression profiles over a set of environmental conditions or different tissue types (1–4). In many cases, functionally linked proteins were shown to be encoded by genes that are conserved in evolution as neighbors (5–7); they were shown to undergo gene fusion events, such that individual genes in one genome were found to be fused into a single gene in another genome (8,9); and they were shown to co-occur in various organisms, either they were preserved together or eliminated together during evolution (10,11). This property of co-occurrence of functionally linked proteins across various organisms traditionally refers only to their presence or absence; however, it is appealing to conjecture that their relative protein abundance levels should be preserved in a coordinated fashion as well. Such a preservation of relative abundance levels across organisms is expected to be important for two types of functional associations: (i) for interacting proteins, where coordinated abundance levels may be needed in order to preserve the stoichiometry of the interaction; and (ii) for proteins within the same metabolic pathway, where a constant balance is needed between the relative abundance of the proteins within the pathway in order to ensure that the flow of the pathway is maintained.

To investigate whether the relative levels of functionally associated proteins are preserved during evolution, measured protein levels are needed. With such data at hand, one can generate a profile for each protein, where each entry contains the abundance of the protein in a specific genome. These profiles can be analyzed in order to determine whether the profiles of functionally linked proteins are correlated. However, since large-scale protein abundance data are scarce and are available only for a subset of proteins in a few organisms (12,13), this analysis is not feasible at present. An alternative approach would be to represent the protein levels by other measures that were shown to correlate with protein abundance, such as the mRNA level of the corresponding genes (13) or the degree of codon bias of the proteins (14–16). As to the former, the nature of available mRNA expression data prevents its use for the intended analysis. The mRNA expression is usually measured relative to background expression levels in a specific condition or set of conditions, and absolute expression levels are scarce. It is therefore impossible to compare the mRNA expression levels of pairs of proteins across different organisms. In addition, gene expression experiments in different genomes are often performed under different conditions, making the comparison of the expression of a gene in different organisms problematic. Another measure that was shown to correlate with the protein abundance level is the degree of codon bias of the protein (14,15,17,18). A common explanation for this phenomenon is based on the observation within several organisms that tRNA molecules corresponding to preferred codons exist at higher levels (19–24). The mRNA sequences of highly abundant proteins have a greater bias toward these preferred codons, enabling their rapid translation

*To whom correspondence should be addressed. Tel: +972 2 6758614; Fax: +972 2 6757308; Email: hanah@md.huji.ac.il

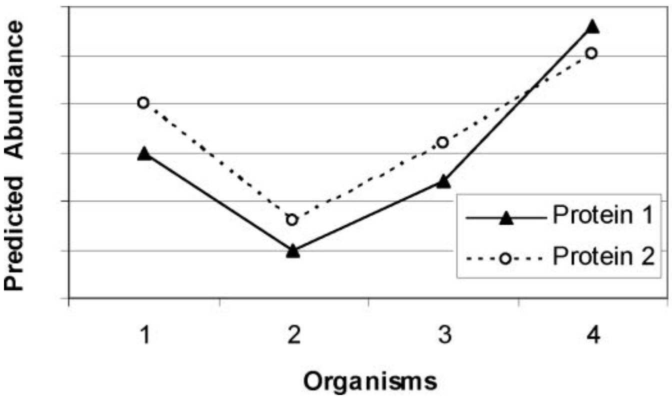


Figure 1. Illustration of protein abundance profiles. The predicted abundance levels of two proteins, each with orthologs in four organisms, are shown. The first is represented by filled triangles, and the second by open circles. The correlation coefficient between these two abundance profiles can be computed. The connecting lines in the plot are for illustration only.

due to the higher availability of the corresponding tRNAs, particularly during exponential cell growth. Several measures of codon bias have been developed, including the Codon Adaptation Index (CAI) (17) and the effective Number of Codons (Nc) (25). These measures were shown to correlate with experimentally measured proteins levels and have therefore been used as predictors of relative protein abundance in the cell (14–16). Each of these measures provides one value of codon bias for a protein encoded within a genome. Therefore, a profile of the codon bias measures, representing the growth-phase protein abundance of a gene in different organisms, can be created.

We, therefore, propose to study the conservation of relative protein abundance levels of functionally related proteins by examining whether their corresponding codon bias measures changed in a coordinated fashion through evolution (Figure 1). For interacting proteins, such a correlation has recently been reported, using four species of yeast (26). In this study, we looked at 48 bacterial, archaeal and fungal genomes. We calculated the correlation coefficients between the codon bias profiles of functionally related proteins of two types, based on data derived from *Escherichia coli*: a protein–protein interaction dataset and a dataset consisting of proteins that are involved in consecutive steps within pathways. For both types of data, we show that functionally associated proteins tend to have correlated protein abundance profiles.

MATERIALS AND METHODS

Calculation of codon bias measures

We used two measures of codon bias, which we calculated using the EMBOSS package (27). (i) The CAI (17) is the most commonly used, and is a measure of the tendency of the coding sequence of a protein to use the codons that are preferentially used by a reference set of highly expressed proteins. We used the ribosomal proteins as this reference set. This measure ranges between 0 and 1, where a higher CAI value indicates a greater level of codon bias, and predicts a higher protein level. (ii) The Nc (25) is a number which ranges between 20 and 61, and measures codon bias by considering the number of various codons used to encode the protein. A lower Nc indicates a greater level of codon bias. This measure was used since it does not require a reference set of abundant proteins.

Genome data

Sequence data and annotations were extracted from GenBank (http://www.ncbi.nlm.nih.gov). The COG database (Clusters of Orthologous Groups of proteins) (28), was used in order to extract orthologs. A COG is defined as the group of orthologs for a given protein. We applied several filters to the organisms and proteins within the COG database. (i) When there existed multiple strains, we kept only one genome (we tried to take the more studied strain). (ii) For certain organisms, it has been shown that codon bias is not a good indicator of protein abundance, because of the existence of other dominating selective pressures, such as GC content and G+C strand bias (29,30). We, therefore, kept only organisms for which we rejected (using a one-tailed Mann–Whitney test with a *P*-value cutoff of 0.01) the null hypothesis that ribosomal proteins do not have lower Nc values than the rest of the proteins in the genome. For this test, hypothetical and mitochondrial proteins were excluded. (iii) Only proteins with a length of at least 100 amino acids were kept since the codon bias measures are not reliable for short proteins.

After these filtering steps, we were left with 48 organisms for the analysis (Supplementary Table 1).

Datasets regarding functional associations between proteins

Two datasets representing different types of functional associations were used, both from *E.coli*: (i) protein–protein interaction data; and (ii) pathway data.

Data regarding protein–protein interactions were extracted from the Database of Interacting Proteins (DIP, version 07/04) (31). Homodimers and interactions involving the chaperone GroL were excluded. For a total of 167 *E.coli* interactions, their interacting pair mates could be assigned to COGs, and these were kept for further analyses (Supplementary Table 2).

Pathway data for *E.coli* were extracted from the EcoCyc database (32), using the PerlCyc interface (33) to the Pathway Tools Software (34). Pairs of neighboring proteins were taken as those conducting consecutive steps within a pathway, or a set of pathways. If one of these proteins was assigned to an additional pathway to which the other protein was not assigned, the pair was discarded. A total of 197 pairs could be assigned to COGs, and were kept for further analyses (Supplementary Table 3).

Creation of codon bias profiles

In order to exclude any bias that may result from the CAI distribution of any of the organisms, and in order to be able to compare the contribution of different groups of organisms, we repeated the analysis for many subgroups of organisms. The rationale for this approach is clarified further in the 'Analysis of significance' section below and in the Results and Discussion. For each combination of four organisms, the CAI was calculated for each protein within each COG. We chose a subgroup size of four, since Spearman correlation coefficients begin to be significant from $n = 4$ (for a one-tailed test). We did not use subgroups of more than four organisms, since as the number of organisms increases the number of positive pairs present in each group of organisms decreases.

In many COGs, there exist paralogs for some of the organisms. Since these paralogs can have very different codon bias measures, we discarded a gene for a given organism in the case that its paralogs had very different scores. Therefore, in such a case, the organism was not considered as represented within this COG. Genes for an organism were discarded if the difference between the highest and lowest CAI values of paralogs exceeded 0.1. For paralogs that were maintained, the score used for this organism within this COG was the average of the paralogs' CAI values. After this step, only COGs having a score for each of the four organisms were kept. The CAI scores of each genome within these COGs were converted to Z-scores, in order to have a common range of values for the various genomes. The normalization was computed per genome by the mean and standard deviation of the proteins included in the analysis.

Calculation of correlation between the predicted abundance profiles of proteins

For each group of organisms, pairs of functionally related proteins (COGs) where both have representatives in each of the four organisms (after filtering, as described in the previous section) were taken. Spearman correlation coefficients were calculated between the normalized CAI measures for these pairs in the various organisms.

Often, COGs include protein domains, and not necessarily entire proteins. This could potentially introduce a bias into the analysis, since a single protein, or a set of proteins, may appear in more than one COG. Since these proteins would have the same codon bias measure in both COGs, this would inevitably contribute to a correlation between these often functionally related COGs. Therefore, for each group of organisms, pairs of COGs with common proteins within this subgroup of organisms were not included.

Analysis of significance

In order to analyze significance for each subgroup of four organisms, a negative set consisting of all possible pairs of proteins from the positive dataset was created (excluding the functionally related pairs). This negative set included only proteins from the positive dataset, since the interacting protein dataset (and to a much lesser extent, the pathway dataset) was found to be biased toward abundant proteins, and we wished to neutralize this bias. The median value of the correlation coefficients in these negative sets was not equal to zero for many subgroups of organisms. This demonstrates that there is a bias in some of the combinations of organisms, which was part of the incentive for our approach of looking at groups of organisms separately (other reasons are detailed above).

For some of the subgroups of genomes, relatively few proteins in the positive dataset had representatives in the four genomes. Therefore, only subgroups of genomes resulting in a negative set of at least 1000 pairs were kept for the final analysis.

For each group of four organisms, a one-sided Wilcoxon test was performed, comparing the correlation coefficients of the positive group with the median of the correlation coefficients of the generated negative set. For this calculation, we used the exactRankTests package (http://cran.r-project.org/doc/packages/exactRankTests.pdf) written for the R project for Statistical Computing (35). This procedure allows for ties in the data, which are common when calculating the Spearman correlation coefficient for $n = 4$. The null hypothesis was that the correlations in the positive group are not greater than the median of the negative set.

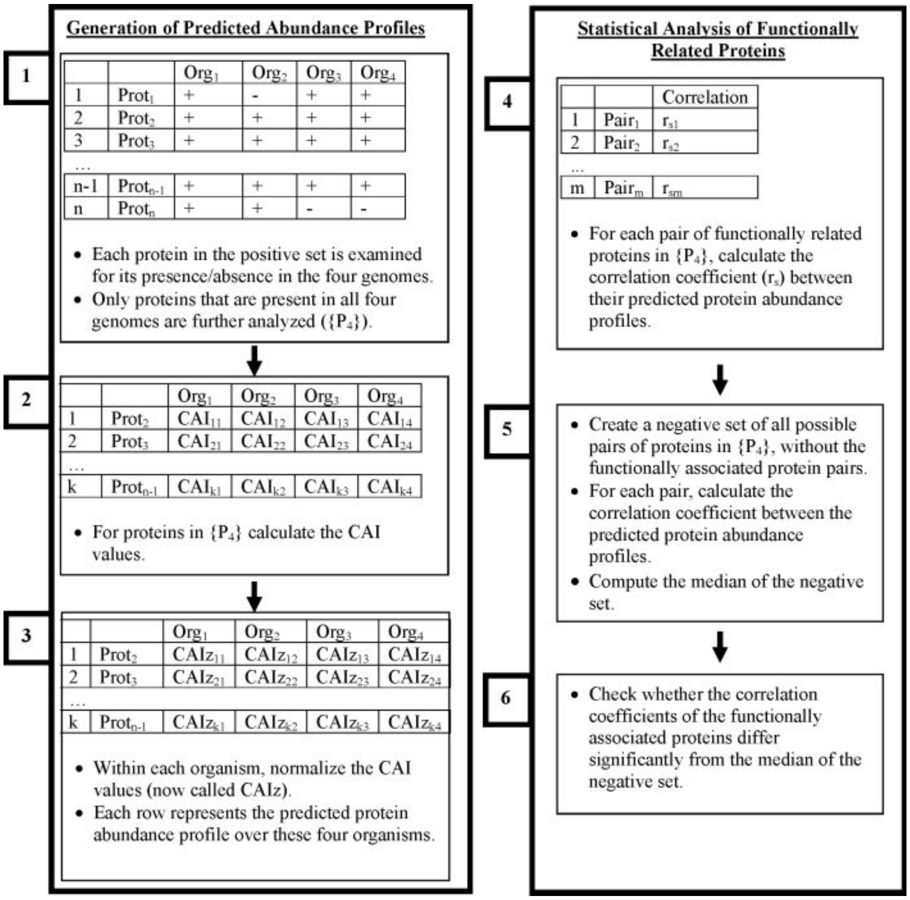
In order to calculate whether the overall results were significant, the observed fraction of groups of organisms with significant Wilcoxon P-values (≤ 0.05) was compared with the expected fraction of 0.05 using the binomial test.

RESULTS AND DISCUSSION

General overview of the approach

We tested whether the predicted protein levels of functionally associated proteins (as predicted by the CAI) change in a correlated fashion across different organisms (Figure 1). We looked at 48 bacterial, archaeal and fungal genomes (Supplementary Table 1), and extracted orthologous proteins within these organisms using the COG database (see Materials and Methods). We looked at two types of functional associations: interacting proteins and proteins that function in consecutive steps within metabolic pathways (we term these 'neighboring proteins'). We refer to protein pairs extracted from these two datasets as 'positive' pairs. For these positive pairs, we calculated the correlation coefficients between the codon bias values of the pair mates in different organisms, and analyzed the significance of these correlations.

Ideally, we would have liked to analyze all the positive pairs at once, for the 48 organisms. However, due to the nature of the data, we used a different approach that allowed us to minimize the loss of data and to analyze the significance of the phenomenon. One characteristic of the data that led us to our approach is that many proteins occur in only a subset of the organisms (different proteins have different phylogenetic profiles). In such cases, the abundance profiles across all organisms will contain many null values that may lead to erroneous correlations. Another characteristic is that for certain combinations of organisms there is background noise, evident in the fact the codon bias values are distributed such that even within a random group of pairs of proteins, most are slightly correlated. Our approach consisted of looking at different subgroups of organisms, and analyzing the statistical significance of the phenomenon within each subgroup separately. We took all subgroups of size four for our analysis (see Materials and Methods), which enabled us to include a large number of positive pairs in many subgroups. With our approach, we could analyze pairs of functionally related proteins even if their phylogenetic profiles across all organisms are very different, while taking background noise into account. An additional advantage of this approach is that it enabled us to look separately at different subgroups of organisms. The main steps of the analysis for each four-organism group are illustrated in Figure 2.


Generation of Predicted Abundance Profiles
1
Org1 Org2 Org3 Org4
1 Prot1 + - + +
2 Prot2 + + + +
3 Prot3 + + + +
...
n-1 Protn-1 + + + +
n Protn + + - -
2
Org1 Org2 Org3 Org4
1 Prot2 CAI11 CAI12 CAI13 CAI14
2 Prot3 CAI21 CAI22 CAI23 CAI24
...
k Protn-1 CAIk1 CAIk2 CAIk3 CAIk4
3
Org1 Org2 Org3 Org4
1 Prot2 CAIz11 CAIz12 CAIz13 CAIz14
2 Prot3 CAIz21 CAIz22 CAIz23 CAIz24
...
k Protn-1 CAIzk1 CAIzk2 CAIzk3 CAIzk4
Statistical Analysis of Functionally Related Proteins
4
Correlation
1 Pair1 rs1
2 Pair2 rs2
...
m Pairm rsm
5
6

- Each protein in the positive set is examined for its presence/absence in the four genomes.
- Only proteins that are present in all four genomes are further analyzed ({P4}).

- For proteins in {P4} calculate the CAI values.

- Within each organism, normalize the CAI values (now called CAIz).
- Each row represents the predicted protein abundance profile over these four organisms.

- For each pair of functionally related proteins in {P4}, calculate the correlation coefficient (r_s) between their predicted protein abundance profiles.

- Create a negative set of all possible pairs of proteins in {P4}, without the functionally associated protein pairs.
- For each pair, calculate the correlation coefficient between the predicted protein abundance profiles.
- Compute the median of the negative set.

- Check whether the correlation coefficients of the functionally associated proteins differ significantly from the median of the negative set.

Figure 2. Main steps of the analysis for a four-organism group. Each step is detailed in Materials and Methods. After performing the analysis on each four-organism group, the number of groups that resulted in significant *P*-values was compared with that expected at random in order to obtain the significance of the overall analysis.

Significant correlation between predicted protein abundance profiles of related proteins

For many four-organism groups, the functionally related proteins had statistically significantly higher correlation coefficients than expected at random (Table 1). The statistical significance was evaluated by comparing these correlation coefficients with those obtained in the negative set, using a significance threshold of 0.05 (see Figure 2 and Materials and Methods). We looked at groups of organisms for which a negative set of at least 1000 pairs of proteins could be created. This resulted in 15 978 groups of four organisms for the protein–protein interaction data and 69 433 groups of four organisms for the pathway data. For the interacting pairs, 58% of the four-organism groups that were analyzed had statistically significantly higher correlation coefficients than expected at random (i.e. 58% of the groups differed from random at a significance cutoff of 0.05). For the pathway pairs, 42% of the four-organism groups had statistically significantly higher correlation coefficients than expected at random (Table 1). These fractions are significantly higher than the fraction of 5% that can be expected at random ($P \ll 1 \times 10^{-10}$, using the binomial test). This demonstrates that for many combinations of organisms, the predicted protein abundance profiles of related proteins are indeed correlated.

When looking at subgroups of organisms (Table 1), the high fraction of subgroups from γ-proteobacteria that have correlated profiles is striking, while for archaea, which are more distantly related to the reference organism *E.coli*, the phenomenon is quite weak. The finding that the majority of groups of organisms from γ-proteobacteria showed that the

Table 1. Correlation between predicted abundance profiles of functionally related proteins for different organism groups

Number of organisms in analysis	Number of organism groups[a]	Number (%) of organism groups with $P \leq 0.05$	*P*-value[b]
Protein–protein interactions			
All (48)	15 978	9253 (58%)	$\ll 1 \times 10^{-10}$
γ-Proteobacteria (7)	35	34 (97%)	$\ll 1 \times 10^{-10}$
Gram-positive[c] (11)	70	45 (64%)	$\ll 1 \times 10^{-10}$
Archaea (8)	0	0	Nr[d]
Pathways			
All (48)	69 433	29 277 (42%)	$\ll 1 \times 10^{-10}$
γ-Proteobacteria (7)	35	30 (86%)	$\ll 1 \times 10^{-10}$
Gram-positive[c] (11)	70	41 (59%)	$\ll 1 \times 10^{-10}$
Archaea (8)	70	8 (11%)	0.02

[a] Groups of four organisms, for which a negative set of at least 1000 protein pairs could be generated, were included in the analyses (see Materials and Methods).
[b] *P*-values, estimating whether the number of organism groups with $P \leq 0.05$ is significant, were calculated using a binomial test, comparing the observed number of groups having *P*-value ≤ 0.05 with the expected fraction (0.05).
[c] Low G+C Gram-positive bacteria.
[d] Not relevant, since there were no groups with a large enough negative set for statistical evaluation.

predicted protein abundance of functionally related proteins co-evolve is reasonable, since the positive data was extracted from *E.coli*, which is a member of this subgroup. Therefore, it is likely that the functional interactions and general cellular mechanisms are more conserved within the γ-proteobacteria. On the other hand, in archaea, the number of groups of organisms having functionally related pairs with significantly higher correlation coefficients than expected at random was much smaller. This may be due to several reasons. (i) The functional relations were based on *E.coli* as a reference, and these protein pairs are not all necessarily functionally related to the same extent in archaea. (ii) Codon bias has been suggested to be less associated with protein levels in archaea since most archaea are thermophiles, and other environmental factors can influence codon usage (36). (iii) Relatively fewer pairs of functionally related proteins occur in archaea, since not all the *E.coli* proteins are conserved in archaea. This results in a smaller sample size that can affect the *P*-values.

In general, CAI and other measures of codon bias are only predictors of protein abundance levels. The correlation of abundant tRNAs with preferred codons in highly expressed genes has long been implied to improve the efficiency of translation (19,20). This led to the development of codon bias measures, which were shown to be correlated with the mRNA expression levels (17,25,37) and with protein abundance levels (14,15). However, it should be noted that these correlations were far from perfect, implying that only some of the variance in protein abundance can be explained by codon bias. Therefore, the results presented here are probably an underestimate of the phenomenon of correlated protein abundance levels in functionally related proteins.

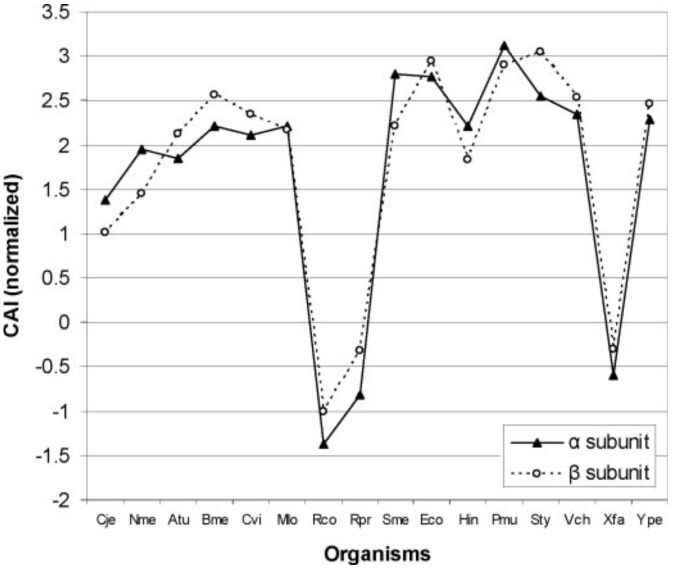


Figure 3. CAI profiles within proteobacteria for the α and β subunits of the F_0F_1–ATP synthase (COG0056 and COG0055, respectively). Normalized CAI values for proteobacterial genomes are shown. For each organism, the normalization was based on the mean and standard deviation of all its proteins that appear in the COG database. The genomes represented here are *Campylobacter jejuni* (Cje), *Neisseria meningitidis Z2491* (Nme), *Agrobacterium tumefaciens* (Atu), *Brucella melitensis* (Bme), *Caulobacter vibrioides* (Cvi), *Mesorhizobium loti* (Mlo), *Rickettsia conorii* (Rco), *Rickettsia prowazekii* (Rpr), *Sinorhizobium meliloti* (Sme), *Escherichia coli* K12 (Eco), *Haemophilus influenzae* (Hin), *Pasteurella multocida* (Pmu), *Salmonella typhimurium* (Sty), *Vibrio cholerae* (Vch), *Xylella fastidiosa* (Xfa) and *Yersinia pestis* (Ype). The connecting lines in the plot are for illustration only.

Examples of functionally related proteins with correlated profiles

We describe here several examples of protein pairs that exhibit correlated profiles. The examples from the interaction dataset are pairs of proteins that have additional evidence of being translationally coordinated. The analysis was carried out for four-organism groups, as described in Figure 2 and in the Materials and Methods. However, for simplicity, the Spearman correlation coefficients (r_s) reported here were computed over groups of relevant organisms, and not over all the possible four-organism groups.

(i) The α and β subunits of the F_0F_1–ATP synthase (COG0056 and COG0055). The F_0F_1–ATP synthase is one of the most highly conserved enzymes (38). It is composed of two components (F_0 and F_1) that use the proton gradient across the membrane in order to synthesize ATP. Three α subunits and three β subunits form a sphere that comprises the main portion of the F_1 component (39). While these two proteins are similar at the protein sequence level (39), this does not necessarily imply that they are similar at the codon level. Indeed, a pairwise alignment of the coding sequences of the two proteins in *E.coli* found no significant similarity. The two proteins are, for many genomes, within the same operon. In addition, they have been found to be translationally coordinated in the chloroplast of *Chlamydomonas reinhardtii* (40), and we would therefore expect that their protein abundance levels be generally correlated. Indeed, the overall correlation of these two CAI profiles (Figure 3) is high (r_s = 0.85 over the proteobacteria and r_s = 0.83 over the 34 organisms in which the subunits are present). Interestingly, the CAI values of both subunits are relatively low in the organisms *Rickettsia conorii*, *Rickettsia prowazekii* and *Xylella fastidiosa*. The former two are obligate parasites, while *X.fastidiosa* is a plant-pathogenic bacterium. Although codon usage has been suggested not to be a good estimator for protein abundance in these organisms (41,42), we did find a significant difference between the codon usage of ribosomal and other proteins in these organisms. A possible explanation for their low CAI values in this example lies is the fact that these organisms have been found to contain an ATP/ADP translocase (COG3202), and thus they can import ATP from their host cells (43). This can explain a low cellular abundance of the F_0F_1–ATP synthase in these organisms. Interestingly, the other organisms within the ATP/ADP translocase COG do not have the α and β subunits.

(ii) The β and β′ subunits of the DNA-directed RNA polymerase (COG0085 and COG0086). The β and β′ subunits are highly conserved, and the genes encoding these subunits are always adjacent (44). It has also been speculated that the assembly of these subunits in *E.coli* occurs co-translationally (45). Thus, it seems that the relative abundance of these proteins should be evolutionarily correlated. Indeed, the correlation coefficient between the

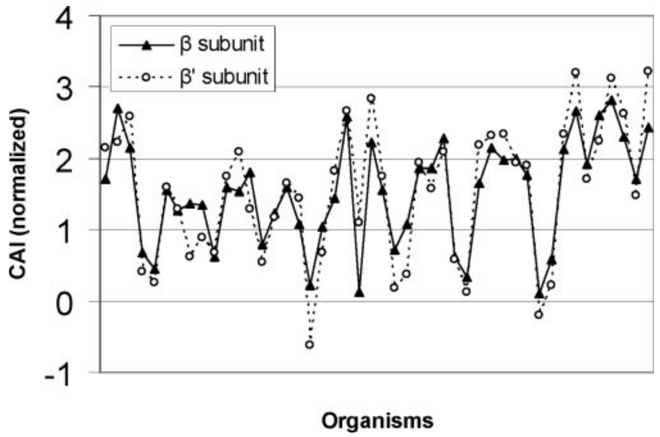


Figure 4. CAI profiles for the β and β′ subunits of the DNA-directed RNA polymerase (COG0085 and COG0086, respectively). The normalized CAI values for 46 genomes are shown here. For each organism, the normalization was based on the mean and standard deviation of all its proteins that appear in the COG database. The two organisms not represented were left out due to a large difference in the CAI values of their paralogs (see Materials and Methods). The connecting lines in the plot are for illustration only.

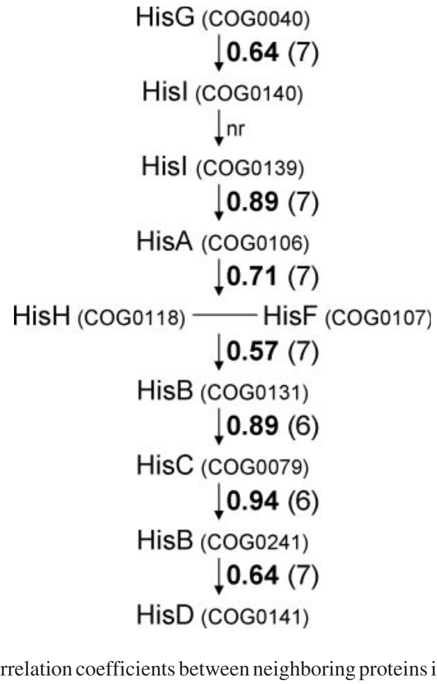


Figure 5. Correlation coefficients between neighboring proteins in the histidine biosynthesis pathway of γ-proteobacteria. Consecutive proteins within the pathway are shown, with the corresponding COG name in parentheses. Next to the arrows, Spearman correlation coefficients between the consecutive proteins are shown, and the number of organisms over which the correlation was calculated is shown in parentheses. HisB and HisI are each composed of two domains, and are thus each represented by two COGs. The correlation between COG0140 and COG0139 is marked 'nr', since the γ-proteobacteria proteins within these two COGs are identical. COG0118 and COG0107 form a heterodimer, and the Spearman correlation coefficient for these proteins, over seven organisms, is one. The correlation coefficient of the preceding protein (COG0106) with both of these proteins is identical, as are the correlation coefficients of the successive protein (COG0131) with both of these proteins.

normalized CAI values is very high (Figure 4, $r_s = 0.91$, over 46 organisms).

(iii) The histidine biosynthesis pathway. The histidine biosynthesis pathway is very conserved, and the genes encoding the pathway's proteins undergo regulation at several stages. This is due to the high metabolic cost of the pathway (36). Indeed, we found that the CAI profiles of the genes encoding the consecutive steps of the pathway within γ-proteobacteria are correlated, as shown in Figure 5. Most of these proteins appear in all seven of the γ-proteobacteria in our data. Therefore, to assess the significance of these correlation coefficients, we compared them with the median of the correlation coefficients calculated for all possible pairs of pathway proteins in our dataset that appear in all seven γ-proteobacteria (not just consecutive proteins), which was 0.107. The correlation coefficients within the pathway are noticeably all much higher ($P = 0.03$ using the Wilcoxon test to compare the correlation coefficients calculated over seven organisms to this median).

CONCLUSION

In this study, we showed that predicted protein abundance levels of functionally related proteins co-evolve, and that this phenomenon is widespread, occurring over many groups of unicellular organisms. Ideally, it would have been best to explore this phenomenon by analyzing experimental protein levels. However, in the absence of this type of data, we demonstrate that putative abundance levels, represented by the CAI, are informative and provide useful insight.

The correlation between predicted abundance levels was demonstrated for two types of functional relations between proteins: interacting proteins and neighboring proteins in metabolic pathways. We suggest that the identification of such correlations between CAI values across organisms can be used for the inference of functional associations between proteins. This approach can then be integrated with other predictive methods for better prediction of such associations.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Norman Grover for useful discussions. We also thank the members of the laboratory, and especially Yael Altuvia, Ruth Hershberg, Esti Yeger-Lotem and Galit Lipsitz for their helpful comments. This study was supported by the Israeli Science Foundation, administered by the Israeli Academy of Sciences and Humanities. Funding to pay the Open Access publication charges for this article was provided by the Israeli Science Foundation.

REFERENCES

1. Grigoriev,A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **29**, 3513–3519.
2. Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, **12**, 37–46.

3. Bono,H. and Okazaki,Y. (2002) Functional transcriptomes: comparative analysis of biological pathways and processes in eukaryotes to infer genetic networks among transcripts. *Curr. Opin. Struct. Biol*., **12**, 355–361.
4. Kemmeren,P., van Berkum,N.L., Vilo,J., Bijma,T., Donders,R., Brazma,A. and Holstege,F.C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell*, **9**, 1133–1143.
5. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.
6. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
7. Yanai,I., Mellor,J.C. and DeLisi,C. (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet*., **18**, 176–179.
8. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
9. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
10. Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.
11. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
12. Link,A.J., Robison,K. and Church,G.M. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis*, **18**, 1259–1313.
13. Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K. and Weissman,J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
14. 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.
15. Lithwick,G. and Margalit,H. (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res*., **13**, 2665–2673.
16. McHardy,A.C., Puhler,A., Kalinowski,J. and Meyer,F. (2004) Comparing expression level-dependent features in codon usage with protein abundance: an analysis of 'predictive proteomics'. *Proteomics*, **4**, 46–58.
17. 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.
18. Ermolaeva,M.D. (2001) Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol*., **3**, 91–97.
19. 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.
20. 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.
21. Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol*., **2**, 13–34.
22. Dong,H., Nilsson,L. and Kurland,C.G. (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol*., **260**, 649–663.
23. Kanaya,S., Yamada,Y., Kudo,Y. and Ikemura,T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**, 143–155.
24. Kanaya,S., Yamada,Y., Kinouchi,M., Kudo,Y. and Ikemura,T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol*., **53**, 290–298.
25. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
26. Fraser,H.B., Hirsh,A.E., Wall,D.P. and Eisen,M.B. (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl Acad. Sci. USA*, **101**, 9033–9038.
27. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
28. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*., **29**, 22–28.
29. Lafay,B., Atherton,J.C. and Sharp,P.M. (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*, **146**, 851–860.
30. Perriere,G. and Thioulouse,J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res*., **30**, 4548–4555.
31. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*., **30**, 303–305.
32. Karp,P.D., Arnaud,M., Collado-Vides,J., Ingraham,J., Paulsen,I.T. and Saier,M.H. (2004) The *E.coli* EcoCyc Database: no longer just a metabolic pathway database. *ASM News*, **70**, 25–30.
33. Mueller,L.A., Zhang,P. and Rhee,S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol*., **132**, 453–460.
34. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**(Suppl. 1), S225–S232.
35. Ihaka,R. and Gentleman,R. (1996) R: A language for data analysis and graphics. *J. Comput. Graph. Statist*., **5**, 299–314.
36. Alifano,P., Fani,R., Lio,P., Lazcano,A., Bazzicalupo,M., Carlomagno,M.S. and Bruni,C.B. (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol. Rev*., **60**, 44–69.
37. Karlin,S., Mrazek,J. and Campbell,A.M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol*., **29**, 1341–1355.
38. Yoshida,M., Muneyuki,E. and Hisabori,T. (2001) ATP synthase—a marvellous rotary engine of the cell. *Nature Rev. Mol. Cell. Biol*., **2**, 669–677.
39. Boyer,P.D. (1997) The ATP synthase—a splendid molecular machine. *Annu. Rev. Biochem*., **66**, 717–749.
40. Drapier,D., Girard-Bascou,J. and Wollman,F.A. (1992) Evidence for nuclear control of the expression of the atpA and atpB chloroplast genes in Chlamydomonas. *Plant Cell*, **4**, 283–295.
41. Andersson,S.G. and Sharp,P.M. (1996) Codon usage and base composition in *Rickettsia prowazekii*. *J. Mol. Evol*., **42**, 525–536.
42. Smolka,M.B., Martins,D., Winck,F.V., Santoro,C.E., Castellari,R.R., Ferrari,F., Brum,I.J., Galembeck,E., Della Coletta Filho,H., Machado,M.A. *et al*. (2003) Proteome analysis of the plant pathogen *Xylella fastidiosa* reveals major cellular and extracellular proteins and a peculiar codon bias distribution. *Proteomics*, **3**, 224–237.
43. Meidanis,J., Braga,M.D. and Verjovski-Almeida,S. (2002) Whole-genome analysis of transporters in the plant pathogen *Xylella fastidiosa*. *Microbiol. Mol. Biol. Rev*., **66**, 272–299.
44. Iyer,L.M., Koonin,E.V. and Aravind,L. (2004) Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. *Gene*, **335**, 73–88.
45. Severinov,K., Mooney,R., Darst,S.A. and Landick,R. (1997) Tethering of the large subunits of *Escherichia coli* RNA polymerase. *J. Biol. Chem*., **272**, 24137–24140.