DNA Research Advance Access published June 6, 2014

DNA RESEARCH pp. 1–15, (2014) doi:10.1093/dnares/dsu017

Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias

RENANA Sabi[1] and TAMIR Tuller[1,2,*]

Department of Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel[1] and The Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel[2]

*To whom correspondence should be addressed. Tel. +972-64058363. Fax. +972-64058363. Email: tamirtul@post.tau.ac.il

Edited by Prof. Hiroyuki Toh
(Received 10 January 2014; accepted 26 April 2014)

Abstract

The tRNA adaptation index (tAI) is a widely used measure of the efficiency by which a coding sequence is recognized by the intra-cellular tRNA pool. This index includes among others weights that represent wobble interactions between codons and tRNA molecules. Currently, these weights are based only on the gene expression in *Saccharomyces cerevisiae*. However, the efficiencies of the different codon–tRNA interactions are expected to vary among different organisms. In this study, we suggest a new approach for adjusting the tAI weights to any target model organism without the need for gene expression measurements. Our method is based on optimizing the correlation between the tAI and a measure of codon usage bias. Here, we show that in non-fungal the new tAI weights predict protein abundance significantly better than the traditional tAI weights. The unique tRNA–codon adaptation weights computed for 100 different organisms exhibit a significant correlation with evolutionary distance. The reported results demonstrate the usefulness of the new measure in future genomic studies.

Key words: codon usage bias; tRNA adaptation index; protein levels; wobble interactions; ribosome

1. Introduction

Allowed by the redundancy of the genetic code, coding regions exhibit non-uniform usage of synonymous codons. This deviation from uniform codon usage is termed codon usage bias (CUB) and is related among others to various aspects of gene translation (and more generally gene expression) and its efficiency;[1–10] specifically, it was suggested that it regulates transcription and translation, but may also be related to recombination rate. Indeed, it is known for over 30 years that in most organisms the degree of CUB of genes correlates with their expression levels.[11–14]

Various approaches for quantifying the CUB of a gene have been suggested in the last decades: the effective number of codons, for instance, measures deviations from equal usage of synonymous codons,[13] while other indices such as the frequency of optimal codons,[15] the codon bias index,[11] and the codon adaptation index (CAI)[16] define a subset of 'optimal' codons and measure the frequency of these codons in the coding region of the gene.

The CUB indices have been employed in hundreds of previous studies. For example, it is known that in many organisms (e.g. *Escherichia coli*) the CAI exhibits a very high correlation with protein levels (similar to the one obtained between mRNA levels and protein levels[17]); thus, CAI has been frequently used as a proxy of expression levels (see, for example,[18–20]). In addition, it has been employed in a vast number of key studies in the recent years.[18,19,21,22]

One disadvantage of measures that are based on a set of reference genes[11,16,23] is the fact that in the case of organisms with poor genomic data and without large scale cellular measurements, creating a meaningful reference set is challenging. Another disadvantage of these measures is the fact that they cannot separate between the various possible causes of CUB in highly expressed genes: some of them may be related directly to the translation process (e.g. co-evolution and/or adaptation to the tRNA pool[8,24,25]) and others may not be related to translation (e.g. GC content and folding[9,26–28]).

In 2004, dos Reis *et al.*[8] proposed the tRNA adaptation index (tAI), which aims to estimate only the adaptation of codons/genes to some aspects directly related to the elongation step occurs in the ribosome via the adaptation to the tRNA pool, wobble interactions, and properties of the ribosome. Specifically, the tAI considers the fact that different tRNA species can recognize a codon with different affinities.[2,8,29] Thus, the tAI is different than CUB-based measurements mentioned above and provides important information related to translation that is not necessarily covered by CUB measures.

Indeed, measures of the adaptation of genes to the tRNA pool (such as the tAI) have extensively been used in the recent years for studying questions in diverse biomedical disciplines such as evolutionary biology, functional genomics, and systems biology (see, for example,[3,30–35]).

Let n_i be the number of tRNA isoacceptors that recognize the ith codon, the absolute adaptiveness value of the ith codon is defined in the following equation:

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij}) \cdot \mathrm{tGCN}_{ij} \qquad (1)$$

where tGCN_{ij} is the gene copy number of the jth tRNA that recognizes the ith codon (a proxy of the tRNA levels[24,29,36]), and S_{ij} is a selective constraint on the efficiency of the interaction between the ith codon and the jth tRNA, which is scored between 0 (perfect interaction) and 1 (no interaction);[8] specifically, the S_{ij} weights can be related to aspects of translation elongation (tRNA, wobble interactions, and properties of the ribosome), as these aspects are expected to affect the efficiency of the codon–anticodon interaction. W_i values are calculated according to Crick's wobble rules for codon–anticodon pairing (Table 1). The codon relative adaptiveness value w_i is obtained by dividing each W_i with the maximum W_i value over all codons.[8] The tAI of a gene is defined as the geometric mean of the w_i values of its codons.

In 1966, Crick[37] suggested that in some cases wobble pairing may occur in the third base of the codon. According to Crick, the pairing at the third codon position has to obey chemical constrains; thus, some of the optional parings will not occur. For example, the unlikely pairing of guanine–adenine is due to a side group of guanine, which cannot make one of its bonds. In addition to the four standard nucleotides, modified nucleosides often occupy the wobble position of the anticodon (usually position 34 of the tRNA). In fact, the wobble position is the most frequently modified nucleoside in tRNA.[38,39] Inosine, for example, is a common modification of adenine that occurs in the wobble position of many tRNA species.[39–43]

Out of the eight S_{ij} weights, four are related to Watson–Crick (WC) interactions and the others are related to wobble interactions. In prokaryotic genomes, $\mathrm{tRNA}^{\mathrm{Ile}}$ has a unique wobble position nucleoside [lysidine (L) in bacteria and agmatidine (agm) in archaea], which recognizes the AUA codon;[44] thus, the prokaryotic S_{ij} set contains one additional wobble S_{ij} weight. The different possible pairs at the wobble position and the current tAI weights of the corresponding interactions are summarized in Table 2.

In the tAI, WC S_{ij} weights are fixed to zero (perfect interactions) under the assumption of no constraint on these interactions. The wobble interaction weights

Table 1. Crick's wobble rules for calculating W_i

	Codon third position		Anticodon first position	W_i
i	U	j	I	$(1 - s_{U:I})\mathrm{tGCN}_{i,j} + (1 - s_{U:G})\mathrm{tGCN}_{i,j+1}$
$i + 1$	C	$j + 1$	G	$(1 - s_{C:G})\mathrm{tGCN}_{i+1,j+1} + (1 - s_{C:I})\mathrm{tGCN}_{i+1,j}$
$i + 2$	A	$j + 2$	U	$(1 - s_{A:U})\mathrm{tGCN}_{i+2,j+2} + (1 - s_{A:I})\mathrm{tGCN}_{i+2,j}$
$i + 3$	G	$j + 3$	C	$(1 - s_{G:C})\mathrm{tGCN}_{i+3,j} + (1 - s_{G:U})\mathrm{tGCN}_{i+2,j+2}$

The W_i values are calculated based on Equation (1). The 64 codons are clustered in the genetic code into 16 groups, each one consists of four codons. The four codons in each group differ only in their third position (the wobble position). The formulas for calculating the W_i values for each of the four codons in the group are given in the table. i denotes the index of the codon in the quartet which ends with U, $i + 1$, $i + 2$, and $i + 3$ denote the three other codons which end with bases C, A, and G, respectively. j denotes the index of the tRNA whose anticodon starts with I; all base pairing between the ith codon and the jth anticodon are WC. $j + 1$, $j + 2$, and $j + 3$ denote the three other tRNAs whose anticodons start with bases G, U, and C, respectively. tGCN_{ij} represents the tRNA gene copy number corresponding to the interaction between the ith codon and the jth tRNA. For each codon, W_i sums over all tRNAs that can pair with the codon. For example, the GCU codon which ends with U can either pair with anticodons that start with I (IGC) and generate a standard WC base pairing, or pair with anticodons that start with G (GGC) and generate a wobble base pairing.

Table 2. The different base-pairings

j	*i*					
	I	G	U	C	A	L
I	—	—	*0*	**0.28**	**0.9999**	—
G	—	—	**0.41**	*0*	—	—
U	—	**0.68**	—	—	*0*	—
C	—	*0*	—	—	—	—
A	—	—	—	—	—	—
L	—	—	—	—	**0.89**	—

S_{ij}*-values* are given to the pairing between the first position of the *j*th anticodon (tRNA) and the third position of the *i*th codon. S_{ij}-values of WC base pairs are shown in italics, wobble values are shown in bold. Interactions which are not included in the calculation of the tAI are marked with hyphens. Lysidine (L) is a bacterial RNA modification of the DNA nucleotide cytidine (c).[44,45]

are inferred by optimizing the correlation between gene expression levels (mRNA levels) and their corresponding tAI in *Saccharomyces cerevisiae*;[8,46] the rationale behind this optimization is based on the following relations (which hold in many organisms): (i) there is correlation between mRNA levels and protein levels; (ii) there is correlation between translation rate and protein levels; and (iii) highly translated genes are under selection to include codons with higher adaptation to the tRNA pool.

The possibility of having different wobble interaction weights across different genomes has not yet been comprehensively studied. Here, we develop a novel generic approach for species-specific estimation of the tAI S_{ij} weights without the need of gene expression measurements; for convenience, we name the new measure: species-specific tAI (stAI). This measure includes different S_{ij} weights for each organism. We show that the correlation between protein levels and stAI is higher than that between protein levels and tAI.

Based on our approach, we infer the wobble S_{ij} weights for a wide variety of organisms from the three domains of life, in order to examine the conjecture that organisms from different domains have significantly different S_{ij} weights and to understand these differences.

2. Materials and methods

2.1. *Computing the S_{ij} weights of the stAI without the need of gene expression measurements*

The tAI weights are based on optimizing the correlation between tAI (Equation 1) and expression levels in *S. cerevisiae* and *E. coli.*[8] However, large scale measurement of mRNA levels and specifically protein abundance (PA) are not available for most of the organisms with sequenced genomes.

To solve this problem we develop an approach that is based on the assumption that highly expressed genes should have both higher adaptation to the tRNA pool (i.e. higher tAI) and higher CUB (i.e. less uniform distribution of codons).[8] Thus, there should be a significant correlation between CUB and tAI. Based on this assumption, we find the S_{ij}-values that optimize the correlation between CUB and stAI. Note that the optimized correlation is at the level of genes while for each gene both measures are based on its codons content. Below we provide additional details about our approach including the CUB measure that we use.

2.2. *Relative codon bias*

In order to infer the S_{ij} weights without the need of expression levels, we used a measure of CUB, which is based solely on the coding sequence. The strength of relative codon bias (RCBS) proposed by Roymondal *et al.*[47] is an example of an index that is based only on the sequence. The RCBS of codon *xyz* is expressed as:

$$d_{xyz} = \frac{f(x,y,z) - f_1(x) \cdot f_2(y) \cdot f_3(z)}{f_1(x) \cdot f_2(y) \cdot f_3(z)} \qquad (2)$$

where $f(x, y, z)$ is the observed frequency of codon *xyz* (where x, y, z denote the first/second/third nucleotides, respectively, of the codon) and $f_1(x)$, $f_2(y)$, and $f_3(z)$ are the observed frequencies of bases x, y, and z at, respectively, positions 1, 2, and 3 of the codon. These frequencies are computed for each gene separately. The RCBS of a gene of length L, in codons, is calculated as:

$$\mathrm{RCBS} = \left(\prod_{i=1}^{L}(1 + d^{i}{}_{xyz})\right)^{1/L} - 1 \qquad (3)$$

RCBS takes into account base compositional bias, to get a more reliable measure of highly favoured codon frequency while controlling for other features of the coding sequence such as GC content bias.

According to Equation (2), rare codons will be given lower d_{xyz} (i.e. a value close to -1) while a very frequent codon will be given a higher d_{xyz} value (e.g. it can be 1). Thus, very rare codons decrease the final RCBS score of the gene and very frequent ones increase its final RCBS score (see Equation 3). However, we believe that (almost by definition) genes with very high CUB should include both very frequent codons and very rare codons. For example, if a hypothetical amino acid A has two codons, one is 'optimal', and the second is 'not optimal', we expect a very highly expressed codon usage biased gene to have a very high d_{xyz} score for the first one and a very low d_{xyz} score for the second one. But, we wish that both cases/codons will contribute to the same direction and increase the RCBS score.

Thus, we employ a modified version of the RCBS, which we term here directional codon bias score (DCBS), as in this measure, both positive and negative codon usage biases contribute (in the same direction) to the total CUB of the gene. We define the directional codon bias (DCB) of a codon triplet *xyz* as:

$$d_{xyz} = \max\left(\frac{f(x,y,z)}{f_1(x)\cdot f_2(y)\cdot f_3(z)}, \frac{f_1(x)\cdot f_2(y)\cdot f_3(z)}{f(x,y,z)}\right) \tag{4}$$

The DCBS of a gene of length *L*, in codons, is the following mean (see example in Supplementary data):

$$\mathrm{DCBS} = \frac{\sum_{i=1}^{L} d_{xyz}}{L} \tag{5}$$

As we later demonstrate, in our framework the DCBS gives better results than the RCBS.

Finally, it is important to emphasize the fact that both RCBS and DCBS control for mutation bias. Specifically, when we compute the DCBS (see above), the frequency of each codon $[f(x,y,z)]$ is normalized by the expected frequency under the assumption that the different nucleotides are independent $[f_1(x)\cdot f_2(y)\cdot f_3(z)]$; the later represents among others the mutation bias. The measure that we use is based on the frequency of the codon normalized by the expected frequency according to the mutation bias, and thus control for mutational bias (see also Supplementary data regarding the way our approach controls for possible factors affecting CUB).

2.3. *Inferring the parameters of the stAI*

The stAI inferred here is based on the same equations of the tAI with an organism-specific S_{ij}-values' set (Equation 1), which is based on a measure of CUB. For every genome used in this study, the unique S_{ij} set was inferred by optimizing the non-parametric (Spearman) correlation between DCBS (Equations 4 and 5) and stAI (Equation 1). To avoid convergence to local maxima point, we used various starting points. Specifically, we included in the set of starting points the original weights of the tAI[8] and also the two extreme values of these weights (all zeros and all ones). In order to choose a set of starting points, we halved the allowed region of the S_{ij} values (i.e. the region: S_{ij} between 0 and 0.5, and the region: S_{ij} between 0.5 and 1) and considered all combinations for sampling values from these two regions (2^4 possibilities for the four eukaryotic wobble S_{ij} and 2^5 for the five prokaryotic wobble S_{ij}); thus, organisms from the same domain shared the same set of starting points. For each specific starting point, we used a hill climbing search method to iteratively optimize the S_{ij} weights using a variable step size (starting with an initial step size of 0.3 and finishing with step size of 0.001). At each step size, when a new optimum was not found, the step size was decreased by a factor of 1.35. Iteration of the hill climbing included a random choice of S_{ij} elements to change and a direction (i.e. increasing and decreasing) that increases the correlation between stAI and DCBS. The final chosen set of S_{ij} was the one exhibited the maximum correlation between the stAI and DCBS. In order to determine whether the chosen set of starting points constituted a sufficient sample of the search space for the algorithm convergence, we added 100 more random starting points. The additional points provided no significant change in the final correlation between stAI and DCBS.

2.4. *Comparison of the hill climbing method to Nedler–Mead search method*

The Nedler–Mead (NM) optimization[48] is the search method used to infer the S_{ij}-values of the original tAI.[8] When considering similar set of initiation points, our heuristic search outperformed the NM in finding the maxima of the objective function (i.e. the correlation between stAI and DCBS) in six of the eight model organisms (and was quite similar in the other two). We do not claim that hill climbing is better than NM; however, in the case of the specific problem analysed here (where the hill climbing explores the search space very well), and when considering the Matlab implementation of NM, the hill climbing was a bit better.

2.5. *The analysed organisms*

Our analysis included 100 different organisms (archaea, bacteria, and eukarya), in which CUB was correlated with the amount of adaptation to the tRNA pool. The correlation between stAI and DCBS/RCBS determined whether or not a tested genome would participate in the analysis. We excluded organisms in which an insignificant positive correlation or a significant negative correlation was observed; in such organisms, the assumptions that connect stAI to CUB do not hold and thus our method cannot be implemented. A detailed list of the excluded organisms is provided in Supplementary Table S1.

2.6. *Generating randomized genes sequences*

Random sequences were generated according to the real genomic codon distribution. For each of the 100 genomes studied in this work, 20 randomizations were performed by randomly drawing codons from the calculated genomic distribution and maintaining the protein content of the original genome.

2.7. *Genomic sequences*

In addition to the model organisms, which were chosen due to their available proteins measurements, we selected the genomes according to the list from ref.[49], while trying to build relatively balanced group size wise (thus since bacteria was significantly larger than other groups in the list, we only included organisms of the three major phylums provided there: *Cyanobacteria*, *Alphprobacteria*, *Gamma-probacteria*).

Coding sequences of all 100 species were retrieved from the NCBI (ftp://ftp.ncbi.nih.gov/genomes/). Genomic tRNA copy numbers of all species except *Aspergillus nidulans*, *Debaryomyces hanasenii*, and *Candida albicans* were obtained from the Genomic tRNA Database (http://gtrnadb.ucsc.edu/). For *A. nidulans*, *D. hanasenii*, and *C. albicans*, we used the tRNA copy number as reported in ref.[30]. A detailed list of all organisms analysed here is provided in Supplementary Table S2.

2.8. *Protein abundance*

Large scale protein abundance (PA) measurements of *S. cerevisiae*, *E. coli*, *Arabidopsis thaliana*, *Shigella dysentariae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Leptospira interrogans* were retrieved from paxdB (http://pax-db.org/#!home). For *S. cerevisiae*, *E. coli*, *S. dysentariae*, and *L. interrogans*, a few datasets were provided. In this case, a weighted average between the different PA values was taken (i.e. we averaged the datasets after normalizing each of them such that they have identical average). *Schizosaccharomyces pombe* expression levels were obtained from ref.[50]. The protein levels of some of the multiple cellular organisms were based on analysis of multiple tissues (*A. thaliana*, *D. melanogaster*, and *C. elegans*) (see details in http://pax-db.org/#!home). Specifically, we analysed all protein levels data that were available in paxdB (http://pax-db.org/#!home) on 2012. Note that in mammals it has been shown that the tRNA levels in various tissues tend to be correlative (the ranking of the tRNA genes abundance remains similar while the average value might change[51]); this is probably the case in many other organisms.

2.9. *Permutation test for comparing two S_{ij} means*

An empirical *P*-value was computed to test the null hypothesis that the means of two S_{ij} distributions do not significantly differ between two groups of organisms; let n and m denote the number of organisms in the two groups, respectively. For each S_{ij} component of the weights vector, we performed the following steps: first, we defined the normalized distance between the S_{ij} means in the two groups of organisms as the absolute difference between the means divided by the sum of the two corresponding standard deviations (SDs). Secondly, we permute the S_{ij} elements of the two groups by randomly drawing n values as the first group and m (non-overlapping) values as the second group. The random permutations were performed 100 times, each time the distance between the two random groups was computed. Finally, the *P*-value was defined as the number of times the random distance was higher or equal to the original distance divided by 100.

2.10. *Spearman correlation as a measure to guide the optimization*

The main advantage of this measure is the fact that it is a non-parametric measure that captures any monotonic relationship between CUB and stAI. Since this measure has been successfully employed in many papers in the field in this context,[18,52,53] we decided to use it also here.

2.11. *The general rational related to evaluating the stAI and demonstrating that stAI outperforms tAI*

In this section, we would like to explain and emphasize the rational related to the analyses reported in this study. First, as mentioned in the section Introduction, CUB measurements such as the CAI quantify different gene expression aspects than the tAI. Here, we aim at improving the tAI (and not the CUB indices such as the CAI) and thus, our major baseline for stAI evaluations is the tAI (and not the CUB indices such as the CAI). Secondly, we use the correlations with PA as an indirect way to evaluate the stAI: we expect that genes with higher translation efficiency will have higher PA; we also expect that a better measure related to the adaptation to the tRNA pool will have higher correlation with translation efficiency; thus, we expect that a better measure related to the adaptation to the tRNA pool will have higher correlation with PA. It is clear that there can be CUB-based measurements with higher correlation with PA than stAI (see, for example,[54])—however, as mentioned, the aim of this study is not to infer PA predictor but to improve the inference of the tAI parameters.

3. Results

3.1. *The correlation between the CUB and tRNA pool varies among different organisms*

A correlation between CUB and stAI is expected; however, the strength of this correlation among different organisms can teach us about the evolutionary forces shaping their genomes.

The correlations between stAI and DCBS obtained in the algorithm vary from a lowest value of 0.1136 (for the archaea *Halomicrobium mukohataei*) to a highest correlation of 0.7626 (for the fungi *Yarrowia* *lipolitica*). The bottom 10% correlations were obtained in prokaryotic genomes (the four archaea: *H. mukohataei*, *Archaeoglobus fulgidus*, *Pyrobaculum aerophilum*, and *Metallosphaera sedula*; and the six bacteria: *Anabaena variabilis*, *Brucella suis*, *Gloeobacter violaceus*, *Prochlorococcus marinus* MIT9313, *Synechococcus elongates*, and *Trichodesmium erythraeum*); thus, in this organisms, selection for CUB is presumably either weak or/and not strongly related to translation elongation and the tRNA pool.

The top 10% of the correlations were obtained mainly in eukaryotic genomes (the eight fungi: *C. albicans*, *C. glabrata*, *Eremothecium gossypii*, *Saccharomyces bayanus*, *S. mikatae*, *S. paradoxus*, *Cryptococcus neoformans*, and *Y. lipolitica*; and the two bacteria: *E. coli* and *Pasteurella multocida*); in these organisms, the selection for CUB is probably strongly related to the tRNA pool and translation elongations. All correlations are reported in Supplementary Table S3.

3.2. *The stAI exhibits better PA predictions than the tAI in non-fungal organisms*

The correlations between stAI and PA are presented in Fig. 1. All eight models showed significant correlations. In six of the eight organisms, the correlation between stAI and PA was higher than that between tAI and PA. This result (Table 3) indicates that stAI outperforms the current tAI as a predictor of PA in all non-fungal organisms. For the two fungi used here (*S. cerevisiae* and *S. pombe*), the original tAI predicted PA better than the stAI. This result is not surprising since the S_{ij}-values in the tAI were inferred based on the optimization of the correlation between tAI and *S. cerevisiae* mRNA expression levels[8] (which strongly correlates with PA in *S. cerevisiae*; Spearman correlation of 0.74, $P < 0.0001$[55]); on the other hand, stAI is based on CUB, which is a less accurate measure of protein levels. However, for most of the sequenced genomes exist to date, expression levels are not available; thus, the stAI is valuable.

We emphasize that although previous studies reported a significant positive correlation between CUB and expression levels in the model organisms studied here,[12,23,56,57] it is not trivial that *Sij* optimization based on CUB improves the correlation with protein levels. Specifically, CUB is correlated with protein levels, but mRNA levels and protein levels in different organisms are also usually correlated;[52,58,59] thus it is not clear that S_{ij} optimized based on the CUB of the organism necessarily have higher correlation with protein levels than the S_{ij} optimized based on mRNA levels of *S. cerevisiae*.

3.3. *Robustness analysis demonstrates that in non-fungal organisms the stAI outperforms the tAI in terms of the correlation with PA*

In order to empirically estimate the organism-specific probability that stAI (which is based on DCBS) improves the correlation with PA, a jack-knifing approach was implemented. One round of it involved the implementation of the algorithm for calculating the stAI on a sample of random subset of 50% of the

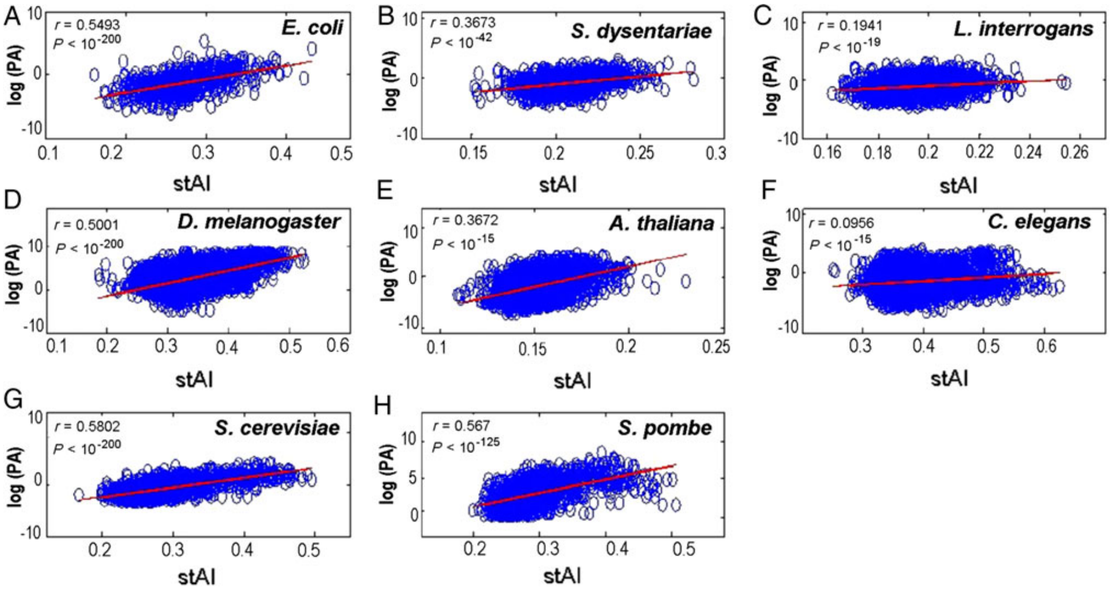

Figure 1. Dot plots of log(PA) vs. stAI and the corresponding Spearman rank correlations between stAI and PA. The correlations (and *P*-values) are calculated for the eight model organisms with PA measurements which include three bacteria (A–C), three non-fungal eukaryotes (D–F), and two fungi (G–H).

Table 3. Spearman rank correlation of the original tAI and the stAI with PA

	Number of genes	Number of proteins	*r* (tAI, PA)	*r* (stAI, PA)	Change (%)
Non-fungal					
E. coli	4145	688	0.5032	0.5493	+8.39
S. dysentariae	4501	1266	0.3574	0.36757	+2.76
L. interrogans	3667	2114	0.0959	0.19408	+50.58
A. thaliana	28,163	8478	0.3328	0.3762	+11.53
C. elegans	22,830	6959	0.0919	0.0956	+3.87
D. melanogaster	10,926	6510	0.4878	0.5001	+2.46
Fungi					
S. cerevisiae	5869	2666	0.6915	0.5802	−19.18
S. pombe	5017	1464	0.6554	0.56715	−15.58

The correlations between tAI and PA vs. the correlations between stAI and PA in eight model organisms with available PA data. The third column refers to the number of genes with available PA measurements in each organism.

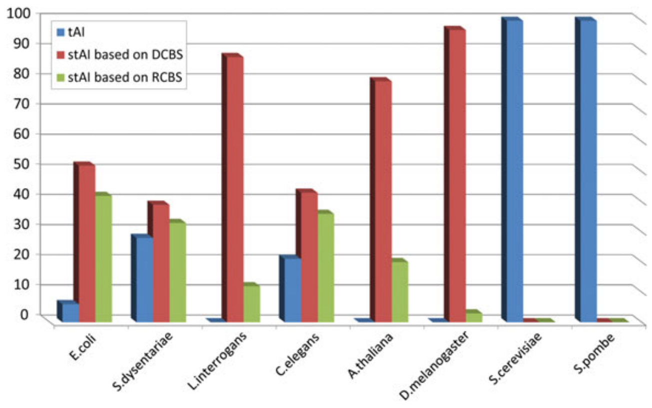

Figure 2. Comparison between stAI and the tAI. The middle bars representing the number of times (based on the jack-knifing analysis) the stAI outperformed the other versions of the tAI; as can be seen, stAI outperforms tAI in all non-fungal organisms.

proteins. Finally, the correlation between stAI and PA was computed for the sample and was compared with the correlation of PA with two related indices: the original tAI and stAI that is based on RCBS (i.e. its S_{ij} were inferred from RCBS and not from DCBS). This procedure was repeated 100 times where each time the index exhibited the highest correlation with PA was counted (Fig. 2).

As can be seen, the results demonstrate again that for non-fungal organisms, the species-specific inference of the S_{ij} tends to predict PA better than the traditional tAI. The 100 S_{ij} sets, their corresponding correlations between stAI and DCBS and the full taxonomy for each organism, are provided in Supplementary Table S3.

3.4. *S_{ij} inferred based on CUB are similar to the S_{ij} inferred based on PA*

In order to check whether the S_{ij} that are inferred based on CUB (i.e. based on the DCBS) converge to similar values as those which are based on expression levels, we computed S_{ij} sets by optimizing the correlation between stAI and PA for the model organisms with available PA measurements. This approach of using expression levels to optimize the tAI was employed in the study of ref.[8]. The Spearman rank correlation between the concatenated vectors of S_{ij}-values (35 points) inferred based on the DCBS and the one inferred based on PA is 0.6902 (P-value $<10^{-5}$; permutation P-value <0.001; 35 points). The Euclidean distance between the two vectors is also significantly lower than the one obtained by random permutation of the two vectors; specifically, when we performed 1000 permutations of these values, all of them had higher Euclidean distance (P-value <0.001). The S_{ij}-values that were obtained via correlation with DCBS and the ones obtained via correlation with PA are provided in Supplementary Table S4.

3.5. *Considering all tRNA–codon pairing possibilities do not improve the performances of the stAI*

There are possible cases of non-standard base pairing that currently are not included in the tAI wobble rules (U–U binding for instance). It is interesting to check whether introducing such additional rules to the model can improve its performances. Using Equation (1), we included in the set of S_{ij} all missing pairing options (U:U, C:U, U:C, C:C, C:A, G:A, I:G, and G:G). An initial weight of 0.5 was given to all non-WC S_{ij} (WC S_{ij} are fixed to zero). Nevertheless, considering all possible pairings in the stAI weights calculation did not improve the correlation of the stAI with DCBS or with PA. The original approach (i.e. WC and wobble only) reached higher maxima values for seven of the eight models. In addition, for five of the eight models, better correlations with PA, were obtained for the original stAI.

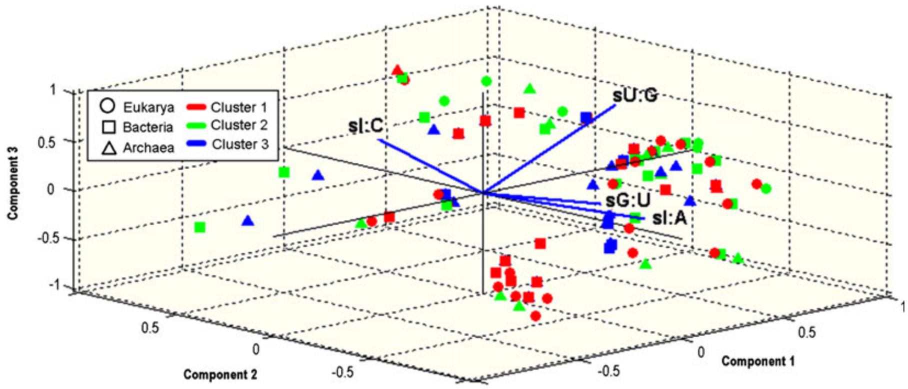

Figure 3. Principal component analysis (PCA) on the 100 different S_{ij} sets demonstrates clustering of S_{ij} according to evolutionary domains. The first three components of the *PCA* are presented. Each point in the figure represents one of the 100 analysed organisms; the shape of the point corresponds to the domain of the organism at the tree of life and the colour corresponds to the cluster the point was classified based on the *k*-means algorithm.

3.6. *Adding constraint on WC interactions do not improve the performances of the stAI*

The first four S_{ij} weights, which represent the constraint on WC interactions, are fixed to zero. Assuming that these interactions might not be perfect and thus allowing them to change during the optimization did not provide further improvement. The starting point for each model was the one exhibited the maximum correlation between stAI and DCBS in the original search. The original search, in which WC S_{ij} are fixed to zero, reached higher maxima values for five of the eight models; in addition, in five of the eight models, the original search exhibited a better correlations with PA than the new search.

3.7. *Distances between the inferred S_{ij}-values correlate with evolutionary proximity*

In the next step, we aimed at understanding if the organism-specific S_{ij}-values reflect evolutionary proximity and if they are biologically meaningful. To this end, the inferred S_{ij}-value sets of the 100 organisms were clustered into three groups using the *k*-means algorithm.[60] We compared the clustering result to the clusters obtained by partitioning the organisms to the three domains of life. The clustering correctly classified 77% of the 26 eukarya, 45% of the 38 bacteria, and 67% of the 36 archaea (Fig. 3). In general, 61% of the total 100 organisms were classified into the correct domain.

Properly randomized genomes that were generated by maintaining the CUB of the genome and its protein content were used to empirically test the significance of this clustering (see section 2). None of the 20 randomizations outperformed the original clustering (with respect to total correct classifications, empirical *P*-value <0.05). This result demonstrates that with high probability the reported clustering cannot be obtained randomly even when considering randomized genomes with similar features to the original ones (global CUB and the same proteome), supporting the conjecture that the obtained S_{ij}-value similarities correlate with the evolutionary distances and thus have biological meaning.

Table 4. The mean inferred wobble S_{ij}-values

	SG:U	SI:C	SI:A	SU:G	SL/agm:A
Eukarya	0.7861	0.4659	0.9075	0.6295	—
Bacteria	0.6294	0.4211	0.8773	0.698	0.7309
Archaea	0.3898	0.3774	0.5015	0.4363	0.6453
Mean	0.6	0.42	0.76	0.588	0.6881

The mean inferred wobble S_{ij}-values strength for each domain of life and for the entire analysed dataset (last row).

Finally, it is important to mention that there is co-evolution between CUB and tRNA levels (see, for example,[24,34,61,62]). Specifically, based on various theories, the CUB should co-evolve with the tRNA pool and the tRNA–codon interaction efficiencies to optimize the organism fitness; thus, CUB, tRNA levels, and tRNA–codon interactions cannot be separated.

3.8. *Similarities among the inferred S_{ij}-values of the analysed organisms*

The mean efficiency of the different inferred codon–anticodon interactions over all the analysed organisms are summarized in Table 4. The results emphasize the similarities among the different organisms and domains.

As mentioned, S_{ij}-values are between 0 and 1. Since these values represent a constraint on the codon–anticodon interactions, interactions with lower values are considered more efficient. For example, it can be seen from Table 4 that the inosine–cytosine interaction has the lowest mean value (*sI:C* $= 0.42$), while

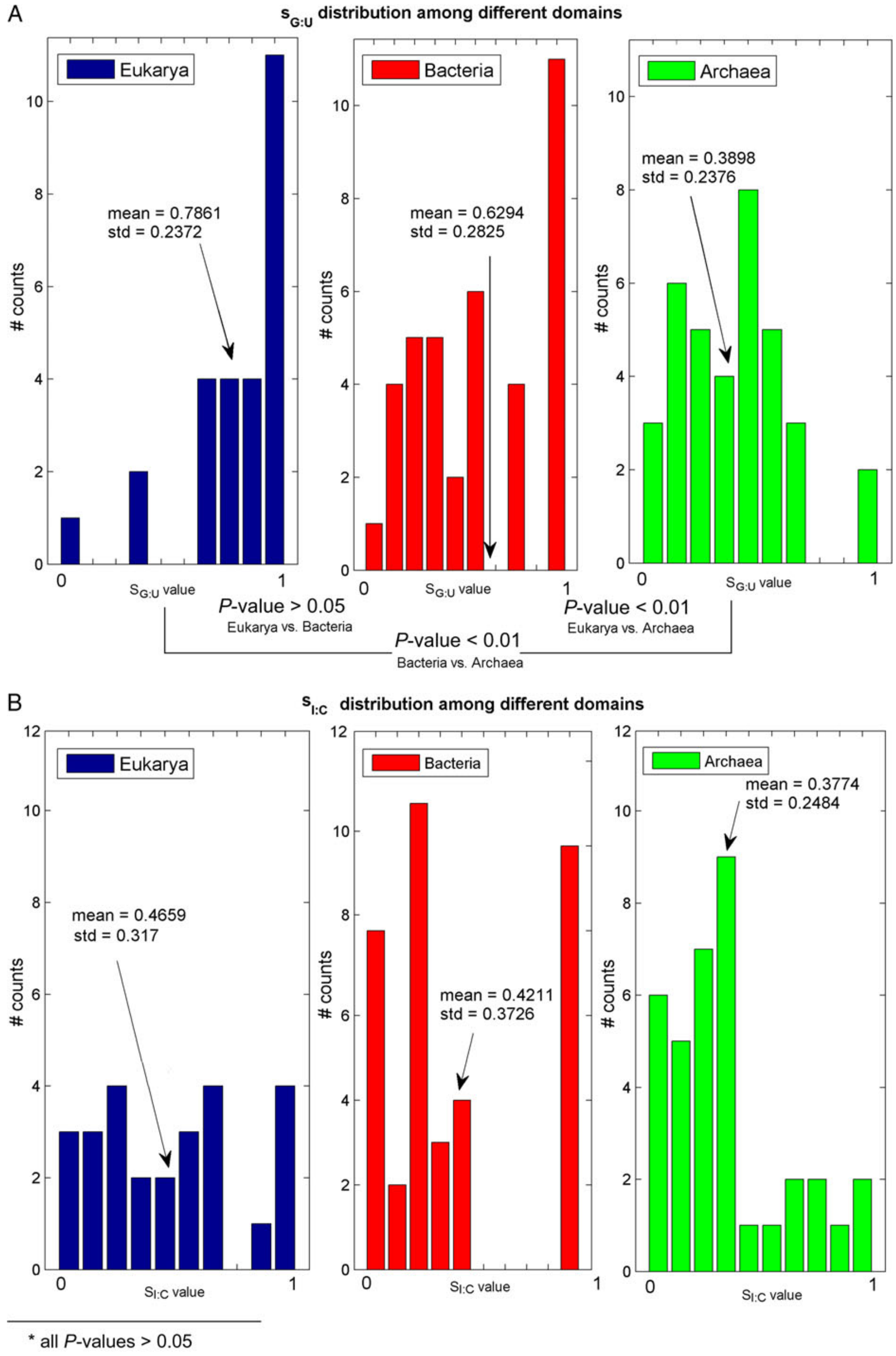

Figure 4. S_{ij} distributions among different domains of life. Each figure contains three histograms representing the S_{ij} in the different domains of life; the mean and SD of the S_{ij}-values in each domain are also reported. The P-values corresponding to the comparison between every two S_{ij} means appear in the bottom of the figure (see section 2 sub-section 'Permutation test for comparing two S_{ij} means').

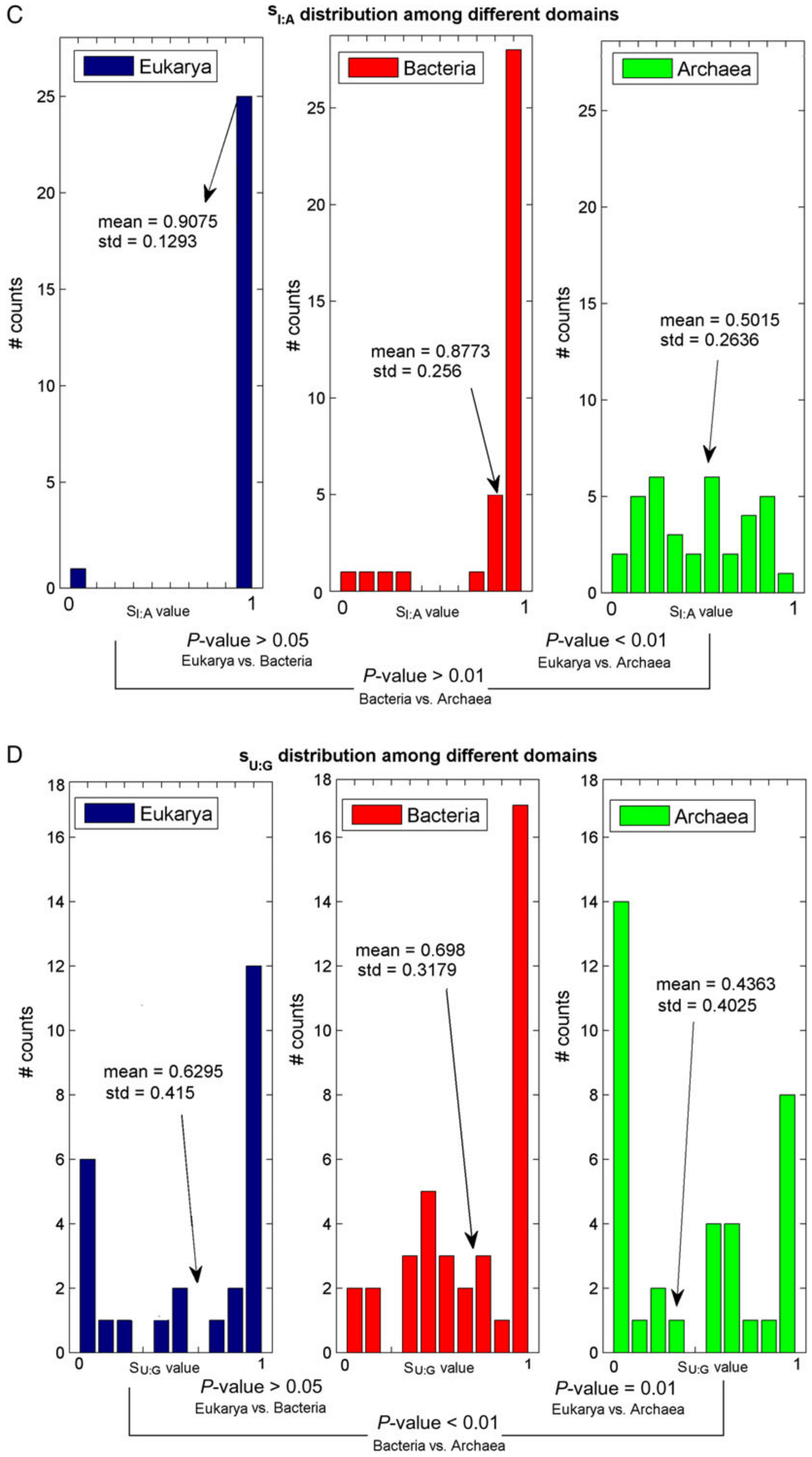

Figure 4. *Continued*

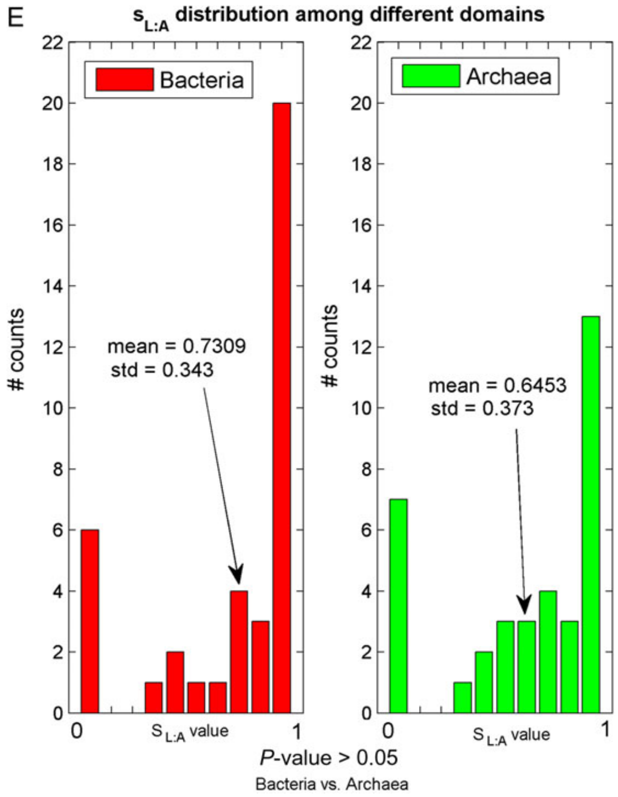

Figure 4. *Continued*

the wobble inosine–adenosine has the largest mean value ($sI{:}A = 0.76$). This suggests a good I:C interaction and an inefficient I:A interaction. These findings are supported by Murphy and Ramakrishnan.[63] where it is stated that the decoding of adenosine-ending codons by inosine is inefficient. It is also mentioned that the inosine–cytosine interaction is very similar to the canonical G:C pair.

SL:A and *Sagm:A* have a similar distribution, and the corresponding *P*-value proves that the mean values are not significantly different (see Fig. 4E). Since agmatidine is in many ways similar to lysidine (see ref.[64]), it makes sense that their S_{ij}-values are similar.

3.9. *Differences among the inferred S_{ij}-values of different groups of organisms*

To test the hypothesis that the S_{ij}-values of different organisms groups (i.e. different domains or different phylums within the same domain) have significantly different means, we computed an empirical permutation *P*-value (see section 2). The S_{ij} distributions and their corresponding *P*-values are presented in Fig. 4.

As can be seen, the *sI:C* distribution is similar between the three domains (Fig. 4B); however, *sU:G*, *sI:A*, and *sG:U* tend to be significantly different among the three domains. An empirical *P*-value was used also for the comparison between the two major phylums within each domain. The only significant difference was obtained for the *sI:A* distribution of eukarya subgroups *Opisthokonta* vs. *Viridiplantae* and bacteria subgroups *Proteobacteria* vs. *Cyanobacteria* (see Fig. 5). All other insignificant S_{ij} distributions among different phylums appear in Supplementary Fig. S5.

4. Discussion

In this study, we describe a new approach for inferring the efficiency of wobble interactions in the tAI without prior knowledge about the expression levels of the analysed organism. The approach is based on the fact that in most organisms highly expressed genes have higher CUB which is, at least partially, due to selection for improved adaptation of the codons to the tRNA pool of the organism. With our approach we infer the efficiency of wobble interactions via optimizing the component of the CUB that is due to adaptation to the tRNA pool (i.e. the correlation between these two measures: CUB and adaptation to the tRNA pool).

Thus, one limitation of our approach (and other CUB-based approaches) is the fact that it will not work in the case of organisms with no strong enough selection for both CUB and the adaptation to the tRNA pool in highly expressed genes; specifically, we assume that the evolutionary selection for this two phenomena tend to be stronger when the gene expression is higher.

In addition, we show that with our approach we are able to infer the efficiency of wobble interactions in non-fungal organisms better than the conventional approach (the tAI that does not optimize these values for each organism separately). In addition, we provide the estimations of these values for 100 organisms and show that they vary among different organism and correlate with evolutionary proximity. We report the similarities and differences among the inferred efficiencies of the analysed organisms.

PA measurements rather than mRNA level measurements are more appropriate for estimating the extent to which a coding sequence feature is related to translation efficiency. Thus, the improved correlation between stAI and PA exhibited for the non-fungal model organisms relatively to the correlation between tAI and PA demonstrates the advantages of our novel approach. Specifically, the improved correlation between stAI and PA indicates a strong association between translation efficiency (and thus PA), and the combined information the stAI provides which includes the co-adaptation of CUB to the tRNA pool, and the efficiency of the different wobble interactions.

Currently, there are less than a few dozen large scale measurements of protein levels, while there are >25,000 sequenced genomes. In addition, in the case of most of the organisms on earth, it is much easier to sequence their genomes, while it is usually impossible

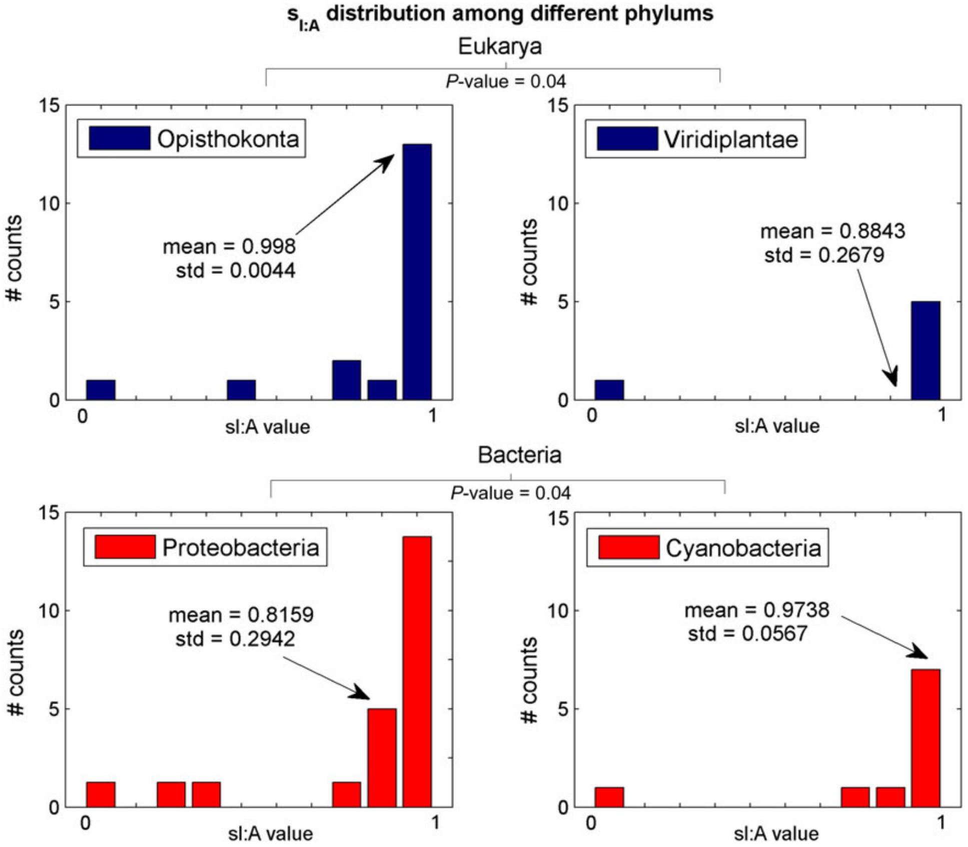

Figure 5. *sI:A* distribution within the major phylums of the eukaryotic and bacterial domains with a significant empirical *P*-value (see details in section 2).

to culture them in order to measure their protein levels (see, for example,[65]). Our approach can improve the study of translation and evolution in such organisms, even if there are no available gene expression measurements.

The idea of different domains having different wobble S_{ij}-values is supportive with the successful significant clustering reported in this study. The differences between the bacterial and eukaryotic ribosomes[66,67] might provide a plausible explanation to this result as specific physical, chemical, and geometrical constraints are imposed on each tRNA–codon interaction. In the budding yeast, for example, the wobble inosine tRNA modification is essential for viability.[40] This result is in line with a recent study[68] that two kingdom-specific tRNA modifications are major contributors that separate archaeal, bacterial, and eukaryal genomes in terms of their tRNA gene composition. Specifically, with our approach, we were able to provide information about the interaction efficiencies that tend to vary among the different domains (*sU:G*, *sI:A*, and *sG:U*) and within some of the domains (*sI:A*); in addition, we show that the efficiencies of some of the interactions are conserved in all the domains (*sI:C*). Combining this information with additional information such as phylogenetic analysis, three dimensional conformations of the ribosome and tRNA molecules and knowledge related to tRNA modifications can provide a better understanding of the exact structure of the ribosome and tRNA molecules, their biochemical interactions, and their evolution.

We further verified that modelling non-conventional interactions between nucleotides does not significantly improve our model. Thus, our analysis supports the conjecture that, in the analysed organisms, wobble/WC interactions/parameters that appear in the original tAI measure should not be updated.

Finally, there has been a debate about the causality of the tRNA adaptation/protein level relations. Some previous studies suggested that increasing the adaptation to the tRNA pool has direct effect on translation rate and thus on protein levels.[34,69] However, other studies have suggested that this relation is not causal: endogenous highly expressed genes have higher adaptation to the tRNA pool via reasons that are not directly related to the translation rate.[62,70] For example, it has been suggested that the adaption of highly expressed genes to the tRNA pool improves the global ribosomal allocation among genes based on the fact that genes with higher adaptation to the tRNA pool consume less ribosomes;[9,70,71] it was also suggested that evolution maintains a balance between codon frequency and the cellular levels of the tRNA genes such that the actual translation elongation speed is constant;[62] highly expressed genes have higher adaptation to the tRNA pool since the effect of these genes on maintaining this balance is higher than in the case of lowly expressed genes.[62] It is important to mention that the success of our approach is robust to the outcome of this debate. The fact that highly expressed genes have higher adaptation to the tRNA pool as reflected by the wobble interactions and the cellular tRNA levels is enough for the success of our approach, the exact biophysical/evolutionary mechanism does not matter.

Acknowledgement: We thank Ms Hadas Zur for helpful discussions and comments.

Supplementary Data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

References

1. Kimura, M. 1968, Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles, *Genet. Res.*, **11**, 247–69.
2. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. 1980, Codon catalog usage and the genome hypothesis, *Nucleic Acids Res.*, **8**, 197–.
3. Plotkin, J.B. and Kudla, G. 2010, Synonymous but not the same: the causes and consequences of codon bias, *Nat. Rev. Genet.*, **12**, 32–42.
4. Behura, S.K. and Severson, D.W. 2013, Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes, *Biol. Rev. Camb. Philos. Soc.*, **88**, 49–61.
5. Behura, S.K. and Severson, D.W. 2011, Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in *Aedes aegypti* and *Anopheles gambiae*, *Insect Mol. Biol.*, **20**, 177–87.
6. Behura, S., Stanke, M., Desjardins, C., Werren, J. and Severson, D. 2010, Comparative analysis of nuclear tRNA genes of *Nasonia vitripennis* and other arthropods, and relationships to codon usage bias, *Insect Mol. Biol.*, **19**, 49–58.
7. Stergachis, A.B., Haugen, E., Shafer, A., et al. 2013, Exonic transcription factor binding directs codon choice and affects protein evolution, *Science*, **342**, 1367–72.
8. dos Reis, M., Savva, R. and Wernisch, L. 2004, Solving the riddle of codon usage preferences: a test for translational selection, *Nucleic Acids Res.*, **32**, 5036–44.
9. Tuller, T., Waldman, Y.Y., Kupiec, M. and Ruppin, E. 2010, Translation efficiency is determined by both codon bias and folding energy, *Proc. Natl. Acad. Sci. USA*, **107**, 3645–50.
10. Arenas, M. and Posada, D. 2010, Coalescent simulation of intracodon recombination, *Genetics*, **184**, 429–37.
11. Bennetzen, J.L. and Hall, B. 1982, Codon selection in yeast, *J. Biol. Chem.*, **257**, 3026–31.
12. Grosjean, H. and Fiers, W. 1982, Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes, *Gene*, **18**, 199–209.
13. Wright, F. 1990, The ‘effective number of codons’ used in a gene, *Gene*, **87**, 23–9.
14. Kurland, C. 1991, Codon bias and gene expression, *FEBS Lett.*, **285**, 165–9.
15. Stenico, M., Lloyd, A.T. and Sharp, P.M. 1994, Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases, *Nucleic Acids Res.*, **22**, 2437–46.
16. 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**, 1281–95.
17. Ishihama, Y., Schmidt, T., Rappsilber, J., et al. 2008, Protein abundance profiling of the *Escherichia coli* cytosol, *BMC Genomics*, **9**, 102.
18. Ghaemmaghami, S., Huh, W.K., Bower, K., et al. 2003, Global analysis of protein expression in yeast, *Nature*, **425**, 737–41.
19. Washburn, M.P., Wolters, D. and Yates, J.R. 2001, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat. Biotechnol.*, **19**, 242–7.
20. Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. and Arnold, F.H. 2005, Why highly expressed proteins evolve slowly, *Proc. Natl. Acad. Sci. USA*, **102**, 14338–43.
21. Blattner, F.R., Plunkett, G., Bloch, C.A., et al. 1997, The complete genome sequence of *Escherichia coli* K-12, *Science*, **277**, 1453–62.
22. Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., et al. 2007, A ‘silent’ polymorphism in the MDR1 gene changes substrate specificity, *Science*, **315**, 525–8.
23. Fox, J.M. and Erill, I. 2010, Relative codon adaptation: a generic codon bias index for prediction of gene expression, *DNA Res.*, **17**, 185–96.
24. 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.
25. Ikemura, T. 1985, Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, **2**, 13–34.
26. Seffens, W. and Digby, D. 1999, mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences, *Nucleic Acids Res.*, **27**, 1578–84.
27. Ohama, T., Muto, A. and Osawa, S. 1990, Role of GC-biased mutation pressure on synonymous codon choice

in *Micrococcus luteus* a bacterium with a high genomic GC-content, *Nucleic Acids Res.*, **18**, 1565–9.
28. 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**, 573–97.
29. Percudani, R., Pavesi, A. and Ottonello, S. 1997, Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*, *J. Mol. Biol.*, **268**, 322–30.
30. Man, O. and Pilpel, Y. 2007, Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species, *Nat. Genet.*, **39**, 415–21.
31. 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**, 1141–53.
32. Chamary, J., Parmley, J.L. and Hurst, L.D. 2006, Hearing silence: non-neutral evolution at synonymous sites in mammals, *Nat. Rev. Genet.*, **7**, 98–108.
33. Shabalina, S.A., Ogurtsov, A.Y. and Spiridonov, N.A. 2006, A periodic pattern of mRNA secondary structure created by the genetic code, *Nucleic Acids Res.*, **34**, 2428–37.
34. Tuller, T., Carmi, A., Vestsigian, K., et al. 2010, An evolutionarily conserved mechanism for controlling the efficiency of protein translation, *Cell*, **141**, 344–54.
35. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. 2011, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes, *Cell*, **147**, 789–802.
36. 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**, 143–55.
37. Crick, F. 1966, Codon-anticodon pairing: the wobble hypothesis, *J. Mol. Biol.*, **19**, 548–55.
38. RajBhandary, U.L. 1988, Genetic code. Modified bases and aminoacylation, *Nature*, **336**, 112.
39. Agris, P. 1991, Wobble position modified nucleosides evolved to select transfer RNA codon recognition: a modified-wobble hypothesis, *Biochimie*, **73**, 1345–9.
40. Tsutsumi, S., Sugiura, R., Ma, Y., et al. 2007, Wobble inosine tRNA modification is essential to cell cycle progression in G1/S and G2/M transitions in fission yeast, *J. Biol. Chem.*, **282**, 33459–65.
41. Wolf, J., Gerber, A.P. and Keller, W. 2002, tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*, *EMBO J.*, **21**, 3841–51.
42. Delannoy, E., Le Ret, M., Faivre-Nitschke, E., et al. 2009, Arabidopsis tRNA adenosine deaminase arginine edits the wobble nucleotide of chloroplast tRNAArg (ACG) and is essential for efficient chloroplast translation, *Plant Cell*, **21**, 2058–71.
43. Maas, S., Gerber, A.P. and Rich, A. 1999, Identification and characterization of a human tRNA-specific adenosine deaminase related to the ADAR family of pre-mRNA editing enzymes, *Proc. Natl. Acad. Sci. USA*, **96**, 8895–900.
44. Ikeuchi, Y., Soma, A., Ote, T., Kato, J., Sekine, Y. and Suzuki, T. 2005, Molecular mechanism of lysidine synthesis that determines tRNA identity and codon recognition, *Mol. Cell*, **19**, 235–46.
45. Lim, V.I. and Curran, J.F. 2001, Analysis of codon: anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure, *RNA*, **7**, 942–57.
46. 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**, 6976–85.
47. Roymondal, U., Das, S. and Sahoo, S. 2009, Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome, *DNA Res.*, **16**, 13–30.
48. Nelder, J.A. and Mead, R. 1965, A simplex method for function minimization, *Comput. J.*, **7**, 308–13.
49. Tuller, T., Birin, H., Gophna, U., Kupiec, M. and Ruppin, E. 2010, Reconstructing ancestral gene content by coevolution, *Genome Res.*, **20**, 122–32.
50. Schmidt, M.W., Houseman, A., Ivanov, A.R. and Wolf, D.A. 2007, Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*, *Mol. Syst. Biol.*, **3**, 79.
51. Mahlab, S., Tuller, T. and Linial, M. 2012, Conservation of the relative tRNA composition in healthy and cancerous tissues, *RNA*, **18**, 640–52.
52. Tuller, T., Kupiec, M. and Ruppin, E. 2007, Determinants of protein abundance and translation efficiency in *S. cerevisiae*, *PLoS Comput. Biol.*, **3**, e248.
53. Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E.M. 2006, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation, *Nat. Biotechnol.*, **25**, 117–24.
54. Zur, H. and Tuller, T. 2013, Transcript features alone enable accurate prediction and understanding of gene expression evolution in *S. cerevisiae*, *BMC Bioinfomatics*, **14** (Suppl 15), S1.
55. Futcher, B., Latter, G., Monardo, P., McLaughlin, C. and Garrels, J. 1999, A sampling of the yeast proteome, *Mol. Cell. Biol.*, **19**, 7357–68.
56. Duret, L. and Mouchiroud, D. 1999, Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, *Proc. Natl. Acad. Sci.*, **96**, 4482–7.
57. Coghlan, A. and Wolfe, K.H. 2000, Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*, *Yeast*, **16**, 1131–45.
58. Greenbaum, D., Colangelo, C., Williams, K. and Gerstein, M. 2003, Comparing protein abundance and mRNA expression levels on a genomic scale, *Genome Biol.*, **4**, 117.
59. de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. 2009, Global signatures of protein and mRNA expression levels, *Mol. BioSyst.*, **5**, 1512–26.
60. MacQueen, J. 1967, Some methods for classification and analysis of multivariate observations. In: *Proceedings of the*

fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, 281–97.

61. Rocha, E.P. 2004, Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization, *Genome Res.*, **14**, 2279–86.
62. Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C. and Zhang, J. 2012, Balanced codon usage optimizes eukaryotic translational efficiency, *PLoS Genet.*, **8**, e1002603.
63. Murphy, F.V. and Ramakrishnan, V. 2004, Structure of a purine-purine wobble base pair in the decoding center of the ribosome, *Nat. Struct. Mol. Biol.*, **11**, 1251–2.
64. Mandal, D., Köhrer, C., Su, D., et al. 2010, Agmatidine, a modified cytidine in the anticodon of archaeal tRNAIle, base pairs with adenosine but not with guanosine, *Proc. Natl. Acad. Sci.*, **107**, 2872–7.
65. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W. and Nielsen, P.H. 2013, Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes, *Nat. Biotechnol.*, **31**, 533–8.
66. Hardesty, B. and Kramer, G., *Structure, function, and genetics of ribosomes*, Springer: NY, 1986.
67. Melnikov, S., Ben-Shem, A., de Loubresse, N.G., Jenner, L., Yusupova, G. and Yusupov, M. 2012, One core, two shells: bacterial and eukaryotic ribosomes, *Nat. Struct. Mol. Biol.*, **19**, 560–7.
68. Novoa, E.M., Pavon-Eternod, M., Pan, T. and Ribas de Pouplana, L. 2012, A role for tRNA modifications in genome structure and codon usage, *Cell*, **149**, 202–13.
69. Frenkel-Morgenstern, M., Danon, T., Christian, T., et al. 2012, Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels, *Mol. Syst. Biol.*, **8**, 572.
70. Kudla, G., Murray, A.W., Tollervey, D. and Plotkin, J.B. 2009, Coding-sequence determinants of gene expression in *Escherichia coli*, *Science*, **324**, 255–8.
71. Supek, F. and Šmuc, T. 2010, On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*, *Genetics*, **185**, 1129–34.