Wei et al. BMC Evolutionary Biology 2013, 13:162
http://www.biomedcentral.com/1471-2148/13/162

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Transcriptional abundance is not the single force driving the evolution of bacterial proteins

Wen Wei, Tao Zhang, Dan Lin, Zu-Jun Yang and Feng-Biao Guo*

Abstract

Background: Despite rapid progress in understanding the mechanisms that shape the evolution of proteins, the relative importance of various factors remain to be elucidated. In this study, we have assessed the effects of 16 different biological features on the evolutionary rates (ERs) of protein-coding sequences in bacterial genomes.

Results: Our analysis of 18 bacterial species revealed new correlations between ERs and constraining factors. Previous studies have suggested that transcriptional abundance overwhelmingly constrains the evolution of yeast protein sequences. This transcriptional abundance leads to selection against misfolding or misinteractions. In this study we found that there was no single factor in determining the evolution of bacterial proteins. Not only transcriptional abundance (codon adaptation index and expression level), but also protein-protein associations (PPAs), essentiality (ESS), subcellular localization of cytoplasmic membrane (SLM), transmembrane helices (TMH) and hydropathicity score (HS) independently and significantly affected the ERs of bacterial proteins. In some species, PPA and ESS demonstrate higher correlations with ER than transcriptional abundance.

Conclusions: Different forces drive the evolution of protein sequences in yeast and bacteria. In bacteria, the constraints are involved in avoiding a build-up of toxic molecules caused by misfolding/misinteraction (transcriptional abundance), while retaining important functions (ESS, PPA) and maintaining the cell membrane (SLM, TMH and HS). Each of these independently contributes to the variation in protein evolution.

Keywords: Evolutionary rates, Bacteria, Multiple features, Transcriptional abundance

Background

Amino acid substitution rates vary considerably among different proteins. Although rapid progress has been made in determining the most important factors that shape protein evolution, the challenge remains to assess the relative importance of various variables, such as gene expression level, essentiality (ESS) and protein interactions [1-10]. One early study [11] proposed a negative correlation between the severity of gene knockout effects and coding sequence evolution, which was dependent upon the notion that purifying selection should be more efficient for essential genes than those that are non-essential. A link has been discovered between protein expression levels and evolutionary rates (ERs) in both unicellular and multicellular organisms [7,12-19].

In general, genes that are highly expressed preferentially use optimal codons to improve translational efficiency. The codon adaptation index (CAI), a measure of synonymous codon usage bias, has been widely used as a proxy for gene expression levels [20]. When CAI values were used as a substitute for actual expression levels in yeast [2] and bacteria [12], only a small proportion of rate variation in protein evolution can be explained by ESS. After replacing CAI values with experimental data and controlling for gene expression levels, ESS still had significant effects on protein ERs, but did not appear to be a major determinant of protein evolution [21,22]. CAI, expression level, and protein abundance can account for most of the variation in yeast protein ERs [13]. Keeping proteins from misfolding or misinteraction result in the slow evolution of highly expressed genes, and impose a general constraint on coding sequence evolution [7,23]. However, by using noiseless variables, protein interactions have explained more ER variation than transcriptional abundance [24]. Results from another study suggest that

* Correspondence: fbguo@uestc.edu.cn
Center of Bioinformatics and Key Laboratory for NeuroInformation of the Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, 610054 Chengdu, China

the molecular evolution of protein-coding genes is affected by both the context of extrinsic translational expression rates and intrinsic structural-functional constraints [25].

Despite the large number of studies evaluating the effects of various mechanisms during protein evolution, the relative importance of these factors compared with transcriptional abundance remains to be elucidated. In this study, we investigated the effects of various biological features on protein-coding sequence ERs for 18 bacterial species. The following genomic variables: CAI; experiment-based expression level (EL); ESS; number of protein-protein associations (PPA); mRNA folding strength (MFS); hydropathicity score (HS); aromaticity score (AS); protein length (LEN); replication strand bias (RSB); number of transmembrane helices (TMH); and subcellular localization [cytoplasm (SLC), cytoplasmic membrane (SLM), periplasm (SLP), outer membrane (SLO), extracellular (SLE), and cell wall (SLW)] have been summarized (Table 1).

Results and discussion

Genomic feature correlates of protein ER

In this study, potential correlation between ER and 16 features (Table 1) for 18 bacterial species (Table 2) was investigated. We observed strong correlations among CAI, EL, PPA, SLC, and ER; and less strong but significant correlations among ESS, SLM, TMH, AS and ER. However, weak relationships were found among MFS, SLP, SLO, SLE, SLW, HS, LEN, RSB and ER.

Table 1 Bacterial features examined in this study

Feature	Abbreviation	Type	Source
Essentiality	ESS	binary	DEG
Evolutionary rate (Ka)	ER	real	PAML
Codon adaption index	CAI	real	Python script
Expression level	EL	real	GEO
mRNA folding strength	MFS	real	ViennaRNA
Number of protein–protein associations	PPA	integer	STRING
Subcellular localization: cytoplasm	SLC	real	PSORTb
Subcellular localization: cytoplasmic membrane	SLM	real	
Subcellular localization: periplasm	SLP	real	
Subcellular localization: outer membrane	SLO	real	
Subcellular localization: extracellular	SLE	real	
Subcellular localization: cell wall	SLW	real	
Number of transmembrane helices	TMH	integer	TMHMM
Hydropathicity score	HS	real	CodonW
Aromaticity score	AS	real	
Length of protein in amino acids	LEN	integer	
Replication strand bias	RSB	binary	DoriC

Transcriptional abundance

A single variable linked to transcriptional abundance (CAI, EL and protein abundance) was found to explain the dominance of observed variation in yeast ERs [13]. CAI and EL are related to transcriptional abundance, while protein abundance is a result of the combined consequences of transcription and translation. Recent studies observed that MFS was strong for more abundant proteins, resulting in stronger evolutionary constraints of more highly expressed proteins [26,27]. We used three variables (CAI, EL and MFS) to highlight the impact of transcriptional abundance on ER.

We used RNAfold (http://www.tbi.univie.ac.at/~ronny/RNA/) to predict the secondary structure of RNA and to compute the strength of mRNA folding. A recent study reported that RNAfold predicted MFS is moderately correlated with experimentally determined MFS [27]. Furthermore, it was also shown that the correlation of the computationally predicted MFS and ER was much weaker than that of the experimentally determined MFS and ER. Similarly, we also observed a low correlation between MFS predicted by RNAfold and ER in bacteria. However, in our study we found that CAI and EL were significantly linked to ERs for most bacteria, depending on rank correlation coefficients (Figure 1). The top absolute coefficients were dominated (12/18 bacterial species) by CAI-ER or EL-ER correlations. The CAI-ER coefficient in *Escherichia coli* was −0.464 ($p = 5.45 \times 10^{-107}$), which is greater than other coefficients. Of these bacteria, *Helicobacter pylori* showed no CAI-ER correlation (rho = −0.039, $p > 0.05$); this species was not subject to periods of competitive exponential growth [28]. As a result, there was a lack of translational selection related to codon usage. In an early study, codon selection for translation was observed to strongly correlate with growth rates [29]. We investigated the effects of growth on CAI-ER correlations, and found that weak correlations could be partially attributed to long-term bacterial generation times (Table 2).

To further investigate the impact of translational selection on CAI-ER correlation, codon usage separation (CUS) was used to measure the strength of codon bias. A greater CUS value indicated a stronger codon bias mediated by translational selection. We confirmed a correlation between CUS and the CAI-ER coefficient (Pearson's r = −0.730, $p = 5.83 \times 10^{-4}$). Significantly different codon usage (CUS = 0.879) was found between ribosomal proteins and other proteins in the *E. coli* genome (Figure 2A); however, this was not observed in the *H. pylori* genome (CUS = 0.148; Figure 2B). The degree to

Table 2 Bacterial species investigated in this study

Organism	Abbreviation	Codon usage separation (CUS)	GC content	Generation times [29]
Acinetobacter ADP1	Abay	0.845	40.4	0.5
Bacillus subtilis 168	Bsub	0.869	43.5	0.43
Bacteroides thetaiotaomicron VPI-5482	Bthe	0.291	42.8	1.47
Caulobacter crescentus NA1000	Ccre	0.754	67.2	1.5
Escherichia coli K-12	Ecol	0.879	50.8	0.35
Francisella novicida U112	Fnov	0.339	32.5	3
Haemophilus influenzae Rd KW20	Hinf	0.877	38.2	0.5
Helicobacter pylori 26695	Hpyl	0.148	38.9	2.4
Mycoplasma genitalium G37	Mgen	0.529	31.7	12
Mycoplasma pulmonis UAB CTIP	Mpul	0.032	26.6	1.5
Mycobacterium tuberculosis H37Rv	Mtub	0.246	65.6	19
Porphyromonas gingivalis ATCC 33277	Pgin	0.800	48.4	2.7
Pseudomonas aeruginosa UCBPP-PA14	Paer	0.915	66.3	0.5
Staphylococcus aureus NCTC 8325	Saur	0.807	32.8	0.4
Streptococcus pneumoniae TIGR4	Spne	0.893	39.7	0.5
Streptococcus sanguinis SK36	Ssan	0.962	43.4	-
Salmonella typhimurium LT2	Styp	0.915	52.2	0.4
Vibrio cholera O1 biovar El Tor N16961	Vcho	0.897	47.4	0.2

which transcriptional abundance influences ERs correlated with the strength of translational selection.

It has been previously demonstrated that translational selection across species is also strongly affected by genomic GC content [30]. We found that CAI-ER coefficients significantly correlated with GC content (Pearson's r = −0.473; $p = 0.045$). CAI-ER coefficients of GC-rich bacteria are significantly greater than those for AT-rich bacteria, as translational selection is often absent in AT-rich organisms [31]. It is also known that mRNAs have a stronger secondary structure if there are more GC-rich codons [32,33]. Moreover, there is stronger selection to improve translation efficiency for weak folding at translation-initiation sites of a gene in GC-rich hosts

Gram-negative

	ESS	CAI	EL	MFS	PPA	SLC	SLM	SLP	SLO	SLE	TMH	HS	AS	LEN	RSB
Abay	-0.341	-0.412		-0.195	-0.480	-0.329	0.093	*0.009*	0.108	0.048	0.136	*0.036*	0.269	-0.089	*-0.015*
Bthe	-0.197	-0.258	-0.489	-0.126	-0.467	-0.386	*0.033*	0.045	0.140	*0.008*	0.131	-0.116	0.307	*0.021*	-0.053
Ccre	-0.310	-0.523		*-0.026*	-0.472	-0.424	0.176	0.069	0.120	0.055	0.238	0.173	0.059	*-0.028*	-0.050
Ecol	-0.160	-0.464	-0.292	-0.096	-0.241	-0.215	0.047	0.062	0.065	0.058	0.082	*0.042*	*0.039*	*-0.008*	-0.078
Fnov	-0.182	-0.148		-0.123	-0.336	-0.280	0.028	0.012	0.061	0.088	0.149	0.058	0.262	-0.076	*-0.039*
Hinf	*0.005*	-0.503	-0.469	-0.118	-0.472	-0.385	0.208	0.075	0.135	*0.048*	0.258	0.113	0.339	*-0.013*	*-0.058*
Hpyl	*-0.014*	*-0.039*	-0.155	-0.083	-0.289	-0.161	*-0.044*	-0.071	*-0.001*	*-0.059*	*0.021*	0.088	0.243	*-0.028*	*0.013*
Paer	-0.150	-0.431	-0.321	0.040	-0.374	-0.294	*0.027*	0.045	0.107	0.056	0.065	*-0.024*	*-0.004*	-0.130	-0.045
Pgin	-0.185	-0.559	-0.395	-0.110	-0.318	-0.304	0.134	*0.023*	0.083	*0.012*	0.196	0.095	0.225	*0.046*	*-0.029*
Styp	-0.107	-0.441	-0.139	-0.126	-0.264	-0.131	-0.069	*0.009*	*0.000*	0.081	-0.056	*0.027*	0.079	*0.024*	*-0.008*
Vcho	-0.208	-0.636		-0.151	-0.387	-0.372	0.089	*0.016*	0.097	*0.024*	0.145	*0.045*	0.189	*-0.039*	*-0.033*

Gram-positive

	ESS	CAI	EL	MFS	PPA	SLC	SLM	SLE	SLW	TMH	HS	AS	LEN	RSB
Bsub	-0.287	-0.396	-0.362	-0.118	-0.317	-0.345	0.196	*0.037*	0.055	0.201	0.056	0.229	-0.064	-0.139
Mgen	*-0.145*	-0.199		*0.145*	-0.329	-0.230	*0.116*			0.210	*0.085*	0.352	*-0.072*	*-0.140*
Mpul	-0.305	-0.241		*-0.047*	-0.511	-0.359	0.222			0.311	0.147	0.508	*0.051*	*-0.048*
Mtub	-0.252	-0.481	-0.271	*0.003*	-0.270	-0.279	0.180	*0.031*	*0.014*	0.182	0.172	*-0.008*	-0.049	-0.103
Saur	-0.392	-0.496	-0.424	-0.150	-0.447	-0.324	0.122	0.084	0.056	0.125	0.105	0.273	*-0.046*	-0.133
Spne	-0.124	-0.549	-0.303	-0.099	-0.398	-0.331	0.200	*0.046*	*0.057*	0.285	0.165	0.268	-0.073	*-0.060*
Ssan	-0.317	-0.429	-0.214	-0.102	-0.511	-0.411	0.304	*0.036*	*0.055*	0.371	0.172	0.351	-0.088	-0.066

Figure 1 Rank correlation coefficient of variables and expression rate (ER). Gray cells indicate the highest absolute coefficients obtained for each bacterium. Blanks indicate that variables are non-existent or defective. Italicized and underlined values are non-significant at $p > 0.05$.

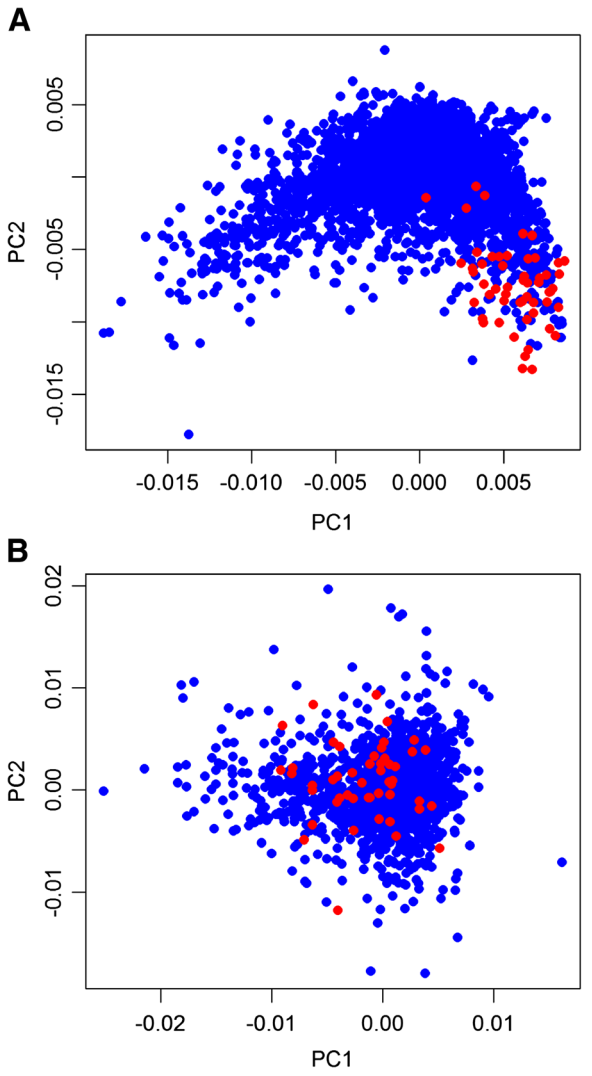

Figure 2 Plots of correspondence analysis of relative synonymous codon usage (RSCU-CA). PC1 is the first principal component and PC2 is the second one. Ribosomal and non-ribosomal genes are represented by red and blue dots, respectively. Significantly different codon usages were observed between ribosomal proteins and other proteins in the genome of **(A)** *E. coli*, but were not observed in the genome of **(B)** *H. pylori*.

[34]. These GC-rich organisms preferentially use GC-rich optimal codons [35]. GC-rich genomes therefore show stronger translational selection compared with AT-rich genomes. Accordingly, we found that transcriptional abundance does not always influence ERs as GC content varies across species.

A recent study found that CAI, microarray-based EL or sequencing-based EL approaches for measuring transcriptional abundance affected the assessment of the importance of transcriptional abundance to ER [9]. We found that bacteria, whose EL-ER correlation was weaker than the CAI-ER correlation, demonstrated greater CUS (0.816 *vs.* 0.220). For those species strongly mediated by translational selection, CAI as opposed to EL likely better explains the variation of ER. Although RNA sequencing (RNA-seq) data could be more accurate than microarray data, there is currently little RNA-seq data available for most bacterial species. In this study, we derived EL from RNA-seq data for *E. coli*, and used microarray data for other bacterial species. Although the sequencing-based EL-ER correlation is weaker than the CAI-ER correlation in *E. coli*, it is stronger than other correlations. With the development of RNA-seq experiments, we believe that the assessment of EL-ER correlations could be more accurate, and the impact of EL on ER could be stronger in certain bacterial species. To compensate for the inadequacy of each single variable to represent expression levels, we used CAI, EL and MFS to describe the impact of transcriptional abundance on ERs.

Functional importance

In an earlier study, it was proposed that ESS and protein interactions were negatively correlated with coding sequence ERs because of the constraints of important physical functions [3,6,11]. We used many types of protein associations (PPA), not only physical protein interactions (PPI), which were directly extracted from the STRING database. As expected, significant correlations between PPA/ESS and ERs were found for almost all the bacteria we investigated in our study. The strength of PPA-ER correlations was even greater than that of CAI-ER/EL-ER correlations in six organisms: *Acinetobacter* ADP1; *Francisella novicida*; *H. pylori*; *Mycoplasma genitalium*; *Mycoplasma pulmonis*; and *Streptococcus sanguinis*. In *F. novicida*, the ESS-ER correlation was also larger than that for CAI/EL-ER. The function of a gene is indeed an important driving force in bacterial protein evolution.

Variation in subcellular localization

Most cellular activities, including many metabolic pathways and processes, occur within the SLC. In this study, we observed significant negative correlations between SLC and ER. For example, the correlation coefficient for *Caulobacter crescentus* was −0.424 ($p = 5.18 \times 10^{-118}$). The SLM surrounds the cytoplasm of living cells, and positive correlations between SLM and ER were also observed in our study. The cell membrane functions as a selective filter, allowing molecules either to be pumped across the membrane by transmembrane transporters, or to be diffused through protein channels. These transmembrane proteins are usually specific; as a consequence, SLM proteins are fast-evolving and well adapted. We also found, as expected, that TMH positively correlated with bacterial protein ERs. The positive correlations are relatively weak between other subcellular localizations (SLP,

SLO, SLE, and SLW) and the ERs of proteins. Secreted proteins located in SLO/SLE for Proteobacteria and SLW/SLE for Firmicutes were found to rapidly evolve [10]. This could be a potential explanation of why SLW, SLO and SLE rapidly evolve.

Limitations of aromatic amino acids

To manufacture proteins, microorganisms must synthesize their aromatic amino acids *via* the shikimate pathway. These amino acids have a limited source that impacts upon the rate at which translation errors can be corrected, and the maintenance of translation efficiency and accuracy. Therefore, the adoption of aromatic amino acids in functional or abundant proteins is not encouraged. In this study, we found that slowly evolving proteins tend to avoid adopting aromatic amino acids. In most of the investigated bacteria, AS positively and significantly correlated with ER (Figure 1).

Head-on conflict

In many bacteria, genes tend to be encoded on the leading strand. The likelihood of a gene being found on the leading strand was weakly, but significantly, associated with ER in most of the studied bacteria. As an example, RSB of *Bacillus subtilis*, whose genome contains over 70% leading proteins, was significantly and positively correlated with ER (Pearson's r = −0.139; $p = 7.55 \times 10^{-11}$). Transcription and replication occur simultaneously in bacterial cells [36-38]. Replication progresses much faster than transcription, and inevitable conflicts occur between DNA and RNA polymerases when they bind to the same template. Co-directional collisions occur when the leading strand is the template for transcription, resulting in head-on collisions taking place when the lagging strand is the template. Head-on collisions have particularly deleterious effects, as replication forks may be arrested and transcription slowed. Over the course of evolution, transcripts are more likely to be retained if they are on the leading strand, which explains why bacterial genes on the leading strand evolve more slowly than those on the lagging strand.

Multiple factors cooperatively dominate ER

In both *E. coli* and *B. subtilis*, CAI has been identified as the most important driving force constraining ERs, through the use of partial correlation and multivariate regression analyses [12]. Drummond et al. first used principal component regression (PCR) analysis, and explained the dominant proportion of variation in yeast protein ERs by transcriptional abundance [13]. This analysis circumvents the problems of partial correlation and multivariate regression, as all principal components are orthogonal and independent. Therefore, it was useful to determine the independent contributions (R^2) of biological features to yeast ERs [13]. In contrast to the situation reported for yeast, *E. coli* and *B. subtilis* [12,13], our PCR results suggest that the contributions of multiple factors are comparable for the determination of bacterial protein evolution (Figure 3). We found that *Staphylococcus aureus* was strongly influenced by codon bias (CUS = 0.915), with ESS, CAI, EL, PPA, SLM, TMH and HS representing 13.1, 8.5, 13.5, 9.7, 12.0, 11.3 and 11.1% of the total rate variation, respectively. These variables are comparable and account for 79.13% of the total variation. In other words, CAI and EL lose their dominance in explaining bacterial protein evolution, even in bacterial genomes with strong codon bias. Of these investigated factors, ESS, CAI, EL, PPA, SLM, THM and HS each represent over 8% of the total variation in 78 (14/18), 67 (12/18), 67 (7/12), 83 (15/18), 72 (13/18), 78 (14/18) and 67 (12/18), respectively, of bacterial genomes.

Based on Correspondence Analysis (CA) results, we observed the universal rule that functional factors (ESS and PPA) and transcriptional abundance (CAI and EL) were roughly grouped together, opposing the ERs in the second principal component (PC2, see Methods) (Additional file 1: Figure S1, Figure 4). Evolutionary constraints on highly transcribed proteins might prevent misfolding [7] or misinteraction [23]. This can hamper functionality and even potentially produce a large quantity of toxic proteins. In contrast, constraints on essential or high connectivity genes possibly operate to avoid the abrogation of important physiological functions. The need for translational accuracy and robustness can help explain the selection exerted on ESS, PPA, CAI, and EL. In a principal component plot obtained for yeast, ESS and PPI were distant from CAI and EL [39]. This suggests a close link between functional factors and transcriptional abundance in some bacteria that is probably dependent on ER in some way.

In all the bacterial species we investigated, SLM, TMH and HS were found to cooperatively affect ER (Additional file 1: Figure S1). These three factors have been grouped in principal component plots (Figure 4). Membrane protein transport takes place *via* helix-dependent protein channels embedded in cell membranes, because of their hydrophobic structure. The need to maintain transmembrane protein function may help explain the relationship among SLM, TMH, and HS.

Different forces drive ER in different species

According to PCR analysis, factors associated with transcriptional abundance (CAI, EL), important functionality (ESS, PPA) and transmembrane protein function (SLM, TMH, and HS) were the main contributors (8%) to protein ER variation in over 50% of bacterial species we studied. Transcriptional abundance is the most dominant factor in yeast [13], but not in mice [40] or bacteria (this study). The extent to which transcriptional abundance

Gram-negative

	ESS	CAI	EL	MFS	PPA	SLC	SLM	SLP	SLO	SLE	TMH	HS	AS	LEN	RSB	Total
Abay	13.77%	7.91%		7.25%	12.86%	6.17%	10.77%	1.32%	6.40%	0.75%	10.02%	11.67%	6.38%	3.23%	1.48%	0.360
Bthe	7.73%	6.28%	9.80%	9.66%	7.20%	4.75%	9.48%	1.34%	6.55%	0.48%	9.02%	11.51%	4.41%	9.24%	2.56%	0.349
Ccre	9.73%	10.47%		6.92%	9.30%	7.01%	9.51%	2.64%	5.93%	2.33%	9.82%	8.73%	5.68%	9.23%	2.70%	0.445
Ecol	12.13%	16.16%	16.83%	2.79%	12.44%	3.52%	5.80%	1.96%	2.79%	1.66%	5.66%	4.76%	3.33%	6.08%	4.10%	0.257
Fnov	15.57%	2.51%		9.62%	13.37%	5.27%	11.09%	2.83%	2.53%	2.96%	9.61%	10.71%	5.22%	3.79%	4.92%	0.209
Hinf	1.98%	11.09%	12.66%	7.49%	6.31%	9.03%	10.70%	2.66%	1.90%	0.26%	10.87%	9.97%	5.06%	5.06%	4.96%	0.479
Hpyl	5.25%	7.51%		7.97%	13.04%	4.89%	4.44%	3.18%	11.51%	3.78%	12.04%	12.48%	5.05%	3.10%	5.75%	0.170
Paer	10.21%	20.51%	10.80%	5.34%	14.97%	2.17%	4.35%	0.85%	3.55%	2.54%	5.05%	3.06%	2.90%	8.97%	4.72%	0.365
Pgin	8.19%	11.69%	12.96%	5.79%	6.21%	6.37%	8.83%	1.19%	2.37%	2.39%	8.75%	8.52%	5.41%	4.87%	6.47%	0.429
Styp	8.60%	8.70%	6.75%	16.40%	9.57%	3.12%	3.58%	2.85%	2.57%	5.62%	3.09%	4.74%	1.60%	14.65%	8.16%	0.301
Vcho	13.68%	9.20%		13.18%	8.00%	5.60%	7.31%	4.90%	3.99%	3.26%	7.04%	7.31%	3.77%	7.92%	4.85%	0.516

Gram-positive

	ESS	CAI	EL	MFS	PPA	SLC	SLM	SLE	SLW	TMH	HS	AS	LEN	RSB	Total
Bsub	10.97%	9.69%	13.24%	2.95%	10.52%	8.57%	11.10%	0.45%	0.32%	10.84%	9.67%	5.54%	1.69%	4.44%	0.359
Mgen	11.43%	9.80%		5.69%	14.14%	9.19%	11.66%			8.42%	7.86%	9.22%	3.10%	9.49%	0.255
Mpul	10.44%	5.98%		3.85%	9.92%	9.44%	13.34%			12.65%	11.20%	11.51%	4.78%	6.89%	0.369
Mtub	12.22%	8.18%	7.23%	3.46%	10.03%	6.71%	8.69%	8.21%	3.49%	8.44%	6.77%	6.32%	4.16%	6.09%	0.328
Saur	13.05%	8.46%	13.51%	4.30%	9.69%	6.64%	12.01%	1.88%	0.78%	11.33%	11.08%	3.93%	0.74%	2.59%	0.498
Spne	2.19%	13.70%	15.73%	2.79%	13.48%	8.38%	11.46%	1.18%	0.98%	10.47%	9.56%	5.38%	1.92%	2.79%	0.496
Ssan	8.35%	7.40%	4.99%	4.21%	9.33%	10.33%	12.63%	1.97%	2.10%	12.17%	10.19%	7.20%	6.32%	2.81%	0.505

Figure 3 Proportion of each independent contribution (R^2) to total contributions. Cells highlighted in gray indicate variables that contributed at least 8% of the total contributions to the indicated species. Blanks indicate that variables are non-existent or defective.

affects ERs correlates with the strength of codon bias. Our PCR analysis indicated multiple factors contribute to the rate of protein evolution in bacteria. We also found that PPA was a common important contributor to bacterial evolution, with greater effects than CAI/EL. Our results were basically identical to those presented by Plotkin and Fraser [24]-PPI appears to be responsible for most of the ER variation in yeast. The deleterious effects of protein misinteractions can affect the optimal protein concentrations and shape functional interaction networks [41].

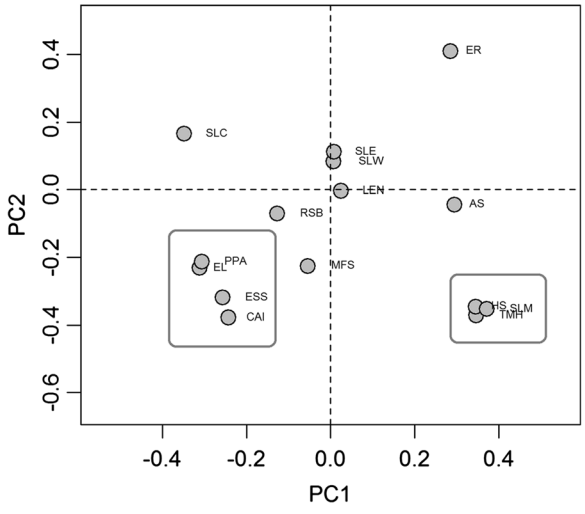

Figure 4 Principal component plot of *S. aureus*. Functional factors (ESS and PPA) and transcriptional abundance (CAI and EL) grouped together and strongly contributed to the PC2 that is opposite to that of the ER. Similarity, SLM, TMH and HS cooperatively and negatively affected the ER at the PC2.

Therefore there is a need to maintain proper interactions among high connectivity proteins as it constrains their evolution. Although ESS does not contribute strongly to yeast ERs, it is still an important factor in determining bacterial protein evolution. Our findings suggest that various forces drive protein sequence evolution in different species.

Conclusions

We have uncovered new relationships among ERs in bacterial genomes related to protein subcellular localization, transmembrane helices, hydropathicity, aromaticity, and replication strand localization. ER had a significant negative correlation with SLC, but a significant positive correlation with SLM. Because of the effects of TMH and HS on SLM, these two variables were also found to positively, although relatively weakly, correlate with bacterial protein ERs. The impact of bacterial SLM/TMH/HS and SLC on ER is independent of functional importance and transcriptional abundance. This is consistent with results from a recent study in mammalian proteins [8]. We also found that proteins that evolved slowly in bacterial genomes tended to avoid adopting aromatic amino acids. Additionally, bacterial genes on the leading strand evolved more slowly than those with genes on the lagging strand. We investigated the independent contributions of biological features to ER, and found that the dominant effect of transcriptional abundance on ER is absent in bacteria. Factors that retain important functionality (ESS, PPA), maintain cell membrane function (SLM, TMH, and HS) and avoid a build-up of toxic molecules caused by misfolding or misinteraction (CAI, EL) influence the ERs of bacterial

proteins. If more RNA-Seq data are available in the future, the correlation of EL-ER could be found to be stronger in certain bacterial species than reported here. However, the influences of PPA, ESS, SLM, TMH, and HS on ER are comparable with the impact of transcriptional abundance on ER in most bacteria.

Methods

Genomic features

Essentiality

We investigated 18 bacterial species (Table 2) in the current version (7.0) of the Database of Essential Genes (DEG; http://tubic.tju.edu.cn/deg/), which hosts records of available essential genes identified by well-known genome-wide experimental techniques from a range of organisms [42]. In each of these experiments, almost all genes were investigated for their ESS scores; therefore datasets were not biased or partial. Complete coding sequences of these bacteria and their gene ESS annotations were obtained from GenBank and DEG databases, respectively.

Evolutionary rates

Orthologous gene pairs between each genome pair were identified based on reciprocal best hits using the Blastp program with criteria of $E < 10^{-5}$, 80% minimum residues that could be aligned, and 30% identity. Protein sequences encoded by identified orthologous gene pairs were aligned with ClustalW [43], and then back-translated into nucleotide sequences based on their original sequences. Numbers of substitutions per non-synonymous site (K_a) were calculated following Yang's definition using the PAML package with default parameters [44]. We retained all ortholog assignments coding for more than 30 amino acids, which were not acquired by horizontal transfer, as determined by the Horizontal Gene Transfer [45] (HGT-DB; http://genomes.urv.cat/HGT-DB/) and DarkHorse [46] (http://darkhorse.ucsd.edu/) databases. Values for ERs were log-transformed after addition of a small constant (0.001).

CAI, expression level and mRNA folding strength

Transcriptional abundance was predicted from CAI, expression levels and mRNA folding strength. CAI is a species-dependent codon bias measurement that has been widely used as an empirical approach for gene expressivity, especially in microbial genomes [20]. With this methodology, dozens of ribosomal protein genes were chosen as a reference set of highly expressed genes for each genome. Our mRNA levels, derived from RNA-seq data for *E. coli* and microarray data for other species, under favorable environmental conditions were extracted from the Gene Expression Omnibus [47] (GEO; http://www.ncbi.nlm.nih.gov/geo/) database. Data were obtained for the following bacteria: *B. subtilis* (GEO Sample Accession Numbers GSM177105–GSM177118); *Bacteroides thetaiotaomicron* (GSM40897–GSM40906); *E. coli* (GSM99211–GSM99216); *Haemophilus influenzae* (GSM114031–GSM114033); *H. pylori* (GSM623401–GSM623404); *Mycobacterium tuberculosis* (GSM71958, GSM71988–GSM71990); *Porphyromonas gingivalis* (GSM590017); *Pseudomonas aeruginosa* (GSM462061–GSM462064, GSM462352–GSM462355); *S. aureus* (GSM724739–GSM724741), *Streptococcus pneumonia* (GSM673840); *Streptococcus sanguinis* (GSM908371–GSM908373); and *Salmonella typhimurium* (GSM874413–GSM874415). Expression level values were scaled using a logarithmic function.

The secondary structures of mRNAs, for a folding temperature under 30°C, were predicted by RNAfold within the ViennaRNA package [48]. Windows comprising 150 nucleotides were slid in 10 nucleotide steps during analysis [26]. At each nucleotide, the probability that it paired was estimated by the number of sliding windows with which it paired, divided by the number of sliding windows that include the nucleotide. We then used the average pairing possibility for an mRNA to estimate its folding strength.

Number of protein–protein associations

Protein-protein association data were obtained from the STRING database [49] (http://string-db.org/). These association data included physical PPIs and other links such as co-expression data. From the original data, we computed the number of associations for each gene using a default confidence score cutoff of 0.4.

Subcellular localization and number of transmembrane helices

We used PSORTb v3.0 [50] (http://www.psort.org/psortb/) to predict subcellular localization of proteins. Four subcellular localization types can be predicted for Gram-positive bacteria and five types can be predicted for Gram-negative bacteria. For a certain localization type, genes were assigned PSORTb prediction scores if they belong to this type, and 0 if they did not. The number of transmembrane helices was predicted from bacterial proteomes using the TMHMM Server v2.0 (http://www.cbs.dtu.dk/services/TMHMM/).

Protein hydropathicity, aromaticity, and length

We used CodonW (http://codonw.sourceforge.net/) to determine hydropathicity, aromaticity and protein length. The general average hydropathicity score for each gene product was obtained by calculating the arithmetic mean of the sum of the hydropathic indices for each amino acid. Aromaticity scores are indices for indicating frequency of aromatic amino acids.

Replication strand bias

Replication origin and terminus positions for each bacterial species were annotated using the DoriC database [51] (http://tubic.tju.edu.cn/doric/index.html). Genes were assigned a value of 1 if these positions were located on the leading strand, and 0 if otherwise.

Statistical analysis

Spearman rank correlation and PCR

Spearman's rank correlation test was used to investigate expected direct correlations between each variable and ER. To further determine the independent contribution (R^2) of each biological feature to ER, we used PCR.

CUS

To assess the impact of translation selection on codon usage, we investigated differences in relative synonymous codon usage (RSCU) between ribosomal proteins and non-ribosomal proteins using correspondence analysis (RSCU-CA). Correspondence analysis is a classical technique to reduce the dimensionality of a dataset by transforming it into its principal components. The first principal component (PC1) maximizes the standard deviation of the derived variable, while the second principal component (PC2) maximizes the standard deviation among axes uncorrelated with the first. The PC1 and PC2 could effectively explain the original 64-D codon datasets. We observed from the PC1-PC2 plot the differential codon usage pattern between ribosomal proteins and non-ribosomal proteins. Then we defined the CUS of ribosomal proteins as the percentage of ribosomal proteins falling outside the non-ribosomal protein cluster on the PC1-PC2 plot. The reference range (90%) of non-ribosomal proteins on the plot was defined as:

$$\bar{X}\text{-}1.64S < X < \bar{X} + 1.64S;\ \ \bar{Y}\text{-}1.64S < Y < \bar{Y} + 1.64S \quad (1)$$

where $\bar{X}$, $\bar{Y}$, and S denote the average PC1 and PC2 values of non-ribosomal proteins, and the standard deviation of the principal component value, respectively. A greater CUS indicates a greater difference in codon usage between ribosomal proteins and non-ribosomal proteins. All statistical analyses were conducted and plots generated using the R package (http://www.r-project.org/).

Additional file

Additional file 1: Figure S1. Principal component plots of other 17 bacteria.

Abbreviations

PCR: Principal component regression; RSCU: Relative synonymous codon usage; CA: Correspondence analysis; PC: Principal component; CUS: Codon usage separation; GEO: Gene expression omnibus; DEG: Database of essential genes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WW participated in the design of the study, compilation of the python script, collection and preparation of data, performed the analyses, and drafted the majority of the manuscript. TZ assisted in writing the python code to implement Blast and PAML. DL double-checked the results. ZJY took part in drafting and editing the manuscript. FBG conceived and guided the study, and also assisted in writing the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

This work was supported by the Program for New Century Excellent Talents in University (NCET-11-0059), the National Natural Science Foundation of China (Grant 31,071,109), and the special fund of the China Postdoctoral Science Foundation (Grant 201,104,687).

Received: 19 March 2013 Accepted: 1 August 2013
Published: 2 August 2013

References

1. Hirsh AE, Fraser HB: **Protein dispensability and rate of evolution.** *Nature* 2001, **411**(6841):1046–1049.
2. Pal C, Papp B, Hurst LD: **Genomic function: rate of evolution and gene dispensability.** *Nature* 2003, **421**(6922):496–497. discussion 497–498.
3. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296**(5568):750–752.
4. Jordan IK, Rogozin IB, Wolf YI, Koonin EV: **Essential genes are more evolutionarily conserved than are nonessential genes in bacteria.** *Genome Res* 2002, **12**(6):962–968.
5. Yang J, Gu Z, Li WH: **Rate of protein evolution versus fitness effect of gene deletion.** *Mol Biol Evol* 2003, **20**(5):772–774.
6. Hahn MW, Kern AD: **Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks.** *Mol Biol Evol* 2005, **22**(4):803–806.
7. Drummond DA, Wilke CO: **Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.** *Cell* 2008, **134**(2):341–352.
8. Liao BY, Weng MP, Zhang J: **Impact of extracellularity on the evolutionary rate of mammalian proteins.** *Genome Biol Evol* 2010, **2**:39–43.
9. Chang TY, Liao BY: **Flagellated algae protein evolution suggests the prevalence of lineage-specific rules governing evolutionary rates of eukaryotic proteins.** *Genome Biol Evol* 2013, **5**(5):913–922.
10. Nogueira T, Touchon M, Rocha EP: **Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria.** *PLoS One* 2012, **7**(11):e49403.
11. Wilson AC, Carlson SS, White TJ: **Biochemical evolution.** *Annu Rev Biochem* 1977, **46**:573–639.
12. Rocha EP, Danchin A: **An analysis of determinants of amino acids substitution rates in bacterial proteins.** *Mol Biol Evol* 2004, **21**(1):108–116.
13. Drummond DA, Raval A, Wilke CO: **A single determinant dominates the rate of yeast protein evolution.** *Mol Biol Evol* 2006, **23**(2):327–337.
14. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158**(2):927–931.
15. Krylov DM, Wolf YI, Rogozin IB, Koonin EV: **Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.** *Genome Res* 2003, **13**(10):2229–2235.
16. Subramanian S, Kumar S: **Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome.** *Genetics* 2004, **168**(1):373–381.
17. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL: **Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions.** *Mol Biol Evol* 2005, **22**(5):1345–1354.
18. Popescu CE, Borza T, Bielawski JP, Lee RW: **Evolutionary rates and expression level in Chlamydomonas.** *Genetics* 2006, **172**(3):1567–1576.
19. Ingvarsson PK: **Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula.** *Mol Biol Evol* 2007, **24**(3):836–844.

20. Sharp PM, Li WH: **The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**(3):1281–1295.
21. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proc Natl Acad Sci U S A* 2005, **102**(15):5483–5488.
22. Zhang J, He X: **Significant impact of protein dispensability on the instantaneous rate of protein evolution.** *Mol Biol Evol* 2005, **22**(4):1147–1155.
23. Yang JR, Liao BY, Zhuang SM, Zhang J: **Protein misinteraction avoidance causes highly expressed proteins to evolve slowly.** *Proc Natl Acad Sci U S A* 2012, **109**(14):E831–E840.
24. Plotkin JB, Fraser HB: **Assessing the determinants of evolutionary rates in the presence of noise.** *Mol Biol Evol* 2007, **24**(5):1113–1121.
25. Wolf MY, Wolf YI, Koonin EV: **Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution.** *Biol Direct* 2008, **3**:40.
26. Park C, Chen X, Yang JR, Zhang J: **Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci U S A* 2013, **110**(8):E678–E686.
27. Zur H, Tuller T: **Strong association between mRNA folding strength and protein abundance in S. cerevisiae.** *EMBO Rep* 2012, **13**(3):272–277.
28. Lafay B, Atherton JC, Sharp PM: **Absence of translationally selected synonymous codon usage bias in Helicobacter pylori.** *Microbiology* 2000, **146**(Pt 4):851–860.
29. Vieira-Silva S, Rocha EP: **The systemic imprint of growth and its uses in ecological (meta) genomics.** *PLoS Genet* 2010, **6**(1):e1000808.
30. Plotkin JB, Kudla G: **Synonymous but not the same: the causes and consequences of codon bias.** *Nat Rev Genet* 2011, **12**(1):32–42.
31. Naya H, Romero H, Carels N, Zavala A, Musto H: **Translational selection shapes codon usage in the GC-rich genome of Chlamydomonas reinhardtii.** *FEBS Lett* 2001, **501**(2–3):127–130.
32. Voges D, Watzele M, Nemetz C, Wizemann S, Buchberger B: **Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system.** *Biochem Biophys Res Commun* 2004, **318**(2):601–614.
33. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in Escherichia coli.** *Science* 2009, **324**(5924):255–258.
34. Gu W, Zhou T, Wilke CO: **A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes.** *PLoS Comput Biol* 2010, **6**(2):e1000664.
35. Hershberg R, Petrov DA: **General rules for optimal codon choice.** *PLoS Genet* 2009, **5**(7):e1000556.
36. Mirkin EV, Mirkin SM: **Mechanisms of transcription-replication collisions in bacteria.** *Mol Cell Biol* 2005, **25**(3):888–895.
37. Pomerantz RT, O'Donnell M: **What happens when replication and transcription complexes collide?** *Cell Cycle* 2010, **9**(13):2537–2543.
38. Kim N, Jinks-Robertson S: **Transcription as a source of genome instability.** *Nat Rev Genet* 2012, **13**(3):204–214.
39. Theis FJ, Latif N, Wong P, Frishman D: **Complex principal component and correlation structure of 16 yeast genomic variables.** *Mol Biol Evol* 2011, **28**(9):2501–2512.
40. Liao BY, Scott NM, Zhang J: **Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins.** *Mol Biol Evol* 2006, **23**(11):2072–2080.
41. Zhang J, Maslov S, Shakhnovich EI: **Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size.** *Mol Syst Biol* 2008, **4**:210.
42. Zhang R, Lin Y: **DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes.** *Nucleic Acids Res* 2009, **37**(Database issue):D455–D458.
43. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673–4680.
44. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24**(8):1586–1591.
45. Garcia-Vallve S, Guzman E, Montero MA, Romeu A: **HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes.** *Nucleic Acids Res* 2003, **31**(1):187–189.
46. Podell S, Gaasterland T: **DarkHorse: a method for genome-wide prediction of horizontal gene transfer.** *Genome Biol* 2007, **8**(2):R16.
47. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, *et al*: **NCBI GEO: archive for functional genomics data sets–update.** *Nucleic Acids Res* 2013, **41**(D1):D991–D995.
48. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL: **ViennaRNA Package 2.0.** *Algorithms Mol Biol* 2011, **6**:26.
49. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, *et al*: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.** *Nucleic Acids Res* 2011, **39**(Database issue):D561–D568.
50. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, *et al*: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26**(13):1608–1615.
51. Gao F, Luo H, Zhang CT: **DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes.** *Nucleic Acids Res* 2013, **41**(Database issue):D90–D93.

doi:10.1186/1471-2148-13-162
Cite this article as: Wei *et al.*: Transcriptional abundance is not the single force driving the evolution of bacterial proteins. *BMC Evolutionary Biology* 2013 **13**:162.