Flagellated Algae Protein Evolution Suggests the Prevalence of Lineage-Specific Rules Governing Evolutionary Rates of Eukaryotic Proteins

Ting-Yan Chang and Ben-Yang Liao*

Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan, Republic of China

*Corresponding author: E-mail: liaoby@nhri.org.tw.

Accepted: April 2, 2013

Abstract

Understanding the general rules governing the rate of protein evolution is fundamental to evolutionary biology. However, attempts to address this issue in yeasts and mammals have revealed considerable differences in the relative importance of determinants for protein evolutionary rates. This phenomenon was previously explained by the fact that yeasts and mammals are different in many cellular and genomic properties. Flagellated algae species have several cellular and genomic characteristics that are intermediate between yeasts and mammals. Using partial correlation analyses on the evolution of 6,921 orthologous proteins from *Chlamydomonas reinhardtii* and *Volvox carteri*, we examined factors influencing evolutionary rates of proteins in flagellated algae. Previous studies have shown that mRNA abundance and gene compactness are strong determinants for protein evolutionary rates in yeasts and mammals, respectively. We show that both factors also influence algae protein evolution with mRNA abundance having a larger impact than gene compactness on the rates of algae protein evolution. More importantly, among all the factors examined, coding sequence (CDS) length has the strongest (positive) correlation with protein evolutionary rates. This correlation between CDS length and the rates of protein evolution is not due to alignment-related issues or domain density. These results suggest no simple and universal rules governing protein evolutionary rates across different eukaryotic lineages. Instead, gene properties influence the rate of protein evolution in a lineage-specific manner.

Key words: expression level, mRNA abundance, gene compactness, protein length, functional density.

Introduction

The general rules governing protein evolutionary rates have been studied not only because of their fundamental importance in molecular evolution but also for their broad implications in genomics, bioinformatics, and systems biology. For example, evolutionary sequence conservation has been widely used in identifying functional coding or noncoding regions in the genome that are important for an organism's fitness (Boffelli et al. 2003; Elnitski et al. 2003; Thomas et al. 2003; Pennacchio et al. 2006). Recent studies have identified several gene properties, including gene essentiality (Hirsh and Fraser 2001; Jordan et al. 2002; Zhang and He 2005; Liao et al. 2006) and mRNA abundance (Pal et al. 2001; Rocha and Danchin 2004; Drummond and Wilke 2008; Slotte et al. 2011), that correlate with protein evolutionary rates in a wide phylogenetic spectrum. However, the relative importance of determinants for protein evolutionary rates varies widely between yeasts and mammals (Liao et al. 2006, 2010; Hudson and Conant 2011). In yeasts, mRNA abundance is the predominant factor determining the rate of protein evolution (Drummond et al. 2006), whereas in mammals, gene compactness, measured by averaged length of introns or untranslated regions (UTRs), has a stronger influence on protein evolutionary rates compared with the abundance of mRNA (Liao et al. 2006, 2010). This discrepancy between mammalian and yeast protein evolution was explained by the specialization of more than 150 cell types in mammals (Vogel and Chothia 2006) that requires additional layers of gene regulation (e.g., tissue specificity and alternative splicing) associated with multicellularity and organismal complexity (Schad et al. 2011) to influence protein evolution (Gu and Su 2007). Additionally, mammalian cells (10–100 μm diameter) are larger than yeast cells (3–4 μm diameter) and can better tolerate cell toxicity of misfolded proteins caused by

mistranslation (Drummond et al. 2005; Liao et al. 2006). Furthermore, although more than 95% yeast genes are intron-less, more than 90% mammalian genes contain at least one intron, which can increase the efficacy of natural selection (Comeron and Kreitman 2002; Liao et al. 2006).

Chlamydomonas reinhardtii and *Volvox carteri* are flagellated algae that diverged approximately 200 Ma (Herron et al. 2009). These two algae have several cellular and genomic characteristics that are intermediate between yeasts and mammals. First, *C. reinhardtii* is a unicellular organism with two cell types, sexual or asexual (Rochaix 1995), whereas *V. carteri* is a colonial/multicellular organism composed of approximately 2,000 *Chlamydomonas*-like somatic cells and approximately 16 asexual reproductive cells called gonidia (Hallmann 2011). Second, *C. reinhardtii* and *V. carteri* cells are 8–10 μm in diameter (Kirk et al. 1993; Rochaix 1995). Third, unlike yeast genes, most *C. reinhardtii* genes (91%) and *V. carteri* genes (92%) have at least one intron (Merchant et al. 2007; Prochnik et al. 2010). Although most genes have an exon–intron structure similar to that of mammals (Waterston et al. 2002), the average intron size of *C. reinhardtii* genes (373 bp) (Merchant et al. 2007) is less than one-tenth of the average intron size (3,888 bp) of mouse genes (Waterston et al. 2002). Because of these intermediate characteristics, it is interesting to examine which gene properties are determinants of protein evolutionary rates in flagellated green algae.

To investigate protein evolutionary rates, we calculate nonsynonymous substitution rates (d_N) and the ratios of d_N to synonymous substitution rates (d_S) of approximately 7,000 *C. reinhardtii–V. carteri* one-to-one orthologous proteins. By examining how mRNA abundance, gene compactness, and other gene features are associated with d_N or d_N/d_S, we address 1) whether factors correlated with d_N (or d_N/d_S) in other eukaryotes are also correlated with d_N (or d_N/d_S) of flagellated green algae with similar trends, 2) whether gene compactness or mRNA abundance plays a greater role in determining evolutionary rates of flagellated algae proteins in terms of d_N or d_N/d_S, and 3) whether either gene compactness or mRNA abundance is the most important factor in determining flagellated algae d_N (or d_N/d_S). Our results indicate that gene properties often influence the rate of protein evolution in a lineage-specific manner. Hence, there is no general rule for interpreting evolutionary rate variation among proteins in a wide range of eukaryotic lineages.

Materials and Methods

Orthologous Genes and the Calculation of Evolutionary Rates

The genomes and annotations of two flagellated green algae *C. reinhardtii* and *V. carteri* were downloaded from Phytozome 7.0 (http://www.phytozome.net/, last accessed April 13, 2011) (Goodstein et al. 2012). Although many *C. reinhardtii* or *V. carteri* genes are likely alternatively spliced (Kianianmomeni et al. 2008; Labadorf et al. 2010), based on the current annotations of these algal genes, each *C. reinhardtii* or *V. carteri* gene only corresponds to one transcript and one protein. In total, information from 17,114 *C. reinhardtii* genes (proteins) and 14,542 *V. carteri* genes (proteins) were obtained. On the basis of the protein sequences of the algae genes, we used reciprocal best hits to define 6,921 *C. reinhardtii–V. carteri* one-to-one orthologous gene pairs using *E* value less than 10^{-10} in BLASTp (v2.2.24, http://blast.ncbi.nlm.nih.gov/, last accessed April 13, 2011) searches (Zhang and He 2005; Wyder et al. 2007). In addition to the BLASTp-based method, we also used InParanoid (version 4.1) (O'Brien et al. 2005) to define and obtained 3,850 *C. reinhardtii–V. carteri* one-to-one orthologous genes. Both sets of orthologs generated similar results, so we present the results using BLASTp-based one-to-one orthologs, unless otherwise noted. To calculate the evolutionary rates of the 6,921 genes after the *C. reinhardtii–V. carteri* divergence, the amino acid sequences of orthologous gene pairs were aligned using ClustalW (v1.83, ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/, last accessed April 15, 2011) (Thompson et al. 1994) or MUSCLE (v3.8.31, http://www.drive5.com/muscle/, last accessed July 7, 2011) (Edgar 2004) with default parameters and back translated to the corresponding nucleotide coding sequences (CDSs). Because results generated based on the two aligners are virtually identical, we only present those based on ClustalW alignments as the main results, unless otherwise noted. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) of each ortholog were estimated using the codeml module of PAML 3.14 (Yang 1997) with following parameters: runmode = −2, seqtype = 1, CodonFreq = 2, model = 0, and NSsites = 0.

Expression and Structural Properties of Genes

Codon adaptation index (*CAI*) (Sharp and Li 1987) ranges from 0 to 1, and a higher *CAI* implies a higher expression level of the gene. To calculate *CAI* of each *C. reinhardtii* and *V. carteri* gene, we used CodonW (v1.4.2, http://codonw.sourceforge.net/, last accessed May 11, 2011). RNA-seq-based transcriptome data (GSM600876 and GSM449827) were obtained from National Center for Biotechnology Information (NCBI) GEO Data Sets (http://www.ncbi.nlm.nih.gov/geo/, last accessed May 19, 2011), which were generated from directly measuring mRNA expression levels of *C. reinhardtii* genes grown in Tris-acetate-phosphate medium. Raw *n*-mer RNA-seq reads of GSM600876 ($n = 75$) and GSM449827 ($n = 35$) were mapped to annotated exons of all *C. reinhardtii* genes by SOAP (v2.21, http://soap.genomics.org.cn/, last accessed January 24, 2013). The number of uniquely mapped RNA-seq reads per gene was divided by the number of unique *n*-

mers per gene to generate normalized mRNA expression signals (Sultan et al. 2008; Qian et al. 2010; Xiong et al. 2010; Chang and Liao 2012), and the signals of different replicates of the same experiment were averaged. A total of 3,480 *C. reinhardtii* genes with one-to-one orthologs in *V. carteri* had detectable expression signals in both *C. reinhardtii* data sets of GSM600876 and GSM449827. For each gene, the expression level in terms of mRNA abundance ($ExpLev_{mRNA}$) was the average expression signal from the two RNA-seq data sets. To examine whether using a direct measure of mRNA abundance changes the importance of expression as a determinant of protein evolutionary rate in mammals or yeasts (Liao et al. 2006, 2010), $ExpLev_{mRNA}$ was also calculated from *n*-mer single read RNA-seq data of mouse (*Mus musculus*, $n = 25$) tissues (Mortazavi et al. 2008) and yeast (*Saccharomyces cerevisiae*, $n = 33$ or 35) (Nagalakshmi et al. 2008). Estimating expression signals of mouse genes in liver, muscle, and brain has been described previously (Qian et al. 2010; Chang and Liao 2012). Expression signals from these three tissues were averaged to yield $ExpLev_{mRNA}$ for each mouse gene. The yeast RNA-seq raw reads (GSM282598) were obtained from NCBI GEO Data Set. Expression signals from three random hexamer-primed samples ($n = 33$, one sample; $n = 35$, two samples) were calculated as described above and were averaged to yield $ExpLev_{mRNA}$ for each yeast gene.

To investigate the potential effect of the protein domain density (*domain%*, see Results and Discussion) on the evolution of algae proteins, we used *C. reinhardtii* protein domain annotations (JGI v.4) downloaded from SUPERFAMILY (v.1.75) (Wilson et al. 2007). Upstream start codons (uAUG) within the 5′-UTR can inhibit protein translation of an mRNA (Calvo et al. 2009; Yun et al. 2012). We defined *#uAUG* as the numbers of AUG codons in the 5′-UTR of the representative transcript for each *C. reinhardtii* gene and *V. carteri* gene. To investigate the independent influence of each gene property (CDS length in nucleotides, *CDS Length*; RNA-seq-based mRNA expression level, $ExpLev_{mRNA}$; *CAI*; average intron length in nucleotides, *Avg Intron Length*; length of UTR in nucleotides, *UTR Length*; number of uAUG motifs, *#uAUG*) on *C. reinhardtii–V. carteri* d_N or d_N/d_S, we performed partial correlation analysis using modules of the "ppcor" package (v.1.0) (Kim and Yi 2007) for R (v2.15.1, http://www.r-project.org/, last accessed June 22, 2012). Information of UTR or intron structures of mammalian genes and yeast genes were obtained from Liao et al. (2010).

Results and Discussion

Orthologous Genes Used for the Correlation Analysis

RNA-seq experiments can directly assess absolute mRNA abundance of genes. This data can be used to address fundamental questions in biology, including questions related to transcriptome evolution (Xiong et al. 2010; Brawand et al. 2011; Chang and Liao 2012). Among the 6,921 *C. reinhardtii–V. carteri* one-to-one orthologous genes, only 3,480 *C. reinhardtii* genes had RNA-seq-based mRNA abundance signals in our analysis (see Materials and Methods). The remaining 3,441 genes either had no unique *n*-mer mappable region in the coding region (three genes) or had no mapped RNA-seq read for the calculation of mRNA abundance due to absent or weak expression (3,438 genes). The variation in expression levels of these 3,438 genes without an estimable $ExpLev_{mRNA}$ can only be observed by deeper sequencing. Because the estimation of protein evolutionary rates was not affected by sequencing coverage of the transcriptome, the inclusion of these 3,438 genes in the correlation analysis could substantially underestimate the contribution of $ExpLev_{mRNA}$ to the variation in d_N or d_N/d_S. Thus, we focused on the 3,480 orthologs with detectable RNA-seq data in *C. reinhardtii* and examined whether this subset of orthologs is representative of all one-to-one orthologs for the correlation analysis.

For each of two data sets, (i) all 6,921 one-to-one orthologs and (ii) 3,480 orthologs with detectable RNA-seq-based mRNA abundance ($ExpLev_{mRNA} > 0$) in *C. reinhardtii*, Spearman's rank correlation analyses were performed for *C. reinhardtii* genes versus *C. reinhardtii–V. carteri* d_N or d_N/d_S on the following gene properties: *CDS Length*, $ExpLev_{mRNA}$, *CAI*, *Avg Intron Length*, *3′-UTR Length*, *5′-UTR Length*, and *#uAUG*. The rank correlation coefficient (ρ) between the examined factor and d_N (or d_N/d_S) was similar for both data sets for nearly all cases (table 1). The results from data set (i) are generally more statistically significant than those of data set (ii) due to the larger sample size in data set (i). Similar patterns were observed when *V. carteri* genes were used to define gene properties (supplementary table S1, Supplementary Material online). The consistent ρ observed between data sets (i) and (ii) suggest that the evolution of the 3,480 orthologs with *C. reinhardtii* RNA-seq data is representative of the genome as a whole. Therefore, we used these 3,480 orthologs for the subsequent analyses that required direct measurement of mRNA abundance ($ExpLev_{mRNA}$).

Gene properties with greater influence on protein evolutionary rates were expected to have higher rank correlations with d_N. Gene properties can affect the mutation rate or the local selection environment, which both comprise protein evolutionary rates. To distinguish between properties that influenced protein evolutionary rates at the selection level from those that influence properties at the mutation level, we calculated both ρ to d_N and ρ to d_N/d_S. If the influence on protein evolutionary rates was at the selection level, rather than at the mutation level, its ρ to d_N and ρ to d_N/d_S should not differ substantially. Because some gene properties were inter-related (fig. 1), partial rank correlation was used to assess the direct influence of a specific gene property on d_N or d_N/d_S by controlling for potential confounding effects of all the other properties (table 2).

Table 1

Spearman's Rank Correlations of *Chlamydomonas reinhardtii* Gene Properties with d_N and d_N/d_S

Gene Properties[a]	All 6,921 Orthologs		3,480 Orthologs[b]	
	ρ (*P* Value[c]) with d_N	ρ (*P* Value[c]) with d_N/d_S	ρ (*P* Value[c]) with d_N	ρ (*P* Value[c]) with d_N/d_S
CDS Length	0.549 ($<10^{-300}$)	0.401 ($<10^{-265}$)	0.552 ($<10^{-275}$)	0.403 ($<10^{-135}$)
#uAUG	0.269 ($<10^{-114}$)	0.191 ($<10^{-57}$)	0.281 ($<10^{-63}$)	0.205 ($<10^{-33}$)
Gene expression level				
$ExpLev_{mRNA}$	NA	NA	−0.546 ($<10^{-268}$)	−0.356 ($<10^{-104}$)
CAI	−0.287 ($<10^{-130}$)	−0.184 ($<10^{-52}$)	−0.281 ($<10^{-63}$)	−0.175 ($<10^{-24}$)
Gene compactness				
Avg Intron Length	−0.055 ($<10^{-5}$)	−0.069 ($<10^{-8}$)	−0.085 ($<10^{-6}$)	−0.075 ($<10^{-5}$)
5′-UTR Length	0.154 ($<10^{-37}$)	0.073 ($<10^{-8}$)	0.147 ($<10^{-17}$)	0.064 ($<10^{-3}$)
3′-UTR Length	−0.020 (0.09)	−0.032 ($<10^{-2}$)	−0.020 (0.23)	−0.038 (0.03)

NOTE.—NA, not applicable.

[a]All gene properties were based on *C. reinhardtii* genes.

[b]The subset of one-to-one orthologs with RNA-seq expression levels in *C. reinhardtii* genes.

[c]*P* values show the probabilities of the observations under the hypothesis of no correlation.

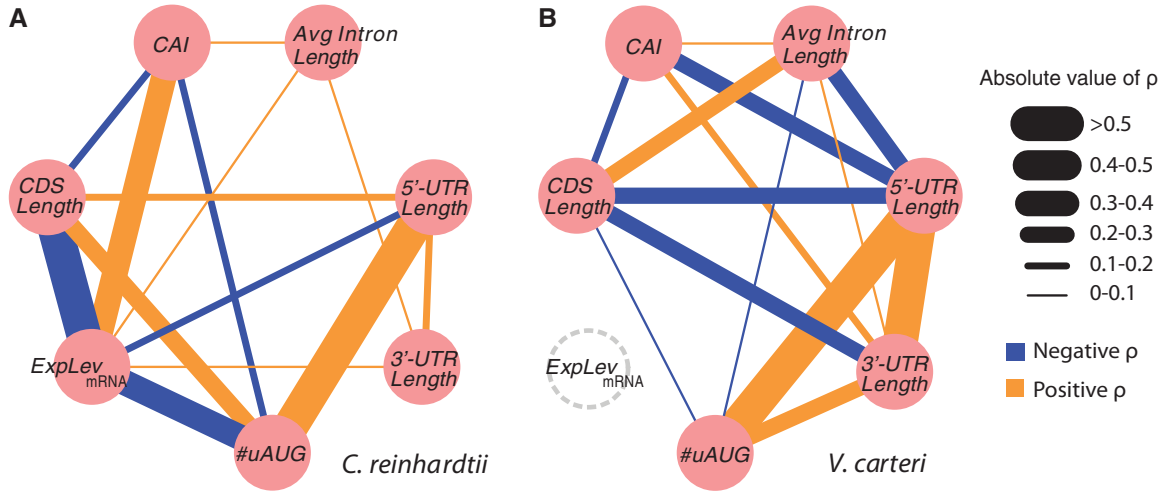

FIG. 1.—The network showing the inter-relatedness of the gene properties in (*A*) *Chlamydomonas reinhardtii* or (*B*) *Volvox carteri*. Gene properties are represented by nodes. Edges between nodes represent a highly significant rank correlation coefficient (ρ) between the two corresponding features. Edge thickness corresponds to the magnitude of ρ; edge color corresponds to the sign of ρ (orange = positive, blue = negative). *Volvox carteri* genes have no $ExpLev_{mRNA}$ data.

Influence of mRNA Abundance and Gene Compactness on the Rate of Algae Protein Evolution

In yeasts, mRNA abundance is the predominant factor determining the rate of protein evolution, as actively transcribed yeast genes evolve slowly (Drummond et al. 2005, 2006). In mammals, gene compactness, measured by average length of introns or length of UTRs, has a stronger influence on protein evolutionary rates than does gene expression level, as mammalian proteins encoded by a more compact gene evolve more rapidly (Liao et al. 2006).

Codon usage bias has been frequently used in predicting mRNA expression levels (Fraser et al. 2004; Goetz and Fuglsang 2005), including that of *C. reinhardtii* (Popescu et al. 2006). Consistent with a previous study on 67 nuclear *Chlamydomonas* genes (Popescu et al. 2006), we found *CAI* to be negatively correlated with d_N (Spearman's correlation coefficient $\rho = -0.281$, $P < 10^{-63}$) and d_N/d_S ($\rho = -0.175$, $P < 10^{-24}$) (based on 3,480 orthologs; table 1). Using direct mRNA abundance from RNA-seq experiments ($ExpLev_{mRNA}$), this negative correlation between mRNA abundance and evolutionary rates became stronger (d_N: $\rho = -0.546$, $P < 10^{-268}$; d_N/d_S: $\rho = -0.356$, $P < 10^{-104}$) (table 1). These results suggest that the influence of mRNA abundance on protein evolutionary rates of flagellate algae was underestimated when indirect

Table 2

Partial Rank Correlations of the Gene Properties with d_N or d_N/d_S After Controlling for All the Other Gene Properties

Gene Properties[a]	*Chlamydomonas reinhardtii* Genes[b]		*Volvox carteri* Genes[b]	
	ρ_p (*P* Value[c]) with d_N	ρ_p (*P* Value[c]) with d_N/d_S	ρ_p (*P* Value[c]) with d_N	ρ_p (*P* Value[c]) with d_N/d_S
CDS Length	0.340 ($<10^{-99}$)	0.247 ($<10^{-50}$)	0.592 ($<10^{-300}$)	0.446 ($<10^{-188}$)
#uAUG	0.041 (0.01)	0.074 ($<10^{-4}$)	0.035 (0.04)	−0.050 ($<10^{-2}$)
Gene expression level				
$ExpLev_{mRNA}$	−0.276 ($<10^{-63}$)	−0.128 ($<10^{-13}$)	NA	NA
CAI	−0.140 ($<10^{-16}$)	−0.067 ($<10^{-4}$)	−0.173 ($<10^{-24}$)	−0.043 (0.01)
Gene compactness				
Avg Intron Length	−0.077 ($<10^{-5}$)	−0.062 ($<10^{-3}$)	−0.016 (0.34)	−0.039 (0.02)
5′-UTR Length	−0.002 (0.93)	−0.063 ($<10^{-3}$)	−0.073 ($<10^{-4}$)	0.025 (0.13)
3′-UTR Length	−0.019 (0.25)	−0.012 (0.47)	−0.099 ($<10^{-8}$)	−0.030 (0.08)

NOTE.—NA, not applicable.

[a]All gene properties were based on *C. reinhardtii* genes.

[b]Gene properties can be defined based on *C. reinhardtii* genes or *Volvox carteri* genes, as indicated.

[c]*P* values show the probabilities of the observations under the hypothesis of no correlation.

expression estimates, such as CAI, are used (Popescu et al. 2006). Direct mRNA abundance data were used to reassess the effect of mRNA abundance on protein evolution of yeast genes and mammalian genes. In yeast genes, mRNA abundance remained the predominant evolutionary rate determinant for yeast proteins (supplementary table S2, Supplementary Material online). In mammalian genes, the correlation between mRNA abundance from oligonucleotide microarrays (Liao et al. 2010) and mammalian d_N or d_N/d_S was previously reported to be statistically insignificant (Liao et al. 2010), but with direct measurements of protein abundance, the correlation between mRNA abundance and protein evolutionary rated was significant and negative (supplementary table S2, Supplementary Material online). Therefore, the direct measurement of RNA-seq experiments clarified how mRNA abundance affects protein evolution. A limitation of microarray data in quantifying mRNA abundance lies in its inability to distinguish between alternative splicing events. This limitation only applied to the mammalian data, as most of the yeasts genes are not alternatively spliced. As a result, the influence of mRNA abundance in yeast protein evolution was the same whether mRNA was measured directly or indirectly.

For flagellated algae proteins, we observed a strong positive correlation between the number of start codons in the UTR (*#uAUG*) and d_N ($\rho = 0.281$, $P < 10^{-63}$) and d_N/d_S ($\rho = 0.205$, $P < 10^{-33}$) (table 1). UTR start codons can inhibit gene activities at both the transcription and the translation levels (Calvo et al. 2009; Yun et al. 2012). Along these lines, we observed significant negative correlations between both *#uAUG* and $ExpLev_{mRNA}$ ($\rho = -0.317$, $P < 10^{-81}$) and *#uAUG* and *CAI* ($\rho = -0.149$, $P < 10^{-18}$) for *C. reinhardtii* genes (fig. 1*A*). In the partial correlation analysis, the influence of *#uAUG* on d_N became only marginally significant ($P = 0.01$, based on *#uAUG* of *C. reinhardtii* genes; $P = 0.04$, based on *#uAUG* of *V. carteri* genes) (table 2), whereas $ExpLev_{mRNA}$ remained significantly negatively correlated with d_N ($\rho_p = -0.276$, $P < 10^{-63}$) and d_N/d_S ($\rho_p = -0.128$, $P < 10^{-13}$) (table 2). Thus, mRNA abundance is an important and independent factor determining the rate of algae protein evolution. Although our analyses on proteins of algae (tables 1 and 2), yeasts (supplementary table S2, Supplementary Material online), and mammals (supplementary table S2, Supplementary Material online) suggested that mRNA abundance has a universal effect on d_N or d_N/d_S, the observed correlation can be caused by avoidance of protein misfolding (Drummond et al. 2005; Drummond and Wilke 2008) or misinteraction (Yang et al. 2012), avoidance of mRNA misfolding (Park et al. 2013), selection on protein function (Cherry 2010), or a combination of all. Whether mRNA abundance affects protein evolution in all three lineages due to a common cause deserves further investigations.

Gene compactness was defined by the lengths of introns and UTRs. Gene compactness was positively correlated with d_N and d_N/d_S in mammals (Liao et al. 2006). In flagellated algae, average intron length (*Avg Intron Length*) was negatively correlated with d_N ($\rho = -0.085$, $P < 10^{-6}$) and d_N/d_S ($\rho = -0.075$, $P < 10^{-5}$) (3,480 orthologs; table 1). Partial correlation analysis confirmed an independent effect of intron length on d_N and d_N/d_S (table 2), although the correlation between *V. carteri*-based intron length and d_N was not statistically significant ($\rho_p = -0.016$, $P = 0.34$). The correlations between lengths of UTRs (*5′-UTR Length* and *3′-UTR Length*) and the rates of protein evolution were more difficult to interpret (table 1). *5′-UTR Length* was positively correlated with d_N ($\rho = 0.147$, $P < 10^{-17}$) and d_N/d_S ($\rho = 0.064$, $P < 10^{-3}$), whereas *3′-UTR Length* was not significantly correlated with d_N ($P = 0.23$) and negatively correlated with or d_N/d_S

($\rho = -0.038$, $P = 0.03$) (3,480 orthologs; table 1). The inconsistency was likely confounded by the inter-relatedness of gene properties (fig. 1). For example, genes with longer 5′-UTRs tend to have more uAUGs (*5′-UTR Length* vs. *#uAUG*: $\rho = 0.651$, $P < 10^{-300}$) and lower mRNA expression levels (*5′-UTR Length* vs. $ExpLev_{mRNA}$: $\rho = -0.165$, $P < 10^{-21}$) (fig. 1). The partial correlations for 5′-UTR Length or 3′-UTR Length versus d_N or d_N/d_S were negative or insignificant (table 2), similar to the results for gene compactness defined by average intron length. It is possible that the lack of partial correlation between 5′-*UTR Length* or 3′-*UTR Length* and d_N or d_N/d_S was due to misannotation of UTRs in algal genes. Nevertheless, together these results suggest that flagellated algal genes with greater gene compactness tend to evolve more rapidly. However, the influence of gene compactness is not as important as expression level in determining the rates of protein evolution after the divergence of *C. reinhardtii* and *V. carteri*. Using *C. reinhardtii* and *V. carteri* orthologous genes defined using the InParanoid algorithm yielded similar results (supplementary table S3, Supplementary Material online).

Protein Length Is the Most Important Gene Property Influencing Algae Protein Evolutionary Rates

Although both mRNA abundance and gene compactness affect d_N (or d_N/d_S) after the divergence of *C. reinhardtii* and *V. carteri*, neither had the most significant correlation with d_N (or d_N/d_S) (tables 1 and 2). Instead, CDS length had the strongest effect on d_N (or d_N/d_S) among all the gene properties examined (*CDS Length* vs. d_N: $\rho = 0.552$, $P < 10^{-275}$; *CDS Length* and d_N/d_S: $\rho = 0.403$, $P < 10^{-135}$) (3,480 orthologs, table 1). This result holds even after controlling for other gene properties (table 2). Using *C. reinhardtii–V. carteri* orthologs defined by the InParanoid algorithm generated a minor difference for *C. reinhardtii*-based gene properties to d_N ($ExpLev_{mRNA}$-d_N ρ_p was slightly stronger than *CDS Length*-d_N ρ_p), not d_N/d_S (supplementary table S3, Supplementary Material online). The role of CDS length in determining the rate of protein evolution has been difficult to discern. Using 363 mouse–rat orthologs, there was a significant, but weak, negative correlation between protein lengths and evolutionary rates (Zhang 2000). However, with a larger data set (~3,500 mouse–rat orthologs), no correlation was found between protein lengths and evolutionary rates (Liao et al. 2006). In the fruit fly, longer proteins were reported to evolve more rapidly (Lemos et al. 2005). In yeasts, the correlation between protein lengths and evolutionary rates depends on the range of protein sizes (Bloom et al. 2006). The widely different effect of protein length effect on d_N or d_N/d_S suggested that any observed correlation was minor or a byproduct caused by confounding factors, such as mRNA expression. However, our findings indicated that *CDS Length* plays a major, independent role in generating the evolutionary rate variation of algae proteins.

To evaluate the validity of the strong positive correlation between *CDS Length* and d_N or d_N/d_S in algae, we examined how potential alignment errors in *C. reinhardtii–V. carteri* orthologs affect this correlation. Errors in sequencing, CDS annotation, and alignment should inflate both d_N and d_N/d_S estimates because these errors do not differentiate between nonsynonymous and synonymous sites (Stoletzki and Eyre-Walker 2011). We repeated our analysis on a subset of *C. reinhardtii–V. carteri* orthologs with "high quality" alignments defined by Heads-or-tails (HoT) scores (Landan and Graur 2007). The HoT column score (or residue score) is the fraction of identical aligned columns (or paired residues that are identical) between the "Heads" alignment, generated from the original sequences, and the "Tails" alignment, generated from the reversed sequences. These scores were calculated for each orthologous pair. The analysis for table 2 was repeated on a subset of 2,923 (or 2,903) orthologs with high HoT column scores ≥ 0.8 (or residue scores ≥ 0.8) and supplementary table S4, Supplementary Material online (or supplementary table S5, Supplementary Material online), was generated. Compared with table 2, values of ρ_p shown in supplementary table S4 or S5, Supplementary Material online, were in general weaker. This can be due to the fact that the orthologs with truly and very diverged CDSs were unavoidably excluded after the data filtering and a bias was introduced. Nevertheless, within both subsets of orthologs, although mRNA abundance had a slightly stronger correlation with d_N than CDS length did, CDS length still had the strongest correlation with d_N/d_S among the factors examined.

We expected that orthologs that prone to alignment errors should have greater differences in protein length and a higher fraction of unalignable residues. To group genes according to their propensity to have inaccurate alignments, we calculated protein length dissimilarity as $\Delta L = |L_C - L_V|/(L_C + L_V)$ and the proportion of unalignable residues as $UnalignRes = 1 - [2 \times N_{aligned}/(L_C + L_V)]$ for each ortholog, where $N_{aligned}$ is the number of aligned codons, and L_C and L_V are lengths of the protein sequence in *C. reinhardtii* and *V. carteri* genes, respectively. CDS length in *C. reinhardtii* genes was positively correlated with both ΔL ($\rho = 0.393$, $P < 10^{-128}$) and *UnalignRes* regardless of alignment tool used (ClustalW: $\rho = 0.412$, $P < 10^{-142}$; MUSCLE: $\rho = 0.481$, $P < 10^{-200}$), indicating in general there was better alignment for orthologs encoding shorter proteins. To evaluate whether an overestimated d_N or d_N/d_S of longer proteins led to the strong correlation between *CDS Length* and d_N or d_N/d_S (tables 1 and 2), we divided all orthologs into two equal-sized groups according to ΔL or *UnalignRes* and examined the relative strengths of correlation for CDS length and mRNA expression level versus d_N (or d_N/d_S) (table 3). These correlations were stronger in the groups of orthologs that predicted to have fewer alignment errors (lower ΔL and lower *UnalignRes*), indicating that high quality annotation and alignment facilitates the identification of protein

Table 3
Effect of Alignment on Rank Correlations of *CDS Length* or $ExpLev_{mRNA}$ with d_N and d_N/d_S

Orthologs	ClustalW		MUSCLE	
	ρ (*P* Value[a]) with d_N	ρ (*P* Value[a]) with d_N/d_S	ρ (*P* Value[a]) with d_N	ρ (*P* Value[a]) with d_N/d_S
Similar *CDS Length* (bottom 50% ΔL)				
CDS Length	0.687 ($<10^{-243}$)	0.483 ($<10^{-101}$)	0.666 ($<10^{-222}$)	0.444 ($<10^{-84}$)
$ExpLev_{mRNA}$	−0.647 ($<10^{-206}$)	−0.361 ($<10^{-54}$)	−0.636 ($<10^{-197}$)	−0.319 ($<10^{-41}$)
Dissimilar *CDS Length* (top 50% ΔL)				
CDS Length	0.390 ($<10^{-63}$)	0.318 ($<10^{-41}$)	0.371 ($<10^{-57}$)	0.295 ($<10^{-35}$)
$ExpLev_{mRNA}$	−0.418 ($<10^{-73}$)	−0.340 ($<10^{-47}$)	−0.410 ($<10^{-70}$)	−0.323 ($<10^{-42}$)
Strongly alignable (bottom 50% *UnalignRes*)				
CDS Length	0.661 ($<10^{-218}$)	0.467 ($<10^{-94}$)	0.592 ($<10^{-163}$)	0.361 ($<10^{-53}$)
$ExpLev_{mRNA}$	−0.630 ($<10^{-192}$)	−0.346 ($<10^{-49}$)	−0.573 ($<10^{-151}$)	−0.236 ($<10^{-22}$)
Poorly aligned (top 50% *UnalignRes*)				
CDS Length	0.382 ($<10^{-60}$)	0.311 ($<10^{-39}$)	0.371 ($<10^{-57}$)	0.319 ($<10^{-41}$)
$ExpLev_{mRNA}$	−0.414 ($<10^{-72}$)	−0.340 ($<10^{-47}$)	−0.416 ($<10^{-73}$)	−0.353 ($<10^{-51}$)

[a]*P* values show the probabilities of the observations under the hypothesis of no correlation.

evolutionary rate determinants. Within the groups with low ΔL or low *UnalignRes*, the correlation between *CDS Length* and d_N (or d_N/d_S) was stronger than that of $ExpLev_{mRNA}$ and d_N (or d_N/d_S). In contrast, the correlation between $ExpLev_{mRNA}$ and d_N (or d_N/d_S) was stronger than that of *CDS Length* and d_N (or d_N/d_S) in the groups of orthologs prone to have alignment errors (table 3), suggesting alignment issues inflate rather than underestimate the relative importance of mRNA expression on protein evolutionary rate in our analysis. Additionally, the partial correlations of *CDS Length* and d_N (or d_N/d_S) after controlling for ΔL (*CDS Length* vs. d_N: $\rho_p = 0.543$, $P < 10^{-300}$; *CDS Length* vs. d_N/d_S: $\rho_p = 0.415$, $P < 10^{-158}$) or *UnalignRes* (*CDS Length* vs. d_N: $\rho_p = 0.528$, $P < 10^{-293}$; *CDS Length* vs. d_N/d_S: $\rho_p = 0.409$, $P < 10^{-153}$) were slightly stronger than those of $ExpLev_{mRNA}$ and d_N (or d_N/d_S) after controlling for ΔL ($ExpLev_{mRNA}$ vs. d_N: $\rho_p = -0.535$, $P < 10^{-300}$; $ExpLev_{mRNA}$ vs. d_N/d_S: $\rho_p = -0.363$, $P < 10^{-115}$) or *UnalignRes* ($ExpLev_{mRNA}$ vs. d_N: $\rho_p = -0.520$, $P < 10^{-281}$; $ExpLev_{mRNA}$ vs. d_N/d_S: $\rho_p = -0.355$, $P < 10^{-110}$), suggesting that alignment quality does not explain the stronger correlation between CDS length and protein evolutionary rate compared with that of mRNA abundance and protein evolutionary rate. Furthermore, the observed influence of CDS length on protein evolutionary rate is real and unlikely an artifact of poor sequence alignment.

Domain Density Does Not Account for the Influence of Protein Length on Evolutionary Rates

It is interesting to consider why CDS length is important to algae protein evolution. Decreased protein length has been associated with increased efficiency of protein synthesis (e.g., Akashi [2003]) that is needed for highly expressed proteins (Coghlan and Wolfe 2000; Jansen and Gerstein 2000), which evolve slowly. However, according to the results of our partial correlation analysis (table 2), mRNA expression levels do not contribute to the influence of CDS length on protein evolutionary rates. In prokaryotes, variation in protein length is caused by variation in the length of linkers (nondomain protein regions) connecting protein domains (Wang et al. 2011). These nondomain regions of the protein were less constrained in sequence evolution, likely due to lower functional importance (Wilson et al. 1977; Steiner et al. 1985) or higher intrinsic disorderness (Orengo and Thornton 2005; Brown et al. 2011). Therefore, we tested the hypothesis that in algae, longer proteins tend to evolve more rapidly because they contain a larger proportion of linker sequences or a smaller proportion of domain sequences (*domain*%).

Among the 3,480 *C. reinhardtii* genes with a one-to-one ortholog in the *V. carteri* genome, there were 626 *C. reinhardtii* genes containing protein domains annotated by SUPERFAMILY. The small number of *C. reinhardtii* genes with SUPERFAMILY domains reflects the incompleteness of current protein domain annotation. Despite the small sample size, we observed a strong negative correlation between *CDS Length* and *domain*%, defined as the percent of amino acids residing in SUPERFAMILY-annotated protein domains of a protein ($\rho = -0.490$, $P < 10^{-38}$; fig. 2*A*), suggesting that longer proteins have a higher proportion of residues outside of protein domains, which potentially explains the strong correlation between *CDS Length* and d_N (or d_N/d_S). Furthermore, when we divided the 626 orthologs into three groups according to *domain*%: *domain*% ≤ 50% ($n = 212$), 50% < *domain*% ≤ 75% ($n = 167$), and *domain*% > 75% ($n = 247$), we found that within each *domain*% group, *CDS Length* had a similar positive correlation with d_N ($\rho = 0.441$–0.551, $P < 10^{-8}$–10^{-20}; fig. 2*B*) and d_N/d_S ($\rho = 0.231$–0.339, $P < 10^{-3}$–10^{-7}; fig. 2*C*). We divided each protein sequence

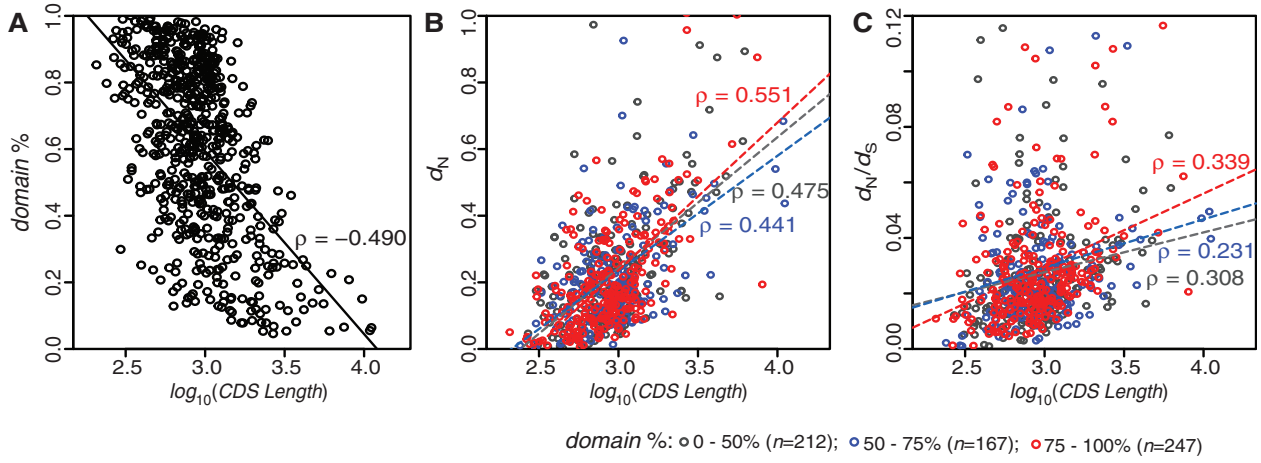

Fig. 2.—(*A*) Genes with longer CDS (larger *CDS Length*) encode proteins with a lower percent of the total sequence length annotated as protein domains (*domain*%). Controlling for domain%, CDS Length remains positively correlated with (*B*) d_N and (*C*) d_N/d_S. The linear regression line and the Spearman's rank correlation coefficient are shown for each domain% bin.

into domain regions and nondomain regions, generated a concatenate domain and nondomain protein for each protein, and calculated the domain-specific or nondomain-specific d_N (or d_N/d_S) for each gene. *CDS Length* was positively correlated with both domain-specific d_N ($\rho = 0.346$, $P < 10^{-14}$) (or d_N/d_S, $\rho = 0.134$, $P < 10^{-2}$) and nondomain-specific d_N ($\rho = 0.411$, $P < 10^{-21}$) (or d_N/d_S, $\rho = 0.118$, $P < 10^{-2}$). These results suggest that domain density does not sufficiently explain the effect of CDS length on protein evolution.

Concluding Remarks

As previously observed in yeasts (Drummond et al. 2006) and mammals (Liao et al. 2006), we found that mRNA abundance and gene compactness influence the evolutionary rates of flagellated algae proteins. However, for algae proteins, gene compactness was only a minor determinant of protein evolutionary rates, and its influence was confounded by the number of start codons in the UTR. Although the hypothesis that compact genes evolve faster because they lack introns that promote the efficacy of natural selection (Liao et al. 2006) cannot be disproven, it is also possible that gene properties associated with regulatory motifs (e.g., uAUG-like motifs and miRNA target sites) better explain the difference in relative importance of gene compactness in determining protein evolutionary rate in yeasts, mammals and algae.

In flagellated algae, CDS length had the greatest independent influence on protein evolutionary rate. This finding was unexpected as genome-wide investigations in yeasts (Drummond et al. 2006) and mammals (Liao et al. 2006) found no correlation between protein length and d_N (or d_N/d_S). Although the underlying cause for this correlation remains to be understood, it is independent from associations with gene expression (Drummond et al. 2005; Yang et al. 2012) and protein domain density, which is similar to functional density (Wilson et al. 1977). Recent studies have shown that the chaperone-mediated protein folding can accelerate protein evolution (Bogumil and Dagan 2010; Warnecke and Hurst 2010). It is unknown whether algae genes encoding longer proteins tend to code for substrates of chaperones. When chaperonin-dependency data or protein–protein interaction data become available for algae, it will be interesting to examine whether chaperone–substrate interaction plays a role in the effect of CDS length on flagellated algae protein evolutionary rates.

Although many studies have searched for universal rules explaining sequence evolution (Li 1997; Wilson et al. 1977; Zeldovich and Shakhnovich 2008; Koonin 2011), our study suggests that although determinants of protein evolutionary rates can be common among multiple eukaryotic lineages (e.g., Drummond and Wilke [2008]), their relative importance can differ between lineages. Consequently, the implications of sequence conservation in a less well-studied lineage are often unpredictable and require further lineage-specific investigation.

Supplementary Material

Supplementary tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank anonymous reviewers for constructive comments. This study was supported by the intramural funding of National Health Research Institutes, Taiwan, and research grant (grant number NSC 101-2311-B-400-001-MY3) from the National Science Council, Taiwan, to B.-Y.L.

Literature Cited

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics 164:1291–1303.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 23: 1751–1761.

Boffelli D, et al. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391–1394.

Bogumil D, Dagan T. 2010. Chaperonin-dependent accelerated substitution rates in prokaryotes. Genome Biol Evol. 2:602–608.

Brawand D, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. 2011. Evolution and disorder. Curr Opin Struct Biol. 21:441–446.

Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 106: 7507–7512.

Chang AY, Liao B-Y. 2012. DNA methylation rebalances gene dosage after Mammalian gene duplications. Mol Biol Evol. 29:133–144.

Cherry JL. 2010. Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol. 2:757–769.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast 16: 1131–1145.

Comeron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics 161:389–410.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102: 14338–14343.

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

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Elnitski L, et al. 2003. Distinguishing regulatory DNA from neutral sites. Genome Res. 13:64–72.

Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A. 101: 9033–9038.

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

Goodstein DM, et al. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40:D1178–D1186.

Gu X, Su Z. 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. Proc Natl Acad Sci U S A. 104: 2779–2784.

Hallmann A. 2011. Evolution of reproductive development in the volvocine algae. Sex Plant Reprod. 24:97–112.

Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. Proc Natl Acad Sci U S A. 106:3254–3258.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046–1049.

Hudson CM, Conant GC. 2011. Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. BMC Evol Biol. 11:89.

Jansen R, Gerstein M. 2000. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res. 28:1481–1488.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12:962–968.

Kianianmomeni A, Nematollahi G, Hallmann A. 2008. A gender-specific retinoblastoma-related protein in *Volvox carteri* implies a role for the retinoblastoma protein family in sexual development. Plant Cell 20: 2399–2419.

Kim SH, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. Genetica 131:151–156.

Kirk MM, Ransick A, McRae SE, Kirk DL. 1993. The relationship between cell size and cell fate in *Volvox carteri*. J Cell Biol. 123:191–208.

Koonin EV. 2011. Are there laws of genome evolution? PLoS Comput Biol. 7:e1002173.

Labadorf A, et al. 2010. Genome-wide analysis of alternative splicing in *Chlamydomonas reinhardtii*. BMC Genomics 11:114.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. 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. 22:1345–1354.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.

Liao B-Y, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 23:2072–2080.

Liao B-Y, Weng M-P, Zhang J. 2010. Impact of extracellularity on the evolutionary rate of mammalian proteins. Genome Biol Evol. 2010: 39–43.

Merchant SS, et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245–250.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5:621–628.

Nagalakshmi U, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349.

O'Brien KP, Remm M, Sonnhammer EL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33: D476–D480.

Orengo CA, Thornton JM. 2005. Protein families and their evolution—a structural perspective. Annu Rev Biochem. 74:867–900.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Park C, Chen X, Yang JR, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 110:E678–E86.

Pennacchio LA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502.

Popescu CE, Borza T, Bielawski JP, Lee RW. 2006. Evolutionary rates and expression level in *Chlamydomonas*. Genetics 172:1567–1576.

Prochnik SE, et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329:223–226.

Qian W, Liao B-Y, Chang AY, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet. 26:425–430.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21:108–116.

Rochaix JD. 1995. *Chlamydomonas reinhardtii* as the photosynthetic yeast. Annu Rev Genet. 29:209–230.

Schad E, Tompa P, Hegyi H. 2011. The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 12:R120.

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

Slotte T, et al. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. Genome Biol Evol. 3:1210–1219.
Steiner DF, Chan SJ, Welsh JM, Kwok SC. 1985. Structure and evolution of the insulin gene. Annu Rev Genet. 19:463–484.
Stoletzki N, Eyre-Walker A. 2011. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. Mol Biol Evol. 28:1371–1380.
Sultan M, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960.
Thomas JW, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. PLoS Comput Biol. 2:e48.
Wang M, Kurland CG, Caetano-Anolles G. 2011. Reductive evolution of proteomes and protein structures. Proc Natl Acad Sci U S A. 108: 11954–11958.
Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. Mol Syst Biol. 6:340.
Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem. 46:573–639.
Wilson D, Madera M, Vogel C, Chothia C, Gough J. 2007. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. 35:D308–D313.
Wyder S, Kriventseva EV, Schroder R, Kadowaki T, Zdobnov EM. 2007. Quantification of ortholog losses in insects and vertebrates. Genome Biol. 8:R242.
Xiong YY, et al. 2010. RNA sequencing shows no dosage compensation of the active X-chromosome. Nat Genet. 42:1043–1047.
Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A. 109:E831–E840.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yun Y, Adesanya TM, Mitra RD. 2012. A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. Genome Res. 22:1089–1097.
Zeldovich KB, Shakhnovich EI. 2008. Understanding protein evolution: from protein physics to Darwinian selection. Annu Rev Phys Chem. 59:105–127.
Zhang J. 2000. Protein-length distributions for the three domains of life. Trends Genet. 16:107–109.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol. 22: 1147–1155.

Associate editor: Tal Dagan