# 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

*Bernardo Lemos,* Brian R. Bettencourt,† Colin D. Meiklejohn,* and Daniel L. Hartl**

*Department of Organismic and Evolutionary Biology, Harvard University; and †Department of Biological Sciences, University of Massachusetts Lowell

Organismic evolution requires that variation at distinct hierarchical levels and attributes be coherently integrated, often in the face of disparate environmental and genetic pressures. A central part of the evolutionary analysis of biological systems remains to decipher the causal connections between organism-wide (or genome-wide) attributes (e.g., mRNA abundance, protein length, codon bias, recombination rate, genomic position, mutation rate, etc) as well as their role—together with mutation, selection, and genetic drift—in shaping patterns of evolutionary variation in any of the attributes themselves. Here we combine genome-wide evolutionary analysis of protein and gene expression data to highlight fundamental relationships among genomic attributes and their associations with the evolution of both protein sequences and gene expression levels. Our results show that protein divergence is positively coupled with both gene expression polymorphism and divergence. We show moreover that although the number of protein-protein interactions in *Drosophila* is negatively associated with protein divergence as well as gene expression polymorphism and divergence, protein-protein interactions cannot account for the observed coupling between regulatory and structural evolution. Furthermore, we show that proteins with higher rates of amino acid substitutions tend to have larger sizes and tend to be expressed at lower mRNA abundances, whereas genes with higher levels of gene expression divergence and polymorphism tend to have shorter sizes and tend to be expressed at higher mRNA abundances. Finally, we show that protein length is negatively associated with both number of protein-protein interactions and mRNA abundance and that interacting proteins in *Drosophila* show similar amounts of divergence. We suggest that protein sequences and gene expression are subjected to similar evolutionary dynamics, possibly because of similarity in the fitness effect (i.e., strength of stabilizing selection) of disruptions in a gene's protein sequence or its mRNA expression. We conclude that, as more and better data accumulate, understanding the causal connections among biological traits and how they are integrated over time to constrain or promote structural and regulatory evolution may finally become possible.

## Introduction

Multiple factors can influence the rate of protein divergence among species, including a gene's level of expression (Subramanian and Kumar 2004), its genomic localization (Williams and Hurst 2000), breadth of tissue expression (Duret and Mouchiroud 2000), functional role (Castillo-Davis et al. 2004), and pattern of sex-biased expression (Zhang, Hambuch, and Parsch 2004). Similarly, a wide range of factors influences the extent of gene expression polymorphism and divergence, including the pattern of sex-biased expression (Meiklejohn et al. 2003; Ranz et al. 2003) and membership in particular functional categories (Rifkin, Kim, and White 2003; Lemos et al. 2005). Because a variety of organismic attributes can similarly constrain both protein and gene expression evolution (e.g., protein-protein interactions; Fraser et al. 2002; Lemos, Meiklejohn, and Hartl 2004), coupling between regulatory and structural evolution may be expected. Furthermore, variation in mutation rate and selection parameters (e.g., strength of stabilizing selection) among genes is expected to result in similar pressures on variation in protein sequences and gene expression, which could thus lead to evolutionary coupling between structural and regulatory evolution.

However, the relationship between evolutionary variation in gene expression and protein sequence is controversial. For instance, Wagner (2000) examined a small number of gene duplicates in yeast and was unable to detect a clear association between divergence in protein-coding sequences and mRNA levels. This led Wagner (2000) to conclude that the evolution of protein sequences and mRNA expression is largely decoupled. More recently, Gu et al. (2002) and Makova and Li (2003) analyzed large samples of gene duplicates in yeast and humans, respectively, and were able to show a positive association between divergence in protein sequences and mRNA levels between duplicates. In addition, more recent studies have examined the hypothesis that protein sequence and gene expression evolution between orthologs may be correlated. A recent analysis of 156 gene orthologs between *Drosophila melanogaster* and *Drosophila simulans* suggested that protein and gene expression evolution may indeed be coupled (Nuzhdin et al. 2004), although the authors acknowledged the possibility that sequence divergence may have contributed to expression divergence as measured with oligonucleotide arrays. On the other hand, another recent study found no association between the rate of protein sequence divergence and the extent of gene expression divergence between human and mouse (Jordan et al. 2004), which led the authors to conclude that these two modes of evolution (i.e., regulatory and structural) were largely decoupled.

Protein-protein interactions may influence protein sequence evolution as well as regulatory evolution if an increased number of interactions results in stronger purifying selection against amino acids substitutions (Fraser et al. 2002) and also stronger stabilizing selection to maintain evolutionarily stable gene expression levels (Lemos, Meiklejohn, and Hartl 2004). Stronger stabilizing selection

Key words: selection, constraint, neutral, network, coupling, correlation, Drosophila.

E-mail: blemos@oeb.harvard.edu.

*Mol. Biol. Evol.* 22(5):1345–1354. 2005
doi:10.1093/molbev/msi122
Advance Access publication March 2, 2005

in highly connected proteins may be expected if these proteins are functionally more important to the cell. Such stronger selection may therefore result in highly connected proteins showing both lower rates of amino acid substitution and lower levels of gene expression polymorphism and divergence. Indeed, an increased number of protein-protein interactions may impose increased stoichiometric demands on protein concentration and therefore result in a stronger pressure for maintaining evolutionarily stable gene expression levels (Veitia 2002). That such highly connected proteins are indeed more important is suggested by the observations that the removal of network hubs results in drastic disruptions in the structure of scale-free networks (Albert, Jeong, and Barabasi 2000; Jeong et al. 2000) and, in mutant yeast, results in larger fitness decreases in the organism (Papp, Pal, and Hurst 2003). Moreover, the multiple interactions of highly connected proteins may play a role in constraining sequence evolution if the multiple interactions are mediated through different domains scattered along the sequence, thus resulting in greater selection pressure to maintain the specificity of protein-protein interactions along the protein's overall length.

Regardless of the specific mechanisms mediating an increased constraint in highly connected proteins, a negative association between the rate of protein divergence and the number of protein-protein interactions is to be expected as long as the strength of purifying selection does increase with the number of interacting partners of a protein. This hypothesis was examined by Fraser et al. (2002), who reported a negative association between the rates of protein evolution, as estimated between *Saccharomyces cerevisae* and *Caenorhabditis elegans*, and the number of interactions within the yeast proteome. Similarly, Teichmann (2002) analyzed the distribution of sequence identities between *S. cerevisae* and *Schizosaccharomyces pombe* and found that proteins involved in complexes are generally more conserved than proteins not known to be involved in complexes.

However, the relevance of protein-protein interactions in protein evolution has remained controversial and the merits of the negative association between the number of protein-protein interactions and the rate of protein evolution have been challenged (Bloom and Adami 2003; Fraser, Wall, and Hirsh 2003; Jordan, Wolf, and Koonin 2003; Bloom and Adami 2004; Fraser and Hirsh 2004; Hahn, Conant, and Wagner 2004). All analyses to date have centered on interactions between *S. cerevisae* proteins as well as on rates of protein evolution estimated by sequence comparisons between *S. cerevisae* and other, often distantly related, organisms. The conclusion that protein interactions constrain sequence evolution has been disputed on the grounds that it may be an artifact caused by estimates of the rate of sequence evolution being confounded with absolute gene expression level (mRNA abundance). This confounding occurs because the rate of protein evolution is negatively associated with absolute transcript abundance (Pal, Papp, and Hurst 2001; Jordan et al. 2004; Subramanian and Kumar 2004), and some yeast data sets reporting protein interactions show a trend toward recording more interactions in highly expressed genes (Bloom and Adami 2003). Nevertheless, Lemos, Meiklejohn, and Hartl (2004) recently showed that protein interactions constrain gene expression polymorphism as well as gene expression divergence and that this effect is independent of gene expression level.

Here we address whether regulatory and structural evolutions are coupled in *Drosophila*. We also carry out a global analysis examining the effects of the number of protein-protein interactions, mRNA abundance (gene expression level), and protein length on structural and regulatory evolutions. We find that protein and nucleotide sequence divergence are positively coupled with gene expression polymorphism and divergence and that while the number of protein-protein interactions has a negative association to both of these modes of evolution, it does not seem to be a significant effect underlying the coupling between protein and gene expression evolution. We find, moreover, that mRNA abundance and protein length have contrasting associations with protein and gene expression evolution.

## Materials and Methods

### Protein and Nucleotide Sequence Divergence

*Drosophila melanogaster–Drosophila pseudoobscura* orthologs ($N = 10{,}987$ pairs) and Drosophila-*Anopheles gambiae* orthologs ($N = 1{,}255$ trios) were identified (Adams et al. 2000; Holt et al. 2002; Richards et al. 2005) by the reciprocal best-hit method (Tatusov, Koonin, and Lipman 1997) using BlastP (e value $< 10^{-5}$). Nucleotide sequences of orthologs were aligned with ClustalW (Thompson, Higgins, and Gibson 1994) using default parameters. To avoid spurious orthology assignment as well as alignment difficulties, genes were removed from the gene set (1) if they showed more than 50% protein sequence divergence between *D. melanogaster* and *D. pseudoobscura*, (2) if alignment gaps corresponded to more than 10% of the protein length in *D. melanogaster*, or (3) if less than 65% of the protein in *D. melanogaster* could be matched with orthologous amino acids in *D. pseudoobscura*. This procedure slightly reduced the number of *D. melanogaster–D. pseudoobscura* gene pairs ($N = 8{,}748$) but substantially improved the reliability of the sequence divergence data. For the set of 8,748 *D. melanogaster–D. pseudoobscura* alignments, maximum likelihood estimates of the number of nonsynonymous nucleotide substitution per nonsynonymous nucleotide site (*dN*) and the number of synonymous substitutions per synonymous site (*dS*) were obtained with yn00 method (Yang and Nielsen 2000) using PAML (Yang 1997, 2002). We note, however, that *dS* is expected to be saturated between *D. melanogaster–D. pseudoobscura* gene pairs, and estimates of this parameter are not reliable. For the set of 1,255 Drosophila-*A. gambiae* orthologs, maximum likelihood estimates of the rate of protein sequence evolution ($\omega$) were obtained with codeml method using PAML (Yang 1997, 2002).

### Gene Expression Polymorphism and Divergence

Genome-wide gene expression data on evolutionary variation in gene expression levels within *D. melanogaster* (gene expression polymorphism) and between *D. melanogaster* and *D. simulans* (gene expression divergence) were reported by Meiklejohn et al. (2003) and Ranz et al.

(2003), respectively. In both data sets, genome-wide gene expression levels were estimated using cDNA microarray platforms. Relative gene expression levels were obtained by a Bayesian procedure using the BAGEL (Bayesian Analysis of Gene Expression Levels) program (Townsend and Hartl 2002). Briefly, BAGEL analyses result in estimates of the fold-change in expression across samples that are normalized so that the lowest observed value is arbitrarily assigned a reference value of 1. Using a Markov Chain Monte Carlo method, BAGEL estimates 95% credible intervals for the expression level of each gene in each strain. Genes whose 95% credible intervals are nonoverlapping in at least two strains show statistically significant evolutionary variation in expression levels. Variances of the BAGEL-normalized data across eight strains of *D. melanogaster* were used as estimates of gene expression polymorphism, whereas the normalized expression difference ($DE_{ij} = |E_i - E_j|/(E_i + E_j)$; where $E_i$ and $E_j$ are gene expression levels in species *i* and *j*, respectively) was used to measure divergence in gene expression between *D. melanogaster* and *D. simulans*.

### Protein Interaction Data

High-confidence protein interaction data for Drosophila were reported by Giot et al. (2003). These authors assigned confidence values to protein interactions (see Bader et al. 2004 for a fuller description of their methods) and identified 4,625 high-confidence interactions (confidence score $> 0.50$) and 5,477 low-confidence interactions (confidence score $< 0.50$). Their statistical analysis substantially reduced the false-positive error rate among protein interactions assigned with high confidence. Interactions were not assigned to proteins in which all interactions had confidence scores less than 0.45. The proteins for which the yeast two-hybrid (Y2H) screen failed to identify interactors were either left out of the analyses (only proteins with $\geq 1$ interactions were counted) or were coded as belonging to a class of proteins with zero interactions. Results for both coding procedures are reported and are in general agreement.

In order to assess the sensitivity of our results to false positives, we carried out several analyses with subsets of the data in which protein interactions were assigned with increasingly larger confidence score. The proportion of false positives decreases in subsets with increasing confidence score. Indeed, in the fly data, interactions with confidence score greater than 0.95 are enriched for experimentally validated protein associations previously reported. A negligible proportion of false positives is therefore to be expected in the subset of the data with the highest confidence score.

### Gene Expression Level (mRNA abundance)

In order to estimate the relationship between the number of protein interactions and "absolute" gene expression level in the fly data, we estimated RNA transcript abundance using the De Gregorio et al. (2001) gene expression data obtained with oligonucleotide arrays. Oligonucleotide arrays were analyzed using Dchip (Li and Wong 2001) and probe match (PM) intensities as a proxy for "absolute" mRNA abundance. PM and average difference (AD = PM – MM; probe match – probe mismatch) values were highly correlated (Pearson correlation coefficient = 0.91; $P < 10^{-10}$), producing similar results. Gene expression levels were estimated from the mean of five replicates of adult mRNA abundance in males of the OREGON R strain.

### Randomization of Protein Interactions

We calculated the absolute normalized difference between metrics of sequence divergence for each protein pair and averaged these values across all pairs of interacting proteins using the following statistic: $1/N \times \sum_{i=1\ to\ i=N}(|\ W_i - W_j\ |\ /W_i + W_j)$, where $W_i$ and $W_j$ are a metric of sequence divergence (*dN*, *%AA*) for protein *i* and *j*, respectively, and *N* is the number of protein-protein pairs. Similarity in levels of expression of interaction proteins was assessed in an analogous way, with $W_i$ and $W_j$ denoting gene expression level for interacting proteins *i* and *j*, respectively.

In order to assess similarity in rates of evolution of interacting proteins without the confounding effect of transcript abundance, divergence values were corrected by the following statistic: $1/N \times \sum_{i=1\ to\ i=N}[(|\ (W_i/E_i) - (W_j/E_j)\ |)/((W_i/E_i) + (W_j/E_j))]$, where $W_i$, $W_j$, and *N* are as above and $E_i$ and $E_j$ are the "absolute" transcript abundance of genes *i* and *j*, respectively.

In order to estimate the null distributions for the above statistics, we generated 10,000 random samples with the appropriate number of interacting protein-protein pairs in each. The average values for the above-described statistics were computed for each sample of randomly matched proteins. The resulting null distributions were used to assess statistical significance.

## Results and Discussion

We analyzed protein and nucleotide sequence evolution in 8,748 pairs of orthologs in *D. melanogaster* and *D. pseudoobscura* for which a single reciprocal best blast hit could be identified and gene models unambiguously aligned, as well as for 1,255 genes for which an ortholog between both Drosophila species and *A. gambiae* could be identified and unambiguously aligned. Metrics of protein and nucleotide sequence evolution (percent amino acid identity [*%AA*], number of nonsynonymous substitution per nonsynonymous codon [*dN*], and in the case of three species alignment, the rate of protein divergence [ω]) were matched to data on gene expression polymorphism and divergence (Meiklejohn et al. 2003; Ranz et al. 2003), data on whole-organism gene expression level (mRNA abundance; De Gregorio et al. 2001), data on protein-protein interactions recorded in *D. melanogaster* (Giot et al. 2003), and data on protein length in *D. melanogaster*.

The statistical analysis of high-throughput Y2H data of Giot et al. (2003) provided confidence scores for individual pairwise interactions and identified 4,625 interactions between *D. melanogaster* proteins with high confidence (confidence score $> 0.50$) and 5,477 interactions with low confidence (confidence score $< 0.50$). We assigned zero interactions to a protein if it was included in the

**Table 1**
**The Number of High-Confidence Protein-Protein Interactions does not Increase with Gene Expression Level in *Drosophila***

| Subset of Genes Included for Analysis | Gene Expression Level Versus Number of Protein-Protein Interactions | |
|---|---|---|
| All interactions included | $\rho = 0.06$, $P < 0.0001$, $N = 6{,}757$ | |
| Confidence score $< 0.45$ | $\rho = 0.07$, $P < 0.0001$, $N = 6{,}291$ | |
| | Including proteins without interactors[a] | Excluding proteins without interactors |
| Confidence score $> 0.50$ | $\rho = -0.03$, $P = 0.02$, $N = 6{,}369$ | $\rho = -0.007$, $P = 0.64$, $N = 4{,}495$ |
| Confidence score $> 0.55$ | $\rho = -0.03$, $P = 0.01$, $N = 5{,}925$ | $\rho = -0.004$, $P = 0.80$, $N = 4{,}051$ |
| Confidence score $> 0.65$ | $\rho = -0.04$, $P = 0.005$, $N = 5{,}147$ | $\rho = -0.004$, $P = 0.83$, $N = 3{,}273$ |
| Confidence score $> 0.75$ | $\rho = -0.04$, $P = 0.005$, $N = 4{,}288$ | $\rho = -0.006$, $P = 0.75$, $N = 2{,}414$ |
| Confidence score $> 0.85$ | $\rho = -0.04$, $P = 0.02$, $N = 3{,}379$ | $\rho = -0.006$, $P = 0.83$, $N = 1{,}505$ |
| Confidence score $> 0.95$ | $\rho = -0.03$, $P = 0.13$, $N = 2{,}667$ | $\rho = -0.02$, $P = 0.52$, $N = 793$ |

[a] Proteins without interactors were defined as those included in the Y2H screen (Giot et al. 2003) for which no single interaction with confidence score greater than 0.45 was detected.

analyses of Giot et al. (2003) but failed to have at least one interaction with confidence score greater than 0.45. Analyses were also done on data sets for which only proteins with $\geq 1$ interaction were included or for which a class of proteins with no interactions was included. In order to assess the sensitivity of our results to false positives, we carried out several analyses with subsets of the data in which only those protein-protein interactions with the largest confidence scores were included.

All associations reported are based on the Spearman rank correlation ($\rho$), a nonparametric metric of association for which no assumptions about the underling distribution of the data are made. Spearman partial rank correlations were used to assess the effects of potentially confounding variables on a particular bivariate association.

## Protein Length is Negatively Associated with the Number of Protein-Protein Interactions and mRNA Abundance

We begin by investigating the associations between mRNA abundance, protein length, and number of protein-protein interactions. It has been suggested that Y2H high-throughput protein-protein interaction data sets may be biased toward counting more interactions for highly expressed genes (Bloom and Adami 2003). For the fly data set herein examined (Giot et al. 2003), we find that counting interactions irrespective of the confidence score or those in the low-confidence subset resulted in protein interaction data sets that are biased toward counting more interactions for highly expressed genes. However, this bias is removed when only high-confidence interactions are considered (table 1). Hence, biases in the fly protein interaction data set are effectively removed by focusing on high-confidence protein-protein interactions.

We also investigated whether protein length may have an effect on the number of protein-protein interactions. We find a significant negative association between protein length and number of protein-protein interactions in the fly (table 2). This effect tends to be stronger in sets of high-confidence interactions, whereas it is weaker when the number of protein-protein interactions are counted irrespective of confidence score and, interestingly, is not present in the set with low-confidence interactions. This suggests that the negative association between number of protein interactions and protein length may have a biological relevance, although we cannot rule out the possibility that it may have been introduced by the statistical analysis to assign confidence to protein interactions.

The negative association between mRNA abundance (gene expression level) and protein length previously observed in yeast (Coghlan and Wolfe 2000; Jansen and

**Table 2**
**Protein Length is Negatively Associated with the Number of High-Confidence Protein-Protein Interactions in Drosophila**

| Subset of Genes Included for Analysis | Protein Length Versus Number of Protein-Protein Interactions | |
|---|---|---|
| All interactions included | $\rho = -0.06$, $P < 0.0001$, $N = 5{,}283$ | |
| Confidence score $< 0.45$ | $\rho = -0.01$, $P = 0.49$, $N = 4{,}931$ | |
| | Including proteins without interactors[a] | Excluding proteins without interactors |
| Confidence score $> 0.50$ | $\rho = -0.08$, $P < 0.0001$, $N = 4{,}976$ | $\rho = -0.06$, $P = 0.0002$, $N = 3{,}475$ |
| Confidence score $> 0.55$ | $\rho = -0.09$, $P < 0.0001$, $N = 4{,}631$ | $\rho = -0.07$, $P < 0.0001$, $N = 3{,}130$ |
| Confidence score $> 0.65$ | $\rho = -0.10$, $P < 0.0001$, $N = 4{,}008$ | $\rho = -0.07$, $P = 0.0002$, $N = 2{,}507$ |
| Confidence score $> 0.75$ | $\rho = -0.12$, $P < 0.0001$, $N = 3{,}347$ | $\rho = -0.07$, $P = 0.002$, $N = 1{,}846$ |
| Confidence score $> 0.85$ | $\rho = -0.15$, $P < 0.0001$, $N = 2{,}639$ | $\rho = -0.08$, $P = 0.01$, $N = 1{,}138$ |
| Confidence score $> 0.95$ | $\rho = -0.18$, $P < 0.0001$, $N = 2{,}111$ | $\rho = -0.06$, $P = 0.11$, $N = 610$ |

[a] Proteins without interactors were defined as those included in the Y2H screen (Giot et al. 2003) for which no single interaction with confidence score greater than 0.45 was detected.

**Table 3**
**Protein Length and mRNA Abundance Show Contrasting Associations with Gene Expression and Protein Evolution**

| | Gene Expression Level[a] | | Protein Length[a] | |
|---|---|---|---|---|
| | Ortholog in *Anopheles* | All proteins | Ortholog in *Anopheles* | All proteins |
| $dN$[b] | $\rho = -0.20$, $P < 0.0001$, $N = 1{,}089$ | $\rho = -0.13$, $P < 0.0001$, $N = 8{,}538$ | $\rho = 0.13$, $P < 0.0001$, $N = 1{,}147$ | $\rho = 0.03$, $P = 0.02$, $N = 7{,}987$ |
| *%AA*[b] | $\rho = -0.20$, $P < 0.0001$, $N = 1{,}258$ | $\rho = -0.12$, $P < 0.0001$, $N = 8{,}333$ | $\rho = 0.14$, $P < 0.0001$, $N = 1{,}328$ | $\rho = 0.01$, $P = 0.48$, $N = 9{,}243$ |
| Gene expression divergence[c] | — | $\rho = 0.07$, $P < 0.0001$, $N = 4{,}851$ | — | $\rho = -0.12$, $P < 0.0001$, $N = 4{,}088$ |
| Gene expression polymorphism[d] | — | $\rho = 0.17$, $P < 0.0001$, $N = 4{,}366$ | — | $\rho = -0.20$, $P < 0.0001$, $N = 3{,}693$ |

[a] Gene expression level and protein length are negatively and identically associated across all *Drosophila melanogaster* proteins ($\rho = -0.23, P < 0.0001, N = 8{,}333$) and across the smaller set of protein for which Drosophila-*Anopheles* orthologs were identified ($\rho = -0.23$, $P < 0.0001$, $N = 1{,}258$).

[b] *dN* and *%AA* calculated between *D. melanogaster* and *Drosophila pseudoobscura*.

[c] Gene expression divergence ($DE_{ij}$) calculated between *D. melanogaster* and *Drosophila simulans* males; only genes with significant ($P < 0.01$) differences between the two species were included.

[d] Gene expression polymorphism calculated for genes with greater than one expression allele across eight *D. melanogaster* strains.

Gerstein 2000) and vertebrates (Urrutia and Hurst 2003; Subramanian and Kumar 2004) is also present in our analysis of the insect data ($\rho = -0.23, P < 0.0001, N = 8{,}333$, all genes; $\rho = -0.23, P < 0.0001, N = 1{,}258$, genes with an ortholog in *Anopheles*). We, therefore, investigated whether the effect of protein length on the number of protein-protein interactions might be mediated by gene expression level. Surprisingly, the effect of protein length on the number of protein-protein interactions remains largely unchanged after the effect of gene expression is removed (e.g., number of protein interactions [confidence score $>$ 0.50] vs. protein length with gene expression level as a covariate: $\rho = -0.08$, $P < 0.0001$, $N = 3{,}231$). It is unclear, therefore, whether the puzzling association of protein length with the number of protein interactions reflects an unexpected technical bias unrelated to the effects of gene expression level or whether it results from a more fundamental and yet unexamined biological phenomenon that is not predicted by any current model of network structure (Barabasi and Oltvai 2004).

## Protein Length and mRNA Abundance Show Contrasting Associations with Gene Expression and Protein Evolution

Next, we investigated the effects of mRNA abundance and protein length on the evolution of Drosophila proteins as well as on levels of gene expression polymorphism and divergence. We find that the gene expression level is negatively associated with *dN* and *%AA* and that these associations hold true irrespective of whether we used the conserved set of genes with orthologs in *Anopheles* or the larger set of genes identified between the two Drosophila species (table 3). These findings are in agreement with previous studies that found a slower rate of evolution in highly expressed genes in bacteria, yeast, and mammals (Pal, Papp, and Hurst 2001; Urrutia and Hurst 2003; Rocha and Danchin 2004; Subramanian and Kumar 2004). The slower rate is likely to result from several causes including higher codon bias of highly expressed genes (Sharp and Li 1989; Moriyama and Hartl 1993), increased functional importance of highly expressed genes (Krylov et al. 2003), and higher bias in amino acid composition (lower protein complexity) of highly expressed genes (Coghlan and Wolfe 2000; Jansen and Gerstein 2000; Urrutia and Hurst 2003).

Conversely, we find that gene expression polymorphism and divergence are both positively correlated with gene expression level (table 3). Presumably, greater accuracy in measuring expression level in highly expressed genes results in increased statistical power to distinguish differences in expression levels. Because the error variance of gene expression estimates decreases with "absolute" mRNA abundance when using oligonucleotide arrays (Lemos et al. 2005), estimates of between-sample variances using oligonucleotide arrays are expected to increase with mRNA abundance. The resulting positive relationship in the between-sample variation in gene expression and absolute transcript abundance underscores the importance of correcting for the effects of mRNA abundance in analyses of gene expression data. It appears that this may be less of a problem with cDNA arrays (Townsend 2003).

We find that all metrics of protein and nucleotide sequence evolution are positively associated with protein length (table 3), and that the positive associations remain largely unchanged when the gene expression level is explicitly taken into account as a covariate. In addition, we find that gene expression polymorphism and divergence are both negatively correlated with protein length (table 3). This effect also remains largely unchanged when the effect of mRNA abundance (gene expression level) is controlled for as a covariate. These results establish protein length as a relevant attribute of both protein and gene expression evolution and indicate that these effects are largely independent of the effects of mRNA abundance. These observations are reminiscent of those of Duret and Mouchiroud (1999), who found a strong negative correlation between codon usage bias and protein length in yeast, worms, and flies, which could not be attributed to the negative association between protein size and mRNA abundance. Our results and those of others (Duret and Mouchiroud 1999; Marais and Duret 2001) suggest that the effects of protein length on molecular evolution may have been underappreciated. Although increasing protein length is typically associated with decreasing the efficiency of protein biosynthesis (e.g., Akashi 2003), the possibility that protein length may be

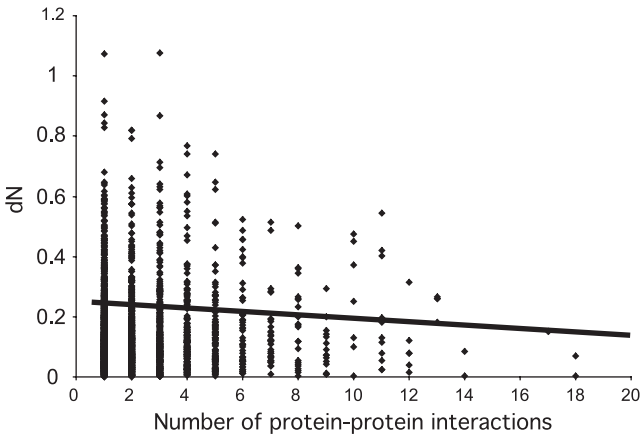


FIG. 1.—Protein divergence (number of synonymous substitutions per synonymous codon, *dN*) is negatively associated ($\rho = -0.10$, $P = 0.01$) with the number of protein-protein interactions (confidence score $> 0.50$) in *Drosophila*.

relevant to other aspects of molecular evolution needs to be more extensively examined.

## Protein and Gene Expression Evolution are Negatively Associated with the Number of Protein-Protein Interactions in Drosophila

In yeast, protein sequence evolution is negatively associated with the number of protein interactions (Fraser et al. 2002), although the merits of such association has been challenged on the grounds that it may be an artifact of the correlations between both protein evolution and number of protein interactions with gene expression level (Bloom and Adami 2003, 2004). We have therefore investigated the relationship between number of protein-protein interactions in *D. melanogaster* and various metrics of sequence evolution. In a preliminary analysis in which the effects of gene expression level and protein length were not considered, for the smaller data set of conserved proteins for which a mosquito ortholog was unambiguously identified and aligned, the estimates of ω, *dN*, and *%AA* are all negatively associated with the number of protein interactions in *D. melanogaster* (fig. 1). These associations are not substantially affected by including or excluding proteins without interactions (i.e., zero interactions) or by including only interactions in subsets of the protein interaction data with increasingly large confidence scores. On the other hand, in view of the effect of gene expression and protein length on metrics of sequence evolution, both of these variables must be explicitly included as covariates when investigating the effect of number of protein-protein interactions on sequence evolution. When this is done we find that ω, *dN*, and *%AA* remain significantly negatively associated with the number of protein interactions. The negative correlation does not depend on gene expression level, on the length of the protein, on the degree of confidence in the protein interactions, or on whether proteins with no interactions are included or excluded (table 4). We note moreover that *dS* is not associated with the number of protein-protein interactions. Our findings about protein sequence evolution in fruit flies parallel those indicating that evolutionary variation in expression levels (expression polymorphism and divergence) is also negatively associated with protein-protein interactions in yeast and fruit flies (Lemos, Meiklejohn, and Hartl 2004).

It has been suggested that highly connected proteins may contribute disproportionally to the negative association between rate of protein evolution and number of protein-protein interactions (Jordan, Wolf, and Koonin 2003) previously reported in yeast (Fraser et al. 2002). We therefore investigated whether the negative association in the fly data may be mostly driven by highly or lowly connected proteins. We find that the negative association holds true even when the analysis is restricted to proteins with a maximum number of protein-protein interactions as low as three or when proteins without interactions or lowly connected proteins are excluded (table 5).

## Protein and Gene Expression Evolution are Coupled

If protein sequence and gene expression evolution are subjected to shared constraints stemming, for instance, from similar stabilizing selection acting upon a gene's

**Table 4**
**Protein Evolution is Negatively Associated with the Number of Protein-Protein Interactions Independently of Gene Expression Level, the Length of the Protein, or the Confidence of the Protein-Protein Interaction Set**

| | Sequence Evolution Versus Number of Protein-Protein Interactions Controlled for Gene Expression Level and Protein Length | | | |
|---|---|---|---|---|
| | Excluding proteins without interactors | | Including proteins without interactors | |
| Subset of Genes Included for Analysis | *dN* | *%AA* | *dN* | *%AA* |
| Confidence score $> 0.50$ | $\rho = -0.10$, $P = 0.01$, $N = 723$ | $\rho = -0.10$, $P = 0.004$, $N = 825$ | $\rho = -0.05$, $P = 0.09$, $N = 1{,}043$ | $\rho = -0.05$, $P = 0.09$, $N = 1{,}200$ |
| Confidence score $> 0.55$ | $\rho = -0.09$, $P = 0.02$, $N = 650$ | $\rho = -0.10$, $P = 0.007$, $N = 743$ | $\rho = -0.06$, $P = 0.08$, $N = 970$ | $\rho = -0.05$, $P = 0.08$, $N = 1{,}118$ |
| Confidence score $> 0.65$ | $\rho = -0.13$, $P = 0.002$, $N = 517$ | $\rho = -0.12$, $P = 0.003$, $N = 596$ | $\rho = -0.06$, $P = 0.06$, $N = 837$ | $\rho = -0.06$, $P = 0.09$, $N = 971$ |
| Confidence score $> 0.75$ | $\rho = -0.11$, $P = 0.02$, $N = 392$ | $\rho = -0.12$, $P = 0.009$, $N = 453$ | $\rho = -0.07$, $P = 0.06$, $N = 712$ | $\rho = -0.07$, $P = 0.05$, $N = 828$ |
| Confidence score $> 0.85$ | $\rho = -0.12$, $P = 0.06$, $N = 250$ | $\rho = -0.12$, $P = 0.05$, $N = 295$ | $\rho = -0.09$, $P = 0.04$, $N = 570$ | $\rho = -0.08$, $P = 0.03$, $N = 670$ |
| Confidence score $> 0.95$ | $\rho = 0.003$, $P = 0.97$, $N = 149$ | $\rho = 0.02$, $P = 0.83$, $N = 170$ | $\rho = -0.07$, $P = 0.13$, $N = 469$ | $\rho = -0.06$, $P = 0.17$, $N = 545$ |

**Table 5**
**The Negative Relationship Between the Number of Protein-Protein Interaction and Protein Divergence is not Driven Disproportionaly by the Effects of Highly or Lowly Connected Proteins**

| Subset of Genes Included for Analysis | Number of Protein-Protein Interactions Versus Protein Evolution (ω) | |
|---|---|---|
| | Including proteins without interactors[a,b] | Excluding proteins without interactors[b] |
| All range of connectivity values | $\rho = -0.08$, $P = 0.003$, $N = 1{,}267$ | — |
| Connectivity $< 10$ | $\rho = -0.08$, $P = 0.005$, $N = 1{,}261$ | $\rho = -0.13$, $P < 0.0001$, $N = 866$ |
| Connectivity $< 8$ | $\rho = -0.08$, $P = 0.006$, $N = 1{,}250$ | $\rho = -0.13$, $P = 0.0001$, $N = 855$ |
| Connectivity $< 6$ | $\rho = -0.07$, $P = 0.02$, $N = 1{,}234$ | $\rho = -0.12$, $P = 0.0008$, $N = 839$ |
| Connectivity $< 4$ | $\rho = -0.05$, $P = 0.11$, $N = 1{,}164$ | $\rho = -0.10$, $P = 0.006$, $N = 769$ |
| Connectivity $> 0$ | $\rho = -0.13$, $P < 0.0001$, $N = 872$ | — |
| Connectivity $> 1$ | $\rho = -0.09$, $P = 0.08$, $N = 389$ | — |
| Connectivity $> 2$ | $\rho = -0.05$, $P = 0.49$, $N = 201$ | — |

[a] Proteins with no interactions are defined as those for which no single interaction scored higher than 0.45.
[b] Only interactions with confidence score greater than 0.55 were counted.

protein sequence and its expression, the structural and regulatory variations are expected to become evolutionarily coupled. Our results, as well as those of Fraser et al. (2002), Teichmann (2002), and Lemos, Meiklejohn, and Hartl (2004), suggest that the number of protein-protein interactions may produce a selection gradient that could, in spite of differences in mechanism, have similar effects on the evolution of protein sequences and gene expression levels. This could in turn lead to coupling of these two modes of evolution. Alternatively, evolutionary pressures on a protein's sequence may be largely independent of, or even in opposition to, evolutionary pressures that act on the expression level of the gene. Recent data concerning the association between the divergence of protein sequences and gene expression levels (Jordan et al. 2004; Nuzhdin et al. 2004) are conflicting. In our Drosophila data, we find that protein sequence evolution and gene expression evolution are indeed positively coupled (table 6).

Because of the opposing bivariate associations between gene expression level and protein length with protein sequence and gene expression evolution (fig. 2 and table 3), gene expression level and protein length are expected to obscure, rather than reinforce, the positive coupling between regulatory and protein sequence evolution. Accordingly, we find that the coupling holds true even when the effects of gene expression level and protein length are simultaneously taken into account. Furthermore, we find that in spite of the negative association between number of protein-protein interactions and regulatory and protein evolution (fig. 2 and table 3), the coupling between regulatory and protein sequence evolution is only slightly lessened when the effect of number of protein-protein interactions is taken into account. Therefore, even though our results and those of others (Fraser et al. 2002; Teichmann 2002; Lemos, Meiklejohn, and Hartl 2004) indicate that an increased number of protein-protein interactions results in an increased stabilizing selection both on gene expression variation and protein sequence, the number of protein-protein interactions does not seem to be a sufficient explanation for the coupling between regulatory and protein sequence evolution. If the number of protein-protein interactions is indeed irrelevant to the positive coupling between gene expression and protein evolution, we predict that the strength of this association should remain the same across proteins with only one interacting partner. In accord with this prediction, we find that the evolution of gene expression and protein sequences remains coupled in this subset of the data (gene expression polymorphism vs. *dN*, $\rho = 0.20$, $P < 0.0001$, $N = 359$; gene expression polymorphism vs. *%AA*, $\rho = 0.27$, $P < 0.0001$, $N = 417$). In conclusion, none of the attributes herein analyzed (mRNA abundance, protein length, and number of protein-protein interactions), whether considered separately or simultaneously, can account for the positive coupling between regulatory and protein evolution (table 6). Nevertheless, the positive coupling between protein sequence and gene expression evolution indicates that they are subjected to similar evolutionary dynamics, possibly because the fitness effects

**Table 6**
**Regulatory and Protein Evolution are Coupled in *Drosophila***

| | Gene Expression Polymorphism[a] | | Gene Expression Divergence[b] | |
|---|---|---|---|---|
| | | Controlled for gene expression level, protein length, and number of protein interactions | | Controlled for gene expression level, protein length, and number of protein interactions |
| *dN* | $\rho = 0.21$, $P < 0.0001$, $N = 1{,}560$ | $\rho = 0.15$, $P < 0.0001$, $N = 851$ | $\rho = 0.30$, $P < 0.0001$, $N = 283$ | $\rho = 0.27$, $P = 0.002$, $N = 130$ |
| *%AA* | $\rho = 0.17$, $P < 0.0001$, $N = 1{,}579$ | $\rho = 0.19$, $P < 0.0001$, $N = 970$ | $\rho = 0.25$, $P < 0.0001$, $N = 278$ | $\rho = 0.25$, $P = 0.002$, $N = 156$ |

[a] Gene expression polymorphism calculated for genes with greater than one expression allele across eight *Drosophila melanogaster* strains.
[b] Gene expression divergence ($DE_{ij}$) calculated between *D. melanogaster* and *Drosophila simulans* males. Only genes with significant ($P < 0.01$) differences between the two species were included.

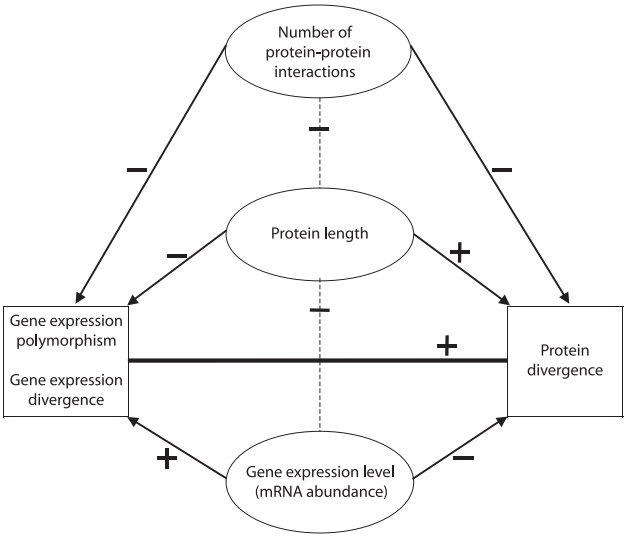


FIG. 2.—Qualitative summary of the sign (positive or negative) of the significant ($P < 0.05$) associations herein reported.

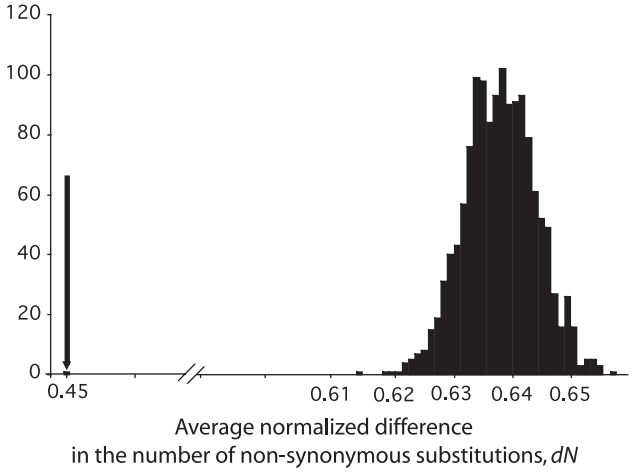


FIG. 3.—Interacting proteins have similar rates of protein divergence and this effect is not due to similar mRNA abundances of interacting proteins. Null distribution of the average normalized difference in the number of nonsynonymous substitutions (*dN*; between *Drosophila melanogaster* and *Drosophila pseudoobscura*) as estimated by 10,000 random samples generated by shuffling the list of interacting partners. *dN* values were corrected by mRNA abundance. The average normalized difference in *dN* between proteins that actually interact is an outlier and is indicated by an arrow ($P \ll 0.0001$).

(i.e., strength of stabilizing selection) of perturbations in the level of expression or in the protein sequence (appropriately scaled) may be similar.

## Interacting Proteins Show Similar Levels of Polymorphism and Divergence

Interacting proteins may show similar patterns of evolution if the members of an interacting protein-protein pair undergo similar evolutionary dynamics, stemming from a number of factors such as coevolution between interacting proteins, similar levels of stabilizing selection pressure in interacting proteins, or similar mutation rates (Fraser et al. 2002; Lemos, Meiklejohn, and Hartl 2004). We have therefore investigated whether proteins that interact in Drosophila show similar rates of protein evolution. Null distributions were calculated from 10,000 samples generated by shuffling the list of interacting partners. In agreement with Fraser et al. (2002), who found that interacting proteins in yeast evolve at similar rates of substitutions, we find that proteins that interact in Drosophila also show similar protein evolution as measured by ω ($P = 0.03$, $N = 191$ protein-protein pairs), *dN* ($P = 0.002$, $N = 2{,}532$ protein-protein pairs), and *%AA* ($P = 0.02$, $N = 2{,}398$ protein pairs). We find, moreover, that interacting proteins also have more similar expression levels than random sets of genes ($P = 0.04$, $N = 4{,}073$ protein-protein pairs). This finding raises the possibility that similarity in rates of evolution of interacting proteins may be a by-product of the similarity in gene expression level between interacting proteins. We find, however, that the similarity in protein divergence holds true when the effect of gene expression level is taken into account (ω, $P = 0.04$; *%AA*, $P = 0.07$; *dN*, $P < 0.0001$, fig. 3). These results, together with those of Fraser et al. (2002) and Lemos, Meiklejohn, and Hartl (2004), indicate that interacting proteins show similar levels of protein divergence and also similar levels of gene expression polymorphism both in yeast and flies.

Fraser et al. (2002) argued that available fitness data would not be sufficient to explain the observed similarity in rates of evolution of interacting proteins and therefore favored the interpretation that the similarity in rates of evolution between interacting proteins arises from coevolution between interacting partners rather than by their sharing of similar strengths of purifying selection. However, in view of the finding of Lemos, Meiklejohn and Hartl (2004) that interacting proteins show similar breadths of population genetic variation in gene expression within species, we emphasize that similarity in the strength of stabilizing selection between interacting proteins may also be relevant. If interacting proteins do indeed share similar stabilizing selection pressures, we predict that levels of sequence polymorphism in interacting proteins should also be similar. In conclusion, we note that coevolution (Fraser et al. 2002, 2004) and similar strengths of stabilizing selection (Lemos, Meiklejohn, and Hartl 2004) are not mutually exclusive explanations to account for the similar divergence (in protein sequence and gene expression) and polymorphism (in gene expression) of interacting proteins and are both likely to contribute to this pattern.

## Concluding Remarks

Organismic evolution requires that variation at distinct hierarchical levels and attributes be coherently integrated, often in face of disparate environmental and genetic pressures. A central part of the evolutionary analysis of biological systems is to learn how protein sequence, expression level, protein length, codon bias, genomic position, and other attributes interact with each other and with mutation, selection, and genetic drift in shaping patterns of evolutionary variation in these attributes. In this regard, synthesizing information from multiple sources including cellular,

tissue, and organismic traits with metrics of evolutionary variation (polymorphism and divergence) is a fundamental task. Although our analyses highlight major genome-wide relationships among biological attributes, we emphasize that protein length, mRNA abundance, and numbers of protein-protein interactions are only a few of a large number of organismic attributes likely to influence protein and gene expression evolution. Indeed, we cannot overstress the relevance of incorporating a number of other biologically important variables in similar analyses and must note, moreover, that even the attributes herein considered can be further refined. For instance, our use of adult whole-organism gene expression information is constrained by available mRNA abundance data, and its validity may therefore be limited by the dynamic nature of gene expression variation across tissues as well as throughout development.

Eventually, a more complete understanding of organismic evolution will require the interplay of multivariate statistical approaches aimed at uncovering causal evolutionary and functional relations between genome-wide (or organism-wide) attributes (e.g., Shipley 2000), modeling approaches based on recent developments in systems biology aimed at understanding organismic functioning and evolution (e.g., Covert, Famili, and Palsson 2003; Price, Reed, and Palsson 2004), and finally the combination of analytical and modeling approaches with creative experimental manipulation (e.g., Forster et al. 2003; Covert et al. 2004). As more and better genome-wide (or organism-wide) data accumulate, understanding the causal connections among biological attributes and how they are integrated within an organism and differ across generations and over evolutionary time may finally become possible.

## Acknowledgments

We are grateful to Rob Kulathinal for discussions and encouragement. We are also grateful to the computational biology group at the Bauer Center for Genomics Research for assistance. This work was carried out with grants from the National Institutes of Health to D.L.H.

## Literature Cited

Adams, M. D., S. E. Celniker, R. A. Holt et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. Nature **406**:378–382.

Akashi, H. 2003. Translational selection and yeast proteome evolutions. Genetics **164**:1291–1303.

Bader, J. S., A. Chaudhuri, J. M. Rothberg, and J. Chant. 2004. Gaining confidence in high-throughput protein interaction networks. Nat. Biotech. **22**:78–85.

Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**:101–113.

Bloom, J. D., and C. Adami. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol. Biol. **3**:21.

———. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. BMC Evol. Biol. **4**:14.

Castillo-Davis, C. I., F. A. Kondrashov, D. L. Hartl, and R. J. Kulathinal. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. **14**:802–811.

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

Covert, M. W., I. Famili, and B. O. Palsson. 2003. Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? Bioeng. **84**:763–772.

Covert, M. W., E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. Nature **429**:92–96.

De Gregorio, E., P. T. Spellman, G. M. Rubin, and B. Lemaitre. 2001. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. Proc. Natl. Acad. Sci. USA **98**:12590–12595.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

———. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–74.

Forster, J., I. Famili, P. Fu, B. O. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res. **13**:244–253.

Fraser, H. B., and A. E. Hirsh. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol. Biol. **4**:13.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science **296**:750–752.

Fraser, H. B., A. E. Hirsh, D. P. Wall, and M. B. Eisen. 2004. Coevolution of gene expression among interacting proteins. Proc. Natl. Acad. Sci. USA **101**:9033–9038.

Fraser, H. B., D. P. Wall, and A. E. Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol. Biol. **3**:11.

Giot, L., J. S. Bader, C. Brouwer et al. (49 co-authors). 2003. A protein interaction map of *Drosophila melanogaster*. Science **302**:1727–1736.

Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **18**:609–613.

Hahn, M. W., G. C. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J. Mol. Evol. **58**:203–211.

Holt, R. A., G. M. Subramanian, A. Halpern et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science **298**:129–149.

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

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. 2000. The large-scale organization of metabolic networks. Nature **407**:651–654.

Jordan, I. K., L. Marino-Ramirez, Y. I. Wolf, and E. V. Koonin. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Mol. Biol. Evol. **21**:2058–2070.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2003. No simple dependence between protein evolution rate and the number

of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. **3**:1.

Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. **13**:2229–2235.

Lemos, B., C. D. Meiklejohn, M. Caceres, and D. L. Hartl. 2005. Rates of divergence in gene expression profiles of primates, mice and flies: stabilizing selection and variability among functional categories. Evolution **59**:126–137.

Lemos, B., C. D. Meiklejohn, and D. L. Hartl. 2004. Regulatory evolution across the protein interaction network. Nat. Genet. **36**:1059–1060.

Li, C., and W. H. Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA **98**:31–36.

Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between duplicate genes. Genome Res. **13**:1638–1645.

Marais, G., and L. Duret. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J. Mol. Evol. **52**:275–280.

Meiklejohn, C. D., J. Parsch, J. M. Ranz, and D. L. Hartl. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proc. Natl. Acad. Sci. USA **100**:9894–9899.

Moriyama, E. N., and D. L. Hartl. 1993. Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics **134**:847–858.

Nuzhdin, S. V., M. L. Wayne, K. L. Harmon, and L. M. McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. Mol. Biol. Evol. **21**: 1308–1317.

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

Papp, B., C. Pal, and L. D. Hurst. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature **424**:194–196.

Price, N. D., J. L. Reed, and B. O. Palsson. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat. Rev. Microbiol. **2**:886–897.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science **300**:1742–1745.

Richards, S., Y. Liu, B. R. Bettencourt et al. (52 co-authors). 2005. Comparative sequencing and analysis of *Drosophila pseudoobscura*. Genome Res. **15**:1–18.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat. Genet. **33**:138–144.

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

Sharp, P. M., and W. H. Li. 1989. On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28**:398–402.

Shipley, B. 2000. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge, UK.

Subramanian, S., and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168**:373–381.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science **278**:631–637.

Teichmann, S. A. 2002. The constraints protein-protein interactions place on sequence divergence. J. Mol. Biol. **324**: 399–407.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTALW: 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.

Townsend, J. P. 2003. Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. BMC Genomics **4**:41.

Townsend, J. P., and D. L. Hartl. 2002. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA levels across multiple strains or treatments. Genome Biol. **3**:RESEARCH0071.

Urrutia, A. O., and L. D. Hurst. 2003. The signature of selection mediated by expression on humans genes. Genome Res. **13**:2260–2264.

Veitia, R. A. 2002. Exploring the etiology of haploinsufficiency. Bioessays **24**:175–184.

Wagner, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc. Natl. Acad. Sci. USA **97**:6579–6584.

Williams, E. J. B., and L. D. Hurst. 2000. The proteins of linked genes evolve at similar rates. Nature **407**:900–903.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

———. 2002. Phylogenetic analysis by maximum likelihood (PAML) version 3.0. (http://abacus.gene.ucl.ac.uk/software/paml.html).

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

Zhang, Z., T. M. Hambuch, and J. Parsch. 2004. Molecular evolution of sex-biased genes in *Drosophila*. Mol. Biol. Evol. **21**:2130–2139.

Takashi Gojobori, Associate Editor

Accepted February 23, 2005