Evolutionary Rate and Duplicability in the *Arabidopsis thaliana* Protein–Protein Interaction Network

David Alvarez-Ponce[1,2,]* and Mario A. Fares[1,2,]*

[1]Department of Abiotic Stress, Integrative and Systems Biology Laboratory, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicias (CSIC-UPV), Valencia, Spain

[2]Department of Genetics, University of Dublin, Trinity College, Dublin, Ireland

*Corresponding author: E-mail: mfares@ibmcp.upv.es; david.alvarez@nuim.ie, david.alvarez.ponce@gmail.com.

Accepted: November 9, 2012

Abstract

Genes show a bewildering variation in their patterns of molecular evolution, as a result of the action of different levels and types of selective forces. The factors underlying this variation are, however, still poorly understood. In the last decade, the position of proteins in the protein–protein interaction network has been put forward as a determinant factor of the evolutionary rate and duplicability of their encoding genes. This conclusion, however, has been based on the analysis of the limited number of microbes and animals for which interactome-level data are available (essentially, *Escherichia coli*, yeast, worm, fly, and humans). Here, we study, for the first time, the relationship between the position of proteins in the high-density interactome of a plant (*Arabidopsis thaliana*) and the patterns of molecular evolution of their encoding genes. We found that genes whose encoded products act at the center of the network are more evolutionarily constrained than those acting at the network periphery. This trend remains significant when potential confounding factors (gene expression level and breadth, duplicability, function, and length of the encoded products) are controlled for. Even though the correlation between centrality measures and rates of evolution is generally weak, for some functional categories, it is comparable in strength to (or even stronger than) the correlation between evolutionary rates and expression levels or breadths. In addition, genes encoding interacting proteins in the network evolve at relatively similar rates. Finally, *Arabidopsis* proteins encoded by duplicated genes are more highly connected than those encoded by singleton genes. This observation is in agreement with the patterns observed in humans, but in contrast with those observed in *E. coli*, yeast, worm, and fly (whose duplicated genes tend to act at the periphery of the network), implying that the relationship between duplicability and centrality inverted at least twice during eukaryote evolution. Taken together, these results indicate that the structure of the *A. thaliana* network constrains the evolution of its components at multiple levels.

Key words: network evolution, *Arabidopsis* interactome, natural selection, rates of evolution, gene duplication, network centrality.

Introduction

Genes and proteins rarely operate in isolation; on the contrary, they often act as parts of complex functional networks of interacting molecules. Understanding the function and evolution of molecular networks is not only a key step toward understanding organisms' function and evolution, but it can also aid applications such as metabolic engineering and drug discovery and design (for review, see Butcher et al. 2004; Korcsmáros et al. 2007; Lee et al. 2011). Furthermore, considering the position of genes and proteins in the networks in which they participate may provide key insight into the evolutionary forces governing their evolution. Indeed, several lines of evidence point to a link between the position of proteins in metabolic and protein–protein interaction networks (PINs) and the patterns of molecular evolution of their encoding genes (reviewed by Cork and Purugganan 2004; Eanes 2011; Wagner 2012). In particular, proteins' network position has an effect on their rate of evolution and on the duplicability of their encoding genes. Several questions, however, remain open.

Proteins' rates of evolution vary across orders of magnitude, as a result of the action of different levels and types of evolutionary forces (Zuckerkandl and Pauling 1965; King and Jukes 1969; Ohta and Kimura 1971; Li et al. 1985). Identifying and understanding the factors responsible for this variability is one of the main open questions in Evolutionary Biology. Purifying selection is expected to act more severely on genes performing functions that are more important for the organism's biological fitness, thus resulting in lower rates of evolution (Wilson et al. 1977; Kimura 1983); however, the relative contribution of genes to fitness remains elusive. Over the last decade, the wealth of genomic and functional data has made feasible to pursue the formulation of a unifying theory that explains the variation of the rates of protein evolution (for review, see Herbeck and Wall 2005; Koonin and Wolf 2006; McInerney 2006; Pál et al. 2006; Rocha 2006; Wolf et al. 2006). Several factors have been shown to correlate with the strength of purifying selection acting on genes, with patterns of expression being the most prominent: more highly or widely expressed genes tend to be more constrained than those expressed at low levels or in a narrow range of tissues (Duret and Mouchiroud 2000; Pál et al. 2001; Krylov et al. 2003; Subramanian and Kumar 2004; Wright et al. 2004; Drummond et al. 2005, 2006; Ingvarsson 2007; Slotte et al. 2011; Yang and Gaut 2011). Other relevant determinants of genes' evolutionary rates include the length (Subramanian and Kumar 2004; Ingvarsson 2007) and function (Castillo-Davis et al. 2004; Alvarez-Ponce and McInerney 2011) of the encoded products. However, these factors are poor predictors of evolutionary rates (e.g., Ingvarsson 2007; Larracuente et al. 2008).

Remarkably, genes acting at the center of the PIN tend to be more selectively constrained than those acting at the network periphery, a pattern that has thus far been observed in *Escherichia coli*, yeasts, worms, flies, and mammals (Fraser et al. 2002; Jordan et al. 2003; Hahn and Kern 2005; Lemos et al. 2005; Davids and Zhang 2008; Alvarez-Ponce 2012). Consistently, proteins involved in complexes tend to be highly conserved (Teichmann 2002). Similar patterns have been observed in metabolic networks, whose most connected enzymes tend to evolve under stronger selective pressures (Vitkup et al. 2006). In addition, genes encoding physically interacting proteins tend to evolve at relatively similar rates (Fraser et al. 2002; Lemos et al. 2005; Alvarez-Ponce et al. 2009, 2011; Cui et al. 2009). Although these trends are significant and consistent across all organisms studied to date, they are often weak, and whether they reflect a direct effect of network position on genes' rates of evolution has been the subject of debate. In particular, some authors have suggested that these trends might be a byproduct of the distribution of confounding factors across the network, such as genes acting at the center of the network being expressed at higher levels or to biases in interactomic data sets (Bloom and Adami 2003; Batada et al. 2006) (but see Fraser et al. 2003; Fraser and Hirsh 2004; Fraser 2005; Lemos et al. 2005).

Genes also widely differ in their duplicability (i.e., the propensity to retain duplicated copies after a gene duplication event), with some genes remaining as single copies (singletons) over long evolutionary periods and others being recurrently duplicated. Genes' duplicabilities are known to be affected by a number of factors, including the position of the encoded proteins in the PIN. However, the direction of the relationship between gene duplicability and network centrality is not universal. In the PINs of *E. coli*, yeast, worm, and fly, singleton genes tend to occupy more central positions than duplicated genes (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009; D'Antonio and Ciccarelli 2011). This trend has been attributed to a fragility of the network to duplications affecting its more connected elements. Indeed, complexes and pathways are thought to perform better with balanced concentrations of their members, and the duplication of a given gene is expected to disrupt the dosage balance of the interactions in which its encoded product is involved (Birchler et al. 2001; Veitia 2002; Papp et al. 2003), which may have more deleterious effects for proteins with a higher number of interactors. Conversely, human proteins encoded by duplicated genes tend to be more central to the PIN than those encoded by singleton genes (Liang and Li 2007; Doherty et al. 2012). This is a derived character resulting from the high duplicability of human hubs originated after the emergence of Metazoans, implying that the relationship between centrality and duplicability underwent modification in the vertebrate lineage (D'Antonio and Ciccarelli 2011). The factors underlying this shift in the relationship between centrality and duplicability remain, however, unclear.

The relationship between network position and genes' patterns of molecular evolution has been hitherto investigated only in the few microorganisms (*E. coli* and yeast) and animals (worm, fly, and human) for which interactome-scale protein–protein interaction data are available. The recent availability of a relatively high-density interactome for *Arabidopsis thaliana* (Arabidopsis Interactome Mapping Consortium 2011; Stark et al. 2011), together with the availability of the genome sequences for this species (Arabidopsis Genome Initiative 2000) and its close relative *A. lyrata* (Hu et al. 2011), allowed us to investigate, for the first time, the relationship between genes' patterns of molecular evolution and their position in a plant network. Consistent with patterns observed in other organisms, we found that genes encoding the most central proteins of the network tend to evolve under stronger levels of selective constraint and that genes encoding physically interacting proteins evolve at relatively similar rates, pointing to general trends across all of life. The relationship between centrality and evolutionary rate is independent of potential confounding factors (gene duplicability, expression level and breadth, and the length of the encoded products), suggesting a direct effect of the network structure on the patterns of molecular evolution of its components. Even though the correlation between the measures of centrality and rates of evolution is in general weak, for some functional categories, it is comparable in strength to (or even stronger than) the correlation between evolutionary rates and expression levels or breadths.

Surprisingly, genes acting at the center of the *A. thaliana* PIN are more likely to be duplicated than those acting at the periphery, implying that the relationship between centrality and duplicability underwent modification not only in vertebrates but also in plants.

Materials and Methods

Protein–Protein Interaction Data

The PIN was obtained by merging the *A. thaliana* networks available from the BioGRID database v3.1.81 (Stark et al. 2011) and from Arabidopsis Interactome Mapping Consortium (2011). Only physical interactions among pairs of *A. thaliana* proteins were considered. All interactions used in the current analysis had been determined experimentally in *A. thaliana* (i.e., the data set does not contain interactions inferred computationally or derived from other species).

Impact of Natural Selection

For each gene in the network, we attempted to identify its 1:1 ortholog in the *A. lyrata* genome (Hu et al. 2011) using a best reciprocal Basic Local Alignment Search Tool (BLAST) approach (using BLASTP and an *E*-value cut-off of 10^{-10}). Each pair of *A. thaliana*–*A. lyrata* orthologous protein sequences was aligned using ProbCons 1.12 (Do et al. 2005), and the resulting alignments were used to guide the alignment of the corresponding coding sequences. Estimates of d_N, $d_{S,}$ and ω were obtained using the one-ratio model M0 from the PAML 4.4 package (Yang 2007).

Paralogs Identification

Each *A. thaliana* protein was used as query in a BLASTP (Altschul et al. 1997) search against the *A. thaliana* proteome. Genes were classified as singleton if no hit was obtained with an E value < 0.1 or as duplicated if at least a homolog was found that met the following criteria: 1) E value $\leq 10^{-10}$; *2*) the aligned region length (L) was $\geq$80% of the length of the query sequence; and 3) amino acid identity was $\geq$30% if $L > 150$ amino acids or $0.06 + 4.8L^{-0.32[1\,+\,\exp(-L/1{,}000)]}$ otherwise (Rost 1999). Duplicated genes were further classified as whole-genome duplication (WGD) genes if classified as such in Blanc et al. (2003); otherwise, they were classified as resulting from small-scale genome duplication (SSD).

Gene Expression Level and Breadth

Expression data from 79 *A. thaliana* tissues were obtained from Schmid et al. (2005). For each gene and tissue, values were averaged across the three replicates, and the gene was considered to be expressed if it was annotated as "present" in at least two of the replicates. For each gene, expression level was computed as the median across the 79 tissues, and expression breadth was computed as the number of tissues in which the gene is expressed. For genes matching multiple probe sets, the set yielding a higher expression level was used. Probe sets matching multiple genes were discarded.

Functional Information

Each *A. thaliana* gene was assigned to one (or sometimes a few) eukaryotic orthologous group (KOG) categories using the eggNOG v2 database (Muller et al. 2010).

Ribosomal Proteins

Proteins were considered to be ribosomal if identified as such in Barakat et al. (2001), if their description contained the text "ribosomal protein" or if they were assigned to the Gene Ontology (Ashburner et al. 2000) terms "large ribosomal subunit" or "small ribosomal subunit."

Age of Genes

Each of the 5,789 *A. thaliana* network proteins was used as query in a BLASTP (Altschul et al. 1997) search against the nr database (obtained from the National Center for Biotechnology Information on January 2012). An *E*-value cut-off of 10^{-6} was used. Only hits whose aligned region length was $\geq$80% the length of the query sequence were considered. Genes presenting hits in prokaryotes or in non-plant eukaryotes (non-Embryophyta) were deemed as "ancient." The remaining genes were considered to be "land plant-specific."

Results

Genes Acting at the Center of the *A. thaliana* PIN Are More Selectively Constrained Than Those Acting at the Network Periphery

We assembled a PIN for *A. thaliana* by merging all interactions available in the BioGRID database (Stark et al. 2011) and in Arabidopsis Interactome Mapping Consortium (2011). The resulting network consisted of 5,789 unique proteins connected by 14,368 physical binary interactions. For each protein in the network, we computed three centrality measures: 1) the number of interactors (degree); 2) the number of shortest paths between all pairs of proteins of which the protein is part (betweenness; Freeman 1977); and 3) the inverse of the average shortest distance to all the other proteins in the network (closeness). Using a best reciprocal BLAST approach, we found 1:1 *A. lyrata* orthologs for 3,868 of the 5,789 *A. thaliana* network genes. For each pair of orthologs, the impact of natural selection was inferred from the nonsynonymous to synonymous divergence ratio ($\omega = d_N/d_S$). Values of $\omega = 1$ are expected for genes evolving neutrally, whereas $\omega < 1$ is indicative of the action of purifying selection preserving the sequence of the encoded proteins, and $\omega > 1$ in a number of codons is indicative of the action of positive selection (adaptive evolution) driving the fixation of nonsynonymous substitutions. The estimated ω values exhibit a

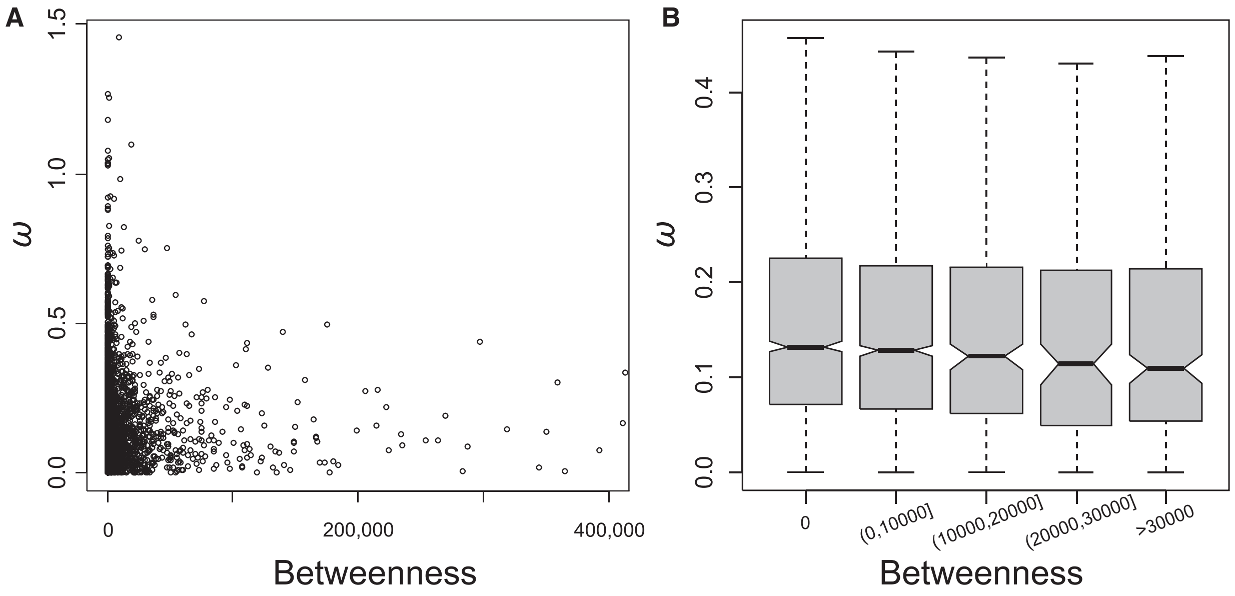

FIG. 1.—Correlation between ω and betweenness.

Table 1

Spearman's Correlations among the Parameters Considered in the Study

	ω	d_N	d_S	Degree	Closeness	Betweenness	Expression Level	Expression Breadth
d_N	0.732***							
d_S	−0.219***	0.310***						
Degree	−0.030	−0.026	−0.019					
Closeness	−0.034*	−0.032*	0.001	0.412***				
Betweenness	−0.053**	−0.051**	−0.017	0.875***	0.436***			
Expression level	−0.331***	−0.342***	−0.049**	−0.013	0.066**	0.046**		
Expression breadth	−0.245***	−0.264***	−0.072**	−0.006	0.024	0.035*	0.692***	
Protein length	0.050**	−0.009	−0.124***	−0.017	−0.042**	−0.023	−0.088***	0.012

*$P < 0.05$.
**$P < 0.01$.
***$P < 10^{-6}$.

median of 0.128, indicating that these genes evolved under relatively high levels of selective constraint.

We found that the 3,868 genes encoding proteins that are represented in the network (i.e., those with described interactors) exhibit lower rates of evolution than those not represented in the network (median ω values of 0.128 and 0.183, respectively; Mann–Whitney test, $P < 10^{-15}$). In addition, among these 3,868 network genes, ω values negatively correlate with betweenness (Spearman's rank correlation coefficient, ρ = −0.053, $P = 0.001$; fig. 1) and closeness (ρ = −0.034, $P = 0.035$): the larger the values of betweenness and closeness for a protein, the stronger are the selective constraints acting on that protein. Although the correlation is only marginally significant for degree (ρ = −0.030, $P = 0.063$), the 1,467 genes with degree >1 show significantly lower ω values than the 2,401 genes with degree = 1 (median ω values of 0.125 and 0.135, respectively; Mann–Whitney test, $P = 0.005$). Taken together, these results indicate that the most central genes in the *A. thaliana* network are subject to stronger levels of purifying selection than those acting at the network periphery. Similar results were obtained for d_N but not for d_S (table 1), indicating that the distribution across the network of selection at the amino acid level is the main responsible for the observed trend.

The Relationship between Centrality and Evolutionary Rate Is Independent of Gene Expression Level and Breadth, Gene Function, and the Length of the Encoded Products

Having established an association between proteins' centralities and their rates of evolution, we sought to determine whether this association reflected a direct effect of network position or, on the contrary, it was the result of the distribution across the network of a number of factors that correlate both with rates of evolution and network centralities.

Levels of selective constraint acting on a gene are known to depend on a number of factors, among which the level and breadth of gene expression seem to be the most important (Duret and Mouchiroud 2000; Pál et al. 2001; Krylov et al. 2003; Subramanian and Kumar 2004; Wright et al. 2004; Drummond et al. 2005, 2006; Ingvarsson 2007; Slotte et al. 2011; Yang and Gaut 2011). In agreement with previous reports, we observed that the ω and d_N values exhibit a strong negative correlation with both expression level and breadth in our data set (table 1). Additionally, we observed that, similar to the PINs of other organisms (Bloom and Adami 2003; Lemos et al. 2005), the most central genes in the *Arabidopsis* PIN are more highly and broadly expressed than those acting at the periphery: expression levels positively correlate with betweenness and closeness and expression breadths with betweenness (table 1). Combined, these observations raise the possibility that the correlation between centrality measures and evolutionary rates could be driven by the higher levels and/or breadths of expression of genes acting at the center of the PIN. However, partial correlation analysis (supplementary table S1, Supplementary Material online) shows that the association between ω and betweenness is independent of expression level and breadth ($\rho = -0.038$, $P = 0.026$).

Another potential confounding factor is protein length, as it correlates positively with ω (i.e., genes encoding shorter proteins tend to be more selectively constrained; Subramanian and Kumar 2004; Ingvarsson 2007; table 1) and negatively with closeness (i.e., genes acting at the center of the network tend to encode shorter proteins, in agreement with previous observations in *Drosophila*; Lemos et al. 2005; table 1). However, the correlation between ω and both betweenness ($\rho = -0.052$, $P = 0.001$) and closeness ($\rho = -0.032$, $P = 0.048$) remains significant when protein lengths are controlled for (supplementary table S1, Supplementary Material online), indicating that the tendency for central genes to evolve under strong levels of selective constraint is also independent of protein length. Furthermore, the correlation between ω and betweenness remains significant when protein length and expression level and breadth are simultaneously controlled for ($\rho = -0.037$, $P = 0.028$; supplementary table S1, Supplementary Material online).

Genes performing different functions are subject to different levels of selective constraint (e.g., Castillo-Davis et al. 2004; Alvarez-Ponce and McInerney 2011) and encode proteins with different network centralities (Kunin et al. 2004; Cotton and McInerney 2010; Alvarez-Ponce and McInerney 2011). Consistently, we found that genes in different categories exhibit different ω and centrality values (Kruskal–Wallis test, ω: $P < 10^{-15}$; degree: $P < 10^{-15}$; betweenness: $P = 1.17 \times 10^{-11}$; closeness: $P = 7.32 \times 10^{-10}$). Therefore, the observed correlation between centrality and evolutionary rate (table 1) could conceivably be the result of these differences. To discard this possibility, the correlation between evolutionary rate and degree was evaluated separately for genes involved in each of the KOG functional categories (Tatusov et al. 2003) (similar results were obtained for betweenness and closeness; data not shown). The analysis was restricted to the 20 KOG categories comprising more than 25 *A. thaliana* network genes. Remarkably, the correlation coefficient was negative for 18 of the 20 categories (table 2). The correlation was significant for only five of these categories, probably as a result of the reduced statistical power resulting from partitioning the data set. These observations indicate that, in spite of the fact that genes with different functions exhibit different ω and centrality values, the negative correlation between centrality and ω is largely independent of gene function.

Remarkably, correlation coefficients are highly variable across the different KOG categories (table 2). First, pairwise comparison shows that 10 pairs of categories exhibit significantly different correlation coefficients (supplementary table S2, Supplementary Material online). Second, for each functional category, we compared the correlation coefficient for genes within that category with the correlation coefficient for all other genes (i.e., all network genes not belonging to that category); this analysis shows that categories "signal transduction mechanisms" ($P = 0.009$) and "chromatin structure and dynamics" ($P = 0.025$) exhibit significantly higher correlation coefficients than the rest of the network (table 2). Taken together, these results indicate that the extent to which sequence evolution is affected by protein centrality is dependent on gene function. In fact, some categories exhibit strong negative ω-degree correlations that are comparable in strength to (or even stronger than) the ω-expression level and the ω-expression breadth correlations (tables 1 and 2).

Finally, because genes encoding ribosomal proteins are highly conserved, and presumably highly connected, the association between centrality and evolutionary rate observed here could potentially be due to these genes. However, when these genes are eliminated, the correlation between ω and betweenness remains significant ($\rho = -0.062$, $P = 1.34 \times 10^{-4}$) and the correlation between ω and degree—which was not significant in the complete data set (table 1)—reaches significance ($\rho = -0.041$, $P = 0.011$). Furthermore, these correlations remain significant when expression level, expression breadth, and protein length are simultaneously controlled for (degree: $\rho = -0.038$, $P = 0.029$; betweenness: $\rho = -0.042$, $P = 0.014$).

Taken together, these observations point to a direct effect of the position of proteins in the *A. thaliana* PIN on the rates of evolution of their encoding genes.

Table 2
Correlation between Degree and ω for Each Functional Category

KOG Category	Description	n	Average ω	Average Degree	Average Expression Level	Average Expression Breadth	Correlation						
							ω–Degree			ω–Expression Level		ω–Expression Breadth	
							ρ	P	Comparison[a]	ρ	P	ρ	P
J	Translation, ribosomal structure, and biogenesis	138	0.122	2.36	1,240.4	76.3	0.086	0.316	0.149	−0.379	2.10×10^{-5}**	−0.254	0.005**
A	RNA processing and modification	148	0.169	5.89	327.4	74.3	−0.162	0.049*	0.105	−0.360	2.06×10^{-5}**	−0.211	0.015*
K	Transcription	313	0.204	6.43	178.8	61.0	−0.118	0.037*	0.143	−0.233	7.71×10^{-5}**	−0.167	0.005**
L	Replication, recombination, and repair	73	0.162	3.85	147.9	62.7	−0.203	0.086	0.138	−0.223	0.074	−0.171	0.173
B	Chromatin structure and dynamics	45	0.167	3.91	357.6	73.2	−0.357	0.016*	0.025*	−0.436	0.003**	−0.198	0.202
D	Cell cycle control, cell division, and chromosome partitioning	107	0.189	10.24	244.6	63.5	−0.092	0.344	0.526	−0.444	5.09×10^{-6}**	−0.487	4.26×10^{-7}***
T	Signal transduction mechanisms	305	0.128	6.47	223.8	65.8	−0.168	0.003**	0.009**	−0.291	5.11×10^{-7}***	−0.314	5.19×10^{-8}***
Z	Cytoskeleton	68	0.138	4.99	472.7	66.3	−0.205	0.094	0.149	−0.312	0.014*	−0.252	0.050*
U	Intracellular trafficking, secretion, and vesicular transport	180	0.117	4.91	369.3	73.8	−0.040	0.592	0.876	−0.359	2.47×10^{-6}**	−0.280	2.99×10^{-4}**
O	Posttranslational modification, protein turnover, and chaperones	342	0.137	6.35	580.3	70.0	−0.069	0.203	0.450	−0.309	4.90×10^{-8}***	−0.181	0.002**
C	Energy production and conversion	121	0.106	3.24	936.5	73.9	−0.197	0.031*	0.067	−0.342	2.72×10^{-4}**	−0.418	5.99×10^{-6}**
G	Carbohydrate transport and metabolism	129	0.099	4.45	803.4	71.0	−0.165	0.062	0.129	−0.461	1.67×10^{-7}***	−0.296	0.001**
E	Amino acid transport and metabolism	91	0.128	2.45	720.2	73.3	−0.069	0.517	0.733	−0.389	5.14×10^{-4}**	−0.332	0.003**
F	Nucleotide transport and metabolism	32	0.103	2.97	697.2	76.0	−0.063	0.733	0.861	−0.130	0.518	0.025	0.900
H	Coenzyme transport and metabolism	26	0.100	3.38	480.5	72.2	−0.030	0.883	0.998	−0.237	0.253	−0.250	0.228
I	Lipid transport and metabolism	63	0.141	2.49	431.0	71.9	−0.165	0.197	0.292	−0.443	4.41×10^{-4}**	−0.249	0.055
P	Inorganic ion transport and metabolism	68	0.155	3.46	465.4	67.3	−0.134	0.275	0.392	−0.044	0.736	−0.047	0.722
Q	Secondary metabolites biosynthesis, transport, and catabolism	47	0.128	2.09	308.7	57.6	−0.190	0.200	0.284	−0.453	0.003**	−0.279	0.077
R	General function prediction only	425	0.160	4.12	297.5	69.4	0.013	0.783	0.364	−0.235	3.56×10^{-6}**	−0.175	6.18×10^{-4}**
S	Function unknown	156	0.158	3.35	306.1	72.1	−0.053	0.512	0.776	−0.185	0.030*	−0.031	0.718

NOTE.—Only functional categories comprising more than $n = 25$ *A. thaliana* network genes are presented.

[a] Comparison of the correlation for genes within the category versus the correlation for all other network genes. Comparisons were conducted using Fisher's z-transformation, followed by standard normal comparison. The suitability of this test for Spearman's correlations is discussed in Myers and Sirois (2006).

*$P < 0.05$.

**$P < 0.01$.

***$P < 10^{-6}$.

Genes Encoding Physically Interacting Proteins Are Subject to Similar Levels of Selective Constraint

We considered whether interacting proteins in the *A. thaliana* network are subject to relatively similar levels of selective constraint compared with random protein pairs. For that purpose, we computed the normalized absolute difference between the evolutionary rates of all pairs of interacting genes in the network (X):

$$X = \frac{1}{m}\sum_{i=1}^{m}\frac{|\omega_{i1}-\omega_{i2}|}{(\omega_{i1}+\omega_{i2})/2}$$

Here, m is the number of interactions in the network, and ω_{i1} and ω_{i2} are the ω values of the two genes involved in interaction i. Self-interactions (i.e., interactions among proteins encoded by the same gene, a total of 383 in the data set) were not considered in this analysis, as PINs are enriched in such interactions, which can inflate the average similarity between interacting genes (Ispolatov et al. 2005; Alvarez-Ponce and McInerney 2011). The observed value ($X = 0.892$) was compared with a null distribution obtained from a collection of random networks with the same proteins, number of interactions, and degree for each node, which we generated from the original network by repeatedly switching pairs of edges (as in Luisi et al. 2012). Out of 10,000 random networks, only 426 showed an X value lower or equal to the observed one, indicating that interacting proteins exhibit rates of evolution that are more similar than expected from a random network ($P = 0.0426$; fig. 3). Random networks exhibit an average X value of 0.900, indicating that the ω values for interacting genes are 0.8% more similar than random pairs of proteins.

Ribosomal proteins might also represent a source of confounding bias, as they are subject to strong levels of purifying selection and are highly connected to each other. However, significant results were obtained when ribosomal proteins were also removed from the network ($X = 0.889$, $P = 0.039$).

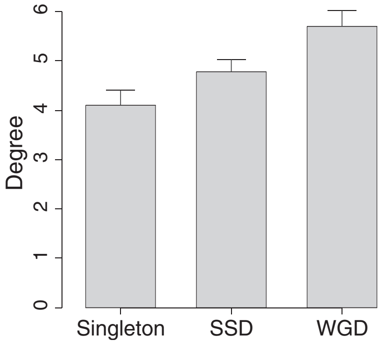

Fig. 2.—Average number of interactors for proteins encoded by singleton, whole-genome duplication (WGD), and small-scale duplication (SSD) genes. Error bars represent the standard error of the means.

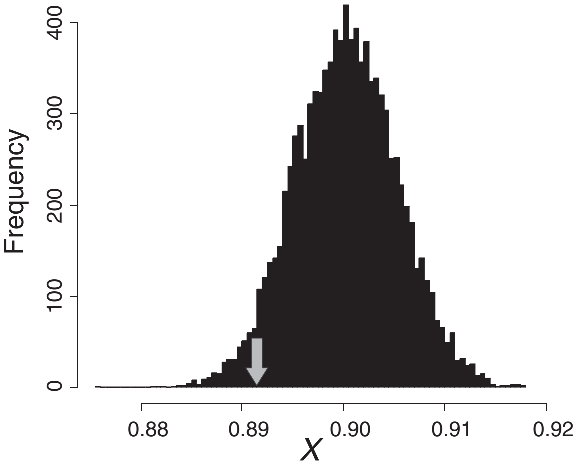

Fig. 3.—Normalized absolute difference between the ω values of pairs of interacting genes in the network (X). The arrow points to the observed value and the histogram represent the null distribution obtained from 10,000 random networks.

Centrality and Duplicability in the *A. thaliana* PIN

Finally, we studied the effect of proteins' network position on the duplicability of their encoding genes. Among the 5,789 genes encoding the *A. thaliana* PIN, 883 were found to be singleton, and 3,532 were deemed as duplicated based on similarity searches (the rest of the genes remained unclassified; see Materials and Methods). Among duplicated genes, 1,540 are the result of one of the WGD events that took place in the *Arabidopsis* lineage (Blanc et al. 2003; De Bodt et al. 2005), and 1,992 were deemed as the result of SSD events. Unexpectedly, proteins encoded by duplicated genes have more interactors in the network (average degree of 5.18) than those encoded by singleton genes (average degree of 4.10) (Mann–Whitney test, $P = 0.008$; fig. 2). These differences remain significant when self-interactions and interactions among proteins encoded by paralogs are removed from the analysis ($P = 0.049$), or when genes are classified as duplicated or singleton based on whether they present annotated paralogs in the Ensembl plants database release 14 (Kersey et al. 2012) ($P = 1.62 \times 10^{-4}$). Furthermore, degree positively correlates with the number of paralogs annotated in Ensembl plants ($\rho = 0.089$, $P = 1.02 \times 10^{-11}$), even when only duplicated genes are considered ($\rho = 0.086$, $P = 5.85 \times 10^{-9}$). This observation is in agreement with the patterns observed in the human interactome (Liang and Li 2007; D'Antonio and Ciccarelli 2011; Doherty et al. 2012) but in contrast with those observed in *E. coli*, yeast, worm, and fly, whose duplicated genes tend to encode lowly connected proteins compared with singleton genes (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009; D'Antonio and Ciccarelli 2011). Among duplicated genes, those resulting from WGDs are more highly connected in the *Arabidopsis* network than those resulting from SSD events (average degree of 5.70 and 4.78, respectively; Mann–Whitney test, $P = 5.63 \times 10^{-6}$; fig. 2).

D'Antonio and Ciccarelli (2011) found that the relationship between centrality and duplicability was different among ancient and metazoan-specific human genes: among genes of premetazoan origin, duplicated genes are less central than singleton genes, whereas among genes that are specific to metazoans, duplicated genes tend to occupy more central positions than singleton genes. We considered whether the relationship between centrality and duplicability observed in *Arabidopsis* also depended on the age of the genes. For that purpose, genes were classified as "ancient" (if they had homologs in prokaryotes or in nonplant eukaryotes) or as "land plant-specific" (otherwise). A total of 3,114 network genes were classified as ancient, and the remaining ones were deemed as plant specific. We found that, among plant-specific proteins, those encoded by duplicated genes are more highly connected than those encoded by singleton genes (average degrees of 5.39 and 3.85, respectively; Mann–Whitney test, $P = 4.73 \times 10^{-6}$). This difference, however, is not significant among ancient genes (average degrees for proteins encoded by duplicated and singleton genes: 5.06 and 4.41 interactions, respectively; $P = 0.484$). Therefore, the relationship between centrality and duplicability observed in *Arabidopsis* seems to be specific to plant-specific genes.

Because duplicated *Arabidopsis* genes are more selectively constrained than singleton genes (Yang and Gaut 2011), the lower evolutionary rates of genes central to the *A. thaliana* PIN could potentially be a by-product of their enrichment in duplicated genes. However, the correlation between ω and both betweenness ($\rho = -0.050$, $P = 0.015$) and closeness ($\rho = -0.049$, $P = 0.016$) remains significant when only duplicated genes are considered. Although these correlations are not significant for singleton genes ($\rho = -0.044$, $P = 0.288$ for betweenness; $\rho = 0.043$, $P = 0.293$ for closeness), this might result from the small number of genes in this category ($n = 883$).

Analysis of a High-Quality Subnetwork Provides Consistent Results

Currently available interactomes are the result, to a high extent, of the application of high-throughput techniques of protein–protein interaction discovery. As a result, interactomic data sets are subject to high rates of false positives and negatives (Bader et al. 2004; Deeds et al. 2006). Given the potential that this could be affecting our observations, we repeated our analyses in a high-quality subset of the *A. thaliana* interactome, containing only reliable interactions. Similar to D'Antonio and Ciccarelli (2011), we filtered our data set, retaining only interactions that had been determined using low-throughput techniques (i.e., more accurate analysis of protein interactions on a one-by-one basis), and those that had been identified by two or more high-throughput analyses independently. The filtered network consisted of 4,808 interactions connecting 2,798 proteins, out of which 1,842 are encoded by genes with 1:1 orthologs in *A. lyrata*. This represents a dramatic decrease in the amount of data when compared with the full data set (only 48.3% of the proteins and 33.5% of the interactions were present in the high-quality subnetwork), which potentially involves a decrease in statistical power but also an increase in the signal-to-noise ratio.

Although the correlation between degree and ω is not significant in this reduced data set ($\rho = 0.003$, $P = 0.907$), ω values significantly correlate with both betweenness and closeness, with correlation coefficients that are higher in magnitude than those observed in the entire data set (betweenness: $\rho = -0.061$, $P = 0.008$; closeness: $\rho = -0.108$, $P = 3.22 \times 10^{-6}$). The correlation between ω and closeness remains significant when the effects of expression level and breadth, and protein length, are simultaneously controlled for ($\rho = -0.061$, $P = 0.013$). Furthermore, genes encoding proteins that are represented in the network are more selectively constrained than those that are not represented in the network (median ω values of 0.120 and 0.176, respectively; Mann–Whitney test, $P < 10^{-15}$).

Consistent with the observations in the full interactome, duplicated genes are more highly connected than singleton genes (average degrees of 3.58 and 2.83, respectively) in the high-quality subnetwork. Although the degrees of both groups are not significantly different (Mann–Whitney test, $P = 0.132$), the duplicated/singleton degree ratio is equivalent to that observed in the full data set (1.27), suggesting that the lack of significance in the analysis of the high-quality subnetwork may result from the reduced statistical power resulting from trimming the data set.

Discussion

Results presented here provide multiple evidences linking the position of proteins in the *A. thaliana* PIN and evolutionary forces acting on their encoding genes. Remarkably, genes acting at the most central positions of the network (measured as degree, betweenness, or closeness) are subject to stronger levels of purifying selection than those acting at the network periphery. The trend is independent of the distribution across the network of potential confounding factors (patterns and levels of gene expression, gene duplicability, function, and protein length), suggesting that network position has a direct effect on levels of selective constraint. The observation that proteins with more interacting partners are subject to stronger levels of selective constraint suggests that direct protein–protein interactions impose constraints on protein sequence evolution. Nonetheless, closeness, and in particular betweenness, seem to be better predictors of evolutionary rates than degree (fig. 1 and table 1). These are global measures of network centrality that take into account not only the immediate network context of a protein (as degree does) but rather its position in relation to the entire network. Therefore, the evolutionary rate of a given protein does not only depend on the number of direct interactors but also on its broader network context. In particular, betweenness is a measure of how the flow of information through the network depends on each individual protein (Jeong et al. 2000; Wagner and Fell 2001). Proteins bridging the gap between parts of the network tend to exhibit a high betweenness (Ravasz et al. 2002). Therefore, our observations suggest that proteins exerting a high degree of control on information flow across the network are particularly relevant for the organism's fitness. These results are in agreement with previous observations in organisms for which dense PINs are readily available (*E. coli*, yeast, worm, fly, and human; Fraser et al. 2002; Teichmann 2002; Jordan et al. 2003; Hahn and Kern 2005; Lemos et al. 2005; Davids and Zhang 2008; Alvarez-Ponce 2012), which show that genes acting at the center of the network, or those participating in protein complexes, are more selectively constrained, thereby suggesting a general trend across all of life. Furthermore, our results are in agreement with previous observations in yeast, worm, and fly that betweenness is a better predictor of rates of evolution than degree (Hahn and Kern 2005)—a pattern that has been also observed in the yeast metabolic network (Lu et al. 2007)—suggesting a general trend at least in Eukaryotes.

Further evidence for the relationship between position of proteins in the *A. thaliana* PIN and rates of evolution is provided by the observation that genes encoding interacting proteins tend to be subject to similar levels of selective constraint (fig. 3), consistent with previous observations in other interactomes (Pazos and Valencia 2001; Fraser et al. 2002; Lemos et al. 2005; Alvarez-Ponce et al. 2009, 2011; Cui et al. 2009). This similarity might in part result from the mutational compensatory dynamics between amino acids involved in protein–protein interactions (Codoñer and Fares 2008; Fares et al. 2011). It should be noted, however, that covariation in the evolutionary rates of proteins can obey to other factors such as these proteins performing shared biological functions or exhibiting correlated patterns of expression (Clark et al. 2012). These alternative explanations—covariation due to mutational compensation, shared function, or coexpression—are not mutually exclusive.

Although our analyses reveal a clear association between network position and evolutionary rates, the trend is generally weak in comparison with other correlates of rates of evolution such as gene expression (Duret and Mouchiroud 2000; Pál et al. 2001; Krylov et al. 2003; Subramanian and Kumar 2004; Wright et al. 2004; Drummond et al. 2005, 2006; Ingvarsson 2007; Slotte et al. 2011; Yang and Gaut 2011)—compare, for instance, the ω-degree correlation ($\rho = -0.030$) with the ω-expression level ($\rho = -0.331$) and the ω-expression breadth ($\rho = -0.245$) correlations (table 1). This is in good agreement with previous analyses linking network position and rates of evolution, which often recover weak effects (for review, see Cork and Purugganan 2004; Rocha 2006). The weakness of these effects may be the result of the relatively low fraction of amino acids participating in protein–protein interactions. It is remarkable, however, that the strength of the correlation between degree and rates of evolution is highly variable across the different functional categories. Indeed, for some categories, the correlation coefficients of the ω-degree correlation attain values that are comparable with (or even surpass) those for the ω-expression level and/or the ω-expression breadth correlations (table 2). This suggests that for certain functional categories, protein–protein interactions strongly constrain protein evolution. The ω-degree correlation is significantly higher for proteins in functional categories "signal transduction mechanisms" ($P = 0.009$) and "chromatin structure and dynamics" ($P = 0.025$) than for the rest of the network (supplementary table S2, Supplementary Material online). Consistently, it has been recently observed that degree is a strong correlate of rates of evolution among genes involved in the human signal transduction network (Alvarez-Ponce 2012).

Despite the weak effect of network position on rates of evolution, a much stronger effect is observed on gene duplicability. On average, proteins encoded by duplicated genes exhibit 26%–27% more interactors than those encoded by singleton genes (fig. 2). This higher connectivity of duplicated genes has been also observed in humans (Liang and Li 2007; D'Antonio and Ciccarelli 2011; Doherty et al. 2012), but the opposite pattern was observed in *E. coli*, yeast, worm, and fly—in which genes acting at the periphery of the network are the ones that tend to undergo duplication (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009; D'Antonio and Ciccarelli 2011).

The differential pattern observed in the human interactome has been shown to be the result of the high duplicability of hubs originated after the emergence of Metazoans (D'Antonio and Ciccarelli 2011). Human ancient genes (those of premetazoan origin), however, exhibit the same tendency as observed in *E. coli*, yeast, worm, and fly: duplications tend to occur in genes acting at the periphery of the network (D'Antonio and Ciccarelli 2011). This indicates that the particular trend observed in the human interactome is a derived character and hence that the relationship between centrality and duplicability shifted in the vertebrate lineage. Our observations that, also among *A. thaliana* novel (i.e., plant specific) genes, duplicated genes are more connected than singleton genes indicate that the relationship between duplicability and centrality shifted not only in vertebrates but also in the *Arabidopsis* lineage.

The higher duplicability of proteins acting at the periphery of the *E. coli*, yeast, worm, and fly PINs has been attributed to the fragility of the network to duplication of its more connected elements. Indeed, the duplication of a given gene is expected to disrupt the dosage balance of the interactions in which it is involved (Birchler et al. 2001; Veitia 2002; Papp et al. 2003), which may have more deleterious effects for genes with a higher number of interactors. On the other hand, highly connected proteins may play a key role in maintaining the robustness of the network, thereby making networks fragile to mutation or loss of these genes (Albert et al. 2000; Jeong et al. 2001). Hence, duplication of highly connected proteins might be favored, as they increase the robustness of the system (Ekman et al. 2006). These and other competing forces probably act with different strengths in different organisms, resulting in the contrasting overall trends observed in available interactomes.

Major evolutionary transitions (from prokaryotes to unicellular eukaryotes, to multicellular eukaryotes, and to land plants and vertebrates) were accompanied by reductions of orders of magnitude in effective population sizes. As a result, vertebrates and land plants exhibit effective population sizes that are smaller than those for *E. coli*, yeast, worm, and fly (Lynch 2007). In populations with a small effective size, natural selection is less efficient, and hence, the fate of mutations is largely determined by random genetic drift. In such populations, natural selection may have small power in removing duplicates of genes involved in a large number of interactions, in spite of the fact that duplication of such genes are expected to be deleterious according to the dosage balance hypothesis (Birchler et al. 2001; Veitia 2002; Papp et al. 2003). This can result in a (partial) suppression of the negative relationship between duplicability and centrality predicted by the dosage balance hypothesis. It is possible that, under such conditions, factors promoting a positive duplicability–centrality relationship can manifest, thereby resulting in the patterns observed in the human (Liang and Li 2007; D'Antonio and Ciccarelli 2011; Doherty et al. 2012) and *Arabidopsis* (current work) interactomes. The future availability of the interactomes of a broader range of organisms, with different effective population sizes, may enable a better understanding of the effect of this factor on the relationship between centrality and duplicability.

Among *A. thaliana* duplicated genes, those resulting from the WGD events that took place in this lineage (Blanc et al. 2003; De Bodt et al. 2005) are more highly connected in the PIN than those resulting from SSD events (fig. 2). This result is in concert with the dosage-balance hypothesis, which predicts that duplication of a gene with interactors may be deleterious unless its interacting partners simultaneously coduplicate (Birchler et al. 2001; Veitia 2002; Papp et al. 2003). Under this scenario, highly connected genes are more likely to retain duplicated copies if they are the result of WGDs, as simultaneous duplication of the entire genome maintains the stoichiometry of all balanced sets (Veitia 2004, 2005).

The current analysis represents, to our knowledge, the first interactome-level evaluation of the relationship between the structure of the *A. thaliana* PIN and the patterns of molecular evolution of its components. Taken together, results presented here indicate that the network imposes constraints on the patterns of molecular evolution of its components at multiple levels, from sequence evolution to gene duplication. Therefore, genes do not evolve independently but as pieces of a more complex system. Currently, the availability of protein–protein interaction data for *A. thaliana* is limited in comparison with other model organisms (Stark et al. 2011). As more data become available, the emergence of a more detailed map of the *A. thaliana* interactome will enable a more detailed understanding of its evolution.

Conclusions

Results presented here provide multiple evidences linking the position of proteins in the *A. thaliana* PIN and evolutionary forces acting on their encoding genes. In agreement with previous observations in other organisms, genes acting at the most central positions of the network are subject to increased levels of selective constraint, and genes encoding interacting proteins exhibit correlated rates of evolution. Duplicated genes tend to be more central than singleton genes, in agreement with the patterns observed in humans, but opposite to those observed in *E. coli*, yeast, worm, and fly, thereby indicating that the relationship between centrality and duplicability underwent modification not only in vertebrates but also in plants. Taken together, these results indicate that the *A. thaliana* PIN impose constraints in the patterns of molecular evolution of its encoding genes at multiple levels.

Supplementary Material

Supplementary tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This study was supported by a grant from the Spanish Ministerio de Ciencia e Innovación (BFU2009-12022) and a grant from the Research Frontiers Program (10/RFP/GEN2685) from Science Foundation Ireland to MAF. D.A.-P. was supported by a Juan de la Cierva postdoctoral fellowship from the Spanish Ministerio de Economía y Competitividad (JCI-2011-11089). We are grateful to two anonymous referees for their helpful comments.

Literature Cited

Albert R, Jeong H, Barabási AL. 2000. Error and attack tolerance of complex networks. Nature 406:378–382.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Alvarez-Ponce D. 2012. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. BMC Evol Biol. 12:192.

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Res. 19:234–242.

Alvarez-Ponce D, Aguadé M, Rozas J. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. Genome Biol Evol. 3:87–101.

Alvarez-Ponce D, McInerney JO. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. Genome Biol Evol. 3: 782–790.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.

Arabidopsis Interactome Mapping Consortium. 2011. Evidence for network evolution in an Arabidopsis interactome map. Science 333: 601–607.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Bader JS, Chaudhuri A, Rothberg JM, Chant J. 2004. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol. 22: 78–85.

Barakat A, et al. 2001. The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome. Plant Physiol. 127:398–415.

Batada NN, Hurst LD, Tyers M. 2006. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2:e88.

Birchler JA, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol. 234:275–288.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 13:137–144.

Bloom JD, Adami C. 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.

Butcher EC, Berg EL, Kunkel EJ. 2004. Systems biology in drug discovery. Nat Biotechnol. 22:1253–1259.

Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. 14: 802–811.

Clark NL, Alani E, Aquadro CF. 2012. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome Res. 22: 714–720.

Codoñer FM, Fares MA. 2008. Why should we care about molecular coevolution? Evol Bioinform Online 4:29–38.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. Bioessays 26:479–484.

Cotton JA, McInerney JO. 2010. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc Natl Acad Sci U S A. 107: 17252–17255.

Cui Q, Purisima EO, Wang E. 2009. Protein evolution on a human signaling network. BMC Syst Biol. 3:21.

D'Antonio M, Ciccarelli FD. 2011. Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol. 7: e1002029.

Davids W, Zhang Z. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. BMC Evol Biol. 8:23.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol. 20:591–597.

Deeds EJ, Ashenberg O, Shakhnovich EI. 2006. A simple physical model for scaling in protein-protein interaction networks. Proc Natl Acad Sci U S A. 103:311–316.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

Doherty A, Alvarez-Ponce D, McInerney JO. 2012. Increased genome sampling reveals a dynamic relationship between gene duplicability and the structure of the primate protein-protein interaction network. Mol Biol Evol. 29:3563–3573.

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.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.

Eanes WF. 2011. Molecular population genetics and selection in the glycolytic pathway. J Exp Biol. 214:165–171.

Ekman D, Light S, Bjorklund AK, Elofsson A. 2006. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biol. 7:R45.

Fares MA, Ruiz-Gonzalez MX, Labrador JP. 2011. Protein coadaptation and the design of novel approaches to identify protein-protein interactions. IUBMB Life. 63:264–271.

Fraser HB. 2005. Modularity and evolutionary constraint on proteins. Nat Genet. 37:351–352.

Fraser HB, Hirsh AE. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol Biol. 4:13.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296: 750–752.

Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol. 3:11.

Freeman LC. 1977. A set of measures of centrality based on betweenness. Sociometry 40:35–41.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803–806.

Herbeck JT, Wall DP. 2005. Converging on a general model of protein evolution. Trends Biotechnol. 23:485–487.

Hu TT, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet. 43:476–481.

Hughes AL, Friedman R. 2005. Gene duplication and the properties of biological networks. J Mol Evol. 61:758–764.

Ingvarsson PK. 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Mol Biol Evol. 24:836–844.

Ispolatov I, Yuryev A, Mazo I, Maslov S. 2005. Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res. 33:3629–3635.

Jeong H, Mason SP, Barabási AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. 2000. The large-scale organization of metabolic networks. Nature 407:651–654.

Jordan IK, Wolf YI, Koonin EV. 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.

Kersey PJ, et al. 2012. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res. 40:D91–D97.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge (United Kingdom): Cambridge University Press.

King JL, Jukes TH. 1969. Non-Darwinian evolution. Science 164:788–798.

Koonin EV, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 17:481–487.

Korcsmáros T, Szalay MS, Böde C, Kovács IA, Csermely P. 2007. How to design multi-target drugs: target search options in cellular networks. Exp Opin Drug Discov. 2:799–808.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13: 2229–2235.

Kunin V, Pereira-Leal JB, Ouzounis CA. 2004. Functional evolution of the yeast protein interaction network. Mol Biol Evol. 21:1171–1176.

Larracuente AM, et al. 2008. Evolution of protein-coding genes in Drosophila. Trends Genet. 24:114–123.

Lee JW, Kim TY, Jang YS, Choi S, Lee SY. 2011. Systems metabolic engineering for chemicals and materials. Trends Biotechnol. 29:370–378.

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 WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2: 150–174.

Liang H, Li WH. 2007. Gene essentiality, gene duplicability, and protein connectivity in human and mouse. Trends Genet. 23:375–378.

Lu C, Zhang Z, Leach L, Kearsey MJ, Luo ZW. 2007. Impacts of yeast metabolic network structure on enzyme evolution. Genome Biol. 8:407.

Luisi P, et al. 2012. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Mol Biol Evol. 29:1379–1392.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. Trends Genet. 25:152–155.

McInerney JO. 2006. The causes of protein evolutionary rate variation. Trends Ecol Evol. 21:230–232.

Muller J, et al. 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species, and functional annotations. Nucleic Acids Res. 38:D190–D195.

Myers L, Sirois MJ. 2006. Spearman correlation coefficients, differences between. In: Kotz S, editor. Encyclopedia of statistical sciences. New York: John Wiley and Sons. p. 7901–7903.

Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. J Mol Evol. 1:18–25.

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

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14:609–614.

Prachumwat A, Li WH. 2006. Protein function, connectivity, and duplicability in yeast. Mol Biol Evol. 23:30–39.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. 2002. Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555.

Rocha EP. 2006. The quest for the universals of protein evolution. Trends Genet. 22:412–416.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94.

Schmid M, et al. 2005. A gene expression map of Arabidopsis thaliana development. Nat Genet. 37:501–506.

Slotte T, et al. 2011. Genomic determinants of protein evolution and polymorphism in Arabidopsis. Genome Biol Evol. 3:1210–1219.

Stark C, et al. 2011. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39:D698–D704.

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

Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

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

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. Bioessays 24: 175–184.

Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics 168:569–574.

Veitia RA. 2005. Paralogs in polyploids: one for all and all for one? Plant Cell 17:4–11.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 7:R39.

Wagner A. 2012. Metabolic networks and their evolution. Adv Exp Med Biol. 751:29–52.

Wagner A, Fell DA. 2001. The small world inside large metabolic networks. Proc Biol Sci. 268:1803–1810.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem. 46:573–639.

Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc Biol Sci. 273:1507–1515.

Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol. 21:1719–1726.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among Arabidopsis genes. Mol Biol Evol. 28:2359–2369.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. J Theor Biol. 8:357–366.

Associate editor: Takashi Gojobori