Determinants of Adaptive Evolution at the Molecular Level: the Extended Complexity Hypothesis

Stéphane Aris-Brosou

Bioinformatics Research Center, North Carolina State University, Raleigh

To explain why informational genes (i.e., those involved in transcription, translation, and related processes) are less likely than housekeeping genes to be horizontally transferred, Jain and coworkers proposed the complexity hypothesis. The underlying idea is that informational genes belong to large, complex systems of coevolving genes. Consequently, the likelihood of the successful horizontal transfer of a single gene from such an integrated system is expected to be low. Here, this hypothesis is extended to explain some of the determinants of the mode of evolution of coding sequences. It is proposed that genes belonging to complex systems are relatively less likely to be under adaptive evolution. To evaluate this "extended complexity hypothesis," 2,428 families and protein domains were analyzed. This analysis found that genes whose products are highly connected, located in intracellular components, and involved in complex processes and functions were more conserved and less likely to be under adaptive evolution than are other gene products. The extended complexity hypothesis suggests that both the mode and the rate of evolution of a protein are influenced by its gene ontology (localization, biological process, and molecular function) and by its connectivity.

Introduction

Finding the determinants of a protein's rate of evolution has proved a difficult task, but recent evidence from the yeast *Saccharomyces cerevisiae* suggests that the rate of evolution of a protein correlates positively with dispensability (Hirsh and Fraser 2001) and negatively with its expression level (Pal, Papp, and Hurst 2001) or with its number of interacting partners (Fraser et al. 2002; Wuchty, Oltvai, and Barabasi 2003). However, for intensely studied model systems such as yeasts and bacteria, this latter correlation is contentious (Fraser, Wall, and Hirsh 2003; Jordan, Wolf, and Koonin 2003). Besides, these recent studies on the determinants of rates of evolution often neglect the important role of a protein's function (Kimura 1983) and use approximate measures of rates of evolution, which do not distinguish mutation rates from fixation probabilities. By focusing exclusively on the pace of molecular evolution, most studies so far overlook some possibly more biologically relevant aspects such as the mode of evolution; that is, whether natural selection discriminates against amino acid changes (purifying selection), is indifferent to them (neutral evolution), or favors them (adaptive evolution).

The idea of a correlation between rates, mode of evolution, function, and the number of interactions (connectivity) with other proteins is actually reminiscent of the complexity hypothesis (Jain, Rivera, and Lake 1999), originally proposed to explain why the horizontal transfer of a gene appears all the more improbable if the connectivity of the protein it encodes is large. Here, we extend the complexity hypothesis from horizontal gene transfer to modes and rates of molecular evolution and primarily propose that adaptive evolution at the molecular level is least likely for proteins with high complexity. To evaluate this new hypothesis, we analyzed a database of gene families and protein domains from a large and diverse set of species. We show that genes whose products are involved in complex functions, such as transcription, translation, and related processes, are significantly less likely to undergo adaptive evolution than are most housekeeping genes. The proposed hypothesis is further assessed by evaluating two specific consequences. First, we show that genes undergoing adaptive evolution interact with significantly fewer proteins, and, second, that rates of evolution correlate negatively with the connectivity of a protein in protein interaction networks. The proposed hypothesis is called the "extended complexity hypothesis." Here, the "complexity" of a protein refers to both its high connectivity and a specific set of terms from the Gene Ontology database, which classifies proteins not only according to their function but also according to their localization and to the biological processes in which they are involved. This extended complexity hypothesis has some consequences that transcend the original question about the determinants of the mode of molecular evolution. We show it can bring together disparate results about the evolution of protein interaction networks, the functioning of bioinformatics tools, and the concept of "speciation genes."

Methods

The Data

The Pandit database of Proteins and Associated Nucleotide Domains with Inferred Trees (Whelan, de Bakker, and Goldman 2003) version 6.2 was obtained from http://www.ebi.ac.uk/goldman-srv/pandit/. The database contains 2,730 families and protein domains with nucleotide sequences corresponding to the manually curated Pfam-A amino acid alignments of homologous protein domains (Bateman et al. 2002). Pfam-A consists of gene families and protein domains inferred using hidden Markov models (Eddy 1998). As a result, each data set can contain clusters of both orthologous and paralogous sequences that can have different functions. These sequences are sampled from 166 organisms covering a large taxonomic range and match 69% of proteins in SWISS-PROT 39 and TrEMBL 14 (Bateman et al. 2002). The Pandit database also contains estimated tree topologies, inferred by neighbor-joining (Saitou and Nei 1987) with maximum-likelihood (ML) estimates of pairwise distances obtained under the

Key words: positive selection, adaptive evolution, gene ontology, connectivity, protein interaction networks.

E-mail: stephane@statgen.ncsu.edu.

Mol. Biol. Evol. 22(2):200–209. 2005
doi:10.1093/molbev/msi006
Advance Access publication October 13, 2004

WAG + F model of amino acid evolution (Whelan and Goldman 2001).

Entries from the Pandit database were filtered to increase power and accuracy of the likelihood methods used below to detect adaptive evolution. Data sets that contained two or fewer sequences or less than 20 aligned codons were discarded. Data sets were also discarded if the product of the number of sequences and the number of codons was less than 150. The remaining 2,589 data sets contained an average of 17 sequences and 773 aligned nucleotides.

Data Analyses

Evidence for adaptive evolution is usually indicated by a ratio of amino acid replacement changes/amino acid silent changes greater than 1 (i.e., $\omega > 1$) (Yang and Bielawski 2000). Likelihood ratio tests (LRTs) and empirical Bayes techniques were used to detect families with codon positions under adaptive evolution, equivalently called positive or diversifying selection in the context of these models (Yang and Bielawski 2000). The null model assumed that the parameter ω follows a beta distribution (M7 of Yang et al. [2000]). Because the beta distribution is bounded between 0 and 1, this model does not allow adaptive evolution. The alternative model allowed adaptive evolution by adding to the previous model a discrete class where ω can be greater than 1 (M8 of Yang et al. [2000]). These two nested models were compared by means of an LRT, where the negative of twice the log-likelihood difference between M7 and M8 was compared with the χ^2 distribution with 2 degrees of freedom.

Evidence for adaptive evolution in a family or protein domain was deemed sufficient when two conditions were met. First, the LRT comparing M7 versus M8 had to be significant at the 1% level with a non-0 frequency of codon sites with an estimated $\omega > 1$ (Nielsen and Yang 1998; Yang et al. 2000). Second, at least one actual site had to be identified to be under positive selection by an empirical Bayes analysis. This site had to have a posterior probability ≥ 0.95 to be in the $\omega > 1$ category (e.g., Yang 2002).

ML computations were performed using PAML (Yang 1997) version 3.13 of August 2002. Convergence of the optimization procedures was systematically checked by running each ML analysis twice, starting from different initial ω values ($\omega = 0.5$ and $\omega = 2$). When results from the two runs of the same codon model differed, the analysis with the largest log-likelihood was kept, in accordance with the ML criterion. In addition, under each model, families for which convergence was dubious (LRT statistic ≤ 0) were discarded.

Gene Ontology and Tests of Functional Genomic Hypotheses

Gene ontologies (GOs) represent a controlled vocabulary (ontology) applicable to all organisms. These GOs correspond to gene annotations that are both unequivocal and consistent across databases. The GO database (Harris et al. 2004) used here has a hierarchical structure, at the top of which are the divisions between cellular components, biological processes, and molecular functions. These three top-level entries are further subdivided into a hierarchical system of ontologies. Ontologies are linked to known genes, which in return are potentially associated with several GO terms, following a representation known as directed acyclic graph. SwissPROT/TrEMBL accession numbers were extracted from Pandit and used to retrieve their associated GO terms.

Families and protein domains were then classified into two categories: those showing evidence for positive selection (hereafter denoted PS^+) and those failing to show evidence for positive selection (PS^-). Fisher's exact test was used to assess whether positive selection was overrepresented or underrepresented within GO terms. Because the same data were used to compare all GO terms at a given hierarchical level of the ontology, P values were adjusted to control the false-discovery rate. This was done via a resampling strategy that offers strong control under arbitrary dependency structure of the test statistics (Benjamini and Yekutieli 2001). All these computations were performed via the FatiGO interface (Al Shahrour, Diaz-Uriarte, and Dopazo 2004).

Testing the Effect of Connectivity and Structural Complexes

Families and protein domains analyzed may form interactions and/or structural complexes with other Pfam members. Both the number of interactions (denoted connectivity hereafter) and the number of structural complexes were extracted from the Pfam database (http://www.sanger.ac.uk/cgi-bin/Pfam/) version 12. The mean connectivity for PS^+ was compared with that for PS^-. This two-sample comparison of means was performed using a bootstrap, conducted by sampling families with replacement within each group. Sample sizes were fixed to the original number of observations within each group. The statistic used was the difference between the mean connectivity for PS^+ and the mean connectivity for PS^-. This procedure was repeated 100,000 times. Mean numbers of structural complexes between PS^+ and PS^- were compared in the same way. No information about connectivity or structural complexes was available for a total of 99 families and protein domains present in Pandit version 6.2 that were removed from Pfam version 12. These data sets were excluded from these comparisons.

Results

Extent of Adaptive Evolution

To test the null hypothesis that the mode of evolution of a protein is independent of its gene ontology and of its connectivity, a large database consisting of 2,428 families and protein domains was mined for evidence of sites under adaptive evolution. Out of these, 799 had a likelihood ratio test significant at the 1% level (1,162 data sets at the 5% level; 494 data sets at the 0.1% level), of which 93 had an ω rate ratio greater than 1 (159 data sets at the 5% level; 44 data sets at the 0.1% level). The empirical Bayes analysis reduced this number to 91 data sets for which at least one site was detected to be under adaptive evolution with a posterior probability ≥ 0.95 (154 data sets at the 5% level;

42 data sets at the 0.1% level). Results from the likelihood (LRT) and the LRT + empirical Bayes analyses were very close. (Observed differences were caused by small data sets.) To remain on the conservative side, further analyses were based on the LRT + empirical Bayes results.

Because the Pandit database contains four data types (protein families, domains, motifs, and repeats), it is important to test whether positive selection was detected differentially among these data categories. The 99 data sets (four in which positive selection was detected) present in Pandit version 6.2 but removed from Pfam version 12 were referenced here as "NA." There was no significant difference in the distribution of positive selection between data types with NA data types included as a fifth type (χ^2_4, $P = 0.3163$) or not (χ^2_3, $P = 0.1991$) so that these four data types were pooled together and no longer distinguished hereafter.

To check that the identification of families under adaptive evolution was robust to the specification of the models of evolution, we performed an independent test based on simpler models of codon evolution. A null model that allowed only conserved ($\omega = 0$) and neutral ($\omega = 1$) sites was compared with a more general model that allowed adaptive evolution by adding to the null model a discrete class where ω can be greater than 1 (M1 versus M2 of Yang et al. [2000]; each model was run twice as above). The LRT + empirical Bayes analyses identified 114 families at the 1% level and 95% posterior probability level. All 91 families identified above by the M7/M8 analyses were also identified in the M1/M2 analyses. The 23 families detected only by the M1/M2 analyses were treated as false positives to remain conservative.

The methods used here have been criticized as unreliable (Suzuki and Nei 2001, 2002). However, recent computer simulations (Wong et al. 2004) demonstrated that claims of excessive false positive rates were based on incorrect results (Suzuki and Nei 2001), optimization problems, or simulation errors (Suzuki and Nei 2002). These recent simulations, furthermore, confirmed over a wide range of parameter values both power and accuracy of likelihood and empirical Bayes methods in detecting positive selection (Wong et al. 2004). Other simulation studies also showed that power and accuracy increase with the number of sequences analyzed and with their sequence length (Anisimova, Bielawski, and Yang 2001, 2002). However, divergence levels potentially affect these methods, because little can be learned by comparing very similar sequences, and overly divergent sequences may be nearly randomized and carry little information about evolutionary processes. In addition, it can be suspected that biased codon composition (measured by G+C content, denoted GC) or differential expression levels (measured by the Codon Adaptation Index [Sharp and Li 1987], denoted CAI as a proxy) can affect the results. To test for these specific effects (number of sequences, sequence length, divergence level, GC, and GC at third codon positions [GC3]) on the detection of positive selection, we performed an analysis of variance. Divergence level was measured by the ratio of tree length and number of branches (twice the number of sequences minus three) to reflect the average amount of evolution separating two nodes in a phylogeny. Five of the factors, namely, sequence length ($F = 0.89$, $P = 0.3458$), divergence level ($F = 0.29$, $P = 0.5922$), GC ($F = 0.06$, $P = 0.8107$), GC3 ($F = 0.62$, $P = 0.4329$), and expression level ($F = 0.09$, $P = 0.7600$), did not show any significant effect on the detection of positive selection. Only the number of sequences included in the analyses showed a significant effect ($F = 21.99$, $P < 0.0001$), positive selection being more likely to be detected in large data sets. Because the power of the methods used increases with the number of sequences (Anisimova, Bielawski, and Yang 2001, 2002), the risk of detecting false positive for large data sets was considered negligible.

Gene Ontology and Adaptive Evolution

Following the hierarchical nature of the GO database (Harris et al. 2004), the distribution of adaptive evolution was studied at three levels: cellular components (localization), biological processes, and molecular functions. A potential limitation of this approach is that the gene ontology structure of the database mined for positive selection affects the power of this analysis. Figure 1 shows that GO terms are not evenly distributed over the database analyzed. These biased distributions suggest that significant differences should be more difficult to detect for sparsely sampled classes such as immunoglobulin complexes, behavioral processes, or chaperone regulator activities, so that no conclusions about these GO terms could be drawn here. Alternatively, greater power is expected for cellular components, mobilizing physiological and cellular processes, and those involving catalytic and binding activities. Conclusions given below were limited to these latter GO terms.

GO terms represented by only a small number of data sets constitute another potential source of bias. For instance, let us consider a GO term represented by only two data sets for which positive selection was detected. Positive selection will be significantly overrepresented in this GO term, but this result may be caused by poor sampling. This caveat was addressed by disregarding GO terms where fewer than five data sets were present (SwissProt accession numbers had to represent at least five data sets both in the PS^+ group and in the PS^- group). This arbitrary cutoff helped eliminate entries such as "postsynaptic and presynaptic membranes" from table 1 (italicized items), entries that each contained sequences from only one data set in the PS^+ group, a neurotoxin, and none in the PS^- group.

Results for localization are presented in table 1 and show that the distribution of adaptive evolution is GO specific. For instance, viral components, which represent obvious targets for adaptive evolution, were identified as such. More generally, most of the cases where evidence for positive selection is overrepresented are restricted to components located on the periphery of the cell. A similar distribution of GO terms is found in terms of biological processes (table 2) and of molecular functions (table 3).

The gene ontology distribution of families and protein domains where evidence for positive selection is underrepresented is also very specific. Conserved cellular

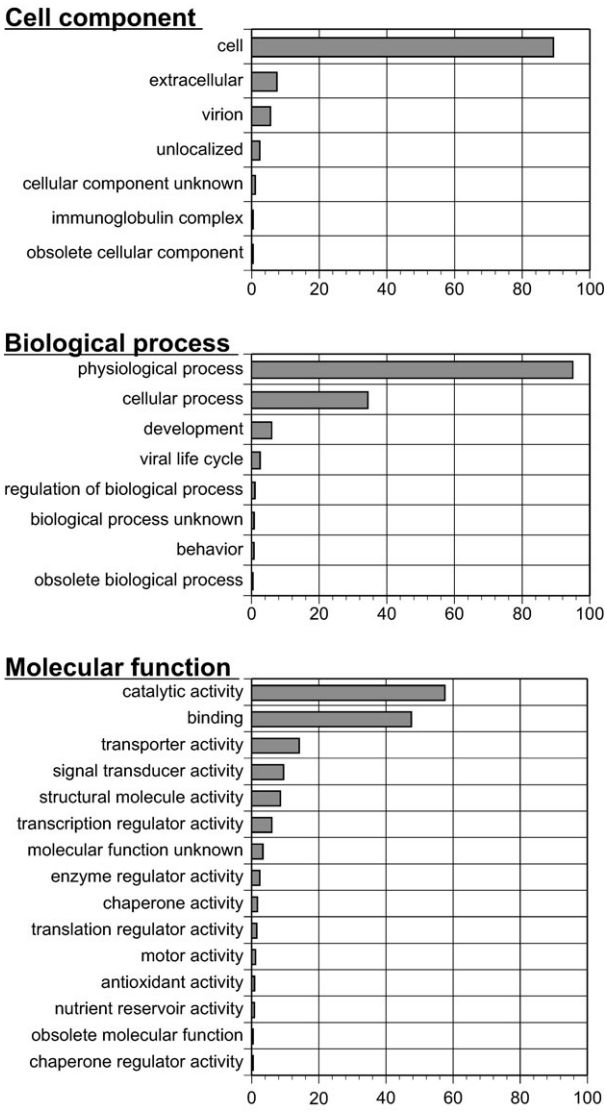

FIG. 1.—Distribution of GO terms among the 2,589 data sets analyzed. All GO terms are taken at the second hierarchical level of GO database (Harris et al. 2004). Frequencies do not sum to 1, because some of the sequences analyzed have more than one GO term at a given hierarchical level.

components are mostly internal to the cell (intracellular, inner membrane, and plasma membrane), involve families and protein domains whose products have an informational role (chromosome, nucleus, and ribonucleoprotein complex), and, thereby, involve complex interactions and/or form complex structures (tables 1–3). Families and protein domains with a role in photosynthesis (thylakoid [table 1]) constitute another example of gene products involved in complex systems.

Taken together, these results demonstrate that the gene ontology of a protein is strongly correlated with its ability to undergo adaptive evolution. In particular, informational genes or genes involved in complex functions appeared to be the most conserved. One reasonable and testable explanation is that the products of these genes form networks of coadapted and highly integrated proteins, wherein adaptive evolution is constrained by the number of interacting partners or connectivity of a protein.

Selective Constraints and Connectivity

The Pfam database version 12 lists known interactions for each member of the database with other members of the database. As a result, interaction counts obtained here underestimate the actual connectivity of each family and protein domain, either because some of the interacting partners can be absent from Pfam or because they are simply unknown. However, whereas Pfam 12 was used for counting interactions, Pfam/Pandit 6.2 was used to detect adaptive evolution. Pfam 12 has a better coverage than version 6.2, as it matches 73% of SWISS-PROT (Bateman et al. 2004). Consequently, the information retrieved about numbers of structural complexes and connectivity includes more potential partners than the number of data sets analyzed.

Each family or protein domain can interact on a transient one-to-one basis with another Pfam member or require the formation of a stable structural complex to bring about certain biological functions. In this latter case, selection is expected to act on all members of the complex, irrespective of the number of complexes formed by each member of the complex. In fact, there is some mild evidence that families and protein domains that form at least one structural complex tend to belong to the PS^- group ($P = 0.0309$), but the PS^- group and the PS^+ group form on average the same number of structural complexes (fig. 2*A*; $P = 0.4817$). However, as predicted by the extended complexity hypothesis, connectivity in the PS^- group was significantly larger than that in the PS^+ group (fig. 2*B*; $P = 0.0045$). Connectivity may, therefore, affect the mode of evolution of a protein by limiting its propensity to undergo adaptive evolution.

Yet, this does not necessarily imply that the rate of evolution is limited by connectivity. To explore this possibility, we first estimated the ω rate ratio for each data set. This estimation was based on the one-ratio model (Goldman and Yang 1994; Yang et al. 2000), in which the ω rate ratio is taken as an average over both codon sites and lineages. As above, convergence of the optimization procedures was checked by running each analysis twice (initial values set to $\omega = 0.5$ and $\omega = 2$). At each connectivity level, the estimated ω rate ratios were averaged. Figure 3 shows these means plotted against connectivity. Although a robust MM regression (Yohai 1987 [fig. 3, solid line]) explained only 11.92% of the variation in response, the slope was significantly different from 0 ($t_{18} = -2.9409$; $P = 0.0087$). For some connectivity levels (10, 18, 19, and 21), only one data set was sampled from the Pandit database and analyzed. A robust MM regression without the data sets at these connectivity levels did not affect the regression line (fig. 3, dotted line) and was still significant at the 5% level ($P = 0.0127$). Although these results should be taken carefully (as the analysis above did not take into account estimation errors about ω), they suggest that the greater the connectivity of a protein, the smaller its ω rate ratio.

Another potential issue is that this relationship between selective constraints and connectivity could also reflect the action of translational selection; that is, the action

Table 1
GO Terms at the Cellular Component Level for Which Families and Protein Domains Under Positive Selection Are Underrepresented or Overrepresented

Level 1	Level 2	Level 3	Level 4	Significance Level	Under (<)/Over (>) Representation
Cellular Component					
	Cell				
		External encapsulating structure		***	>
			Cell envelope	***	>
		Intracellular		***	<
			Chromosome	***	<
			Cytoplasm	***	<
			Fimbria	***	>
			Inclusion body	***	>
			Nucleus	***	<
			Ribonucleoprotein complex	***	<
			Thylakoid	***	<
		Membrane		***	>
			Inner membrane	**	<
			Integral to membrane	***	>
			Outer membrane	***	>
			Plasma membrane	**	<
			Postsynaptic membrane	***	>
			Presynaptic membrane	***	>
	Extracellular			***	>
		Extracellular matrix		***	>
	Unlocalized			***	<
	Virion			***	>
		Viral capsid		***	>
		Viral envelope		***	>
		Vral procapsid		**	>

NOTE.—"Level" headings indicate the hierarchical levels in the GO nomenclature. Significance levels are at the 5% (*), 1% (**), or 0.1% (***) levels. GO terms containing fewer than five families are italicized. Underlined boldface indicates underrepresentation. Boldface indicates overrepresentation.

of selection to optimize the speed and accuracy of translation. Under this type of selection, highly expressed genes are expected to present strong codon usage biases, which have been documented for many prokaryotes and for a number of eukaryotes (Akashi 2001). Such a correlation was found here (Pearson's coefficient $\text{corr}_{\text{CAI,GC3}} = 0.4124$) and was highly significant ($t_{2435} = 23.02$, $P < 0.0001$). Besides, it is also known that highly expressed genes tend to evolve slowly (Pal, Papp, and Hurst 2001). Given the correlation between connectivity and rate of evolution (fig. 3), expression levels are expected to correlate positively with connectivity if the hypothesis of translational selection holds. Pearson's correlation coefficient was found to be low ($\text{corr}_{\text{CAI},\omega} = 0.0642$) but highly significant ($t_{2435} = 3.1748$, $P = 0.0015$). However, although Pal, Papp, and Hurst (2001) argued this was the result of translational selection, they did not rule out other explanations.

Discussion

The Extended Complexity Hypothesis

The complexity hypothesis (Jain, Rivera, and Lake 1999) was originally proposed to explain why housekeeping genes appear more likely to be horizontally transferred than informational genes. In this context, the authors hypothesized that the complexity of the protein interaction networks formed by informational gene products limits the likelihood of the transfer of these genes relative to that of housekeeping genes.

We proposed here to extend this concept to explain some of the determinants of the mode of evolution of proteins. We showed that evidence for adaptive evolution can be found mostly in certain GO contexts (cellular components, biological processes, and molecular functions) associated with a significantly small connectivity. More specifically, evidence for the existence of at least a site under adaptive evolution was found principally for gene products with membrane or extracellular localizations, mostly involved in pathogenesis, belonging to different aspects of cell communication or to viral life cycles. Alternatively, the most conserved proteins were found to have the highest connectivity and were mostly intracellular components involved in informational processes and functions. The number of families and protein domains analyzed (2,428) was limited, thereby, restricting the power of the present analysis. Yet, the significance of the results obtained suggests a general tendency, embodied by the concept of the "extended complexity hypothesis." The gene ontology of a protein and its connectivity were found to be correlated, which is consistent with a previously found correlation between the structure of protein networks, its function, and the localization of its constituents (Yook, Oltvai, and Barabasi 2004). This suggests that whereas connectivity is important for the evolution of protein networks (e.g., Fraser et al. 2002; Wuchty, Oltvai, and Barabasi 2003), natural selection may play an important role depending on the localization, process, and function in which the constituents of a network are involved.

The extended complexity hypothesis has some important consequences that are readily testable and are discussed below. The extended complexity hypothesis

Table 2
GO Terms at the Biological Process Level for Which Families and Protein Domains Under Positive Selection Are Underrepresented or Overrepresented

Level 1	Level 2	Level 3	Level 4	Significance Level	Under (<)/Over (>) Representation
Biological Process					
	Cellular process				
		Cell communication			
			Cell adhesion	***	>
			Cell-cell signaling	***	>
		Cell differentiation			
			Sporulation	***	>
		Cellular physiological process			
			Cell proliferation	***	<
			Cell motility	**	>
	Development				
		Pigmentation		***	>
	Obsolete biological process				
		Pyrimidine-dimer repair, DNA damage excision		**	<
		Pyrimidine-dimer repair, DNA damage recognition		**	<
	Physiological process			***	<
		Cellular physiological process			
			Cell growth and/or maintenance	***	<
			Sporulation	***	>
		Coagulation		**	>
			Blood coagulation	**	>
		Metabolism		***	<
			Amine metabolism	***	<
			Amino acid and derivative metabolism	***	<
			Biosynthesis	***	<
			Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	***	<
			Organic acid metabolism	***	<
			Oxidative phosphorylation	***	<
			Phosphorus metabolism	***	<
			Pigment metabolism	**	>
			Vitamin metabolism	*	<
		Pathogenesis		***	>
		Response to stimulus			
			Response to endogenous stimulus	**	<
			Response to external stimulus	***	>
			Response to stress	***	<
		Secretion		***	<
			Protein secretion	**	<
	Viral life cycle				
		Viral infectious cycle			
			Viral assembly, maturation, egress, and release	***	>

NOTE.—"Level" headings indicate the hierarchical levels in the GO nomenclature. Significance levels are at the 5% (*), 1% (**), or 0.1% (***) levels. GO terms containing fewer than five families are italicized. Underlined boldface indicates underrepresentation. Boldface indicates overrepresentation.

may help understand and link disparate results for which a common ground may not be obvious otherwise.

First Implications: Determinants of a Protein's Rate of Evolution

Previous studies suggested that rates of evolution are correlated with a protein's essentiality and its fitness effect (Hirsh and Fraser 2001). Under the assumption that protein evolution is largely caused by slightly deleterious amino acid substitutions (Ohta 2003), rates of evolution are expected to be lower in genes with the largest individual fitness contributions (but see Pal, Papp, and Hurst [2003]). One prediction of the extended complexity hypothesis is that genes with the largest individual fitness contributions are those whose products are highly connected in interaction networks. This prediction was shown to hold for *S. cerevisiae* (Jeong et al. 2001; Fraser et al. 2002), and the present results suggest it can be extended to the larger number of organisms contained in Pfam; that is, to a number of viruses, bacteria, higher plants, fungi, and animals.

We also showed that the rate of protein evolution is significantly negatively correlated with connectivity (fig. 3). This result complements that presented for *S. cerevisiae* (Wuchty, Oltvai, and Barabasi 2003), in which the authors showed that proteins forming complex interaction networks tend to be more conserved than those forming simpler networks. Yet, unlike this aforementioned study (Wuchty, Oltvai, and Barabasi 2003), which used the

Table 3
GO Terms at the Molecular Function Level for Which Families and Protein Domains Under Positive Selection Are Underrepresented or Overrepresented

Level 1	Level 2	Level 3	Level 4	Significance Level	Under (<)/Over (>) Representation
Molecular Function					
	Binding			**	<
		Carbohydrate binding		***	>
			Polysaccharide binding	***	>
			Sugar binding	***	>
		Metal ion binding			
			Calcium ion binding	***	>
		Nucleic acid binding		***	<
			DNA binding	***	<
			RNA binding	***	<
		Receptor binding		***	>
	Catalytic activity			***	<
		Hydrolase activity		***	>
			Hydrolase activity, acting on acid anhydrides	***	>
			Hydrolase activity, acting on ester bonds	***	>
		Kinase activity		***	<
			Protein kinase activity	***	<
		Ligase activity			
			Ligase activity, forming carbon-nitrogen bonds	***	<
		Oxidoreductase activity		***	<
			Oxidoreductase activity, acting on NADH or NADPH	**	<
		Transferase activity		***	<
			Transferase activity, transferring alkyl or aryl (other than methyl) groups	**	<
			Transferase activity, transferring one-carbon groups	***	<
	Enzyme regulator activity				
		Caspase regulator activity		**	>
	Motor activity			**	<
	Obsolete molecular function				
		Other lyase activity		***	>
	Signal transducer activity			***	>
		Receptor activity		***	>
			Transmembrane receptor activity	***	>
	Structural molecule activity			**	>
		Structural constituent of ribosome		**	<
	Transcription regulator activity			***	<
		Transcription factor activity		***	<
	Translation regulator activity			**	<
	Transporter activity				
		Carrier activity			
			Primary active transporter activity	***	>
		Channel/pore class transporter activity			
			Porin activity	***	>
		Ion transporter activity		***	<
			Cation transporter activity	**	<
			Metal ion transporter activity	**	<
		Protein transporter activity		***	<

NOTE.—"Level" headings indicate the hierarchical levels in the GO nomenclature. Significance levels are at the 5% (*), 1% (**), or 0.1% (***) levels. GO terms containing fewer than five families are italicized. Underlined boldface indicates underrepresentation. Boldface indicates overrepresentation.

degree of conservation of proteins as a proxy for rates of evolution, the results presented here are in terms of rates of evolution at the codon (DNA) level, which account for gene-specific mutation rates. However, whereas adaptive evolution did not appear to be influenced by expression levels, connectivity and expression levels were found to be significantly correlated, so that it may be difficult to disentangle the effects of connectivity and of translational selection on rates of evolution. Although the confounding effect of translational selection would question a large body of evidence favoring a direct relationship between connectivity and rates of molecular evolution (Jeong et al. 2001; Fraser et al. 2002; Wuchty et al. 2003), future studies should not overlook the potential effects of translational selection.

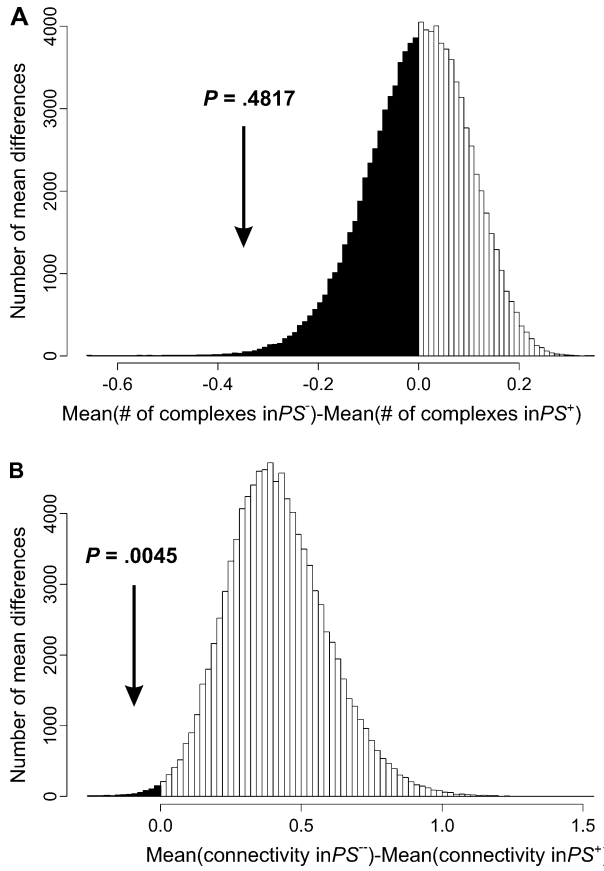

FIG. 2.—Bootstrap distributions for the difference in mean number of complexes formed (*A*) and mean connectivity (*B*) between database members failing to show evidence for positive selection (PS^-) and those showing such evidence (PS^+). Black-filled tails of the distributions give an estimate of the *P* value for the null hypothesis of no difference. The 95% confidence intervals are for structural complexes (−0.2174, 0.1792), which does not exclude the value 0, and for connectivity (0.0962, 0.8009), which excludes the value 0.

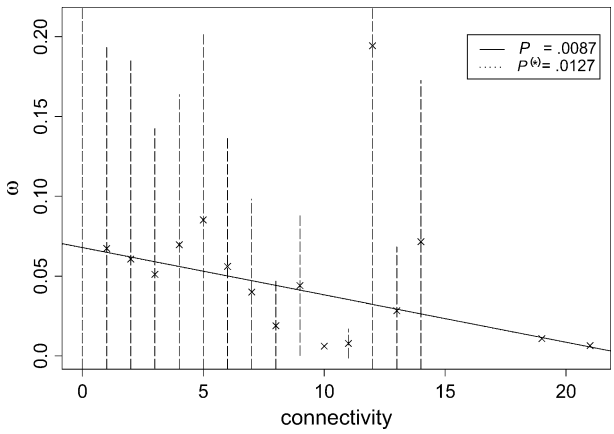

FIG. 3.—Mean rates of evolution and connectivity. The $dN/dS = \omega$ rate ratio was estimated as an average over both codon sites and lineages (one-ratio model [Goldman and Yang 1994; Yang et al. 2000], run twice to check convergence as done with other models). Vertical bars represent 1 standard deviation (SD) of mean estimated ω rate ratios, averaged within each connectivity class. This SD of the means does not encompass estimation errors on ω. The slope of the robust MM regression (solid line [Yohai 1987]) is significantly different from zero ($P = 0.0087$) and suggests that rates of molecular evolution significantly decrease with increasing number of interactions. The MM regression on the data pruned for connectivity values where only one data set was analyzed (at connectivity 10, 18, 19, and 21) is practically unchanged (dotted line) and still significant at the 5% level ($P^{(*)} = 0.0127$).

Other Implications: Interaction Networks, Bioinformatics, and Speciation Genes

The extended complexity hypothesis also has some implications with respect to the evolution of interaction networks. For instance, in *S. cerevisiae*, the connectivity of proteins in metabolic networks is found to follow a power-law distribution (Jeong et al. 2000). This means that metabolic networks have a highly inhomogeneous topology, called scale-free, in which a small number of proteins form a large number of interactions and constitute the network's "hubs." These hubs have high connectivity and, in the context of the extended complexity hypothesis, are expected to be more conserved than proteins occupying a peripheral position, because they form fewer interactions. This prediction is consistent with a study finding that less connected proteins in scale-free networks are less essential and, therefore, less conserved (see above) than are highly connected ones (Jeong et al. 2001).

Another implication of the extended complexity hypothesis is directly relevant to gene finding and functional annotation algorithms. It has been proposed that proteins might preferentially bind within their functional class (von Mering et al. 2002). As a result, if a gene of high connectivity is already annotated, but some of its binding partners still have an "unknown" function, current gene finding and functional annotation algorithms could make use of this prior information as an annotation guide (see also Wuchty, Oltvai, and Barabasi [2003]).

Alternatively, genes of low connectivity are expected be more difficult to annotate for two reasons. First, because of their expected high rate of evolution, they are more difficult to "Blast" (Altschul et al. 1990). Second, because they are coding for proteins with small connectivity, the strategy outlined above cannot be used. This consequence is partly confirmed by a recent study, which found that proteins of "unknown" function in the Database of Interacting Proteins (Xenarios et al. 2002) have low connectivity (Kunin, Pereira-Leal, and Ouzounis 2004).

An ultimate consequence in bioinformatics is that the quality of annotations (e.g., among the recently sequenced genomes) should be better for highly connected genes, which are more conserved during evolution, than for most housekeeping genes. Databases are, therefore, expected to be nonrandom and exhibit some sampling bias toward genes of moderate to high connectivity. Although this should facilitate the reconstruction of hubs in scale-free networks, external nodes are expected to be more difficult to reconstruct. With the development and extensive use of data-mining strategies, the extent and consequences of such biases warrant further investigation.

Last but not least, if the most-connected genes are the most conserved across species (see above), then "speciation genes" (Coyne 1992) should preferentially be found among those of low connectivity, which are expected to be the predominant targets of diversifying selection (see

Results). This is consistent with a study of 43 organisms from the three domains of life from the What-Is-There database (Overbeek et al. 2000), in which species-specific differences emerged for the less-connected genes (Jeong et al. 2000). On the other hand, the most-connected genes were found to be shared by all organisms (Jeong et al. 2000). The extended complexity hypothesis would then imply that not all genes in the genome potentially participate in the speciation process, a view not widely accepted (Shaw 2001) but which could easily be tested today.

Acknowledgments

I thank Steffen Heber, Tae-Kun Seo, Jiaye Yu, Xuhua Xia, and, especially, Jeffrey Thorne for discussions and two reviewers for critical comments. I am also grateful to Wendy Wong for sharing a manuscript before its publication. This work was funded by NSF grant DEB-0120635 and a Japanese Science and Technology grant.

Literature Cited

Akashi, H. 2001. Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11**:660–666.

Al Shahrour, F., R. Diaz-Uriarte, and J. Dopazo. 2004. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics **20**:578–580.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18**:1585–1592.

———. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. **19**: 950–958.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. 2002. The Pfam protein families database. Nucleic Acids Res. **30**:276–280.

Bateman, A., L. Coin, R. Durbin et al. (13 co-authors). 2004. The Pfam protein families database. Nucleic Acids Res. **32**: D138–D141.

Benjamini, Y. and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. **29**:1165–1188.

Coyne, J. A. 1992. Genetics and speciation. Nature **355**:511–515.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics **14**:755–763.

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., 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–

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Harris, M. A., J. Clark, A. Ireland et al. (58 co-authors). 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. **32**:D258–D261.

Hirsh, A. E., and H. B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **411**:1046–1049.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA **96**:3801–3806.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. Nature **411**:41–42.

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., 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

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kunin, V., J. B. Pereira-Leal, and C. A. Ouzounis. 2004. Functional evolution of the yeast protein interaction network. Mol. Biol. Evol. **21**:1171–1176.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

Ohta, T. 2003. Origin of the neutral and nearly neutral theories of evolution. J. Biosci. **28**:371–377.

Overbeek, R., N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov Jr, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. **28**:123–125.

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

———. 2003. Genomic function: rate of evolution and gene dispensability. Nature **421**:496–497.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sharp, P. M., and W. H. Li. 1987. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281–1295.

Shaw, K. L. 2001. The genealogical view of speciation. J. Evolution. Biol. **14**:880–882.

Suzuki, Y., and M. Nei. 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **18**:2179–2185.

———. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **19**:1865–1869.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**:399–403.

Whelan, S., P. I. de Bakker, and N. Goldman. 2003. PANDIT: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics **19**:1556–1563.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**:691–699.

Wong, W. S. W., Z. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics (in press).

Wuchty, S., Z. N. Oltvai, and A. L. Barabasi. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat. Genet. **35**:176–179.

Xenarios, I., L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30**:303–305.

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

———. 2002. Inference of selection from multiple species alignments. Curr. Opin. Genet. Dev. **12**:688–694.

Yang, Z. and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496–503.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yohai, V. J. 1987. High breakdown-point and high-efficiency robust estimates for regression. Ann. Stat. **15**:642–656.

Yook, S. H., Z. N. Oltvai, and A. L. Barabasi. 2004. Functional and topological characterization of protein interaction networks. Proteomics **4**:928–942.

Brian Golding, Associate Editor

Accepted October 4, 2004