Impacts of protein–protein interaction domains on organism and network complexity

Kai Xia, Zheng Fu, Lei Hou, et al.

Genome Res. 2008 18: 1500-1508 originally published online August 7, 2008
Access the most recent version at doi:10.1101/gr.068130.107

Letter

Impacts of protein–protein interaction domains on organism and network complexity

Kai Xia,[1] Zheng Fu,[1] Lei Hou, and Jing-Dong J. Han[2]

Chinese Academy of Sciences, Key Laboratory of Molecular Developmental Biology, Center for Molecular Systems Biology, Institute of Genetics and Developmental Biology, Beijing 100101, China

It has been a puzzle that genome or proteome sizes are not correlated with the complexity of the organisms. Although alternative splicing and noncoding and regulatory elements explain some of the differences, the complexity of the protein interaction network and regulatory network may provide additional explanations. Here, we collected 642 domains that mediate protein–protein interactions (PPIs) and examined the evolution of the PPI domains and its impact on organismal complexity and PPI network complexity. In agreement with previous more general studies of protein domains, a significant expansion of PPI domains per proteome was found in metazoa. We also found both the number and coverage of PPI domains per protein increased. However, a better correlation with complexity was seen with increasing PPI domain coverage per protein, so that proteins in complex organisms are more compact and specialized in PPI. Such a structural adaptation of the proteins is correlated with the number of interactions that the proteins can make in PPI networks, and seems to be a more favorable way to increase network connectivity than other structural adaptations.

[Supplemental material is available online at www.genome.org.]

Genome size is not always proportional to the genetic complexity. For example, *Xenopus laevis* and human have essentially the same genome size while they have a magnitude of difference in complexity, estimated by the number of cell types in an organism. This puzzling phenomenon, dubbed as the "C-value paradox," refers to the lack of correlation between the complexity of an organism and its DNA content (C-value) (Futuyma 2005). In terms of gene number, human has about 20,000 ~ 25,000 genes, just a little more than the worm *Caenorhabditis elegans*, whereas *C. elegans* has about 29 cell types and human 169. These phenomena lead to the question "why do humans have so few genes?" (Pennisi 2005)

One explanation is that alternative splicing makes up the difference in gene number, and therefore in complexity, by providing alternative isoforms to carry out different functions. However, so far, most alternatively spliced isoforms have been found to have similar, if not identical, functions (Lopez 1998; Smith and Valcarcel 2000; Graveley 2001). A second factor, which is considered to be a major factor contributing to the paradox is the "dark matter," or the noncoding and regulatory elements in the genome (Gerstein et al. 2007; Prasanth and Spector 2007). Besides these two, a third alternative answer to this question is that humans have a much more complex molecular interaction network (Koonin and Galperin 2003), that is, the connections in the network are greater in number and much more intricate and dynamic in pattern, despite a similar number of nodes to lower organisms. Here, we present our study from the perspective of protein–protein interaction (PPI) networks, or the interactome networks.

[1]These authors contributed equally to this work.
[2]Corresponding author.
E-mail jdhan@genetics.ac.cn; fax 86-10-64845797.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.068130.107.

PPI functions have been found to be enriched in domain superfamilies whose abundance in a proteome (the number of proteins containing a superfamily of domains in a proteome) correlates with organismal complexity (Vogel and Chothia 2006). In this study, we specifically examined whether expansion of PPI domains in general is a major factor contributing to organismal complexity. Further are changes at the individual protein level involved, which might directly link organismal complexity to network complexity. Specifically, does an increase in PPI domain number, length, or coverage per protein also contribute to organismal complexity, and could these structural adaptations increase organismal complexity through increasing PPI network complexity?

We collected 642 protein domains that are involved in PPI and compared the domain compositions of proteins in 19 different organisms ranging from *Kluyveromyces lactis* to *Homo sapiens*, as well as two plants, *Oryza sativa* and *Arabidopsis thaliana*. We also included a slime mold, *Dicytostelium discoideum,* as a transition point between unicellular organisms and metazoan (Supplemental Table 1). Surprisingly, we found that specialization and compaction of PPI domain-containing proteins through increasing coverage of PPI domains per protein are potentially the most significant structural changes associated with organism complexity (measured as number of cell types in an organism) and network complexity (measured as number of PPIs of a protein).

Results

Collection of PPI domains

A protein domain is a structural unit that often folds independently of the rest of the protein (David and Nelson 2005). We determined each protein's domain architecture as annotated by InterPro (Mulder et al. 2005), an integrated domain database that

18:1500–1508 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

combines a number of detection methods and annotations currently from 11 databases. PPI domains are those that can recognize exposed sites on their binding partners. Here, we obtained a list of PPI domains through four sources: (1) annotated to be involved in protein domain–domain interactions by the Database of Domain Interaction & Bindings Database (DDID), (2) have a GO term of "protein (any protein) binding" in the InterPro domain gene ontology (GO) annotation database, (3) labeled as "experimentally proven to be involved in PPI" in the InterPro Interactions field, (4) have protein binding or interaction function described in the InterPro domain description or literature. This results in 642 PPI domains that are found in the 19 organisms. The 642 PPI domains are selected from all protein families listed in the InterPro database without any prior selection for a particular family or representatives of a small number of families. PPI domains are annotated solely based on whether they are experimentally determined or annotated by DDIB and InterPro as domains participating in PPI. These domains, together with the source of evidence supporting their PPI function, are listed in Supplemental Table 2. Because the SwissProt database contains only experimentally validated polypeptide sequences, it may be more biased than proteins predicted from genomic sequences, such as those in the TrEMBL database, which is developed to complement the SwissProt entries. We therefore based the analyses on UniProt proteins, which is a combination of SwissProt and TrEMBL protein entries.

We first looked for any evolutionary trend of PPI domains among the 19 species (Supplemental Table 1). Compared with unicellular organisms, the percentage of PPI domain-containing proteins increases dramatically in multicellular animals, indicating a big expansion of PPI domains at the proteome level upon the transition from unicellular to metazoan organisms (Fig. 1A). We listed the proportions of proteins unannotated for any domain to examine whether a particular organism is significantly under- or overannotated for domain structures as compared with *Saccharomyces cerevisiae* or human. If so (proportions test $P < 0.05$), the organism's name is marked with an asterisk or a pound sign in Figure 1A. It indicates that rice, *Candida albicans*, *Neurospora crassa*, and *Dictyostelium discoideum* are significantly underannotated, and *Arabidopsis*, fission yeast, zebrafish, frog, rat, mouse, and cow are significantly overannotated. To control for the annotation differences, we excluded all of the proteins with no domain annotation from our analysis. For example, the percentage of proteins having PPI domains versus the total proteins that have at least one annotated domain is shown in Figure 1A. Different accuracy of different genome and proteome sequences and annotations is bound to remain one cause of variation between organisms, even when significantly reduced by considering only the domain-annotated portion of the proteomes. Therefore, we studied not only the change between yeast and human, but also those in 17 other different species to derive statistically valid conclusions that are independent of annotation variations. In addition, wherever possible, we included the non-PPI domains as controls for any potential domain annotation biases. If the PPI and non-PPI domains have the same trend, bias must be considered.

Thirty-seven percent of the PPI domains are specific to metazoa ("metazoan specific"), whereas only 8% of the PPI domains are specific to unicellular organisms. Ten percent of them are "expanded" proteome-wide in metazoa, that is, an increased fraction of a metazoan proteome contains these domains. Measured by the d_N/d_S of each domain (Methods), both "metazoan specific" and "expanded" PPI domains seem to evolve faster than non-PPI domains and other PPI domains (Supplemental Data; Supplemental Fig. 1; Supplemental Tables 3, 4). The abundance distribution and the evolutionary rates are consistent with the previous findings that PPI functions are enriched among protein superfamilies whose expansions correlate with organismal complexity. Protein domains with signaling and regulatory functions have been repeatedly found highly expanded in various metazoan proteomes (Kirschner and Gerhart 1998; Rubin et al. 2000; Lander et al. 2001; Pawson and Nash 2003; Vogel and Chothia 2006). We also found that PPI domains specific to or expanded in metazoa preferentially participate in signaling and regulation (Supplemental Data; Supplemental Table 5).

Expansion of PPI domains in individual proteins

PPI domain content of individual proteins directly dictates what proteins they interact with and what PPI network they construct. We therefore examined whether the PPI domains are also expanded at the level of individual proteins through three types of structural changes: domain number, domain length, and domain coverage (see below for definition). We examined the whole collection of PPI domains on a protein, not just any particular domains. We also analyzed non-PPI domains as a whole collection to compare with PPI domains and control for study bias in higher organisms.

Different domains often have distinct functions, so that a protein with multiple domains may have more opportunities to interact with other proteins or small molecules to carry out additional functions. We first calculated the percentage of proteins with single or multiple PPI domains in different proteomes (Fig. 1B). We found that through evolution, multicellular organisms tend to have larger fractions of proteins with multiple PPI domains compared with the fractions of proteins with a single PPI domain, but that the fraction of proteins containing multiple non-PPI domains does not increase relative to that of single non-PPI domain-containing proteins (Fig. 1C). We then examined the relationship of PPI domain number increase on individual proteins to organismal complexity, estimated by the number of cell types in an organism (Vogel and Chothia 2006).

To ensure the robustness of the analysis and to rule out the possibility that only small proportions of proteins are responsible for the results, we randomly divided the proteins in each organism into 10 nonoverlapping groups, and carried out leave-one-out analysis and boxplot analysis. By leaving one group out and examining the remaining nine groups as a combination, we obtained one correlation to organismal complexity for each nine-group combination (Fig. 1D). This would detect differences in the correlations obtained using different datasets. By boxplot, the variations among all of the 10 different datasets were visualized and considered when examining a correlation between two variables. The average non-PPI domain number per protein is slightly anticorrelated with organismal complexity, or the number of cell types in an organism (Supplemental Fig. 2A), while the PPI domain number per protein is highly correlated with organismal complexity (Fig. 1D). Similar results can be seen if all 10 groups were analyzed together with the variations among different groups taken into account (Supplemental Fig. 3A,B).

Increased coverage by PPI domains on proteins

In addition to an increase in PPI domain number per protein, we also found that in complex organisms PPI domain-containing

proteins become more compact or packed with PPI domains. By "compactness," we refer to proteins with smaller fractions of domain-free regions that lack any domain. These domain-free regions are not necessarily the disordered regions, which can frequently appear inside DNA-binding and protein-binding domains, especially on signaling proteins (Dunker et al. 2005).

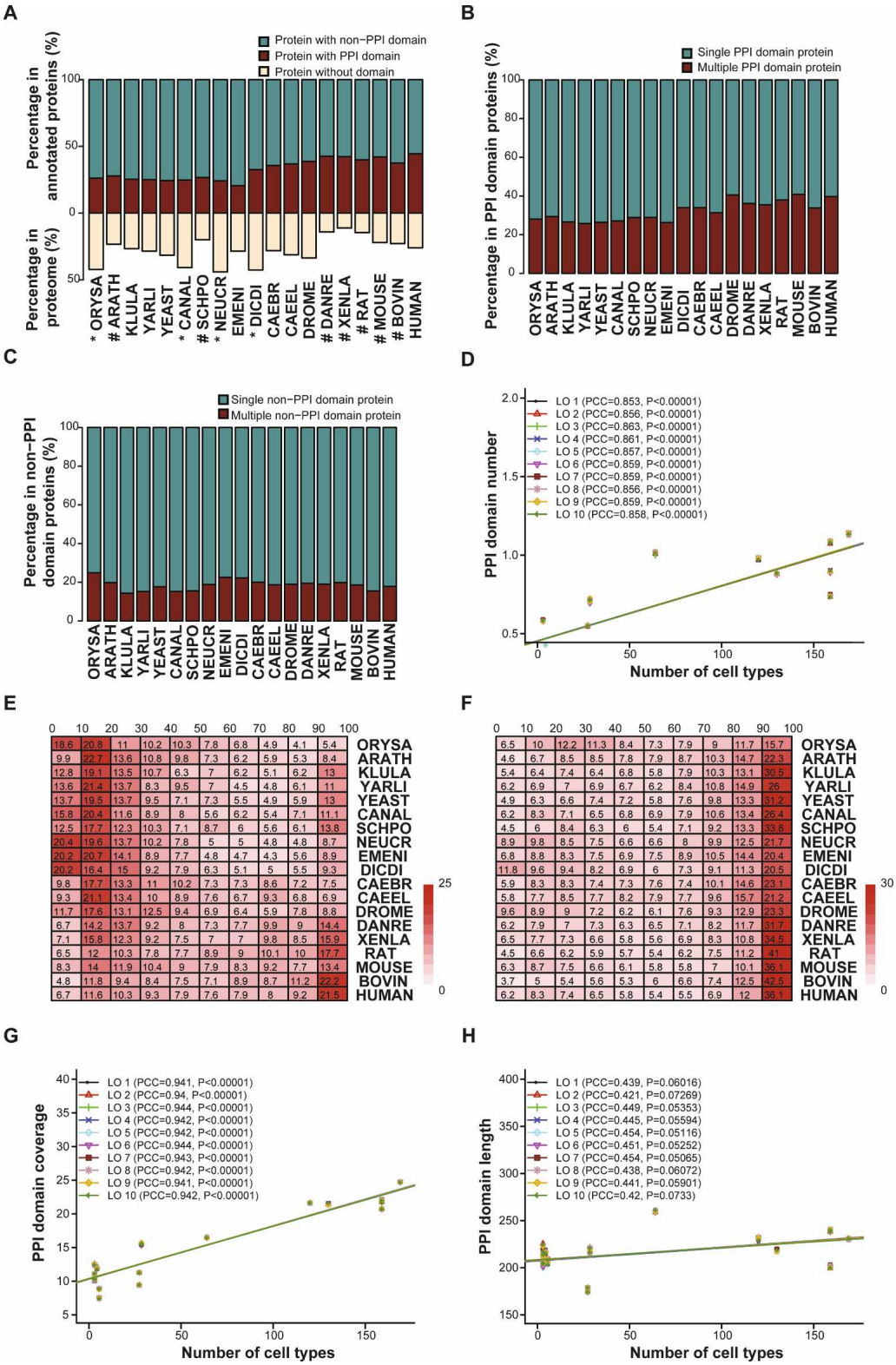

Figure 1. (Legend on next page)

Thus, compactness does not necessarily correspond to the tightness in proteins' three-dimensional packing. To examine such "compactness" quantitatively, we defined a metric "domain coverage" to quantify the functional compactness of protein domain architecture. It is defined as the percentage of a protein covered by certain protein domains over its entire length. Domain coverage then represents the percentage of the functional regions within the full length of a protein. It can be calculated by the following formula:

$$\text{Domain Coverage} = \frac{\sum \text{Each domain length}}{\text{Protein length}} \times 100$$

Higher domain coverage value, when all of the domains are included, implies a compact protein with fewer nonfunctional regions over its primary structure. PPI domain coverage, that is, the percentage of the entire length of a protein occupied by PPI domains, may also indicate whether a protein is more specialized with fewer non-PPI domains.

We used heatplots to visualize the evolutionary trend of the PPI and non-PPI domain coverage (Fig. 1E,F). In addition to the rise of multi-PPI domain proteins in higher organisms, proteins shifted from being enriched at low PPI domain coverage in lower organisms and plants to being enriched at high coverage in animals (Fig. 1E). In cow and human, >20% of the annotated proteins have PPI domain coverage near 100 (Fig. 1E). In contrast, the non-PPI domain coverage does not change (Fig. 1F). In addition, the shift is not gradual; instead, it is from one extreme (0%–20%) to the other (90%–100%). The fractions with medium coverage remain almost constant (Fig. 1E), whereas the fractions between 0%–20% and 90%–100% change. This suggests that if a protein acquires a PPI domain or function, it tends to become specialized, so that it contains few other domains and domain-free sequences (Fig. 1E). For example, a whole category of adaptors and scaffold proteins specialized for PPI, such as GRB2, has high PPI domain coverage and is expanded greatly in higher organisms. Non-PPI domain coverage is not correlated with organismal complexity (Supplemental Fig. 2B), whereas PPI domain coverage is highly correlated with organismal complexity, even more so than the average PPI domain number (Fig. 1G). Both the leave-one-out and boxplot analyses produced similar results (Fig. 1G; Supplemental Fig. 3C,D).

We also examined whether an increase in average domain length can explain the increase in PPI domain coverage. As shown in Figure 1H, the average length of PPI domain is slightly correlated to organismal complexity, whereas that of the non-PPI domain is negatively correlated to organismal complexity (Supplemental Fig. 2C), indicating that the PPI domain length increase is small, but not attributable to annotation bias, and may contribute to the increase in domain coverage and organismal complexity to a small extent (Supplemental Fig. 3E,F).

How does a protein increase its domain coverage or compactness through evolution? We selected 2629 groups of orthologous proteins that have orthologs in at least five metazoan species and explored the change of protein structure within each group (Fig. 2; Methods). Here, we put forward four possible explanations for the increase of PPI domain coverage on proteins through evolution. The first is PPI domain length increase (Figs. 1H, 2A). The second is that some orthologous proteins gradually lose the domain-free sequences (not occupied by known protein domain) on their N termini (Fig. 2B), C termini (Fig. 2C), or in the middle of the protein during evolution. The third explanation is the loss of non-PPI domains. The fourth is the replacement of the domain-free regions with new functional domains (Fig. 2D). Among the 191 ortholog groups that have PPI domains in at least five metazoan species and are found to increase in PPI domain coverage during evolution (Spearman rank correlation coefficient [RCC] of PPI domain coverage to organismal complexity > 0.6), loss of non-PPI domains or decrease in non-PPI domain length (RCC to organismal complexity < −0.55) together accounts for a minimal 3.7% of the cases. Whereas trimming the domain-free sequences (RCC to organismal complexity > 0.55, RCCs are the same below) contributes the most to PPI domain coverage increase (51.3%), followed by increasing PPI domain length (31.4%) and PPI domain number (10.5%).

Other proteins bear no clear sequence similarities across many species. We found that for all of the proteins in the 19 organisms, reduced domain-free regions best explain an increase of PPI domain coverage on a protein than any other factors (linear regression $R^2 = 0.724$). Increased PPI domain length contributes slightly to PPI domain coverage increase (linear regression $R^2 = 0.119$). All other factors contribute very little to it (linear regression $R^2 < 0.1$). Hence, removing the sequences outside of any functional domains to make the protein become more functionally compact is the most common way to enrich for PPI domains and to increase PPI domain coverage at the protein level.

Structural changes at the protein level are also consistent with the expansion at the proteome level. The proteins containing "Expand" or "Metazoan Specific" PPI domains apparently have higher PPI domain number and coverage, but not longer PPI domains than those having the PPI domains that are "Common" to unicellular organisms, metazoa, and plants, or "Shrink"

Figure 1. Evolutionary structural adaptations toward PPI domain expansion at the individual protein level. (*A*) The percentage of proteins with PPI domains (maroon blocks), other non-PPI domains (cyan blocks) among all proteins with domain annotation in each organism (*above X*-axis), and those without any annotated domains (beige blocks) in the proteomes of different organisms (*below X*-axis). The organisms that are significantly under- or overannotated for domains are indicated by asterisks or pound signs before their names. (*B*) Percentage of proteins with single (cyan) or multiple (maroon) PPI domains in different organisms among PPI domain-containing proteins. (*C*) Percentage of proteins with single (cyan) or multiple (maroon) non-PPI domains in different organisms among non-PPI domain-containing proteins. (*D*) The relationship of the number of PPI domains per protein to the number of cell types in an organism in each of the leave-one-out nine-group combinations (LO1 ~ 10). The 10 regression lines, PCCs, and linear regression slope *P*-values result from 10 leave-one-out analyses, one for each nine out of the 10 random protein groups (same for *G* and *H*). (*E*) Distribution of PPI domain coverage. Proteins in each organism are divided into 10 groups based on their PPI domain coverage. The boundaries of each PPI domain coverage interval are shown on the *top* of the plot. The numbers in the grid give the percentage of the total domain-annotated proteins in each organism that belong to a certain PPI domain coverage interval. The color intensity in each cell is proportional to the relative percentages within each organism (row). (*F*) Distribution of non-PPI domain coverage. The color intensity and number inside each grid are denoted as in *E*, except that PPI domains are replaced by non-PPI domains. (*G*) The relationship of average PPI domain coverage to the number of cell types in an organism in each of the leave-one-out nine-group combinations (LO1 ~ 10). (*H*) The relationship of average length of PPI domains to the number of cell types in an organism in each of the leave-one-out nine-group combinations (LO1 ~ 10).

in multicellular organisms, with the proteins having "Metazoan Specific" PPI domains display the highest PPI domain coverage overall (Table 1; *P*-values are listed in Supplemental Table 6).

These results indicate that domain-free regions are more dispensable (David and Nelson 2005). In complex organisms, there is a selection for new, more, and slightly longer PPI domains, and

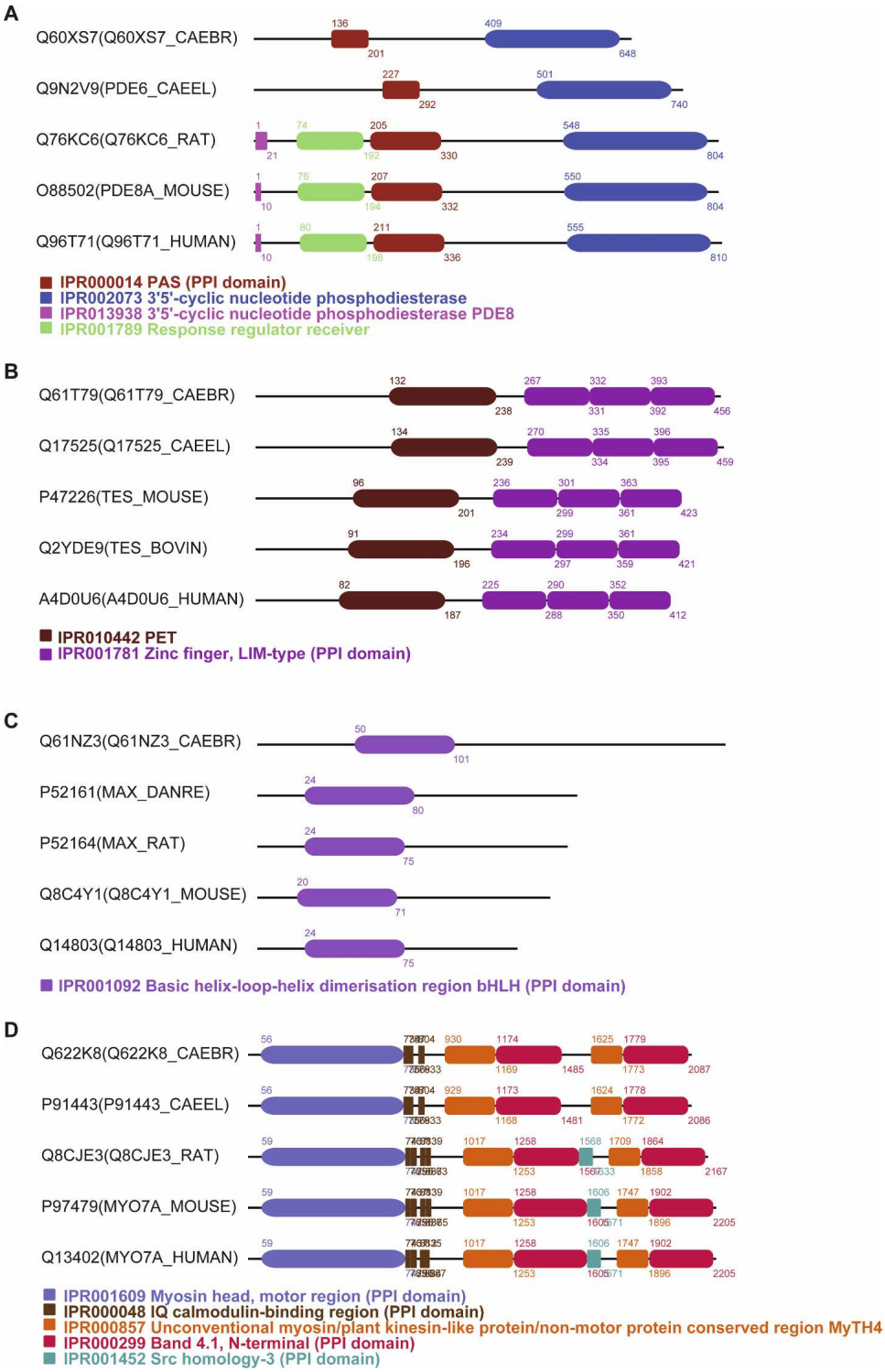

Figure 2. Examples of increasing domain coverage on orthologous proteins through evolution. PPI domain length increase (*A*), loss of the domain-free sequences at orthologous proteins' N termini (*B*), C termini (*C*), and PPI domain insertion (*D*) that contribute to increased PPI domain coverage through evolution. Each protein is labeled with its SwissProt identifier. A protein's name suffixed by the abbreviation of its species name is included inside the parentheses. Within each group, orthologous proteins are ordered by taxonomy. Colored blocks stand for different domains and their annotations are listed at the *bottom* of each figure. PPI domains are indicated in the block-color legend.

Table 1. Structural and network properties of proteins containing PPI domains of different evolutionary profiles

Metric	Expand	Common	Shrink	Metazoan specific	Multicellular	UM	Metabolic	Non-PPI
PPI DN	3.754	1.965	1.733	2.315	5.098	3.952	0.125	0.746
PPI DC	0.519	0.465	0.491	0.757	0.602	0.559	0.018	0.068
PPI DL	70.050	117.898	133.142	116.156	72.637	99.311	103.400	68.916
Non-PPI DN	0.861	0.594	0.495	0.419	0.944	0.641	1.284	1.486
Non-PPI DC	0.094	0.077	0.089	0.046	0.089	0.087	0.783	0.653
Non-PPI DL	79.406	101.602	159.567	94.671	96.895	115.016	242.589	171.742

(DN) Domain number; (DC) domain coverage; (DL) Domain length; (UM) for unicellular and metazoan.

fewer domain-free sequences in the metazoan PPI domain-containing proteins.

PPI domain coverage and PPI network complexity

At the molecular level, the correlation between protein structural complexity and organismal complexity might correlate with complexity of protein–protein or protein–nucleic acid interaction networks. To investigate this hypothesis, the relationship of PPI domain arrangements in individual proteins to the complexity of their PPI network was examined. The most straightforward way to increase network complexity is to increase the interaction degrees (k) of proteins, that is, the number of interactions a protein makes in a network. Interaction degree is therefore a most simple measurement of network complexity. We studied the high-quality PPI networks constructed from literature-based human PPIs collected by the Human Protein Reference Database (HPRD) (Peri et al. 2003) and from the core yeast PPI data set curated by the Database of Interacting Proteins (DIPcore) (Xenarios et al. 2000).

We examined whether the expansion of PPI domains in a protein is correlated with the protein-interaction degrees of the protein in PPI networks. We compared a protein's interaction degree against each of the following variables: PPI and non-PPI domain number, PPI and non-PPI domain coverage, and PPI and non-PPI domain length. Again, we applied both leave-one-out and boxplot analyses to ensure the robustness of results and to exclude the effects due to a smaller number of outliers (Fig. 3; Supplemental Figs. 4, 5). Among the six variables, PPI domain coverage has the highest correlation with protein interaction degree (Fig. 3A,B; Supplemental Figs. 4, 5), more than between PPI domain number and degree (Fig. 3C,D; Supplemental Fig. 5), and between PPI domain length and degree (Fig. 3E,F). Testing the significance of the association by linear regression indicates that only PPI domain coverage per protein is significantly correlated with the PPI degree of the protein. This is in agreement with PPI domain coverage being the factor most correlated to organismal complexity (Fig. 1). When the proteins in PPI networks were separated into hubs ($k \geq 5$) and non-hubs ($k < 5$), it is obvious that compared with non-hubs, hubs have small shifts to higher PPI domain number (Fig. 4A,B) and large shifts to higher PPI domain coverage (Fig. 4C,D) in both yeast and human PPI networks.

An increase of PPI interaction degree with an increase in the number of PPI domains on a protein would be naturally expected. But why does PPI domain coverage rather than the number of PPI domains per protein significantly correlate with PPI degrees? We wondered whether it was due to the constraints (e.g., structural constraints) for a protein to gain more domains than to increase coverage. Consistent with this hypothesis, we found that proteins with high domain number are rare in a proteome (5.2% proteins in human and 1.1% in yeast with domain number > 3), and the proportion of proteins decreases sharply as the number of PPI domains on the proteins increases, suggesting that there are constraints for the protein to gain a domain. In contrast, although the proteins with low PPI domain coverage are more common, the distribution of PPI domain coverage is much flatter than the distribution of domain number (Fig. 4), suggesting relaxed constraints. Consequently, there can be a larger number of hubs having high PPI domain coverage than having high PPI domain number (Fig. 4).

Discussion

Our main finding is that PPI domain coverage provides the best genome-wide predictor of organismal complexity yet reported. PPI domain coverage is also highly correlated with PPI degree. This suggests that complexity of PPI may be required for organismal complexity.

Why does PPI domain coverage correlate better with organismal complexity than with PPI network degree if network complexity is required for organismal complexity? It cannot be excluded that PPI domain coverage might be able to affect organismal complexity independently of PPI network complexity. For example, PPI domains could also be associated with and contribute to the complexity of the regulatory network and metabolic networks, albeit probably more indirectly. However, we note that correlations to PPI network complexity might be underestimated because the PPI network is made noisy by false-positive PPIs, because the network datasets are incomplete, or because PPI degree is an oversimplified measurement for network complexity. These factors could also explain why all of the parameters we measure correlate less significantly with network complexity than with organismal complexity. If such limitations to the network statistics could be overcome, association of other factors to PPI degree might surface.

Our results suggest that, although an increase in interaction degree could be achieved by an increase in either PPI domain number or PPI domain coverage, the increase in domain coverage is a more frequent means to increase PPI degree. The preference for increasing PPI domain coverage versus PPI domain number suggests protein structure constraints on network connectivity. It will be important to explore the molecular relationship between network complexity and PPI domain coverage in the future.

Our findings argue that, in addition to the possible contribution of alternative splicing, increased complexity in the higher organisms may also be attributable to protein structures that allow a protein to be more compact, specialize in PPI, and achieve more interactions among the same number of network nodes. From the network point of view, an increase in alternative splicing forms is a means of adding nodes to the network, whereas PPI

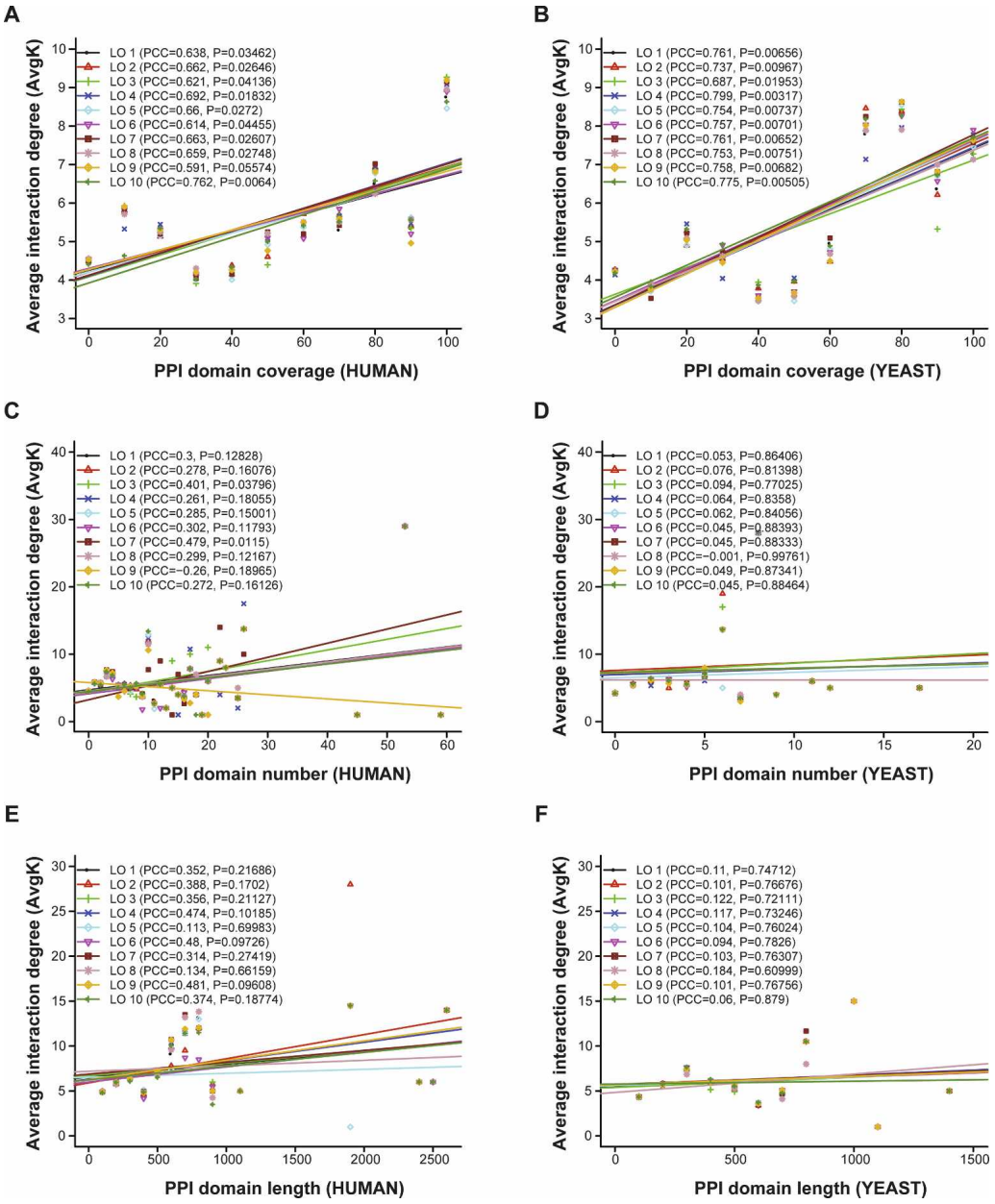

Figure 3. Relationships of PPI domain coverage, number, and length to PPI degrees by leave-one out analysis. The average human and yeast PPI degrees of proteins in each of the nine out of 10 group combinations within each interval of PPI domain coverage (*A*,*B*), PPI domain number (*C*,*D*), or PPI domain length (*E*,*F*) are plotted against their average values within the intervals. The 10 regression lines, PCCs, and linear regression slope *P*-values result from 10 leave-one-out analyses (LO1 ~ 10), one for each nine out of the 10 random protein groups.

domain expansion on individual proteins is a means of adding edges to the network. Increasing edges is a much more efficient way to increase complexity compared with increasing nodes, as the number of potential edges is the square of the number of nodes in a network (Papin et al. 2005).

Methods

Datasets

All domain information was downloaded from the InterPro ftp site (ftp://ftp.ebi.ac.uk/pub/databases/interpro/) on September 5, 2007. After parsing the XML file, we obtained the entire structural information of each protein, including its family, domain, repeat, binding site, active site, post-translational modification site, and annotated Gene Ontology information. An InterPro domain is an independent structural unit, which can be found alone or in conjunction with other domains or repeats and an InterPro repeat is a region that is not expected to fold into a globular domain on its own (Mulder et al. 2005) (ftp://ftp.ebi.ac.uk/pub/databases/interpro/user_manual.txt). To avoid excluding useful structural information of proteins, we define the word "Domain" in this work as a functional and structural unit that is confirmed by an InterPro domain entry, InterPro family entry, or InterPro repeat entry. InterPro combines a number of databases, and we chose PANTHER, Pfam, PIRSF, PRINTS,

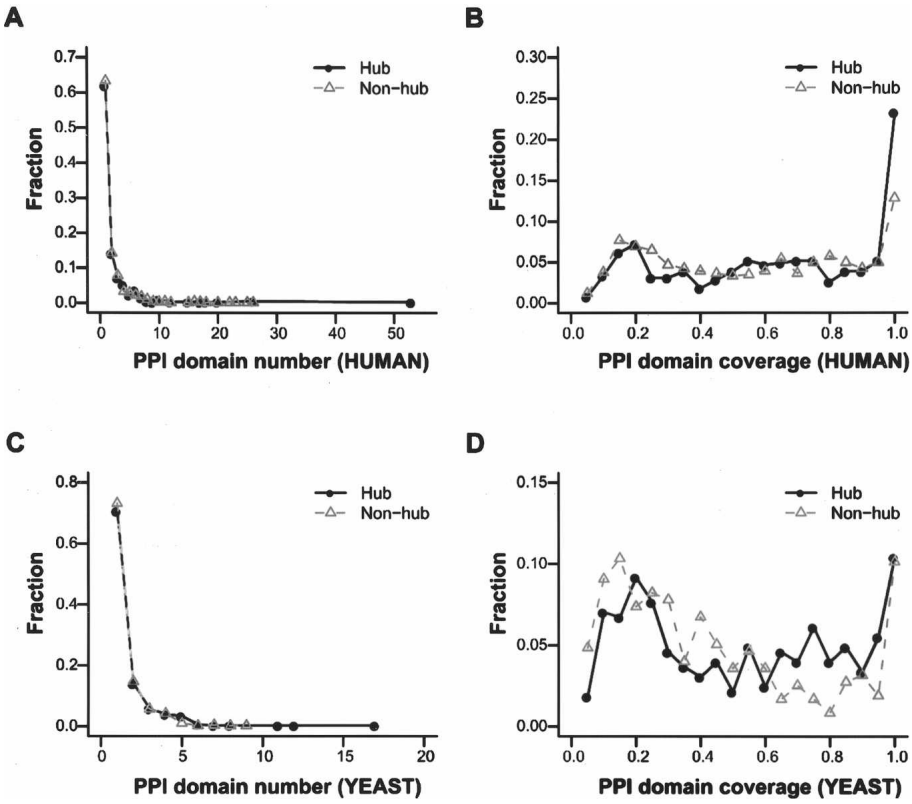

Figure 4. Distribution of PPI domain number and PPI domain coverage among human and yeast protein hubs and non-hubs. Proteins in human (*A*,*B*) and yeast (*C*,*D*) PPI networks are divided into hubs ($k \geq 5$) and non-hubs ($k < 5$). The fraction of hubs and non-hubs with a certain number of PPI domains or PPI domain coverage are plotted against the PPI domain number (*A*,*C*) and average PPI domain coverage (*B*,*D*).

ProDom, PROSITE, SMART, and TIGRFAMs as the domain architecture data sources. These databases use different methodologies to identify proteins' domain architecture, and they may give rise to different domain lengths and domain numbers at the same stretch of sequence on a protein. We used the longest functional or structural segment within an InterPro domain/family/repeat annotation to represent the boundary of the domain.

Domain coverage depends on the correctness of domain boundaries as given by InterPro, which might not be always accurate. But, there is no apparent bias toward a larger boundary for higher organisms or for higher degree proteins, as indicated by the lack of correlation of non-PPI domain length or coverage with organismal complexity or PPI degree. Comparatively, it is easier to find scenarios where the domain number bias is introduced by study and annotation bias, but again, both can be controlled by non-PPI domains.

All of the protein data were downloaded from UniProt (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/) on September 5, 2007.

HPRD data set was downloaded from www.hprd.org on March 7, 2005.

The yeast core data set (DIPcore) was downloaded from http://dip.doe-mbi.ucla.edu on June 22, 2006.

We combined the DDIB (http://www.biosino.org/DIDWeb/index.htm) PPI domains, Pawson's collection of protein interaction domains (http://pawsonlab.mshri.on.ca/), the InterPro domain description, the InterPro experimentally confirmed PPI domain annotation, and the GO annotations (ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go) for the InterPro domain, family, and repeat to arrive at the 642 protein–protein interaction domains (Supplemental Table 2).

The orthologous proteins were downloaded from Inparanoid database (http://inparanoid.sbc.su.se).

Taxonomy

InterPro has 13,383 organisms, and most of them only have a very limited number of proteins in the database. We selected the 19 eukaryotic organisms based on the following criteria: (1) nearly intact proteome; (2) model animals or organism for genetics analysis; (3) the full genome sequences are due to finish soon. *Dictyostelium discoideum* is a slime mold that grows up with an independent unicellular form; however, they aggregate together by releasing cAMP to signal each other and generate a multicellular structure in response to an unfavorable environment such as starvation (Postma et al. 2004). Hence, we chose *D. discoideum* as a transitional species between unicellular organisms and multicellular organisms. The identifier, scientific name, and common name of the 19 organisms selected are shown in Supplemental Table 1. The phylogeny is determined by NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root) and Tree of Life web project (http://tolweb.org/tree/).

Leave-one-out and boxplot analysis

Proteins in each organism were randomly divided into 10 nonoverlapping groups of an approximately equal number of proteins. The average domain number (DN), domain coverage (DC), or domain length (DL) was calculated 10 times for each organism (to test association to organismal complexity) or within each interval of DN, DC, or DL (to test association to PPI degree, whose average was also calculated), using a different nine out of the 10 groups (leaving one group out) each time. The boxplots were generated by using all of the average values in each of the 10 groups for an organism (to test association to organismal complexity) or within an interval of DN, DC, or DL (to test association to PPI degree). Same procedures were applied to both PPI domains and non-PPI domains.

d_N/d_S calculation

d_N/d_S was calculated by the YN00 program (Yang and Nielsen 2000) using the DNA sequences of PPI domains in orthologous pairs of proteins between human and mouse. Ortholog pairs are obtained from Inparanoid database (http://inparanoid.sbc.su.se). Only the highest ranking orthologous pair in each ortholog group was considered as orthologs, so that each ortholog protein belongs to one and only one ortholog group. For each shared PPI domain on orthologs, we first extracted the amino acid sequences of the human and mouse domains, and then aligned them by ClustalW with default parameters to match the domain

sequences. After replacing the amino acid sequences by their corresponding DNA sequences defined in Ensembl Genome Database (www.ensembl.org), the human and mouse DNA sequences of a domain on a pair of orthologs between the two organisms were used as input to the YN00 program with default parameters.

PCC, RCC, and regression calculation

Pearson correlation coefficients (PCC), Spearman rank correlation coefficients (RCC), linear regressions, or boxplots were calculated or plotted using R (http://www.r-project.org/). PCC and RCC both measure the degree of association between two variables and are defined as $PCC = (\sum_{i=1}^{n}(Xi - \overline{X})(Yi - \overline{Y})/(n - 1)S_xS_y)$, where S_x and S_y are the standard deviations of variables X and Y, respectively, n is the length of the vector, and $RCC = 1 - 6\sum(d^2/n(n^2 - 1))$, where d^2 is the difference in statistical rank of corresponding variables and is an approximation to the exact PCC. Their values range between -1 and 1, with 1 signifying perfect correlation, –1 perfect anticorrelation, and 0 no association. For linear regression, $y = ax + b$, P-values for slope measures the significance of the association between two variables, R^2 for the regression tests the goodness of fit of the data to the regression model.

GO enrichment test

Enrichment was determined by Fisher exact test, followed by Benjamini-Hochberg correction as described previously (Xia et al. 2006) for multiple hypothesis testing on all the annotated InterPro ID in the InterPro ID to GO term mappings.

Acknowledgments

We thank Tony Pawson (Mount Sinai Hospital) for a discussion that inspired this project. We thank Nicholas Baker (Albert Einstein College of Medicine) for editing the manuscript, Nansheng Jack Chen (Simon Fraser University) for critical reading of the manuscript, Ziheng Yang (University College London) and Rasmus Nielson (UC Berkeley) for helpful suggestions on evolutionary rate analysis, and Liang Du and Jine Li for domain annotations. We also thank the anonymous reviewers for their valuable suggestions. This work was supported by grants from the China National Science Foundation (grant nos. 30588001 and 30620120433), National Basic Research Program of China (2006CB910700), and funds from the Chinese Academy of Sciences to J.-D.J.H.

References

David, L. and Nelson, M.M.C. 2005. The three-dimensional structure of proteins. In *Lehninger principles of biochemistry,* 4th ed., pp. 116–156. W.H. Freeman & Co, New York.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272:** 5129–5148.

Futuyma, D.J. 2005. *Evolution*, pp. 458–459. Sinauer Associates, Sunderland, MA.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17:** 669–681.

Graveley, B.R. 2001. Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet.* **17:** 100–107.

Kirschner, M. and Gerhart, J. 1998. Evolvability. *Proc. Natl. Acad. Sci.* **95:** 8420–8427.

Koonin, E.V. and Galperin, M.Y. 2003. *Sequence–evolution–function: Computational approaches in comparative genomics*. Kluwer Academic, Boston, MA.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lopez, A.J. 1998. Alternative splicing of pre-mRNA: Developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* **32:** 279–305.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. 2005. InterPro, progress and status in 2005. *Nucleic Acids Res.* **33:** D201–D205.

Papin, J.A., Hunter, T., Palsson, B.O., and Subramaniam, S. 2005. Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* **6:** 99–111.

Pawson, T. and Nash, P. 2003. Assembly of cell regulatory systems through protein interaction domains. *Science* **300:** 445–452.

Pennisi, E. 2005. Why do humans have so few genes? *Science* **309:** 80.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13:** 2363–2371.

Postma, M., Bosgraaf, L., Loovers, H.M., and Van Haastert, P.J. 2004. Chemotaxis: Signalling modules join hands at front and tail. *EMBO Rep.* **5:** 35–40.

Prasanth, K.V. and Spector, D.L. 2007. Eukaryotic regulatory RNAs: An answer to the "genome complexity" conundrum. *Genes & Dev.* **21:** 11–42.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.

Smith, C.W. and Valcarcel, J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends Biochem. Sci.* **25:** 381–388.

Vogel, C. and Chothia, C. 2006. Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2:** e48. doi: 10.1371/journal.pcbi.0020048.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., and Eisenberg, D. 2000. DIP: The database of interacting proteins. *Nucleic Acids Res.* **28:** 289–291.

Xia, K., Xue, H., Dong, D., Zhu, S., Wang, J., Zhang, Q., Hou, L., Chen, H., Tao, R., Huang, Z., et al. 2006. Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. *PLoS Comput. Biol.* **2:** e145. doi: 10.1371/journal.pcbi.0020145.

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

Received June 26, 2007; accepted in revised form May 16, 2008.