Organization and function acquisition in protein-protein interaction networks

Jesse Farnham

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Adviser: Mona Singh

November 2013

Abstract

Protein-protein interaction (PPI) networks enable the transmission of biological information throughout cells, allowing cells to respond to environmental stimuli. PPI networks can be represented as graphs, and graph analysis techniques have been applied in order to determine the topological roles played by individual proteins in PPI network structure. However, more complex analysis is needed to study the *functional* organization of PPI networks. In addition, the proteins that make up PPI networks change and evolve new functions over time.

In the first part of this thesis, we introduce a metric, *functional insularity*, to measure the degree to which proteins physically interact with functionally related proteins. Proteins in PPI networks exhibit significant variation in insularity values, suggesting the presence of a tradeoff between network modularity and connectivity. Low-insularity proteins—those that interact with many functionally unrelated proteins—are more crucial than high-insularity proteins to maintaining network connectivity, are less likely to be essential, and have more regulators. Furthermore, we show that between-species homologs tend to have similar levels of functional insularity. Low-insularity proteins are found between topological network modules as well as within them. We find that functional and topological network modules contain proteins with a range of insularity values, including low-insularity proteins that might may function as "interfaces" to other modules. Finally, we show how functional insularity analysis can be applied to improve network clustering analyses.

In the second part of this thesis, we study the acquisition of new functions by proteins and their integration into the PPI network. We first use a maximum parsimony-based approach to infer the ages of human proteins. We then determine various function-related traits for each age group, such as protein-protein interaction count, expression ubiquity, and number of unique domains. We find that young proteins in human have fewer protein-protein interactions, have fewer unique domains, are expressed in fewer tissues, and are less likely to be essential than older proteins. In addition, we find that proteins tend to physically interact mainly with other proteins of similar age. Finally, we find that younger pairs of paralogs are more coexpressed and share more common regulators than older pairs.

In sum, this thesis advances our understanding of PPI networks by showing that the dual requirements of modularity and connectivity are balanced using “connector” proteins and “module” proteins, which have distinct biological traits, and by uncovering differences between young and old proteins that suggest that proteins gain functions and integrate into networks over time.

Acknowledgments

First and foremost, I'd like to thank my advisor, Prof. Mona Singh, for an extraordinary level of support and commitment in helping me to complete my Ph.D. program. Without her commitment in meeting with me consistently every week to discuss problems and next steps, and in repeatedly helping me to revise my paper drafts, I would not have been able to complete my program. I would also like to express my appreciation for her support as my interests and career plans changed over the course of my grad school career. Whether I planned to stay in academia, go into government research, or go into industry, she was always willing to provide advice for how best to achieve my goals. I especially appreciate her willingness to let me spend a summer interning in the software industry, even though it slowed my research progress.

Secondly, I would like to thank my undergraduate advisor, Prof. Eric Aaron, for his continued support, mentoring, and friendship throughout my graduate school career. His advice on research, graduate school, and career choices has been invaluable over the past five years. In addition, he provided advice and support from a source outside the "graduate school bubble," while simultaneously understanding the graduate school process in a way that only someone who has also been through the process can understand.

I would also like to thank the other members of the Princeton community who have helped out over the past five years. Thanks to all the members of the Singh lab, past and present, for answering my questions and giving me so much useful feedback on my research projects. Thanks to Peng Jiang for assistance with the LaTeXformatting of this thesis. Thanks also to Young-Suk Lee for valuable guidance on the processing of microarray data. Thanks to the faculty members who served on my general exam and thesis committees: Tom Funkhouser, Andrea LaPaugh, Olga Troyanskaya, and Vivek Pai.

I'd like to thank my all Princeton undergraduate friends (names intentionally unlisted so I don't forget anyone) for making graduate school worth all the effort and frustration, for listening to my constant complaints about research and paper-writing, for letting me pretend to be an undergrad, and most importantly, for making the past five years the best of my life so far. I will never forget all the memories of anime club, juggling club, Katsucon, Kpop music videos, long AIM/MSN conversations, group dinners, lifting, basketball, squash, Colonial parties, room parties, Dbar, Houseparties, Lawnparties, Winter Formals, Reunions, dance shows, midnight USG movies, Teriyaki Boy and movie outings, Sushi Palace, the Boston trip, and the Japan trip. I'm sure there are plenty of other epic memories that I can't remember right now.

I'd like to thank my friends from Smith College, Zin Min and Sanita, for always keeping in touch and inviting me to visit even after they were no longer students at Princeton. My visits to Smith, our experiences at Princeton, our trip to Six Flags, and our Jersey Shore beach trips were some of the highlights of my graduate school career. Thanks also to everyone else from the Smith community whom I had the pleasure of meeting; I'm proud to be an honorary Smithy.

Thanks to the members of Colonial Club from 2010 through 2013, for always making me feel welcome there, and even giving me the opportunity to become an official member of the club, despite this not usually being an option for grad students. Thanks also to Margaret Lo and the members of her karate class, for being the first to welcome me to Princeton.

I'd also like to thank all the non-University locals whom I've met over the past five years, for reminding me that there is life outside the Orange Bubble. Thanks especially to the community of Air Mods Flight and Service Center at Trenton-Robbinsville Airport for giving me the ability to pursue my flying dreams while at Princeton.

Very special thanks to Katrina for all the great memories from the past two years, and for always being supportive and sympathetic as I tried to figure out my next steps after graduate school–all while she was doing the same with her own graduate program. Spending time in New York or Princeton was a great way to end the week, and our trips to Washington DC, Lake Tahoe, and the Poconos were some of my favorite grad school experiences.

Finally, I'd like to thank my family for their support throughout grad school, and especially for always welcoming me back to my home in Pomfret, CT when I needed to take a break from the grad student life. Those visits are the best way I know of to recover from too much programming or paper-writing.

My PhD study was supported by Princeton University, NSF grant ABI-0850063, NIH/NIGMS R01 grant GM076275, the Quantitative and Computational Biology Program NIH grant T32 HG003284, and the NIH Center of Excellence grant P50 GM071508.

In memory of Ishboo.

“Dogs are not our whole life,
but they make our lives whole.”
—Roger Caras

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Proteins are large molecules that perform most of the functions necessary to life. They are responsible for molecular transport, muscle contraction, and catalysis of most chemical reactions necessary for life. Proteins are synthesized by a process known as the “central dogma of biology.” First, subsequences of DNA known as genes are transcribed to messenger RNA, which is then translated to proteins. Thus, all of the information necessary to encode every protein in an organism is contained in that organism’s DNA.

1.1 Proteins and protein-protein interaction networks

Proteins rarely perform their functions alone; they often cooperate with other proteins that perform the same or similar functions. This cooperation is often accomplished by way of physical interactions between specific proteins, in which two or more proteins make physical contact while performing their functions [1]. Physical interactions may occur in a binary fashion between two specific proteins [2, 3], or groups of proteins may interact to form large *protein complexes.* Examples of such complexes are the

Species	Number of proteins	Number of interactions
C. elegans	3,207	5,656
D. melanogaster	8,058	36,420
H. sapiens	17,808	138,087
S. cerevisiae	6,354	81,839

Table 1.1: Sizes of the largest PPI networks from BioGRID [6] as of August 2, 2013.

ATP synthesis complex, which produces the energy storage molecule ATP, and the ribosome, which translates RNA into proteins.

Over the past decade, biologists have developed high-throughput methods, including the yeast two-hybrid method [4] and affinity capture [5], to identify protein-protein physical interactions on a large scale. As a result, tens of thousands of specific protein-protein interactions have been identified for certain species. This data can be represented as a graph, in which nodes represent proteins and edges represent physical interactions between two individual proteins. As shown by Table 1.1, these graphs are very large, making their structure difficult to analyze.

Protein-protein interaction (PPI) networks are of interest to biologists because they represent the means by which information may be transmitted throughout the cell. Signaling pathways such as the one shown in Figure 1.1 are specific examples of information transmission. A protein on the cell surface might bind a specific molecule, which then triggers a cascade of protein-protein interactions inside the cell, ultimately resulting in a change in gene expression or cell behavior. While specific signaling pathways such as the one shown in Figure 1.1 represent relatively small sets of protein-protein interactions that are understood by biologists, the global structure of PPI networks remains an area of active study.

Graph analysis techniques have been used to study the global structure of PPI networks. Early work discovered that PPI networks are “scale-free,” meaning that there are a large number of low-degree nodes, with a much smaller number of high-degree nodes [8]. Additional work found that PPI networks are *modular*, meaning that

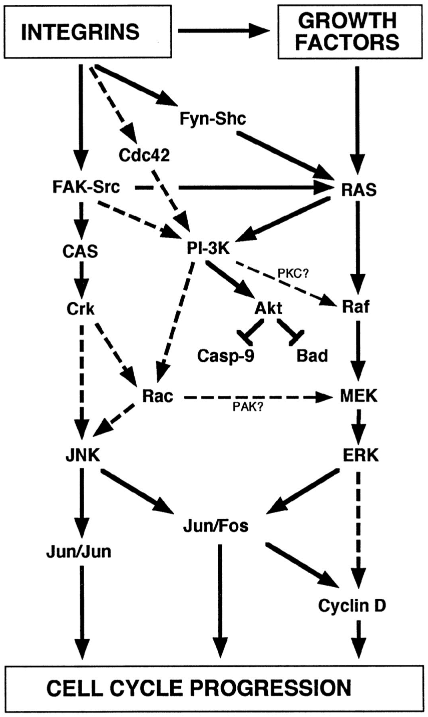


Figure 1.1: An example of a signaling pathway. Figure taken from [7].

they contain locally dense subgraphs [9]; these *topological modules* often correspond to *functional modules*, or groups of proteins that perform the same or similar functions in the organism. Thus, PPI network structure shows two correlated types of modularity: topological modularity and functional modularity.

Previous work has also found that different proteins play different roles in maintaining the overall structure of the PPI network. Centrality measures are metrics for measuring the importance of a specific protein to the global network structure. Two such measures are degree, a local centrality measure, and betweenness centrality [10], a global measure. Betweenness centrality of a node is the fraction of all shortest paths in a network that pass through the node. Some proteins participate in many more interactions than others (i.e., they have high degree); these proteins are more likely to be essential to the organism's survival than proteins that participate in fewer interactions [11]. Similarly, proteins with high betweenness centrality are also more likely to be essential than other proteins [12].

The work cited above uses graph theoretic measures such as degree and betweenness centrality to take a topological view of network structure in order to identify differences in individual proteins' contributions to the global structure. However, PPI networks may also be viewed as functional, rather than purely topological networks. In this paper, we propose a method by which to study proteins' roles in organizing the functional as well as the topological structure of the PPI network.

1.2 Protein evolution and function acquisition

Organisms evolve, and speciation events result in the genesis of new species, over extremely long timespans. The root cause of evolution is the random mutation of DNA, which can result in altered protein sequences, structures, and functions. Some of these mutations are deleterious to the survival of the organism, some have no effect, and others are beneficial.

Due to the extremely long timespans (millions of years) over which evolution occurs, it is rarely possible to directly observe the process of protein evolution. Nonetheless, researchers can reconstruct the evolutionary tree of life by observing similarities in the DNA sequences of present-day organisms (Figure 1.2). For example, if the DNA of two organisms, such as human and chimpanzee, have high sequence similarity, then researchers can infer that those organisms are closely related, that is, they have a recent common ancestor in the tree of life. Species whose genomes show less sequence similarity, such as humans and bacteria, are assumed to be less closely related; that is, they have no recent common ancestors in the tree of life. Due to tech-

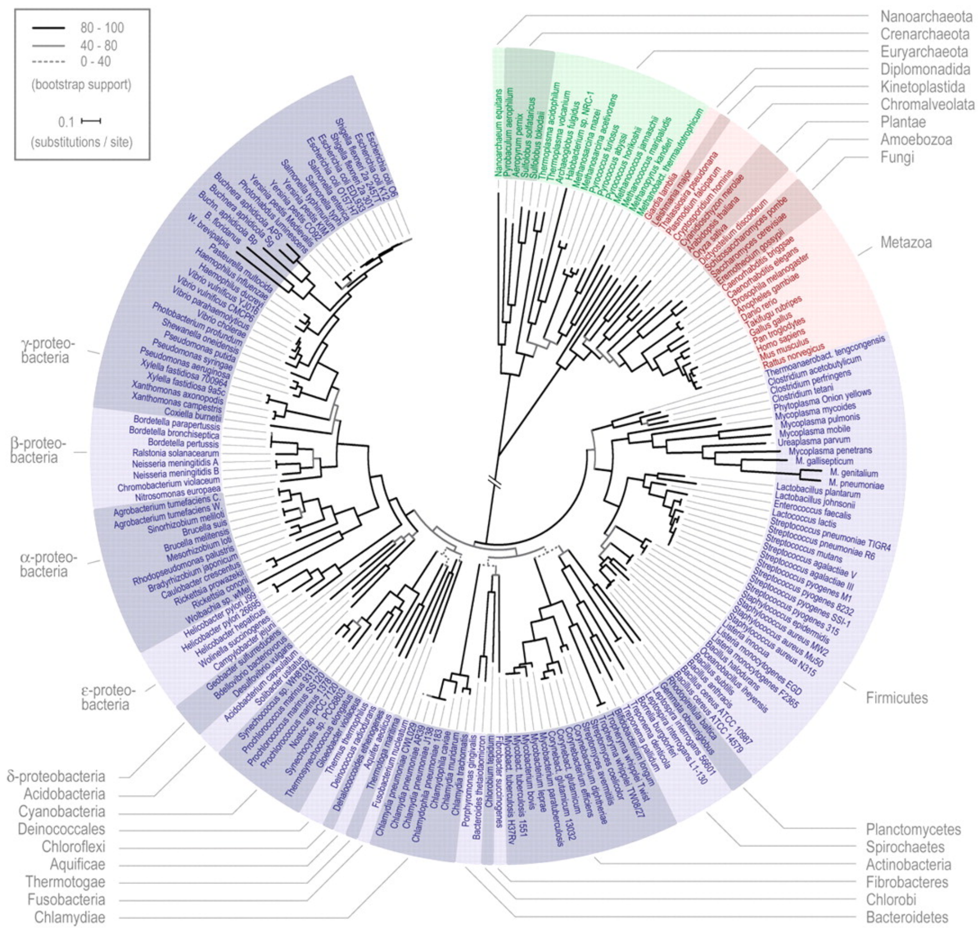


Figure 1.2: A reconstruction of the evolutionary history of 191 species. Figure taken from [13].

nological advances in DNA sequences, complete genome sequences of many species are now available, enabling this type of analysis. Due to the large amount of data, computational techniques must be used in these analyses.

Because of the central role of proteins within organisms, the mechanisms and principles by which proteins evolve and acquire new functions over time is of interest to researchers. Gene duplication events, in which one or more genes are duplicated in a species's genome, have been implicated as a major driver for the acquisition of new protein functions [14, 15, 16]. In contrast to duplication events, which create copies of existing genes, novel proteins can also be formed from formerly non-coding DNA sequence, as well as other mechanisms [17]. These novel proteins have also been shown to evolve and acquire new functions over time [18].

Protein interaction networks also evolve over time [19]. It has been shown that proteins with many interactions evolve more slowly than proteins with few interactions and that interacting proteins evolve at similar rates [20]. In this thesis, we present an analysis of protein function acquisition over time, including the integration of new proteins into existing PPI networks; our findings suggest that new functional modules are added to PPI networks over time.

1.3 Our contributions

In this thesis, we provide two main contributions that advance our current understanding of protein-protein network structure, function acquisition by proteins, and integration of new proteins into existing PPI networks. First, we provide a metric, called *functional insularity*, that measures the degree to which a protein physically interacts with functionally related vs. functionally unrelated proteins. We apply this measure to a large set of proteins in yeast and human PPI networks and draw several conclusions about the roles played by different proteins in organizing the functional network of the cell. We show that proteins with high functional insularity scores have different biological properties from proteins with low scores; furthermore, we show that functional insularity is conserved between evolutionarily related proteins in yeast and human.

Second, we take a dynamic view of protein function and PPI network structure over time. We use a sequence-based evolutionary analysis to infer an approximate time of origin for every human protein. We then identify differences between young and old proteins in several function-related biological features. Our findings suggest several ways by which proteins may gain functions over time. For example, we find that younger proteins tend to have lower degree than older proteins, suggesting that proteins may gain physical interactions and integrate into the existing protein-protein interaction network over time.

Chapter 2

Functional analysis of the modularity-connectivity tradeoff in protein interaction networks

2.1 Introduction

Over the past decade, large-scale protein-protein interaction (PPI) networks have been determined for a diverse set of organisms. These networks, in which nodes represent proteins and edges represent physical interactions between proteins, provide a global view of the relationships among the components and processes that enable organisms to function.

The topological organization of PPI networks has been extensively analyzed in order to improve our understanding of cellular functioning [1]. Topological analysis has proven to be a powerful technique, because it enables the use of well-known algorithmic concepts and topological measures that can be applied to networks in general. Topological techniques have been used to provide insights into protein essentiality [11, 21, 12] and pleiotropy [22], and to understand organizational patterns such as network motifs [23], information flow [22], and network bottlenecks [12].

Among the key findings of topological network analysis is that cellular networks are *modular*. Modularity [24, 25] is an organizational principle in which large systems are composed of smaller, relatively self-contained subcomponents. In the context of PPI networks, modularity typically refers to the presence of preferentially interacting sets of proteins (*topological* modularity) or to groups of proteins involved in the same biological process (*functional* modularity); in practice, there is overlap between modules defined either topologically or functionally. Previous work has identified both types of modules in PPI networks [26, 27].

Despite the presence of modularity in PPI networks, there are many crosstalk interactions between proteins in different modules [28, 29], and modules are therefore not completely separate entities. Whereas modularity is helpful in order for biological processes to function efficiently without interference, connectivity is required for interprocess cooperation. We set out to analyze the balance between modularity and connectivity in protein-protein interaction networks.

It is clear from analysis of topological network structure that proteins play different roles in balancing the modularity-connectivity tradeoff in a *topological*, as opposed to *functional*, sense; these roles can be quantified through the use of various local and global topological network measures. For example, proteins vary widely in local clustering coefficient (LCC), a local network centrality metric that measures the connectivity among a protein's neighbors [30]. Proteins also vary in betweenness centrality [10, 12], a global network centrality measure that indicates the importance of the protein in maintaining shortest paths between other network proteins. Low-LCC and high-betweenness proteins can be viewed as supporting topological connectivity, while high-LCC and low-betweenness proteins can be viewed as supporting topological modularity, thereby providing further evidence of the connectivity/modularity tradeoff.

While general-purpose graph analysis techniques are powerful, they do not make full use of the cellular network, because they do not explicitly consider known functional information. Instead, topological analysis uses network topology as a proxy to study the key item of interest: functional organization. Here, we propose a method that uses PPI networks to study functional organization directly, rather than indirectly through general-purpose topological methods.

In this chapter, we make several contributions to advance our understanding of the functional organization of the cell. First, we provide a measure, *functional insularity*, that quantifies the degree to which proteins perform biological processes similar to their interacting partners. Proteins with low functional insularity interact with proteins of varying biological processes, whereas proteins with high functional insularity interact largely with proteins that share biological processes. Second, we compute the functional insularity of yeast and human proteins to show that proteins vary greatly in their degree of functional insularity. Further, we show that functional modules tend to contain both high-insularity and low-insularity proteins. Third, we show that low-insularity proteins are more important to maintaining network connectivity as defined by both betweenness centrality [10, 31] and two clustering-based topological measures. Fourth, we show that low-insularity proteins have more regulators and are less likely to be essential than high-insularity proteins. Finally, we give evidence that functional insularity is evolutionarily conserved between proteins in yeast and human.

2.2 Methods

2.2.1 Data Sources

PPI networks. We performed our analysis on three yeast and three human networks. For our primary analysis, we focus on version 3.2.95 of the BioGRID [6] PPI network for yeast; for this network we used all evidence types indicating the presence of a physical protein-protein interaction. We also used BioGRID version 3.2.95 as our primary human network. In addition, we repeated the analysis on the high-throughput co-complex and high-throughput binary networks from [32] for both yeast and human.

Networks were preprocessed by removing the 1% of proteins with highest degree in order to eliminate potentially "sticky" proteins that appear to have many interactions, but that may be due to experimental artifacts. This resulted in the removal of 52 proteins from the BioGRID yeast network with degrees ranging from 148 to 1,912, producing a final network with 4,825 nodes and 40,371 edges. For the human BioGRID network, 133 proteins were removed with degrees ranging from 108 to 8,959, producing a final network with 9,694 proteins and 39,859 interactions.

All single-network results in the main body of this chapter are reported for the yeast BioGRID network; for alternate network results, see the Appendix of this thesis.

Functional annotations. The June 2013 version of the GO term hierarchy and protein annotations were used [33]. Annotations with only the evidence code of "inferred by electronic annotation" (IEA) and "reviewed computational analysis" (RCA) were ignored. In order to avoid circularity in the definition of functional insularity, we also ignored annotations with only the evidence code of "inferred by physical interaction" (IPI). In order to remove possible confounding effects from very specific or broad GO terms, we did not consider terms that annotate fewer then 4 or greater than 1,000 yeast proteins, leaving 2,129 Biological Process terms for yeast.

For human, we ignored terms that annotate fewer than 4 or greater than 5,000 human proteins, leaving 5,204 Biological Process terms.

Other data Yeast essentiality data were downloaded from the SGD [34]. Yeast regulatory data were obtained from the YEASTRACT regulatory network dataset [35, 36].

Gene homology data were downloaded from the Princeton Protein Orthology Database (P-POD) [37]. Three datasets are available, each generated using a different algorithm: OrthoMCL [38], MultiParanoid [39], and Jaccard clustering. The results described in this chapter were obtained using the OrthoMCL dataset; running the same analysis on the other two datasets produced similar results (see Section A.2).

2.2.2 Computing the Functional Insularity Measure of Proteins

We define a measure called *functional insularity*, which quantifies the degree to which proteins interact with functionally related vs. functionally unrelated neighbors. Let $|R|$ be the number of proteins that are annotated with at least one GO Biological Process term under consideration. For a given term t, let $R(t)$ be the set of proteins annotated by term t and $|R(t)|$ be the number of such proteins. For a given protein p, let $A(p)$ be the set of terms under consideration annotating protein p, let $\mathcal{N}(p)$ be the set of annotated proteins that interact with protein p, and let $|\mathcal{N}(p)|$ be the number of annotated proteins that interact with p. First, functional similarity $f(p, q)$ was computed for each pair of proteins (p, q) using an approach similar to that of [9] and described below.

Then, the functional insularity $i(p)$ of a protein p is defined as the mean functional similarity of p vs. each of its neighbors:

$$i(p) = \frac{\sum_{q \in \mathcal{N}(p)} f(p, q)}{|\mathcal{N}(p)|} \tag{2.1}$$

The functional similarity of a pair of proteins is defined as the Jaccard coefficient of the terms annotating each protein, weighted by the information score of each term:

$$f(p,q) = \frac{\sum_{t \in A(p) \cap A(q)} s(t)}{\sum_{t \in A(p) \cup A(q)} s(t)} \tag{2.2}$$

where $s(t) = -\log \frac{|R(t)|}{|R|}$ is the information content for each term under consideration, defined as in [40]; it scores the specificity of each term, such that terms that annotate few proteins are scored higher than terms that annotate more proteins.

Functional insularity scores were only computed for proteins with degree at least 3, and the analyses described in the rest of this chapter were run only on these proteins. Out of the scored proteins, we classified the top third as "high-insularity" and the lower third as "low-insularity."

2.2.3 Topological Protein Roles

We determined the overlap between high-insularity and low-insularity proteins with sets of proteins that serve as "module connectors" in the following ways. First, we ran the graph clustering algorithm SPICi [41] in order to obtain a topological modular representation of the network. SPICi has three parameters: the minimum cluster size, the density threshold, and the support threshold. We fixed the minimum cluster size at 3, then systematically varied the support and density thresholds between 0 and 1 in increments of 0.1 and selected the combination that produced the clustering with maximal modularity by Newman's method [42]. We also repeated the analysis using Stijn van Dongen's MCL clustering algorithm [43, 44], available from `http://micans.org/mcl/`. Default parameters were used for MCL, and clusters with size less than 3 were considered to be unclustered proteins.

Using the clustered network, all proteins were categorized as "module connectors" if they had one of the two topological features:

(A) **"Between-cluster connectors."** Unclustered proteins with degree at least 3 that connect to proteins in at least two distinct clusters.

(B) **"Within-cluster connectors."** Clustered proteins, at least half of whose neighbors are outside the protein's own cluster.

The hypergeometric test was used to compute the overlap between high-insularity or low-insularity proteins and module connectors to determine the relative roles of high-insularity and low-insularity proteins in maintaining network connectivity.

Additionally, we used protein betweenness centrality and local clustering coefficient to evaluate the topological roles of high-insularity and low-insularity proteins. Betweenness centrality is the fraction of all shortest paths between a pair of nodes that pass through the node of interest, summed over all pairs of nodes; betweenness centrality values are normalized to a range of 0–1. Local clustering coefficient is the number of edges between a protein's neighbors, compared to the number of such edges that could possibly exist. Betweenness centrality and local clustering coefficient were computed using the NetworkX graph analysis package [45].

2.2.4 Functional insularity values of homologs in yeast and human

We ran the following analysis to investigate the evolutionary conservation of functional insularity. We built a network in which nodes represent human or yeast proteins, and edges represent inter-species homology relationships. We then categorized the nodes into four groups: human low-insularity, human high-insularity, yeast low-insularity, and yeast high-insularity. Functional insularity is considered conserved if there are significantly fewer edges from high-insularity to low-insularity proteins than expected by chance. To compute P-values for conservation of insularity, the networks were randomized 1,000 times using the stub-rewiring approach from [23], and the number of edges from high-insularity to low-insularity proteins was counted for each randomization and compared to the number of such edges in the original graph. In stub-rewiring, the edges of the graph are randomized in a way that preserves the degree of all nodes; it therefore allows network-based results to be compared to null distributions generated on random networks, while controlling for the possible effects of high-degree nodes on the results.

2.2.5 Identification of enriched GO terms in high-insularity and low-insularity proteins

We used an approach similar to that of GO Termfinder [46] in order to identify GO terms that are enriched in the sets of high-insularity or low-insularity proteins. For each term, we obtained a hypergeometric P-value of enrichments and applied a false discovery rate correction by running 100 trials in which the sets of high-insularity and low-insularity proteins were randomized. Terms that annotate at least 5% of the high-insularity or low-insularity proteins with an FDR of less than 0.05 were considered to be enriched.

2.2.6 Data visualization

Q-Q plots, histograms, Figure 2.8, and the similar figures in the Appendix of this thesis were generated using the Python graphing library Matplotlib [47]. Q-Q plots are used to visualize differences in high-insularity vs. low-insularity proteins relative to a certain measure, e.g., betweenness centrality. They are obtained by plotting the quantile of one population against the same quantile of the other; if the two populations have unequal sizes, linear interpolation is used to infer matching percentiles between the populations. In the Q-Q plots shown in this chapter, the x-coordinate of a given point represents the value of a given property (e.g., betweenness centrality) in

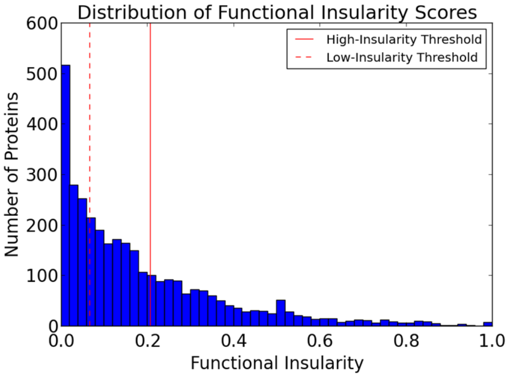


Figure 2.1: Distribution of functional insularity scores of all scored proteins in the yeast PPI network. Low scores indicate low-insularity proteins that connect primarily to unrelated proteins; high scores indicate high-insularity proteins that connect primarily to related proteins. Proteins with insularity scores to the left of the dashed line are classified as "low-insularity;" proteins with insularity scores to the right of the solid line are classified as "high-insularity."

the set of low-insularity proteins at a given percentile. The y-coordinate of the same point represents the value of the property in the set of high-insularity proteins at the same percentile. Thus, if low-insularity and high-insularity proteins tend to have similar values of the property, the plotted points will lie on the diagonal. If low-insularity proteins tend to have higher (respectively, lower) values than high-insularity proteins, the points will lie below (respectively, above) the diagonal.

2.3 Results and Discussion

2.3.1 Number of scored proteins

The distribution of functional insularity values is shown in Figure 2.1; see Figure A.1 for the equivalent histogram in human. Analysis of the yeast network resulted in 1,119 low-insularity and 1,119 high-insularity proteins, with a total of 3,359 scored

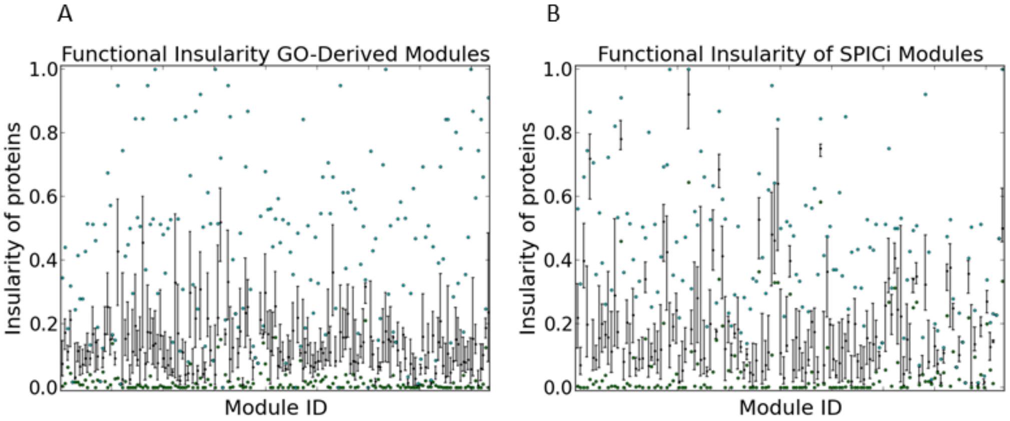


Figure 2.2: Distribution of functional insularity scores by functional module **(A)** and SPICi cluster **(B)**. Functional modules are obtained from [48]. Teal and green markers indicate the maximum and minimum functional insularity value present in the module. Black markers and bars indicate the median and interquartile range of the functional insularity values in the module. There is much variation in the functional insularities of proteins within a given module, and many modules contain proteins with functional insularity values significantly above and below those of the rest of the module.

proteins. In human, 4,809 proteins were scored, for a total of 1,603 low-insularity and 1,603 high-insularity proteins.

2.3.2 Functional modules contain a large range of insularity values

We used our functional insularity measure to investigate the functional organization of topological and functional modules. Here, functional modules are defined as protein sets annotated with a particular GO term from the list of functionally relevant GO terms presented in [48]. Topological modules are defined as clusters obtained from a network clustering using SPICi with optimized parameters, as discussed in Section 2.2.3. We reasoned that there are at least two possible hypotheses regarding the connectivity of these modules to the overall network. In one model, intermodule interactions tend to be diffuse, such that all proteins in a module participate equally in communicating with extramodular proteins. Alternatively, a subset of the the proteins in a given module might serve as a physical "interface" to the rest of the network, performing most of the intermodular interactions, while the other proteins in the module perform few intermodular interactions.

We used our functional insularity measure to find evidence to distinguish between these possibilities. Figure 2.2 shows the mean and standard deviation insularity of GO-derived functional modules and SPICi-derived topological modules, along with outlier proteins. Many topological and functional modules have a few proteins with insularity much lower than that of other proteins in the module. As evidenced by their low insularities, these proteins form more intermodular connections than do the other proteins in the same module, suggesting that the "interface" model of intermodular communication is prevalent in the network. This suggests that certain (low-insularity) proteins in a given module serve to support network connectivity, while other (high-insularity) proteins in the module serve to support network modularity.

2.3.3 High-insularity and low-insularity proteins play different roles in network connectivity

Proteins play different roles in maintaining the overall connectivity of the PPI network. By definition, high-insularity proteins connect mainly to functionally related proteins; therefore, one might expect them to be primarily responsible for local connectivity within complexes and modules. In contrast, low-insularity proteins might be expected to provide connectivity between separate modules. We therefore hypothesized that low-insularity proteins play a larger role in maintaining global network connectivity than do high-insularity proteins.

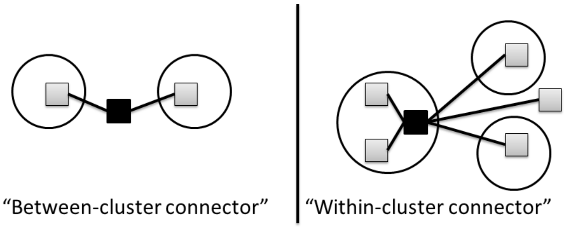


Figure 2.3: Illustration of two topological roles expected to be fulfilled by low-insularity proteins. In the "between-cluster connector" role, an unclustered protein (black) connects several clusters. In the "within-cluster connector" role, a clustered protein connects to many proteins outside of its own cluster.

Betweenness centrality and local clustering coefficient. The different topological roles of low-insularity and high-insularity proteins are supported by comparing their betweenness centrality and local clustering coefficient (LCC) (Figure 2.4). Betweenness centrality [10] is a global measure of network centrality that takes into account the entire network structure, rather than the local environment of the protein in question. As shown by Figure 2.4, the betweenness centrality of each quantile of the low-insularity proteins is higher than the equivalent quantile of the high-insularity proteins, indicating that low-insularity genes have higher betweenness centrality (mean 8.82×10^{-4}) than high-insularity genes (mean 3.79×10^{-4}). Thus, if the PPI network is seen as a large cellular communication network, then low-insularity proteins are more important than high-insularity proteins in managing information transmission throughout the network. Furthermore, low-insularity proteins do not have uniformly higher degree than high-insularity proteins, suggesting that the betweenness centrality result may not be due to any correlation between degree and betweenness centrality; in fact, the partial Spearman correlation of insularity and betweenness given degree is -0.278, indicating that degree is not a confounding factor.

We note the presence of a small set of high-degree proteins in which low-insularity members have higher degree than high-insularity members. The rightmost points in Figure 2.4A fall below the diagonal, indicating that these for these quantiles, about 5% of the proteins in the analysis, low-insularity proteins have higher degree than high-insularity proteins. This is in contrast to the finding that low-insularity proteins have lower degree than high-insularity proteins overall. In order to confirm that these proteins were not biasing the above results, we removed them from the betweenness centrality analysis. The modified set of low-insularity proteins still had significantly higher betweenness centrality than the modified set of high-insularity proteins ($p = 8.53 \times 10^{-4}$), supporting the robustness of the above result.

In addition, each quantile of low-insularity proteins has lower local clustering coefficient (LCC) than the equivalent quantile of high-insularity proteins (Figure 2.4C), indicating that low-insularity proteins have lower LCC than (mean 0.095) than high-insularity proteins (mean 0.370). This suggests that low-insularity proteins connect to multiple topological modules, while high-insularity proteins connect mainly to proteins in a single module. We found similar results after applying the functional insularity analysis to a human network (see Section A.1.2).

Effect of targeted node removal on network structure. If low-insularity proteins connect multiple modules, then one would expect targeted removal of those proteins to have a strong negative effect on network connectivity. In contrast, one might expect targeted removal of high-insularity proteins to have a smaller effect on network connectivity. We tested this by selectively removing proteins in increasing order of functional insularity (i.e., we removed the most low-insularity proteins first) and measured the effect on the size of the network's largest connected component, a measure of global network connectivity. We compared this to the effects of removing proteins in decreasing order of functional insularity (i.e., the most high-insularity proteins first), and to the effects of removing scored proteins in random order. Figure 2.5A shows that random removals have the greatest effect on network structure initially; however, as larger numbers of proteins are removed, targeted removal of the most low-insularity proteins caused a larger decrease in network average clustering

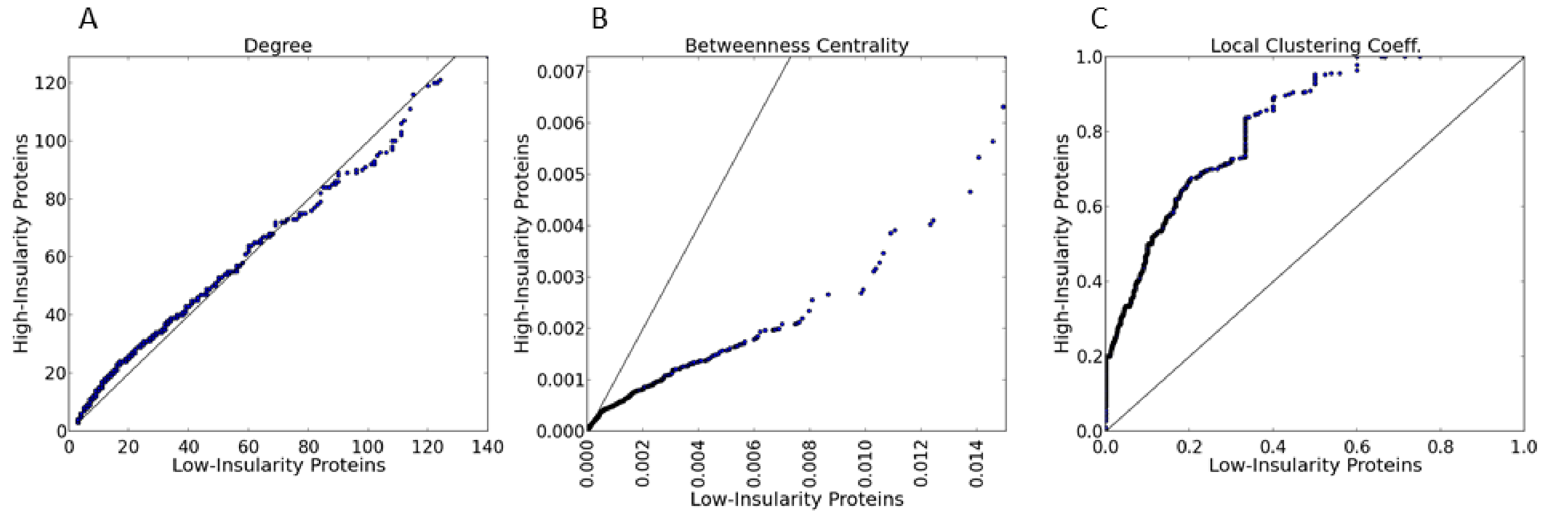


Figure 2.4: Q-Q plots comparing **(A)** degree, **(B)** betweenness centrality, and **(C)** local clustering coefficient of low-insularity and high-insularity genes. Equivalent quantiles of low-insularity and high-insularity genes are plotted; values consistently above or below the diagonal indicate differences between the high-insularity and low-insularity protein sets. While high-insularity proteins have higher degree than low-insularity proteins overall ($p = 6.62 \times 10^{-11}$, Wilcoxon rank sum test), the Q-Q plot shows that this is not uniformly true, as the highest percentiles of low-insularity proteins have higher degree than the highest percentiles of high-insularity proteins. Low-insularity proteins have higher betweenness centrality ($p = 5.44 \times 10^{-4}$) and lower local clustering coefficient ($p = 3.47 \times 10^{-204}$) than high-insularity proteins.

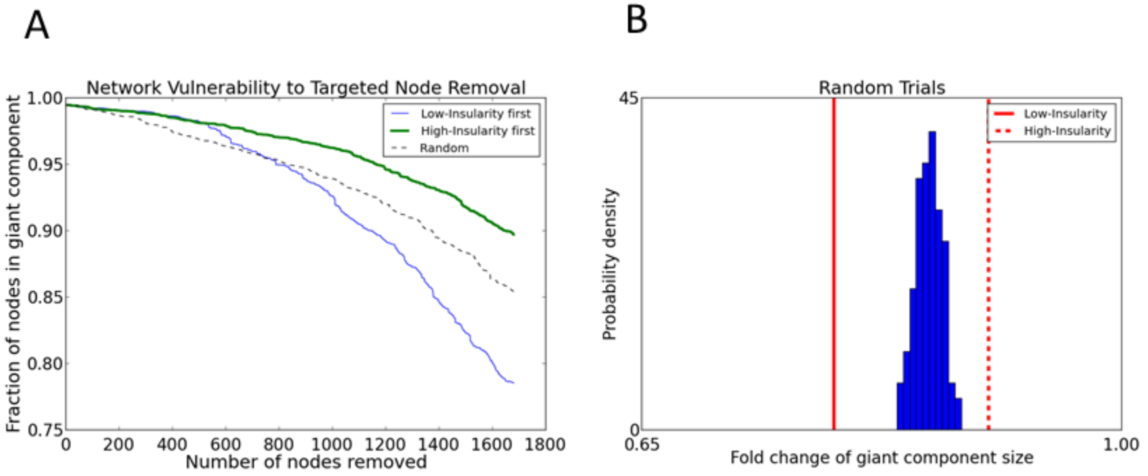


Figure 2.5: **(A)** Effect on largest connected component size of removing genes in order of increasing functional insularity (low-insularity first), decreasing functional insularity (high-insularity first), and random order of scored proteins. Removing a sufficiently large number of the most low-insularity proteins has a larger effect on network connectivity than removing an equal number of the most high-insularity proteins. **(B)** Effect of removing the top 50% of low-insularity and high-insularity proteins compared to a series of trials in which an equal number of random scored proteins are removed, indicating that low-insularity proteins are more critical to maintaining network connectivity than high-insularity proteins.

coefficient, as expected. As an additional test, we compared the effect of removing the 50% of scored proteins with lowest functional insularity, and the 50% of proteins with highest functional insularity, each at once, to a series of random trials in which an equal number of randomly chosen proteins were removed; only proteins that could be given an insularity score were chosen for random removal. Whereas removal of low-insularity proteins had a larger effect than the random trials, the effect of removing high-insularity proteins was less than the effect of random removals (Figure 2.5B). This confirms our hypothesis that low-insularity proteins are more important than high-insularity proteins in maintaining network connectivity. We found similar results on a human network (Section A.1.4).

Topological roles of low-insularity proteins. The difference in local clustering coefficient between low-insularity and high-insularity proteins suggests that low-insularity proteins might connect to multiple topological modules, while high-insularity proteins connect mainly to proteins in a single module. In order to test this hypothesis more directly, we defined two local topological roles that low-insularity proteins might be expected to fulfill in the PPI network: "within-cluster connectors" and "between-cluster connectors" (Figure 2.3, Section 2.2.3). These definitions are purely topological in nature and do not consider protein function, in contrast to our definition of functional insularity.

We find that when proteins in yeast fulfill the two topological roles, there is significant overlap with low-insularity proteins and significant lack of overlap with high-insularity proteins (Figure 2.6). Furthermore, the topological roles cover a large fraction, 80%, of the low-insularity proteins. Thus, proteins with low functional insularity tend to connect multiple topological modules, whereas high-insularity proteins are less likely to perform this role and more often connect proteins within the same topological module. We found similar results using a human network (Section A.1.3).

We note the existence of a small number of high-insularity proteins that fulfill one of the topological roles. While these proteins connect separate topological modules as determined by network clustering, their high functional insularity indicates that they share biological functions with their interaction partners. Thus, these proteins are "module connectors" in a topological, but not functional, sense. This suggests that functional insularity is more accurate than purely topological methods for identifying proteins that connect distinct functional modules.

We repeated the analysis using an alternate clustering algorithm [43, 44] to define the topological roles and found similar results (Section A.3).

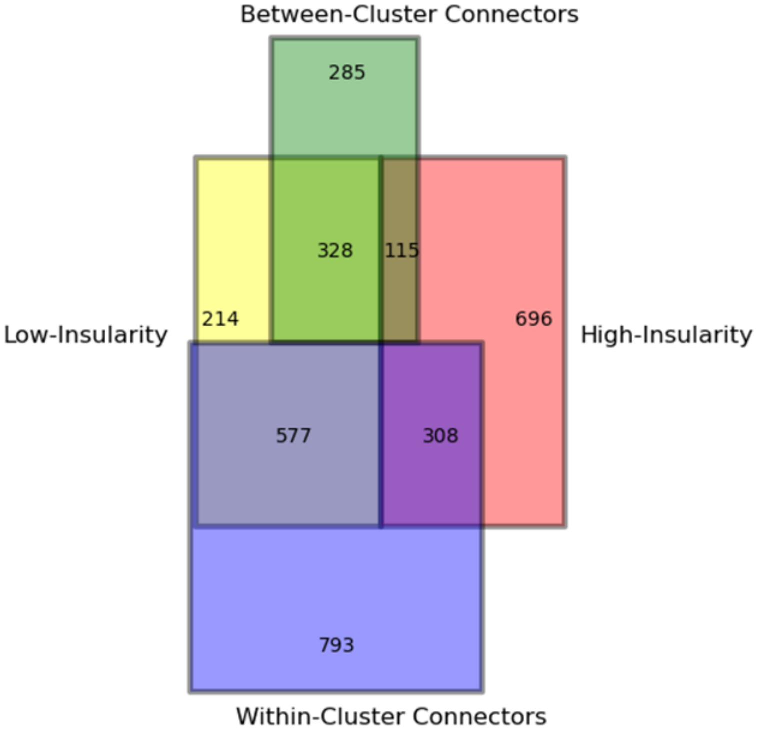


Figure 2.6: Overlap among the topological roles from Figure 2.3, high-insularity proteins, and low-insularity proteins. Between-cluster connectors are significantly enriched in low-insularity genes ($p = 5.83 \times 10^{-22}$, hypergeometric) and has significant lack of enrichment with high-insularity genes ($p = 9.97 \times 10^{-23}$). Within-cluster connectors are also significantly enriched with low-insularity genes ($p = 9.70 \times 10^{-8}$) and has significant lack of enrichment with high-insularity genes ($p = 2.00 \times 10^{-46}$).

2.3.4 High-insularity and low-insularity proteins have different biological traits

Using the YEASTRACT regulatory network dataset [35, 36], we found that low-insularity proteins have more regulators than equivalent quantiles of high-insularity proteins, indicating that low-insularity proteins tend to have more regulators (mean 8.85) than high-insularity proteins (mean 6.56). This suggests that low-insularity proteins, which tend to connect to multiple modules as shown above, are under the control of multiple modules' regulators.

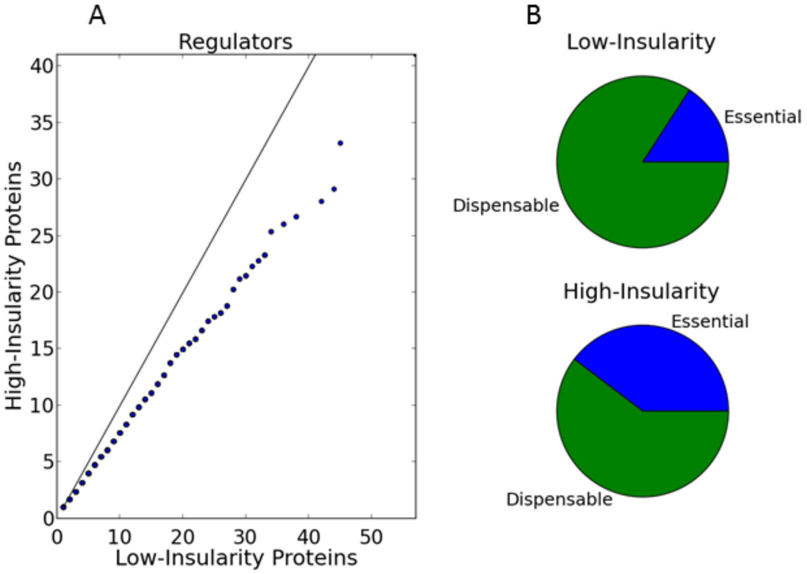


Figure 2.7: Number of regulators and essentiality for high-insularity and low-insularity genes. Low-insularity genes have more regulators ($p = 2.05 \times 10^{-15}$, Wilcoxon rank sum test) and are less likely to be essential than low-insularity genes ($p = 2.48 \times 10^{-37}$, Fisher's exact test).

High-insularity proteins are significantly more likely to be essential (39.7% essential) in yeast than low-insularity proteins (15.8% essential) (Figure 2.7). This result, combined with our finding that high-insularity proteins tend to connect mostly to proteins within the same module, is consistent with the finding that high-insularity proteins tend to be crucial members of essential complexes [21], whereas low-insularity proteins serve to connect different complexes and may not be crucial to maintaining the structure of any individual complex. We found similar results in a human network, using essentiality of mouse orthologs as a proxy for protein essentiality (see Section A.1.5).

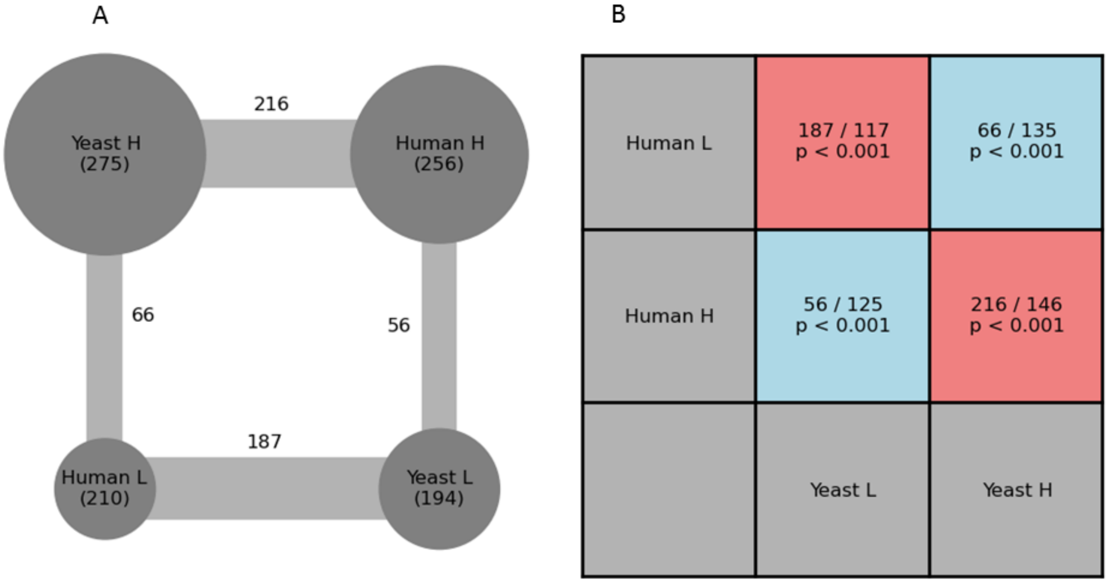
2.3.5 High-insularity and low-insularity proteins are involved in different biological functions

As suggested by the network topology analysis, high-insularity proteins may serve as crucial components of protein complexes, allowing specific biological processes to be efficiently completed. This hypothesis is supported by identifying enriched GO terms in the set of high-insularity and low-insularity proteins. To accomplish this, we used an FDR-corrected hypergeometric analysis similar to that of the GO Termfinder software described in [46]. High-insularity genes are enriched in the following Cellular Component terms, among others, in both human and yeast: "macromolecular complex," "intracellular organelle part," "protein complex," and "nuclear part" (see Section A.4 for complete results). These terms suggest that high-insularity proteins tend to be present in protein complexes and other large cellular structures.

In contrast, low-insularity proteins have no enriched Cellular Component or Molecular Function terms that are shared across human and yeast. Two Biological Process terms are enriched in both species: "small molecule metabolic process" and "single organism metabolic process," suggesting that some low-insularity proteins are enzymes. The general lack of term enrichment in low-insularity proteins suggests that low-insularity proteins fulfill a variety of roles in the cell, and that their roles might differ across species.

2.3.6 Yeast-human homologs exhibit similar functional insularity values

We investigated the evolutionary relationships between high-insularity and low-insularity proteins in yeast and human, using the Princeton Protein Orthology Database (P-POD) [37] to determine evolutionary relationships between proteins. We created the homology graph structure described in Section 2.2.4, a summary



Human L	187 / 117 p < 0.001	66 / 135 p < 0.001
Human H	56 / 125 p < 0.001	216 / 146 p < 0.001
	Yeast L	Yeast H

Figure 2.8: **(A)** A summary representation of the graph structure used to determine the relationship between homology and functional insularity. The network summarized in this figure was created by first removing all proteins other than low-insularity and high-insularity proteins in human and yeast from the ortho-groups. An edge was then drawn between every pair of yeast and human proteins in the same ortho-group, but intraspecies edges were not included. Nodes in the figure represent four protein categories: yeast high-insularity, human high-insularity, yeast low-insularity, and human low-insularity proteins. Numbers and node sizes represent the number of proteins in each category. Edges represent homologous relationships; edge sizes represent the number of such relationships between proteins in the applicable categories. **(B)** Homology relationships between low-insularity and high-insularity proteins are less common than expected by chance. Each cell contains the number of homology relationships between two protein categories, followed by the expected number of homology relationships based on 1,000 stub-rewiring randomizations. Each cell contains an empirical P-value representing the enrichment or de-enrichment of homologies in the real network compared to the random networks; red cells represent enrichment in the real network, and blue cells represent de-enrichment. These results show that homology relationships between high-insularity and low-insularity proteins are less common than expected by chance.

version of which is shown in Figure 2.8. To determine whether or not homologous proteins are likely to have similar functional insularity values, we counted the number of edges that occur between low-insularity and high-insularity proteins and computed P-values by running 1,000 stub-rewired trials (Figure 2.8). The random trials had significantly more edges between high-insularity and low-insularity proteins than did the real graph. Therefore, homologous proteins across yeast and human tend to have similar functional insularity.

We investigated further to determine whether the link between homology and functional insularity is due primarily to conservation of low-insularity proteins, high-insularity proteins, or both. To accomplish this, we counted the number of high-to-high insularity edges and low-to-low insularity edges in the real graph and compared to the randomizations. As shown by Figure 2.8B, there are more high-to-high and low-to-low edges than expected by chance, indicating that insularity values of both low-insularity and high-insularity proteins are conserved across human and yeast. In addition, there is a strong correlation between the functional insularity scores of yeast-human homologs (Spearman correlation of 0.540), confirming that homologous proteins tend to have similar insularities.

2.3.7 Low-insularity proteins decrease quality of network clustering results

The applications of functional insularity are not limited solely to revealing general principles of functional organization in PPI networks; it may also be used to improve the quality of network clustering results. Network clustering is a well-known method for analysis of large biological networks [49]. Network clustering algorithms are used for at least two complementary purposes. First, they are used to separate a network into locally dense subgraphs. Secondly, they are used to identify functional modules of proteins that physically interact. We determined that the presence of low-insularity proteins in a PPI network causes network clustering to perform less reliably at both of these tasks.

In order to determine the effect of functional insularity on network clustering's ability to separate a network into locally dense subgraphs, we ran the graph clustering algorithm SPICi [41] with parameter optimization (see Section 2.2.3) on the original network and on the network that remained after removing all low-insularity proteins (one third of the scored proteins), then measured the topological quality of the resulting clusterings using the modularity measure from [50]. We found that the clustering of the original network had lower modularity (0.371) than the clustering of the low-insularity-free network (0.492).

To verify that this result is not due merely to the effect of removing a large number of proteins from the network, we compared the modularity of the low-insularity-free clustering to that of 100 networks where an equal number of random scored proteins had been removed. For these networks we ran SPICi using the optimized parameters determined using the low-insularity-free network. We found that the clustering of the networks with random removal had lower modularity (mean 0.370, standard deviation 1.58×10^{-2}) than the clustering of the low-insularity-free network. Thus, the presence of low-insularity proteins results in a lower quality topological clustering of the network. Using MCL [43, 44] rather than SPICi as the clustering algorithm yielded similar results, as did running the analysis on a human PPI network (see Section A.3).

In order to determine the effect of functional insularity on network clustering's ability to recover known protein functional modules, we used the semantic density evaluation method described in [49]. The clustering of the low-insularity-free network scored higher (0.154) than that of the original network (0.105) and the 100 networks with random node removal (mean 9.06×10^{-2}, standard deviation 4.80×10^{-3}), indicating that the presence of low-insularity proteins interferes with the ability of clustering algorithms to recover known functional modules. Thus, if proteins with low functional insularity can be identified in a network, researchers might choose to remove such proteins from the network before running a clustering analysis.

2.3.8 Discussion

Related work has explored the roles played by proteins in organizing network structure [1]. Initial work focused on attributes of individual proteins, such as degree, and PPI networks were shown to have power-law degree distributions, in which a small number of high-degree hubs connect a larger number of low-degree nodes [51]. Significant previous work has focused on identifying intermodular proteins as connecting separate modules [27, 52, 53, 50]. Our work adds two key contributions to build on this previous work.

First, we show that intermodular connector proteins can themselves be within modules. This is shown by Figure 2.2, which shows the presence of low-insularity proteins within functional and topological modules. In addition, the significance of the overlap between low-insularity proteins and "within-cluster connectors" shown in Figure 2.6 indicates the existence of proteins within topological modules that serve as interfaces to other modules.

Secondly, our work adds a functional dimension to previous work. A previous topological study suggested the existence of "module organizer" and "module connector" proteins [27]. In this study, the authors identified proteins located between network clusters ("connectors") and proteins that were centrally located within clusters ("organizers"). Our work reveals the existence of "connectors" (low-insularity proteins) and "organizers" (high-insularity proteins) on a functional as well as topological level. Our work also provides a functional view of two expression-based studies [52, 53], which combined PPI network data with gene expression data to identify two classes of hubs [52] and modules [53] that differ in their expression patterns.

We note that functional insularity could be applied to protein networks other than physical interaction networks. For example, [50] identified several topological roles that metabolites play in metabolic networks; this work could be expanded by computing functional insularities of proteins in a metabolic network and identifying differences between high-insularity and low-insularity proteins.

When speciation events occur, the results in Figure 2.8 suggest that pairs of homologous genes tend to remain both high-insularity or both low-insularity. This may indicate the presence of an evolutionary pressure for certain genes to retain high or low functional insularities. As suggested by the biological differences between high-insularity and low-insularity proteins, the biological requirements of a high-insularity protein that organizes a macromolecular complex are likely different from those of a low-insularity protein that serves as an interface among multiple complexes. Once the required biological traits of a highly high-insularity or low-insularity protein have evolved, further modification may be evolutionarily infeasible, and homologous proteins may therefore retain similar levels of functional insularity.

Our functional insularity measure has several potential uses. First, functional insularity has important ramifications for the use of topological clustering-based network analysis [27]. We find that proteins with low functional insularity cause clustering algorithms to produce a less modular network partition and to recover known functional modules less effectively. Thus, researchers might obtain more useful clustering results by removing low-insularity proteins from the network. In addition, functional insularity allows study of the modularity/connectivity tradeoff on a module-based level, rather than globally, as required by purely topological techniques such as betweenness centrality. Finally, functional insularity allows identification of intermodular connector proteins in low-betweenness network locations that would be missed by a purely topology-based identification.

2.4 Conclusions

We propose that in order to best elucidate the relationships between the functions of life, PPI networks are best viewed in a functional as well as a topological sense. We define functional insularity, a local functional/topological measure by which to determine proteins' roles in maintaining the functional modularity/connectivity tradeoff. We also show that proteins very widely in their functional insularity and that low-insularity and high-insularity proteins differ in biological traits such as essentiality and regulation. These results suggest the presence of a tradeoff between modularity and connectivity in PPI networks, with high-insularity proteins supporting network modularity and low-insularity proteins supporting network connectivity. Studies of network topology have often been used as a proxy by which to gain insight on the functional organization of organisms. We suggest that functional organization be studied directly, through the use of measures such as functional insularity, rather than indirectly, through the use of purely topology-based measures.

Chapter 3

Evidence for function acquisition in human proteins over evolutionary time

3.1 Introduction

Genomes evolve through changes in their genetic makeup; in addition to changes in noncoding regions, genes are gained, lost, and mutated over time. New genes may form via several mechanisms [17, 54, 55], including duplication [14], *de novo* creation [56, 57], exon shuffling, retroposition, and gene fusion or fission [17]. When a gene is created with no sequence similarity to existing genes, it can become the progenitor to a new *ortho-group* or *gene family*, which grows through duplication and speciation events. New gene families may also be formed when two ancient duplicates diverge to the point where they no longer have detectable sequence similarity.

In order to understand the mechanisms of evolution, it is necessary to understand the ways in which the genetic makeup of organisms changes over time. As stated above, gene gains, losses, and mutations are three important mechanisms that drive evolution. There is a large body of previous work investigating gene gain and its effects on evolution, through both duplication [14, 58, 59, 60, 61, 62] and *de novo* creation [18]. In this paper, we aim to study changes in gene function over time, by identifying functional differences between younger and older genes.

Previous work has identified several functional differences between young and old genes, largely in yeast. For example, in yeast it has been observed that young genes tend to have fewer protein-protein physical interactions than older genes in [19, 18]. Previous studies in human have mainly focused on expression and disease. In human, young genes tend to be expressed in fewer tissues [63], have fewer regulators [64], and are less likely to be Mendelian disease genes [65] than older genes. The presence of these functional differences between young and old genes suggests that the functional properties of genes may change over evolutionary time.

In this paper, we aim to build on the work cited above to identify trends in the acquisition of new functions by human genes over evolutionary time. We categorize human proteins into 8 age groups using 12 genomes spanning from human to *E. coli*. Here, we take several views of functional integration, including protein sequence and structure [66], protein-protein physical interactions, and gene expression. We find systematic differences in the functional properties of young vs. old genes. First, we find that younger genes in human have shorter sequence length than older genes, and that younger genes have fewer unique domains than older genes. Second, we find that younger human genes participate in fewer protein-protein interactions than older genes. Third, we find that human genes tend to interact mainly with similar-aged genes, extending the result found in yeast [18]. Fourth, we find that younger genes are less likely to be essential than older genes. Finally, we identify differences in the expression of young and old human genes. We confirm that older genes are expressed in more tissues than younger genes, but we also show that younger pairs of paralogs are more coexpressed and share more regulators than older pairs of paralogs.

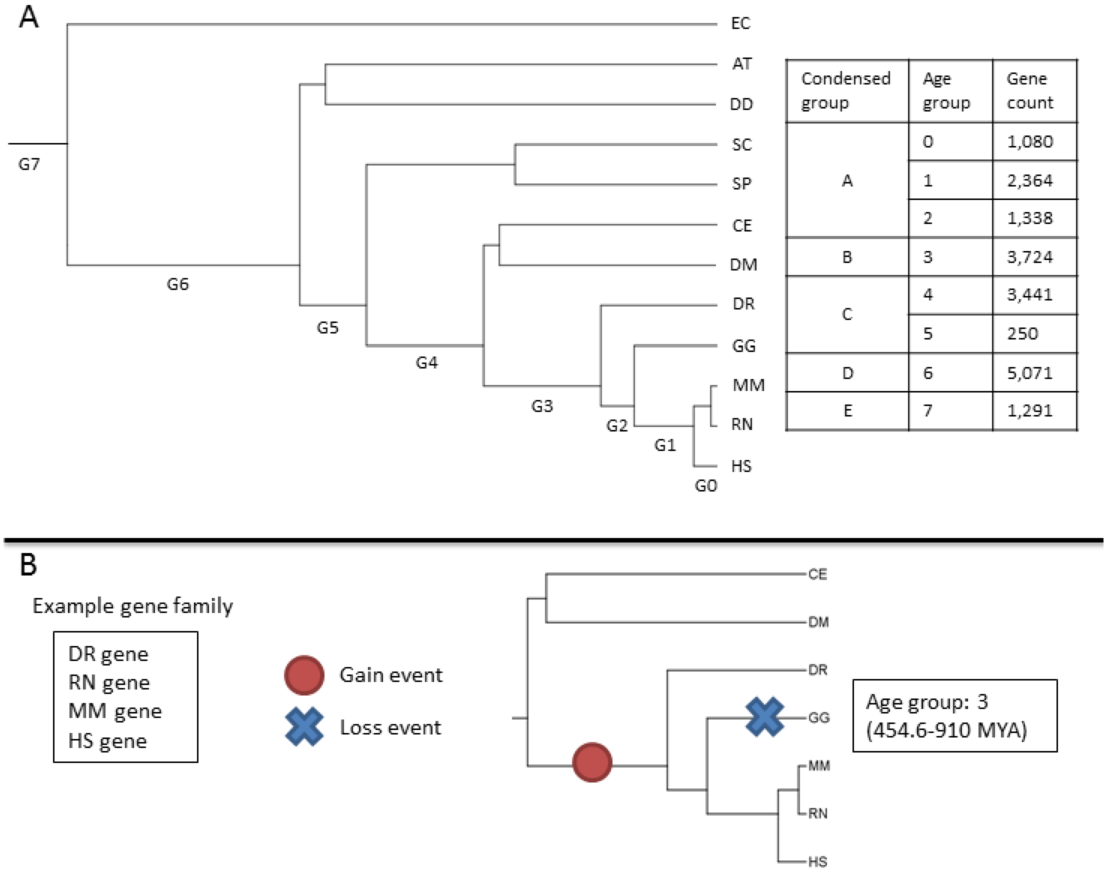
The results found in this paper and in previous work suggest that the functional characteristics of genes change over time. An alternative explanation for these results is that older genes are involved in biological processes that require different functional characteristics (e.g., physical interactions, expression patterns, etc.) from those required by the biological processes typically carried out by other genes. In order to distinguish between these alternatives, we extended the previous work by incorporating a new type of analysis, in which we investigate the functional features of young vs. old genes involved in specific biological processes. With few exceptions, the results shown in this paper hold for specific biological processes as well as in aggregate, suggesting that genes change their functional characteristics over time.

3.2 Methods

3.2.1 Gene families

Gene homology data were obtained from the Princeton Protein Orthology Database (P-POD) [67]. This dataset contains non-overlapping sets of evolutionarily related proteins called "ortho-groups." A set of duplicate genes, or paralogs, was needed for several of the analyses in this chapter; for this, we used the list of paralogs provided by P-POD. This list contains pairs of paralogs; both genes in each pair are members of the same ortho-group. Several PPOD datasets are available; we used the "naive ensemble" dataset of PPOD version 4.

3.2.2 Determining the age of gene families



Condensed group	Age group	Gene count
A	0	1,080
	1	2,364
	2	1,338
B	3	3,724
C	4	3,441
	5	250
D	6	5,071
E	7	1,291

Figure 3.1: **(A)** The species tree and age groups used to infer gene age using the Dollo Parsimony Principle. Species abbreviations are as follows: EC–*Escherichia coli*, AT–*Arabidopsis thaliana*, DD–*Dictyostelium discoideum*, SC–*Saccharomyces cerevisiae*, SP–*Schizosaccharomyces pombe*, CE–*Caenorhabditis elegans*, DM–*Drosophila melanogaster*, DR–*Danio rerio*, GG–*Gallus gallus*, MM–*Mus musculus*, RN–*Rattus norvegicus*, HS–*Homo sapiens*. **(B)**Example computation of the age of an ortho-group using the Dollo parsimony principle. The ortho-group contains genes from mouse, rat, zebrafish, and human. By the Dollo parsimony principle, only one gene gain event is allowed per ortho-group. Thus, the most parsimonious solution is for the initial formation of the ancestral gene to occur 300-400 MYA (millions of years ago), between the divergences of yeast and fly. Because there is no chicken gene in the ortho-group, a gene loss event must be assigned to the diverged chicken branch of the species tree. The initial formation event cannot be assigned to occur any earlier than 300-400 MYA, because then a second gene gain event would be required to explain the presence of the zebrafish gene in the ortho-group. Nor can the initial formation event be assigned to occur any later than 300-400 MYA, because these solutions would require more than one loss event and are therefore less parsimonious than the chosen solution. Thus, the ortho-group is assigned to the 300-400 MYA age group.

We used a maximum parsimony analysis, based on the Dollo parsimony principle [68], of the homology data to classify genes into one of 8 age groups (see Figure 3.1). In a maximum parsimony analysis, a set of evolutionarily related genes (e.g., a PPOD ortho-group) is compared to an existing species tree showing known evolutionary relationships among species. In our analysis, we used the species tree from the NCBI Taxonomy database [69] and labeled the internal nodes of the tree with approximate timepoint data from the TimeTree database [70]. The set of species with genes in the ortho-group is compared to the topological structure of the species tree in order to determine the origin point of the ortho-group on the species tree that implies the lowest possible number of gene loss events. Our analysis is similar to the gene age prediction functionality of the ProteinHistorian program [71].

3.2.3 Determining the age of duplication events

It is also possible to use maximum parsimony to determine the age groups of individual duplication events in an ortho-group's evolutionary history. To accomplish this, we used the reconstructed evolutionary history trees of the ortho-groups; this data is generated by the NOTUNG program [72] and provided with PPOD. Specifically, a tree is provided for each ortho-group, in which leaves represent the ortho-group's genes and internal nodes represent speciation or duplication events. To compute the age group of internal duplication nodes, we ran the maximum parsimony analysis on the set of genes in the subtree of the duplication node in question.

3.2.4 Defining the age of genes and paralogs

After calculating the ages of ortho-groups and duplication events, we then defined two metrics for gene age, each of which can be applied to individual genes or to pairs of paralogous genes. First, *family age* is defined as the age of a gene's ortho-group. Note that for a pair of paralogs, both genes must be part of the same ortho-group and therefore have the same family age. Secondly, for genes that have at least one paralog, *duplication age* is defined as the age of a gene's most recent ancestral duplication event in its ortho-group's phylogenetic tree reconstruction (i.e., the age of the gene's lowest ancestor node that represents a duplication event, not a speciation event). Duplication age of a pair of paralogs is defined as the age of their lowest common ancestor node in the ortho-group's phylogenetic tree; note that the lowest common ancestor node must be a duplication event, not a speciation event. Genes without paralogs have no ancestral duplication event, so their duplication age is defined to be equal to their family age.

In the analysis of paralog coexpression and regulator overlap, the original age groups were condensed to five coarser age groups in order to obtain a larger sample size for each age group. The original age groups were merged as follows:

- Condensed age group A consists of original age groups 0 to 2 (amniotes).
- Condensed age group B consists of original age group 3 (bony vertebrates).
- Condensed age group C consists of original age groups 4 and 5 (opisthokonts; ancestor of metazoa and fungi).
- Condensed age group D consists of original age group 6 (eukaryotes).
- Condensed age group E consists of original age group 7 (cellular life).

3.2.5 Expression data

Human expression data from Su et. al [73] were downloaded from the Gene Expression Omnibus [74, 75]. This dataset contains 79 tissues, each of which has 2 biological replicates. Moreover, a given gene may be detected by more than one probeset; therefore, there are multiple expression values for a given gene in a given tissue. To obtain a consensus expression value for each gene in each tissue, we first computed the average of all probeset expression values for each biological replicate, thereby obtaining one value for each replicate. We then averaged the biological replicates to obtain the final expression value of a given gene in a given tissue.

The Spearman rank correlation coefficient was used to compute coexpression between genes. However, before computing the correlation, the gene expression values in each tissue were normalized, so that all tissues had the same mean and standard deviation of expression values. This was done in order to control for tissues that display systematically high or low gene expression.

For the analysis of gene expression ubiquity, we defined a binary metric for whether or not a gene is expressed in a given tissue. For this, we used a simple threshold value of 200, as was used in previous work [76]. If a gene's non-normalized consensus expression value in a given tissue is greater than 200, the gene is considered to be expressed in that tissue; otherwise, it is considered to be absent.

3.2.6 Other data

Human protein-protein interaction data was obtained from BioGRID [77, 78]. All evidence types indicative of a physical interaction were considered: "affinity capture–luminescence," "affinity capture–MS," "affinity capture–RNA," "affinity capture–Western," "biochemical activity," "co-crystal structure," "co-fractionation," "co-localization," "co-purification," "far Western," "FRET," "PCA," "Protein-peptide," "protein-RNA," "proximity label–MS," "reconstituted complex," and "two-hybrid." The 1% of proteins with highest degree were removed from the network in order to address the presence of "sticky" proteins that seem to interact with many other proteins due to experimental artifacts. Network data were processed using the NetworkX graph analysis package [45].

Human regulatory network data was obtained from Gerstein, et. al [79]; the "Enets2 proximal filtered" network was used. This network contains promoter-proximal regulatory interactions, filtered to avoid false positives using a probabilistic model. In this network, nodes represent transcription factors and targets, and the presence of a directed edge from a transcription factor to a target indicates that the transcription factor binds upstream of the target's promoter.

Human gene essentiality data were inferred by mapping human genes to their mouse orthologs. If a human gene has an essential mouse ortholog, it is considered to be essential; if it has a dispensable mouse ortholog, it is considered to be dispensable. Essentiality of mouse genes was obtained from MGI [80].

Protein disorder data were produced by running the IUPred [81, 82] disorder prediction program on the amino acid sequence of every human protein; the "long" disorder prediction option was used. IUPred outputs a predicted disorder between 0 and 1 for each residue. We considered all residues with score greater than or equal to 0.5 to be disordered, which results in approximately 26% of all residues in the human proteome being classified as disordered. We defined the final disorder score of each protein as the fraction of disordered residues in the protein's sequence.

Human protein sequence data were obtained from Ensembl [83]. In order to determine the number of unique domains in human proteins, PFAM [66] domain predictions were run on the human protein sequences. Repeated domains in a given protein were only counted once, as additional copies of the same domain presumably do not add new functionality to the protein.

Spearman rank correlations between gene age and various characteristics (number of PPI interactions, sequence length, etc.) were computed. Empirical P-values were obtained by running 1,000 random trials, in which Spearman rank correlations were computed after randomly shuffling the gene age values. Partial Spearman correlations and P-values of various characteristics vs. family age given duplication age (and vice versa) were computed using the R statistical software package [84].

3.2.7 Figures

Figures displaying mean and standard deviation of various protein characteristics, as well as the table from Figure 3.5, were generated with the Matplotlib graphing package for Python [47]. To create the plots in part B of Figures 3.2, 3.3, 3.4, 3.6 3.7, and 3.8, we first filtered the generic Gene Ontology slim to include only terms that annotate at least one protein from each age group. We computed the Spearman correlation between age and the feature in question for the sets of proteins annotated with each term. We then arranged the terms in the generic GO slim in decreasing order of Spearman correlation and plotted each term's correlation.

3.3 Results

In order to investigate the means by which human genes acquire functions and integrate into cellular networks, we investigated several function-related gene traits, including sequence length, domain counts, protein-protein interactions, and expression data, and tested for correlation with gene age. We computed age groups for human genes using the Dollo parsimony principle [85] as described in Section 3.2.4. There are several possible definitions for age of a gene, including "family age" and "duplication age" (see Section 3.2.4). For the results described below, we used family age as our definition of gene age for both individual genes and pairs of paralogs. After running the maximum parsimony analysis described in Section 3.2.2, human genes were assigned to the age groups shown in Figure 3.1.

3.3.1 Sequence and structure properties of young vs. old genes

As a first step in investigating function acquisition over time, we investigated the relationship between protein sequence length and unique domain count with age.

Because protein structure (and therefore function) is determined by its sequence, features such as sequence length can be informative in regards to protein function. In particular, sequence length places a physical constraint on the number of domains a protein may contain. Protein domains are structural protein subunits that perform specific functions; thus, a protein's function is determined by the set of domains it contains. Features such as unique domain count are therefore informative as to the number and complexity of functions that may be performed by a protein. In addition, we also investigated the relationship between age and protein disorder, a structural feature which has been implicated in cellular organization and recruitment of binding partners [86]. Disordered regions in a protein are amino acid subsequences that do not form a unique three-dimensional structure *in vivo* [86], existing instead as random coil or adopting a variety of different conformations.

Younger genes have shorter sequence length and fewer domains than older genes. In order to investigate the reasons for the difference in degree between older and younger genes, we identified differences in sequence length and domain count between older and younger genes. We found that younger genes tend to have shorter protein sequence length and fewer unique domains than older genes, although they do not have fewer domains than older genes if repeated domains are counted (Figures 3.2A and 3.2C). There is a slight reversal of the trend in the oldest age groups; however, the overall correlations between gene age and sequence length, and between gene age and unique domain count, remain positive.

In addition, we investigated the correlation of gene age vs. unique domain count and sequence length for individual Gene Ontology [33] Biological Process (BP) terms, using the Gene Ontology Generic Slim term set. For each of the 70 terms in the set, we computed the Spearman correlation between gene age and domain count (as well as sequence length). We found that most gene sets annotated with an given slim term show a positive correlation between gene age and domain count, as well as gene age

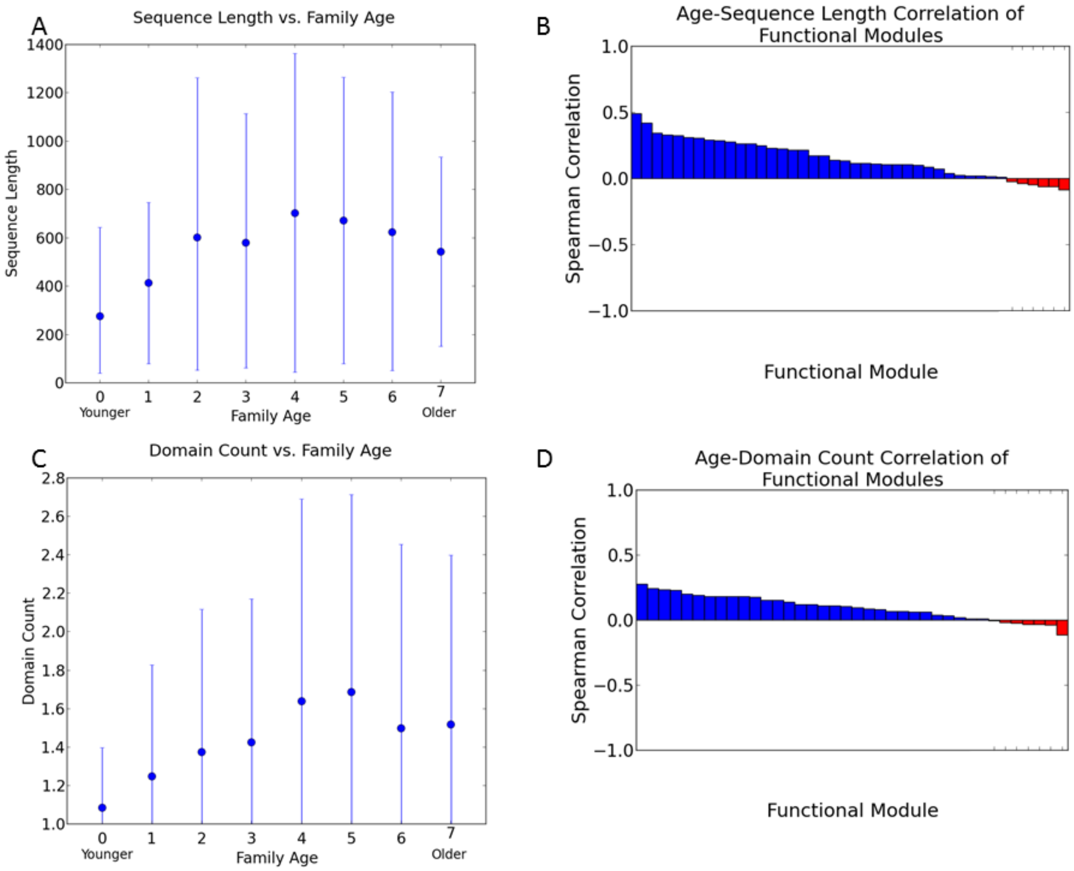


Figure 3.2: **(A)** Older genes tend to have longer sequence length than younger genes (Spearman correlation 0.167; $p < 0.001$). **(B)** Spearman correlations between age and sequence length of individual GO terms tend to be positive. **(C)** Older genes tend to have more unique domains than younger genes (Spearman correlation 0.079; $p < 0.001$). **(D)** Spearman correlations between age and unique domain count of individual GO terms tend to be positive.

and sequence length, indicating that the aggregate correlations seen in parts A and C of Figure 3.2 also hold for individual biological functions (Figures 3.2B and 3.2D).

Young genes have more structural disorder than old genes. We predicted structural disorder for human genes using IUPred [81, 82] and found that younger genes tend to have more structural disorder than older genes (Figure 3.3). In addition, the negative correlation between disorder and age holds for individual biological functions as well (Figure 3.3B). This suggests that younger genes are initially formed with little structure, and that they form more structured domains as they gain biological functions over time.

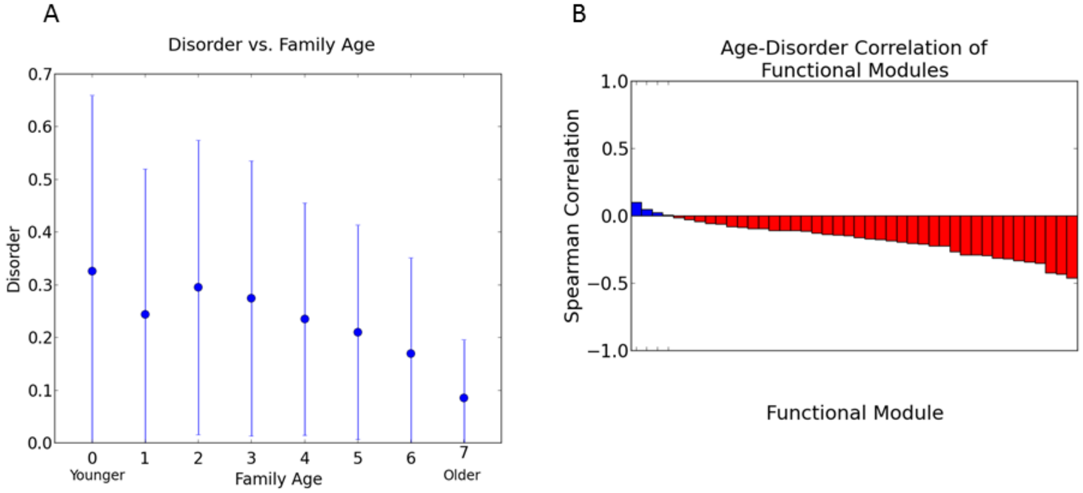


Figure 3.3: **(A)** Younger genes have more structural disorder than older genes (Spearman correlation -0.150, $p < 0.001$). **(B)** The Spearman correlations of gene age vs. disorder of individual GO terms tend to be negative.

3.3.2 Physical interactions of young and old genes

Proteins rarely function in isolation; they often physically interact with other proteins in order to fulfill their functions. Together, these interactions combine to form a large *protein-protein interaction network* (PPI network) [78] with tens of thousands of interactions. In order to further investigate protein function acquisition, we measured the integration of young and old proteins in the human PPI network in two ways. First, we determined the average number of physical interactions in which young and old proteins participate. Secondly, we investigated the enrichment of physical interactions within and between individual protein age groups.

Younger genes have fewer protein-protein interactions than older genes. For each age group, we computed the distribution of protein-protein interaction count of proteins in the age group. As shown in Figure 3.4A, younger genes tend to have fewer protein-protein interactions than older genes; a similar result was found in yeast by previous work [18, 87]. We note a reversal of the trend in the oldest age group, but there is an overall positive correlation between gene age and interaction count. In

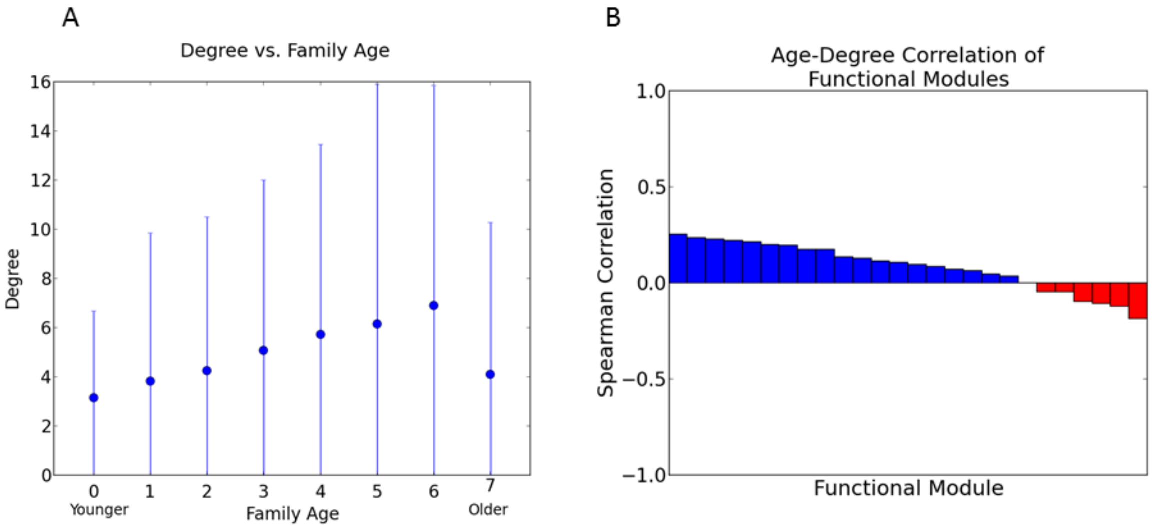
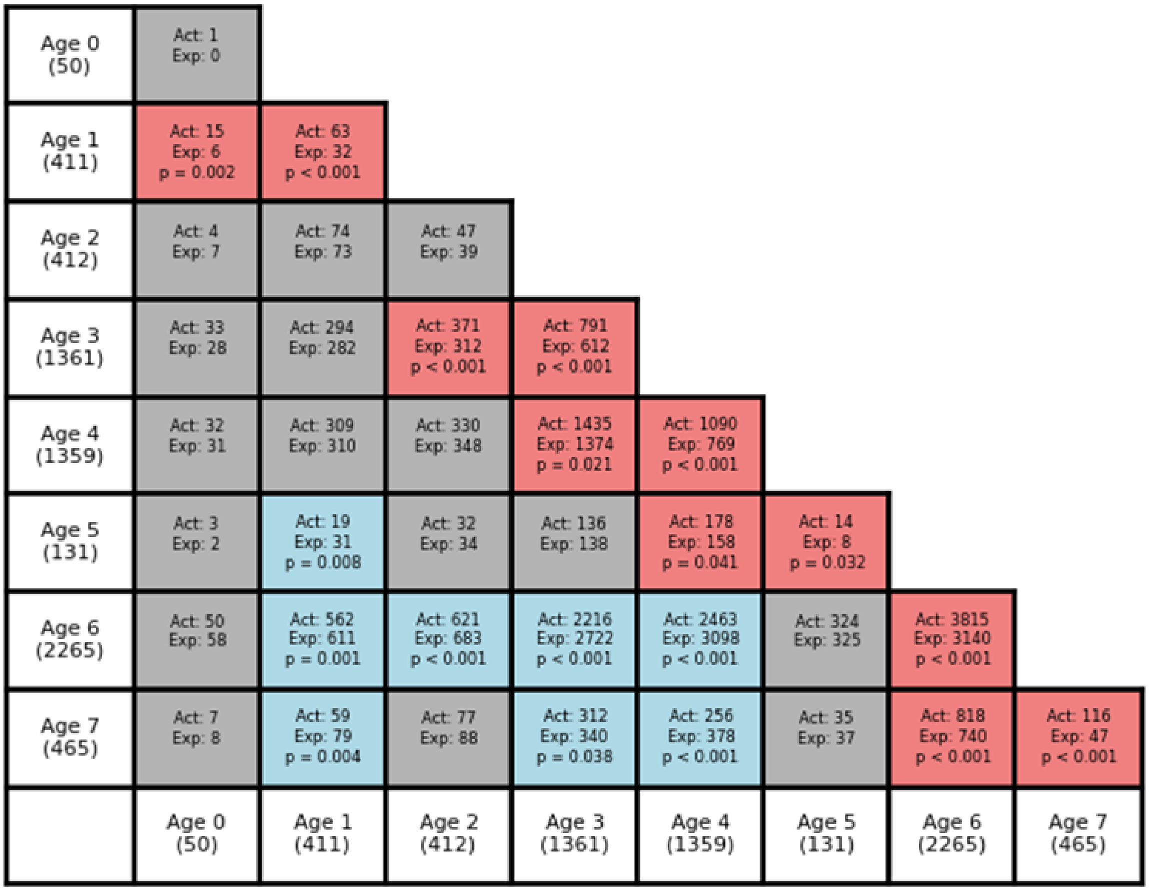


Figure 3.4: **(A)**Younger genes tend to have fewer protein-protein interactions than older genes (Spearman correlation of gene age group vs. interaction count 0.093, $p < 0.001$). **(B)** The Spearman correlations of gene age vs. number of protein-protein interactions of individual GO terms tend to be positive.

addition, the correlation between protein-protein interaction count and age holds for individual biological functions (Figure 3.4B). Because proteins domains can provide interfaces by which to interact with another protein [88], the presence of more unique domains in older proteins might explain the higher interaction count of older proteins.

Genes of similar age physically interact often. If proteins gradually acquire physical interactions over time, this integration into the PPI network could occur in several ways. Young proteins might tend to form interactions primarily with older proteins, therefore integrating mainly into existing functional modules. Alternatively, younger proteins might tend to form interactions primarily among themselves. We tested for evidence of these alternatives by investigating the degree to which proteins interact with other proteins of similar age. For this analysis, we counted the number of protein-protein interactions within and between the age groups. We then compared this result to a series of random trials run on stub-rewired [23] randomizations of the PPI network to determine significance of the edge counts. As shown in Figure 3.5,



Age 0 (50)	Act: 1 Exp: 0							
Age 1 (411)	Act: 15 Exp: 6 p = 0.002	Act: 63 Exp: 32 p < 0.001						
Age 2 (412)	Act: 4 Exp: 7	Act: 74 Exp: 73	Act: 47 Exp: 39					
Age 3 (1361)	Act: 33 Exp: 28	Act: 294 Exp: 282	Act: 371 Exp: 312 p < 0.001	Act: 791 Exp: 612 p < 0.001				
Age 4 (1359)	Act: 32 Exp: 31	Act: 309 Exp: 310	Act: 330 Exp: 348	Act: 1435 Exp: 1374 p = 0.021	Act: 1090 Exp: 769 p < 0.001			
Age 5 (131)	Act: 3 Exp: 2	Act: 19 Exp: 31 p = 0.008	Act: 32 Exp: 34	Act: 136 Exp: 138	Act: 178 Exp: 158 p = 0.041	Act: 14 Exp: 8 p = 0.032		
Age 6 (2265)	Act: 50 Exp: 58	Act: 562 Exp: 611 p = 0.001	Act: 621 Exp: 683 p < 0.001	Act: 2216 Exp: 2722 p < 0.001	Act: 2463 Exp: 3098 p < 0.001	Act: 324 Exp: 325	Act: 3815 Exp: 3140 p < 0.001	
Age 7 (465)	Act: 7 Exp: 8	Act: 59 Exp: 79 p = 0.004	Act: 77 Exp: 88	Act: 312 Exp: 340 p = 0.038	Act: 256 Exp: 378 p < 0.001	Act: 35 Exp: 37	Act: 818 Exp: 740 p < 0.001	Act: 116 Exp: 47 p < 0.001
	Age 0 (50)	Age 1 (411)	Age 2 (412)	Age 3 (1361)	Age 4 (1359)	Age 5 (131)	Age 6 (2265)	Age 7 (465)

Figure 3.5: Genes from similar age groups tend to physically interact more often than expected by chance. This table contains empirical p-values indicating enrichment or de-enrichment of protein-protein interactions between and within the different age groups. P-values were obtained by running stub-rewiring randomizations of the original PPI network. Each cell indicates the number of actual edges found between the indicated age groups, and the expected number of such edges based on randomized trials. Red cells indicates presence of significantly more edges than expected; blue indicates the presence of significantly fewer edges than expected. Parenthesized numbers in edge cells indicate the number of proteins in the PPI network that are categorized into each age group. Red cells tend to be near the table's diagonal, indicating enrichment of physical interactions between similar-aged genes.

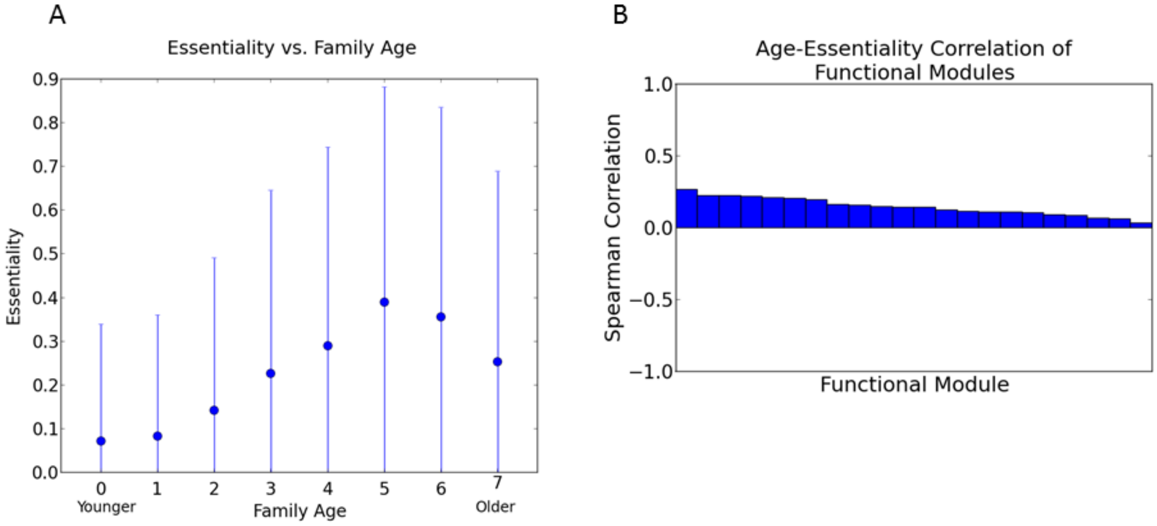


Figure 3.6: **(A)** Older genes are more likely to be essential than younger genes (Spearman correlation of gene age group vs. essentiality 0.155, $p < 0.001$). Here, essentiality is interpreted as a binary number, with 0 indicating that a gene is dispensable and 1 indicating that it is essential. The mean and standard deviation of the essentiality of each age group is plotted. **(B)** The Spearman correlations of gene age vs. essentiality of individual GO terms tend to be positive.

we found that proteins tend to interact with similar-aged proteins more often than expected by chance.

3.3.3 Essentiality of young and old genes

We computed the average essentiality of each age group, using essentiality of mouse orthologs as a proxy for essentiality of human genes (see Section 3.2.6). In this analysis, dispensable genes were assigned a value of zero, and essential genes were assigned a value of 1; the mean and standard deviation of these values are displayed for each age group in Figure 3.6A. As shown in Figure 3.6, older genes are more likely to be essential than younger genes. The trend reverses in the two oldest age groups; however, it is still the case that older genes are more likely to be essential than younger genes overall, and this is true of individual biological functions as well (Figure 3.6B).

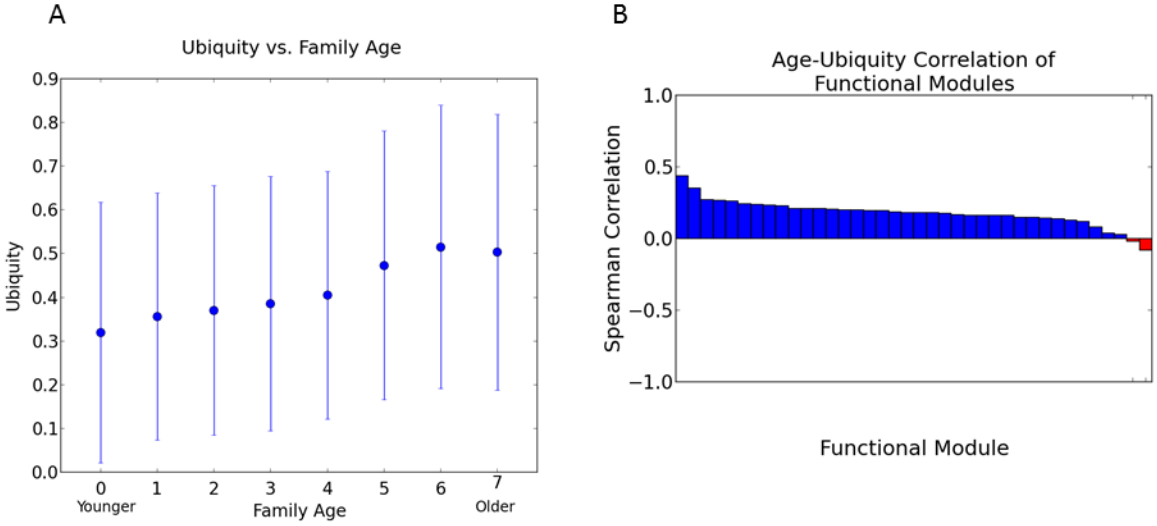


Figure 3.7: **(A)** Older genes have higher expression ubiquity than younger genes (Spearman correlation of gene age group vs. ubiquity 0.192, $p < 0.001$). **(B)** The Spearman correlations of age vs. ubiquity of individual GO terms tend to be positive.

3.3.4 Expression of young and old genes and paralogs

In addition to investigating the relationships of sequence, structure, and physical interactions to gene age, we also investigated three properties related to gene expression: expression ubiquity, coexpression of paralogs, and shared regulators between paralogs. Ubiquity, the fraction of tested tissues in which a protein is expressed, is informative with regard to gene function because a protein can only perform its function where it is expressed. Similarly, coexpression is informative when comparing the functions of two related genes. Because gene regulation is an important factor in determining gene expression patterns, regulator data can provide insight into expression patterns and gene function. We observed correlations between gene age and these three properties.

Older genes are expressed in more tissues than younger genes. We computed the ubiquity of human genes as described in Section 3.2.5 and plotted the ubiquity distribution of the different age groups. As shown in Figure 3.7A, we found that younger genes tend to be less ubiquitous (i.e., more tissue specific) than older genes, meaning that younger genes are expressed in fewer tissues than older genes.

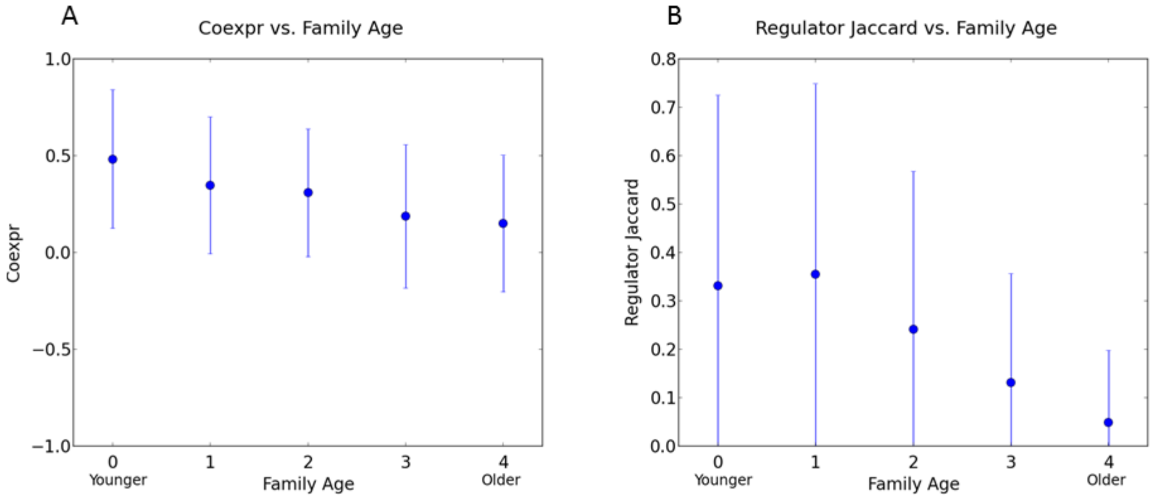


Figure 3.8: **(A)** Younger pairs of paralogs have higher coexpression than older pairs of paralogs (Spearman correlation of paralog pair age group vs. pair coexpression -0.182, $p < 0.001$). **(B)** Younger pairs of paralogs tend to share more common regulators than older pairs of paralogs (Spearman correlation of ortho-group age vs. Jaccard coefficient of the two sets of regulators for each gene: -0.262, $p < 0.001$).

In addition, the sets of genes annotated with individual GO Biological Process terms tend to show positive correlations between age and ubiquity (Figure 3.7B).

Younger pairs of paralogs are more coexpressed and share more regulators than older pairs. The expression patterns of pairs of paralogs can be used to investigate the change in gene function over time. Gene duplication has been implicated as a driving factor in the acquisition of new functions over time [89, 90, 91]. Two potential mechanisms for this process are neofunctionalization, in which a duplicated gene evolves a function distinct from that of the original gene, and subfunctionalization, in which duplicate genes split the function or functions of the original gene. In both of these mechanisms, the functions, and therefore presumably the expression patterns, of the two genes diverge.

In order to find evidence of the functional divergence of paralogs over time, we computed the coexpression of pairs of paralogs as described in Section 3.2.5, then plotted these coexpression values against the age groups of the paralog pairs (both

Characteristic	vs. family age given duplication age	vs. duplication age given family age
PPI interactions	0.066 ($p = 1.00 \times 10^{-7}$)	0.024 ($p = 0.054$)
Sequence length	0.070 ($p = 4.52 \times 10^{-20}$)	0.097 ($p = 2.69 \times 10^{-37}$)
Unique domains	0.065 ($p = 6.74 \times 10^{-15}$)	-0.003 ($p = 0.738$)
Ubiquity	0.155 ($p = 2.62 \times 10^{-60}$)	0.006 ($p = 0.537$)
Essentiality	0.082 ($p = 5.42 \times 10^{-7}$)	0.086 ($p = 1.48 \times 10^{-7}$)
Disorder	-0.158 ($p = 1.29 \times 10^{-97}$)	0.067 ($p = 1.91 \times 10^{-18}$)
Paralog coexpression	-0.110 ($p = 6.05 \times 10^{-29}$)	-0.053 ($p = 6.84 \times 10^{-8}$)
Paralog regulators	-0.064 ($p = 0.002$)	-0.279 ($p = 1.30 \times 10^{-44}$)

Table 3.1: Partial correlations and p-values for the eight functional characteristics investigated in this paper vs. family age given duplication age, and vs. duplication age given family age.

genes involved in a paralogy relationship have the same age, because they belong to the same gene family). In this analysis, we condensed the 8 original age groups into 5 larger groups in order to increase the sample size of each group (see Section 3.2.4). As shown in figure 3.8A, younger pairs of paralogs tend to be more coexpressed than older pairs.

Because gene expression is controlled mainly by gene regulation, expression differences between two genes could be explained by regulatory differences. Therefore, in order to determine the reason for the decrease in paralog coexpression with age, we computed the Jaccard overlap of regulators for each pair of paralogs as described in Section 3.2.6, then plotted the overlap distribution for each age group. As shown in figure 3.8B, younger pairs of paralogs tend to have more regulators in common than older pairs.

3.3.5 Function acquisition as a function of family age vs. duplication age

Throughout this chapter, age of a gene or pair of genes has been defined as the estimated time range at which the gene's ortho-group originated, based on application of the Dollo Parsimony Principle to the ortho-group in question. This is a reason-able definition of gene age with respect to function acquisition because genes will continually evolve, diverging from the family's initial ancestor gene.

However, family age is not the only possible definition of a gene's age. New genes within a family are formed by duplication events, which can also be dated using a method similar to that used to date entire gene families. Thus, age of a single gene can be defined as the estimated time range of its most recent duplication event, and age of a pair of paralogs can be defined as the estimated time range of their source duplication event (see Section 3.2.4). Defining age in this manner allows functional changes inherited from ancestral genes in the ortho-group to be distinguished from functional changes that occurred since the genesis (through duplication) of an individual gene. When studying the functional divergence of a pair of paralogs, we can determine whether the extent of divergence is more dependent on age of the family, or on the time since the actual divergence event of the two genes.

In order to answer these questions, we attempted to determine whether family age or duplication age is more correlated with the views of function acquisition used in the above analyses. One difficulty in this investigation is that family age and duplication age are correlated, because family age must be greater than or equal to duplication age. Thus, it is necessary to remove the effect of this correlation when computing correlations between a particular definition of gene age vs. a view of gene function. In order to accomplish this, we used a partial correlation analysis (Table 3.1). We found that family age is the most important factor for number of interactions, number of unique domains, disorder, ubiquity, and coexpression of paralogs, but that duplication age was the most important factor for sequence length and paralog regulator overlap. We found essentiality to be equally correlated with duplication age and family age.

We found that number of protein-protein interactions, domain count, and ubiquity were correlated only with family age (after controlling for duplication age); there was no significant correlation with duplication age after controlling for the effects of family age. This indicates that genes from more ancient families tend to be more ubiquitous, have more domains, and have more protein-protein interactions than genes from more recent families, regardless of the time of their most recent duplication event.

In contrast, we found that sequence length and essentiality are correlated with both family age (after controlling for duplication age) and duplication age (after controlling for family age). Thus, this suggests that while gene families as a whole become more essential and increase in sequence length over time, these changes also occur in individual duplicate genes. In particular, genes with a recent ancestral duplication event are less likely to be essential than genes with an ancient duplication event.

We found that the coexpression of paralogs is correlated both with family age and duplication age. The correlation with duplication age indicates that paralogs originating from more ancient duplication events tend to have more diverged expression, suggesting that divergence of paralog expression occurs gradually over time after the source duplication event. This may suggest the presence of neofunctionalization or subfunctionalization. The correlation with family age, even after controlling for duplication age, indicates that paralogs from older families tend to have more diverged expression, regardless of the time of their source duplication event.

We found that the degree of regulator overlap for paralogs is correlated with both family and duplication age, but the correlation is much more significant for duplication age (after controlling for the effect of family age). This indicates that regulatory divergence of paralog pairs is mainly a function of the time elapsed since divergence from their root duplication event, suggesting that the regulators of paralog pairs diverge gradually over time after their source duplication event.

3.4 Discussion

We have shown that various functional features of genes are correlated with gene age. This suggests at least two hypotheses relating to function acquisition. First, the functional feature in question (e.g., sequence length), might change in individual proteins over evolutionary time. Younger proteins have shorter sequence length than older proteins, suggesting that newly formed proteins might tend to have short sequences that increase in length over time. Alternatively, it may be the case that the functional feature in question does not change in individual proteins, and that older proteins were initially formed with different functional features than younger proteins. The positive correlation of sequence length with age could also be due to older proteins performing functions that require longer sequence lengths, while younger proteins perform functions that require shorter sequence lengths. In this model, the sequence lengths of individual proteins do not change over evolutionary time.

In order to distinguish between these two hypotheses, we investigated the correlation between the various functional features and gene age for individual functional modules. If the first hypothesis were true, we would expect to find similar correlations within individual functional modules as we did for all genes in aggregate. If the second hypothesis were true, we would expect the driving factor in the aggregate age-feature correlation to be the function that the gene performs. In the case of sequence length, we would expect functional modules that perform ancient functions (such as DNA replication) to contain genes with longer sequence length than younger functional modules; thus, we would not expect to find significant age–sequence length correlation within most functional modules.

As shown by part B of Figures 3.2, 3.3, 3.4, 3.6, 3.7, and 3.8, we found that the feature–age correlations within individual functional modules recapitulate the aggregate correlations found in part A of those figures, supporting the first hypothesis. This suggests that the sequence length, domain count, protein interaction count, and expression ubiquity of individual genes may evolve over the large evolutionary timescales investigated in this chapter.

We found two key results that are informative with regards to the evolution of protein-protein interaction networks (Figure 3.4 and Table 3.1). First, young proteins tend to participate in fewer physical interactions than older proteins. Secondly, proteins tend to interact mainly with other proteins of similar age. These results suggest a hypothesis for the addition of new functional modules to the PPI network. Initially, newly formed groups of young proteins may participate in few physical interactions. Over time, a group of young proteins might begin to interact and perform a new function, adding a new functional module to the existing PPI network. In this interpretation, these groups form modules in both a topological and functional sense; they both interact more with each other than with the rest of the network, and they cooperate to perform shared or similar functions. The data shown in Figure 3.5 support this hypothesis over the alternative model, in which young proteins tend to interact primarily with older, existing modules.

If the increase in degree with age is interpreted to indicate that proteins acquire interactions over time, the increase in sequence length and domain count with age suggests a mechanism for this gradual integration into the PPI network. Upon initial formation, new proteins might have few functions, few domains, and a relatively short sequence. Over time, mutations might occur to add sequence, and eventually entire domains, to these proteins. The new domains may then allow the proteins to perform new functions and physically interact with functionally related protein. In this way, younger genes might acquire more domains and physical interactions as they functionally integrate into the cellular network over time.

We found that younger genes tend to have lower expression ubiquity than older genes (Figure 3.7). From the perspective of function acquisition, one possible interpretation of this result is that younger genes have few functions upon initial formation, and therefore are not expressed in many tissues, because they are not needed in those tissues. Over time, they may gain new functions in certain tissues and therefore become expressed in those tissues.

We found that younger pairs of paralogs are more coexpressed and share more regulators than older pairs. This suggests the following hypothesis regarding the functional divergence of paralogs. Immediately after a duplication event, the two duplicates have identical sequences and regulatory regions; they therefore have equivalent functions and are expressed in the same tissues. Over time, their regulatory regions diverge, so that they have fewer regulators in common, and their expression patterns therefore diverge as well. During this process, the two genes begin to perform different functions in different tissues. This hypothesis suggests that neofunctionalization and subfunctionalization occur over time.

3.5 Conclusions

Organisms have generally evolved to become more complex over evolutionary time; while addition of genomic material is one mechanism for this increase in complexity, individual genes are generally believed to evolve over time as well. In this chapter, we have categorized genes into 8 age groups and shown that there are consistent differences between younger and older genes with respect to several functional features of proteins. Together, these findings suggest a model of function acquisition in which young genes gradually integrate into the organism's protein interaction network over time, while they concurrently add domains and become expressed in more tissues. In addition, our results suggest that new duplicates gradually diverge in expression over time, and that this divergence might be produced by changes in regulators over time.

Chapter 4

Conclusion

In this thesis, we presented a metric, functional insularity, that quantifies the degree to which a protein physically interacts with functionally similar proteins. By taking a functional perspective of PPI networks, we were able to reveal the presence of a modularity/connectivity tradeoff and showed that intermodular proteins can exist within network modules. We found that PPI networks as a whole, as well as individual topological and functional modules, contain proteins with a wide range of functional insularity values. We found that low-insularity proteins have different biological properties from high-insularity proteins and that functional insularity is conserved across homologous genes in human and yeast. We also determined that the presence of low-insularity proteins in the PPI network can decrease the ability of graph clustering algorithms to identify biologically meaningful network modules, suggesting a possible application of functional insularity in network pre-processing before running a clustering analysis.

In the second part of this thesis, we took a dynamic view of protein function and PPI network structure over time by quantifying the age of human proteins and determining trends in various function-related features for each age group. We found evidence suggesting that proteins acquire functions gradually over time, adding domains and physical interactions, and becoming expressed in more tissues over time. We also found evidence suggesting that gene duplicates diverge in expression, and presumably function, over time, and that this divergence may be driven by changes in the regulation of the duplicates over time. In addition, we found that proteins tend to interact mainly with other proteins of similar age, suggesting a possible model of PPI network evolution, in which young proteins begin to interact with other young proteins, creating new functional modules that gradually integrate into the PPI network.

Protein-protein interaction networks may be seen as the "wiring" by which information is transmitted through cells. In addition, along with other cellular networks, they may be seen as the functional network of the cell. Thus, understanding their structure is crucial to understanding the mechanisms by which life operates. However, PPI networks, as well as other types of biological data such as genome sequences, can be difficult to interpret due to the large size of the datasets. The methodologies and results presented in this thesis illustrate that techniques from computer science can be applied to these data in order to further elucidate principles of the structure and evolution of protein-protein interaction networks, enabling further understanding of the relationships among the functions of life.

Appendix A

Functional insularity results with alternate datasets

A.1 Alternate Network Results

A.1.1 Network statistics

133 high-degree nodes were removed from the human Biogrid network, with degrees ranging from 108 to 8,959. Analysis resulted in in 1,603 low-insularity and 1,603 high-insularity proteins, with a total of 4,809 scored proteins.

45 high-degree nodes were removed from the yeast HINT Combined network, with degrees ranging from 65 to 737. Analysis resulted in 766 low-insularity and 766 high-insularity proteins, with a total of 2,300 scored proteins.

68 high-degree nodes were removed from the human HINT Combined network, with degrees ranging from 58 to 333. Analysis resulted in 736 low-insularity and 736 high-insularity proteins, with a total of 2,210 scored proteins.

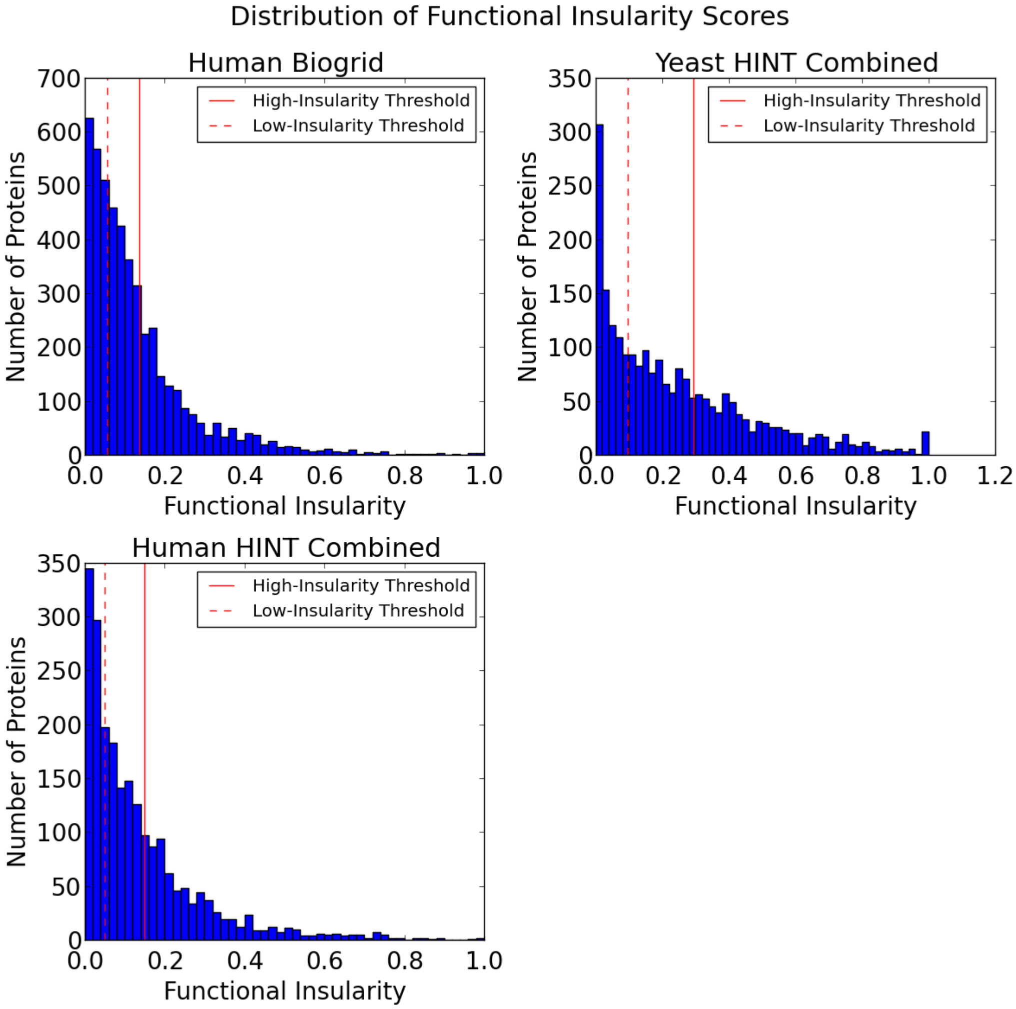


Figure A.1: Distribution of functional insularity scores of all scored proteins in the human BioGRID, yeast HINT Combined, and human HINT Combined networks. Proteins with scores to the left of the dashed line were classified as low-insularity; proteins with scores to the right of the solid line were classified as high-insularity.

A.1.2 Topological Measures

In the human BioGRID network, low-insularity proteins have lower degree (mean 10.3737) than high-insularity proteins (mean 13.238927), p = 1.25558e-22, and low-insularity proteins have lower local clustering coefficient (mean 0.0882532) than high-insularity proteins (mean 0.213403), p = 1.59834e-110. The difference in betweenness centrality between high-insularity and low-insularity proteins is statistically insignificant; however, the partial correlation of betweenness and insularity given degree is -0.191978. This indicates that, after controlling for degree, there is a negative correlation between betweenness and insularity, as found in the other networks. The discrepancy is likely due to the fact that low-insularity proteins have significantly lower degree than high-insularity proteins; given the positive correlation between degree and essentiality, this confounds the trend of low-insularity proteins tending to have higher betweenness centrality than high-insularity proteins.

In the yeast HINT Combined network, low-insularity proteins have lower degree (mean 10.3329) than high-insularity proteins (mean 11.304178), p = 1.61849e-08. Low-insularity proteins have higher betweenness centrality (mean 0.00171632) than high-insularity proteins (mean 0.000716), p = 3.83494e-22. The effect holds if the top 5% of high-degree proteins are removed: p = 3.305e-24. Low-insularity proteins have lower local clustering coefficient (mean 0.112094) than high-insularity proteins (mean 0.481805), p = 4.17959e-138. The partial correlation of betweenness and insularity given degree is -0.393677.

In the human HINT Combined network, low-insularity proteins have lower degree (mean 7.61413) than high-insularity proteins (mean 8.570652), p = 3.19192e-07. Low-insularity proteins have slightly higher betweenness centrality (mean 0.00114638) than high-insularity proteins (mean 0.001025); while this result is not statistically significant (p = 0.104022), the general trend of low-insularity proteins having higher betweenness centrality matches that seen in other networks . The effect holds if the top

5% of high-degree proteins are removed, but it is still not statistically significant: p = 0.0930863. Nonetheless, the partial correlation of betweenness and insularity given degree is -0.206513, indicating that after controlling for the effects of degree, there is a negative correlation between betweenness centrality and functional insularity, matching the trend seen in other networks. Low-insularity proteins have lower local clustering coefficient (mean 0.0580448) than high-insularity proteins (mean 0.212471), p = 8.90233e-72.

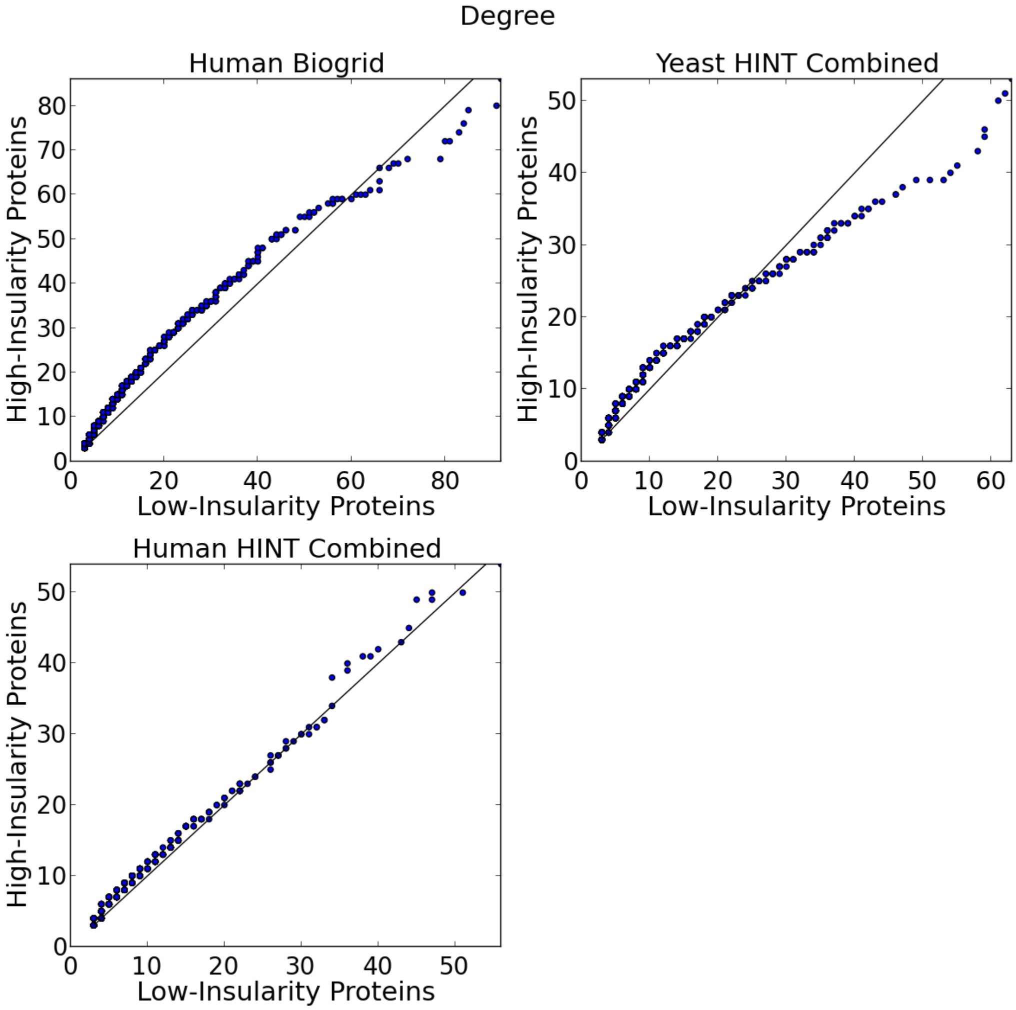


Figure A.2: Degree of low-insularity and high-insularity proteins in alternate networks.

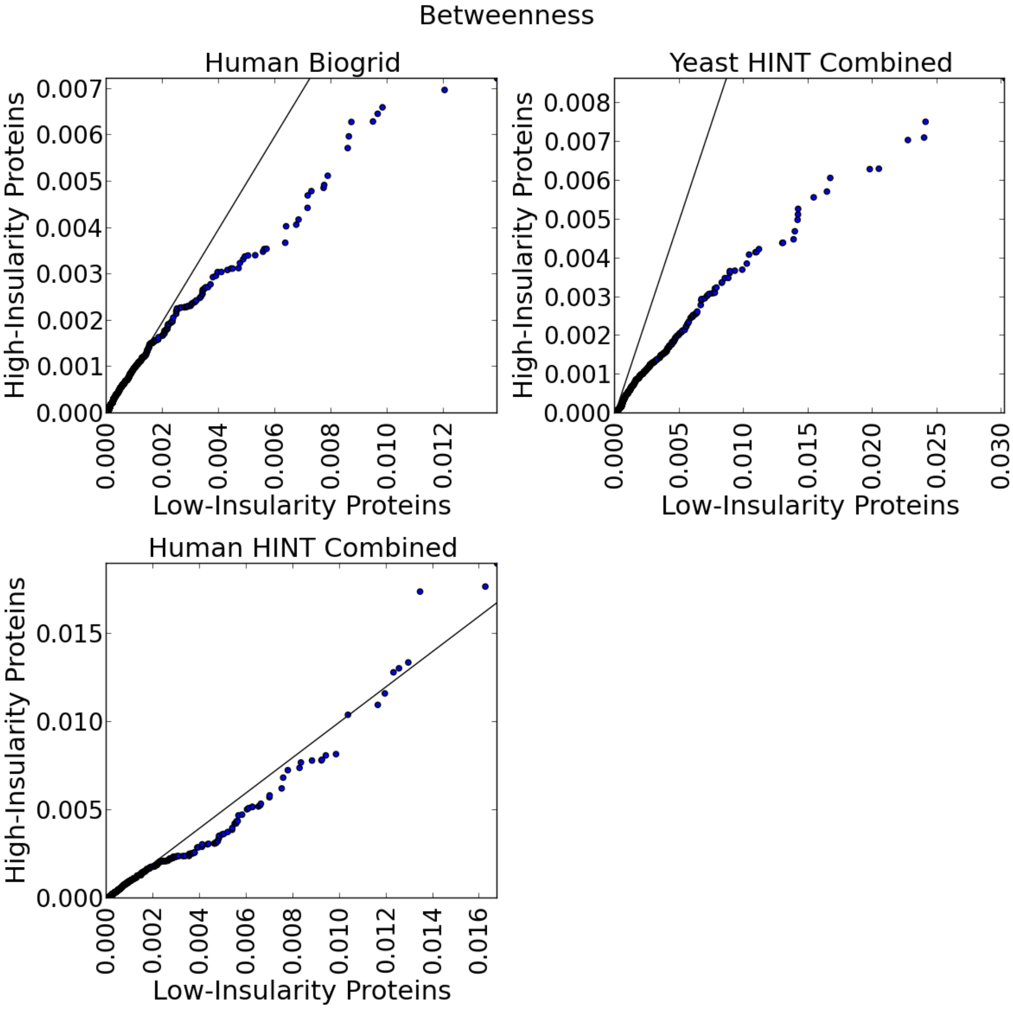


Figure A.3: Betweenness centrality of low-insularity and high-insularity proteins in alternate networks.

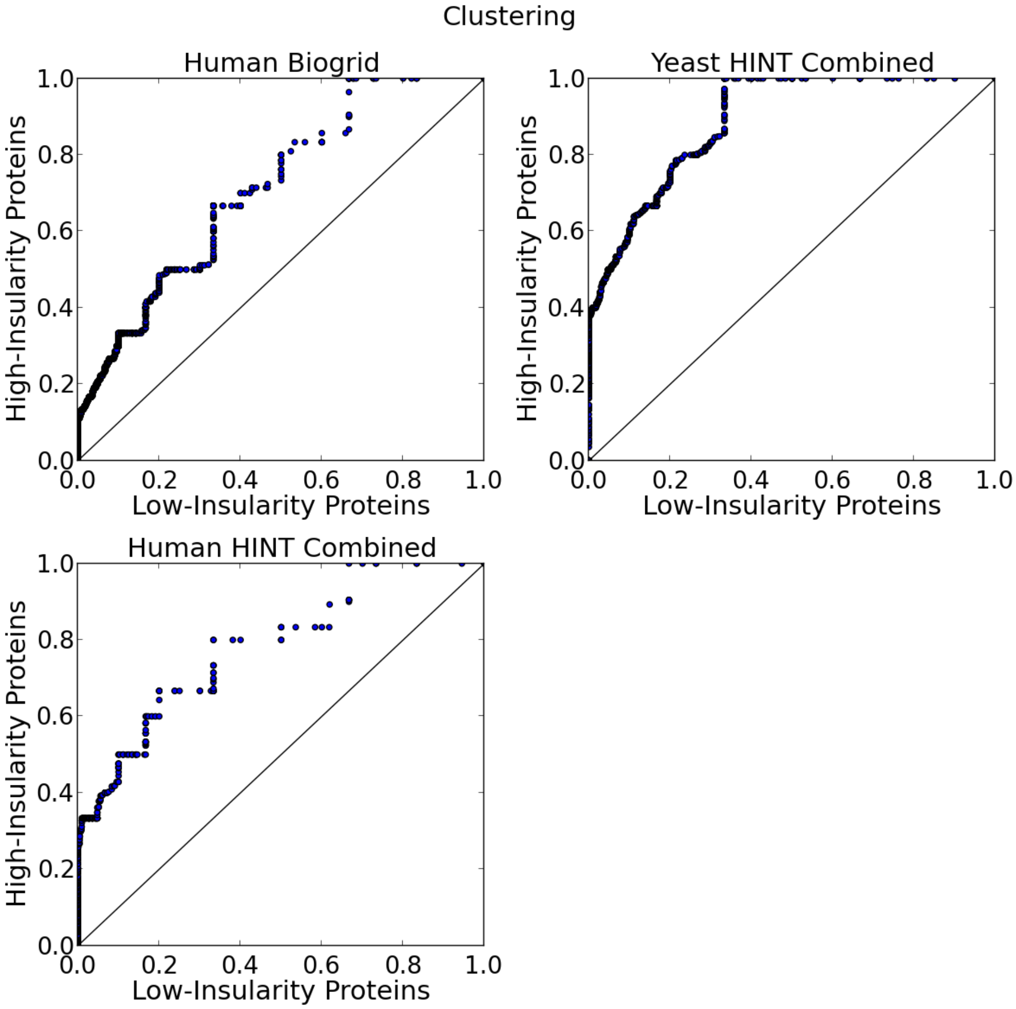


Figure A.4: Local clustering coefficient of low-insularity and high-insularity proteins in alternate networks.

A.1.3 Topological Protein Roles

In the human BioGRID network, low-insularity genes overlapped significantly with between-cluster connectors (p = 1.91407e-08) and within-cluster connectors (p = 0.0334395). High-insularity genes were significantly unenriched in between-cluster connectors (p = 3.4105e-16) and within-cluster connectors (p = 1.97445e-08). 70.804741% of the low-insularity proteins are covered by the two topological roles.

In the yeast HINT Combined network, low-insularity genes overlapped significantly with between-cluster connectors (p = 3.22373e-14) and within-cluster connectors (p = 7.37317e-16). High-insularity genes were significantly unenriched in between-cluster connectors (p = 3.45762e-13) and within-cluster connectors (p = 2.15748e-24). 56.527415% of the low-insularity proteins are covered by the two topological roles.

In the human HINT Combined network, low-insularity genes overlapped significantly with between-cluster connectors (p = 0.000500317); while the overlap with within-cluster connectors was statistically insignificant (p = 0.0520113), the enrichment matches the trend seen in other networks. High-insularity genes were significantly unenriched in between-cluster connectors (p = 5.89876e-07) and within-cluster connectors (p = 0.000114639). 53.668478% of the low-insularity proteins are covered by the two topological roles.

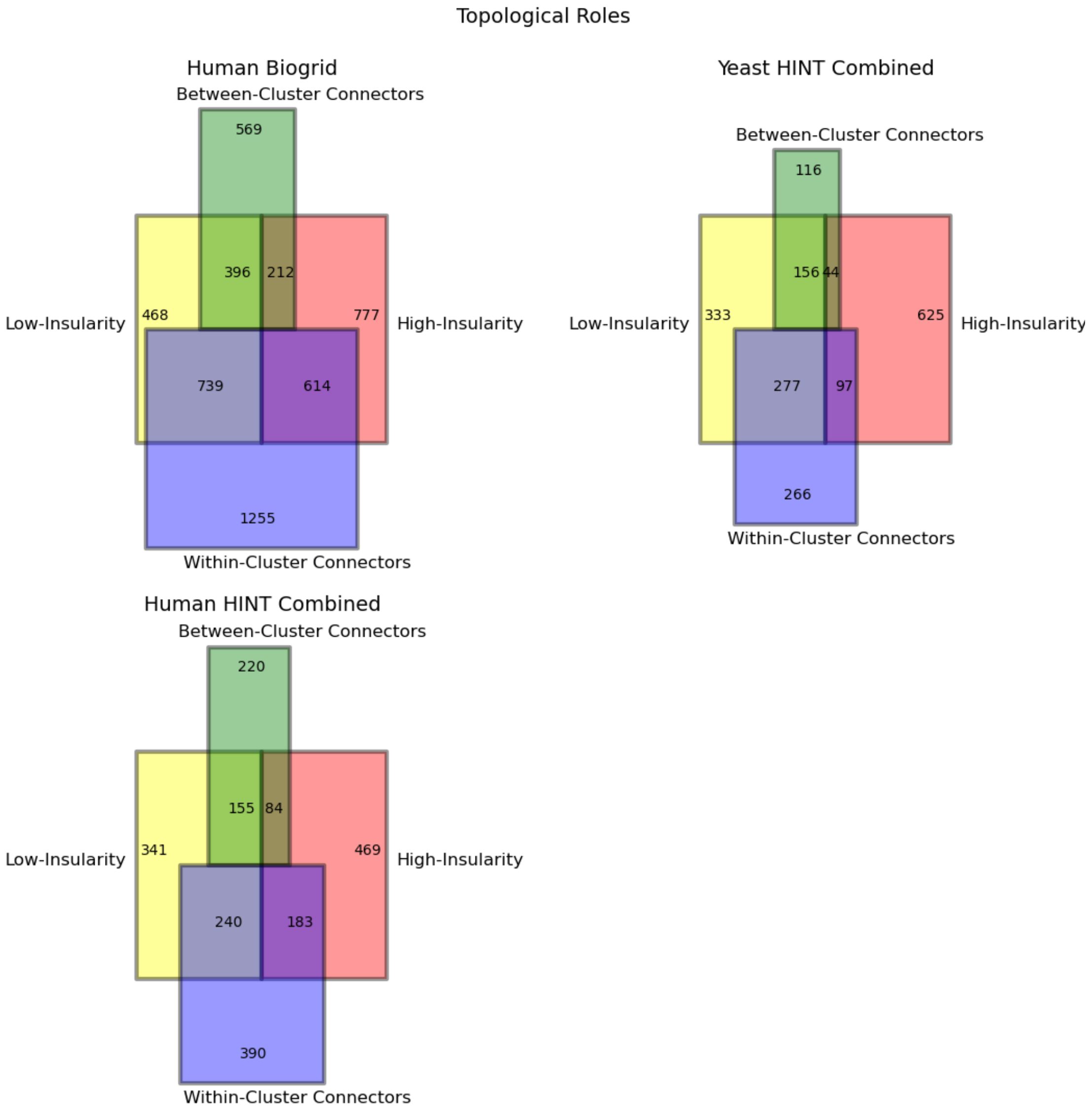


Figure A.5: Overlap between low/high-insularity proteins and the two topological roles defined in Figure 2.3 for alternate networks.

A.1.4 Targeted node removal

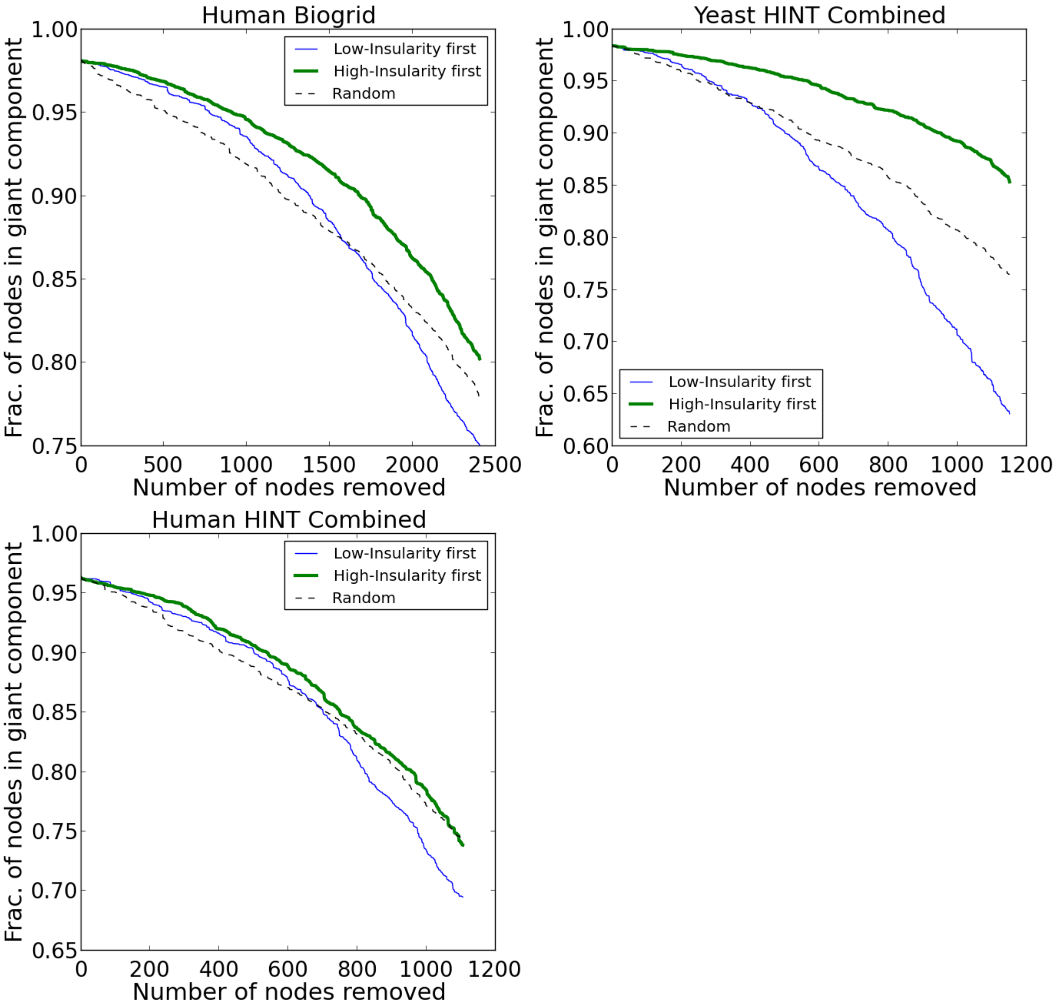


Figure A.6: Effect on largest connected component size of removing proteins in increasing order of insularity, decreasing order of insularity, and random order, for alternate networks. As in the network analyzed in the main body of this chapter, removal of low-insularity proteins has the largest effect on network connectivity.

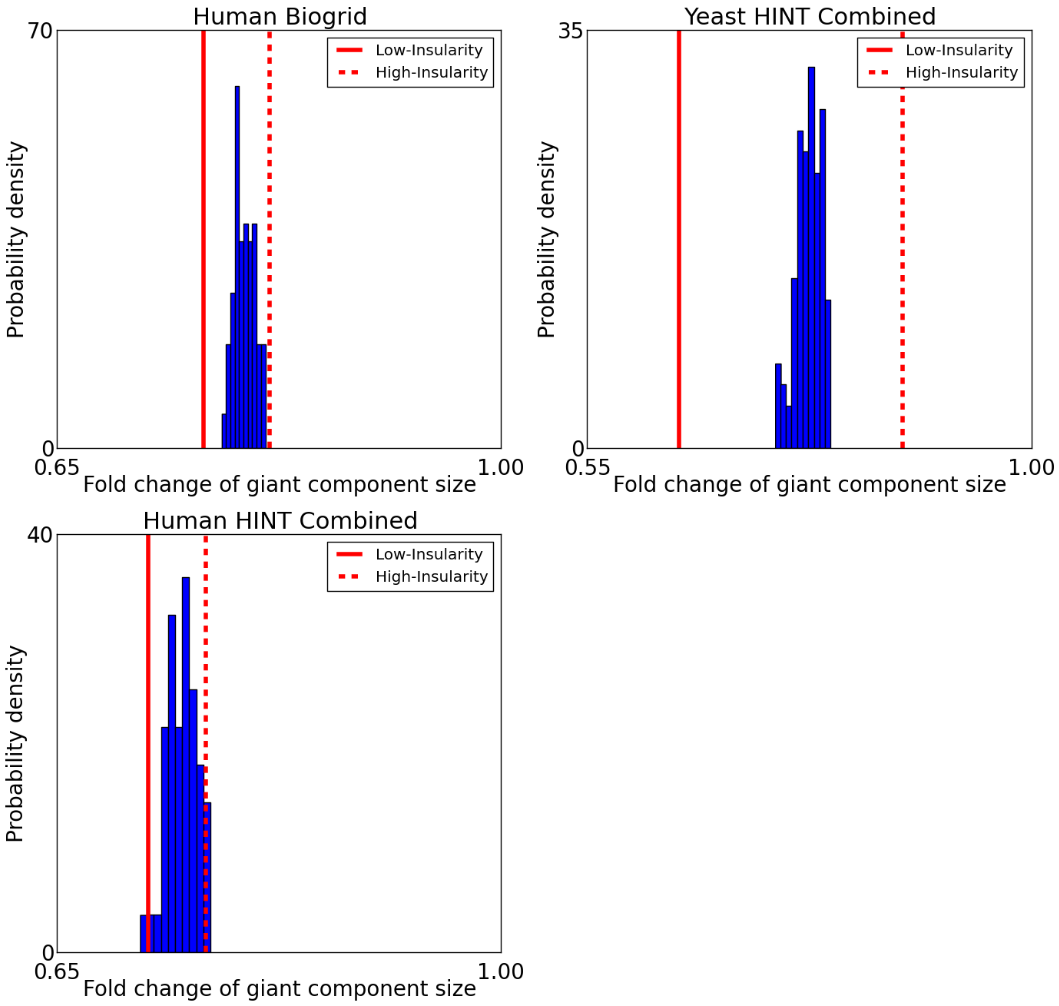


Figure A.7: Random trials, in which the 50% of scored proteins with highest (and lowest, respectively) insularities are removed, and the fold change of largest connected component size is compared to that resulting from removal of an equal number of random proteins. As expected, removal of low-insularity proteins has a larger effect than removal of high-insularity or random proteins.

A.1.5 Essentiality

In the human BioGRID network, low-insularity genes are less likely to be essential (26.178010% essential) than high-insularity genes (39.804241% essential) (p = 4.01062e-07, Fisher's exact test).

In the yeast HINT Combined network, low-insularity genes are less likely to be essential (19.843342% essential) than high-insularity genes (47.911227% essential) (p = 7.93432e-32, Fisher's exact test).

In the human HINT Combined network, low-insularity genes are less likely to be essential (31.538462% essential) than high-insularity genes (47.800587% essential) (p = 3.81196e-05, Fisher's exact test).

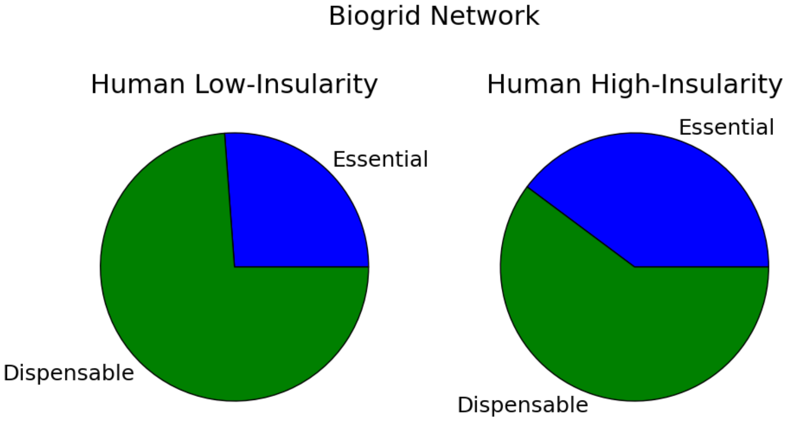


Figure A.8: Essentiality of low-insularity and high-insularity proteins in the human Biogrid network.

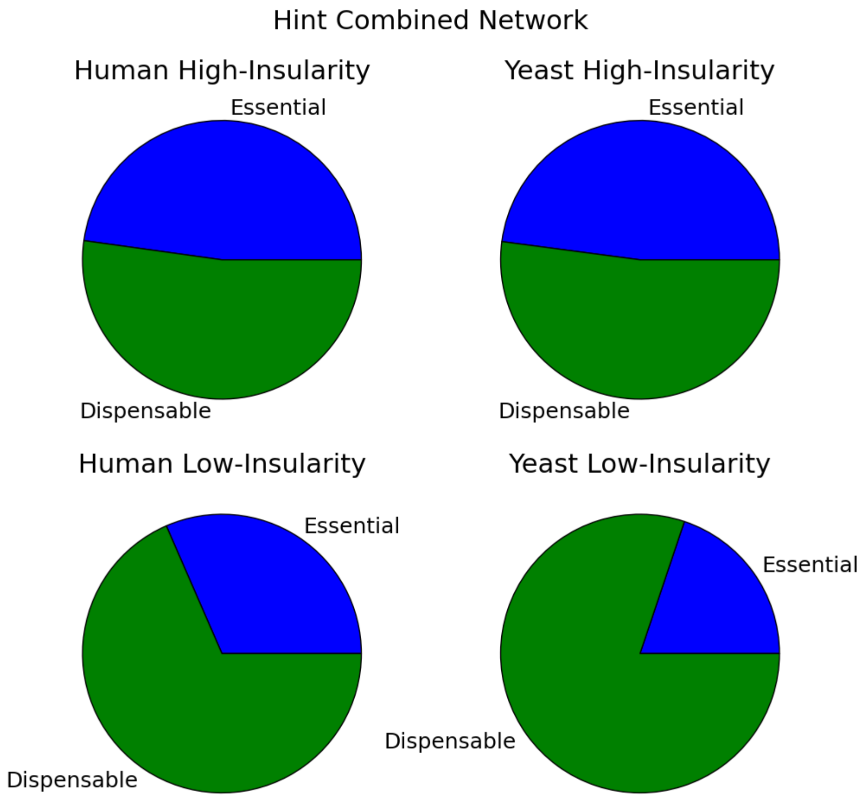


Figure A.9: Essentiality of low-insularity and high-insularity proteins in the yeast and human HINT Combined networks.

A.1.6 Regulators

In the yeast HINT Combined network, low-insularity genes had more regulators (mean 8.939073) than high-insularity genes (mean 5.958501) (p = 6.97321e-23).

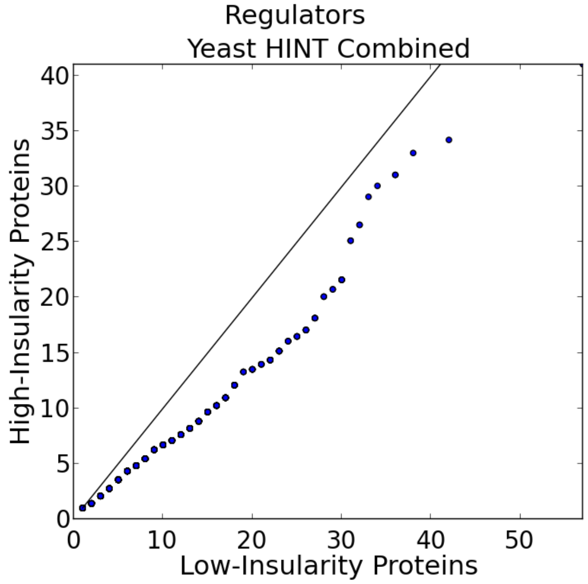


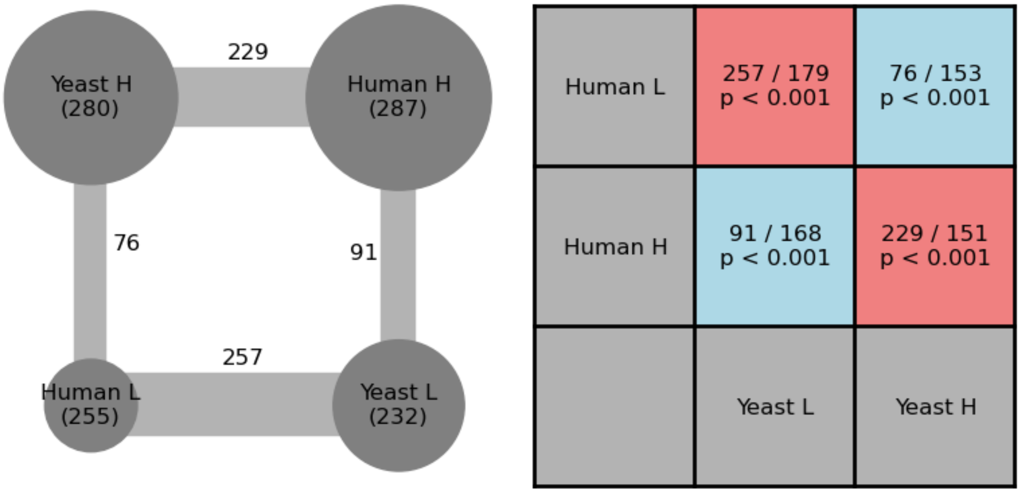
Figure A.10: Number of regulators of low-insularity and high-insularity genes in the yeast HINT Combined network.

A.2 Homology results with alternate PPOD homology datasets

When using the MultiParanoid homology dataset, the Spearman correlation between the functional insularities of homologous genes in yeast and human was 0.504179. There were significantly fewer cross-class homology relationships than expected by random chance ($p < 0.001$).

When using the Jaccard homology dataset, the Spearman correlation between the functional insularities of homologous genes in yeast and human was 0.193708. There were significantly fewer cross-class homology relationships than expected by random chance ($p < 0.001$). As shown by Figure A.12, there are a larger number of homology relationships between human high-insularity and yeast low-insularity proteins than in the other two homology datasets. However, the stub-rewiring analysis shows that there are still significantly fewer of these relationships than expected by random chance, indicating that the large number of these relationships is due mainly to the large number of human high-insularity and yeast low-insularity proteins in the analysis.

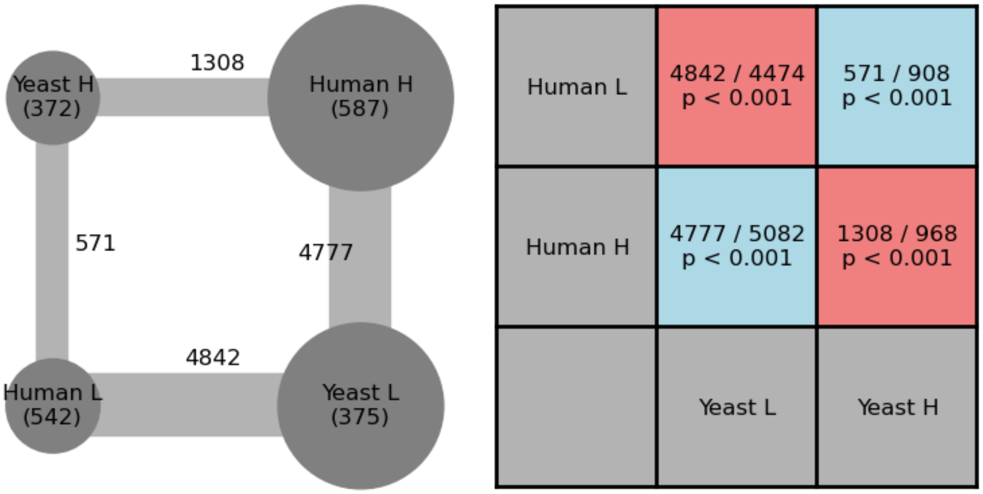
Orthology: Multiparanoid Dataset



Human L	257 / 179 p < 0.001	76 / 153 p < 0.001
Human H	91 / 168 p < 0.001	229 / 151 p < 0.001
	Yeast L	Yeast H

Figure A.11: Yeast-human homology relationships between high-insularity and low-insularity proteins when using the MultiParanoid homology dataset from PPOD.

Orthology: Jaccard Dataset



Human L	4842 / 4474 p < 0.001	571 / 908 p < 0.001
Human H	4777 / 5082 p < 0.001	1308 / 968 p < 0.001
	Yeast L	Yeast H

Figure A.12: Yeast-human homology relationships between high-insularity and low-insularity proteins when using the Jaccard homology dataset from PPOD.

A.3 Alternate Clustering Evaluation Results

In the human Biogrid network using the topological modularity evaluation, a SPICi clustering of the original network had modularity 0.310217, while the clustering of the low-insularity free network had modularity 0.350689. Clusterings of networks with random node removal had a mean modularity of 0.332389, and a standard deviation of 0.00721574.

In the human Biogrid network using the semantic density evaluation with GO, a SPICi clustering of the original network had modularity 0.0350689, while the clustering of the low-insularity free network had modularity 0.0454128. Clusterings of networks with random node removal had a mean modularity of 0.0277216, and a standard deviation of 0.00163421.

In the yeast Biogrid network using the topological modularity evaluation, an MCL clustering of the original network had modularity 0.260249, while the clustering of the low-insularity free network had modularity 0.397892. Clusterings of networks with random node removal had a mean modularity of 0.281492, and a standard deviation of 0.00941433.

In the yeast Biogrid network using the semantic density evaluation with GO, a MCL clustering of the original network had modularity 0.0636584, while the clustering of the low-insularity free network had modularity 0.0987579. Clusterings of networks with random node removal had a mean modularity of 0.0563868, and a standard deviation of 0.00276532.

In the human Biogrid network using the topological modularity evaluation, an MCL clustering of the original network had modularity 0.256511, while the clustering of the low-insularity free network had modularity 0.302612. Clusterings of networks with random node removal had a mean modularity of 0.322006, and a standard deviation of 0.00559525. This is not consistent with results for other networks and clustering algorithms, as the networks with random node removal had higher modularity than the low-insularity-free network. However, it matches the trend of the low-insularity-free network having higher modularity than the original network, again suggesting that low-insularity proteins decrease the quality of network clustering results.

In the human Biogrid network using the semantic density evaluation with GO, an MCL clustering of the original network had modularity 0.0154808, while the clustering of the low-insularity free network had modularity 0.0200427. Clusterings of networks with random node removal had a mean modularity of 0.0137017, and a standard deviation of 0.00106182.

A.4 Full GO TermFinder results

Table A.1: Enriched Biological Process terms in yeast low-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
single-organism metabolic process	24.6%	14.1%	2e-33
small molecule metabolic process	18.9%	10.2%	1.6e-30
carboxylic acid metabolic process	11.0%	5.4%	1.2e-22
organic acid metabolic process	11.3%	5.6%	4.9e-22
oxoacid metabolic process	11.3%	5.6%	4.9e-22
organonitrogen compound metabolic process	14.7%	8.1%	1.6e-21
phosphate-containing	15.7%	9.6%	2.1e-16

compound metabolic process			
phosphorus metabolic process	16.2%	10.1%	1.1e-15

Table A.2: Enriched Biological Process terms in yeast high-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
gene expression	41.6%	24.1%	1.5e-60
RNA metabolic process	30.7%	18.6%	5.6e-35
RNA processing	21.0%	11.6%	8.9e-32
cellular macromolecule metabolic process	62.2%	48.7%	1.3e-28
macromolecule metabolic process	62.7%	49.4%	2.8e-28
nucleic acid metabolic process	39.0%	27.4%	1.6e-25
cellular macromolecule biosynthetic process	29.0%	19.1%	5.1e-24
macromolecule biosynthetic process	29.1%	19.2%	6e-24
ncRNA processing	13.9%	7.6%	1.4e-20
rRNA processing	11.0%	5.7%	2.7e-19
nucleobase-containing compound metabolic process	40.3%	30.2%	3.1e-19
rRNA metabolic process	11.4%	6.0%	3.5e-19

cellular aromatic compound metabolic process	40.8%	31.0%	6.4e-18
cellular nitrogen compound metabolic process	41.8%	32.0%	1.2e-17
organic cyclic compound metabolic process	42.2%	32.4%	1.4e-17
heterocycle metabolic process	40.9%	31.3%	1.9e-17
ribonucleoprotein complex biogenesis	16.9%	10.4%	2.9e-17
ncRNA metabolic process	14.0%	8.7%	4e-14
cellular component biogenesis	29.1%	21.5%	4.6e-14
cellular biosynthetic process	34.6%	26.5%	8.2e-14
translation	13.3%	8.2%	1.3e-13
cellular metabolic process	71.4%	62.8%	1.5e-13
primary metabolic process	68.4%	59.7%	1.9e-13
mRNA metabolic process	11.0%	6.5%	3.6e-13
organic substance metabolic process	69.8%	61.4%	6.3e-13
nitrogen compound metabolic process	43.3%	35.0%	1.1e-12

ribosome biogenesis	13.4%	8.5%	2.4e-12
protein targeting	10.8%	6.5%	6.4e-12
metabolic process	72.0%	64.2%	8.5e-12
organic substance biosynthetic process	34.7%	27.4%	1.9e-11
mitochondrion organization	10.6%	6.7%	2e-10
cellular component organization or biogenesis	52.8%	45.2%	2.3e-10
biosynthetic process	34.9%	27.9%	2.4e-10
cellular macromolecular complex assembly	13.4%	8.9%	2.8e-10
macromolecular complex assembly	14.8%	10.2%	5.5e-10
intracellular protein transport	11.4%	7.5%	2.7e-09
cellular process	94.1%	90.1%	1.2e-08
protein transport	11.5%	7.8%	1.5e-08
nucleobase-containing compound biosynthetic process	11.6%	7.9%	1.8e-08
establishment of protein localization	11.9%	8.2%	8.6e-08
protein localization to organelle	11.3%	7.8%	2.3e-07

organic cyclic compound biosynthetic process	13.5%	9.8%	4e-07
aromatic compound biosynthetic process	12.1%	8.6%	5.3e-07
cellular nitrogen compound biosynthetic process	12.4%	9.1%	2.6e-06
cytoplasmic transport	16.4%	12.7%	3.4e-06
heterocycle biosynthetic process	12.2%	9.0%	4.2e-06
cellular component assembly	17.2%	13.6%	1.5e-05
macromolecular complex subunit organization	18.2%	14.6%	2.2e-05
intracellular transport	18.6%	15.4%	0.0002
cellular protein localization	12.6%	10.0%	0.00025
cellular component organization	41.8%	38.1%	0.00092
cellular macromolecule localization	13.0%	10.6%	0.0012
protein localization	13.1%	11.1%	0.0058
organic substance transport	13.0%	11.1%	0.0063
biological process	97.7%	96.6%	0.0087
establishment of	19.0%	16.9%	0.01

localization in cell			

Table A.3: Enriched Biological Process terms in human low-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
small molecule metabolic process	15.8%	11.6%	1.6e-10
single-organism metabolic process	17.7%	13.6%	6.6e-09

Table A.4: Enriched Biological Process terms in human high-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
nucleic acid metabolic process	38.2%	21.4%	1.7e-86
RNA metabolic process	29.2%	15.3%	1.6e-75
gene expression	31.1%	16.7%	6.2e-75
nucleobase-containing compound metabolic process	39.5%	24.1%	2.3e-67
heterocycle metabolic process	39.6%	24.5%	4.6e-64
cellular aromatic compound metabolic process	39.5%	24.5%	3e-63
organic cyclic compound metabolic process	39.8%	25.2%	4.3e-59
cellular nitrogen	39.9%	25.4%	5.5e-58

compound metabolic process			
macromolecule metabolic process	58.4%	42.4%	6.6e-57
cellular macromolecule metabolic process	56.4%	40.5%	2.9e-56
nitrogen compound metabolic process	40.4%	26.7%	5.3e-51
cellular macromolecule biosynthetic process	24.1%	13.7%	8.7e-48
transcription, DNA dependent	16.1%	7.9%	1.7e-47
macromolecule biosynthetic process	24.1%	14.0%	1.4e-44
mRNA metabolic process	14.5%	7.0%	3.8e-43
RNA biosynthetic process	17.2%	9.0%	8.4e-42
transcription from RNA polymerase II promoter	13.6%	6.6%	2.5e-40
primary metabolic process	60.8%	47.9%	4.1e-37
nucleobase-containing compound biosynthetic process	17.7%	9.8%	6.4e-37
heterocycle biosynthetic process	17.7%	9.9%	3.1e-35
organic substance metabolic process	61.6%	49.0%	3.2e-35
aromatic compound	17.7%	10.0%	9e-35

biosynthetic process			
metabolic process	63.5%	51.1%	1.1e-34
cellular nitrogen compound biosynthetic process	17.7%	10.1%	6e-33
cellular metabolic process	60.4%	48.3%	8.2e-33
organic cyclic compound biosynthetic process	17.8%	10.2%	8.7e-33
RNA processing	11.5%	5.7%	4.6e-32
cellular biosynthetic process	25.0%	16.3%	1.4e-29
organic substance biosynthetic process	25.2%	16.8%	2.6e-27
biosynthetic process	25.3%	17.0%	4.4e-26
cellular macromolecule catabolic process	13.2%	8.2%	1.3e-18
DNA metabolic process	11.9%	7.3%	6.7e-17
regulation of RNA metabolic process	28.8%	21.8%	8.4e-17
regulation of macromolecule biosynthetic process	30.8%	23.8%	7.4e-16
regulation of macromolecule metabolic process	41.4%	33.7%	1.5e-15
regulation of RNA	27.7%	21.1%	3.2e-15

biosynthetic process			
positive regulation of macromolecule metabolic process	25.6%	19.3%	3.8e-15
macromolecule catabolic process	13.4%	8.8%	7.6e-15
regulation of nitrogen compound metabolic process	34.4%	27.4%	1.1e-14
regulation of cellular macromolecule biosynthetic process	29.5%	22.9%	1.4e-14
positive regulation of cellular metabolic process	26.0%	19.7%	1.7e-14
positive regulation of RNA metabolic process	15.5%	10.6%	1.8e-14
regulation of cellular biosynthetic process	31.3%	24.7%	5.8e-14
chromosome organization	10.5%	6.7%	1.2e-13
regulation of primary metabolic process	42.0%	34.8%	1.4e-13
regulation of gene expression	31.3%	24.7%	1.7e-13
regulation of biosynthetic process	31.4%	24.9%	2e-13
response to DNA damage	10.4%	6.6%	3.4e-13

stimulus			
cellular process	88.1%	82.7%	4.2e-13
positive regulation of nucleobase-containing compound metabolic process	16.3%	11.6%	8.5e-13
positive regulation of metabolic process	26.3%	20.4%	8.7e-13
regulation of transcription, DNA dependent	26.4%	20.6%	2.5e-12
positive regulation of nitrogen compound metabolic process	16.3%	11.7%	5e-12
regulation of cellular metabolic process	42.1%	35.5%	9.8e-12
positive regulation of macromolecule biosynthetic process	16.0%	11.6%	1.3e-11
regulation of nucleobase containing compound metabolic process	32.1%	26.3%	1.3e-10
positive regulation of cellular biosynthetic process	16.3%	12.0%	2.4e-10
positive regulation of	16.3%	12.2%	1.2e-09

biosynthetic process			
regulation of metabolic process	44.8%	38.9%	2.2e-09
positive regulation of transcription, DNA dependent	13.3%	9.8%	3.9e-09
multi-organism process	12.9%	9.4%	5.7e-09
positive regulation of gene expression	13.8%	10.3%	1.8e-08
cellular response to stress	15.1%	11.5%	3.6e-08
cellular catabolic process	15.5%	11.9%	6.4e-08
regulation of transcription from RNA polymerase II promoter	15.1%	11.6%	7.7e-08
positive regulation of protein modification process	10.8%	8.3%	1.2e-05
catabolic process	15.6%	12.8%	2.8e-05
organic substance catabolic process	14.9%	12.2%	3.4e-05
positive regulation of cellular protein metabolic process	11.7%	9.3%	3.6e-05
negative regulation of macromolecule metabolic	16.7%	13.8%	4.9e-05

process			
negative regulation of cellular metabolic process	16.5%	13.8%	0.00014
macromolecule modification	21.0%	18.1%	0.00015
positive regulation of biological process	35.8%	32.4%	0.00021
positive regulation of cellular process	33.4%	30.2%	0.00034
positive regulation of protein metabolic process	11.9%	9.8%	0.00037
cellular protein modification process	20.0%	17.4%	0.00054
protein modification process	20.0%	17.4%	0.00054
negative regulation of metabolic process	16.9%	14.5%	0.00057
cell cycle	13.0%	10.9%	0.00058
protein metabolic process	28.3%	25.4%	0.00068
negative regulation of RNA metabolic process	10.1%	8.3%	0.00069
cellular protein metabolic process	26.6%	23.8%	0.00087
negative regulation of gene expression	10.4%	8.5%	0.001
negative regulation of	10.9%	9.1%	0.0017

nucleobase-containing compound metabolic process			
negative regulation of cellular macromolecule biosynthetic process	10.9%	9.2%	0.002
negative regulation of nitrogen compound metabolic process	10.9%	9.2%	0.0025
cell cycle process	10.9%	9.2%	0.003
positive regulation of catalytic activity	11.9%	10.2%	0.0033
negative regulation of macromolecule biosynthetic process	11.1%	9.5%	0.0047
organelle organization	18.6%	16.7%	0.0081

Table A.5: Enriched Cellular Component terms in yeast low-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
cell periphery	13.0%	8.9%	6.6e-09
cytoplasm	72.7%	66.3%	1.3e-08
mitochondrion	21.6%	18.1%	0.00011

Table A.6: Enriched Cellular Component terms in yeast high-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value

macromolecular complex	77.6%	51.0%	1.4e-109
organelle part	74.3%	54.4%	4.2e-62
intracellular organelle part	74.2%	54.3%	6.7e-62
protein complex	53.2%	37.6%	3.3e-39
ribonucleoprotein complex	27.4%	15.6%	8.6e-39
nuclear part	40.5%	26.6%	5.1e-37
ribosomal subunit	13.0%	6.3%	8.3e-27
organelle lumen	35.5%	24.4%	1.3e-25
intracellular organelle lumen	35.5%	24.4%	1.3e-25
membrane-enclosed lumen	36.4%	25.3%	6.1e-25
nuclear lumen	28.6%	19.6%	6.1e-20
nucleoplasm part	13.0%	7.2%	6e-19
nucleoplasm	13.7%	7.9%	2.6e-17
ribosome	13.9%	8.2%	4.5e-17
intracellular organelle	87.4%	79.5%	8.8e-17
organelle	87.4%	79.5%	1.1e-16
nucleolus	11.2%	6.4%	1.7e-14
membrane-bounded organelle	79.2%	71.2%	1.3e-13
intracellular membrane bounded organelle	79.2%	71.2%	1.3e-13
mitochondrial part	13.9%	9.4%	6.3e-10
intracellular part	95.3%	91.8%	7.1e-08
non-membrane-bounded	35.0%	29.1%	8.8e-08

organelle			
intracellular non membrane-bounded organelle	35.0%	29.1%	8.8e-08
intracellular	95.4%	92.1%	1.4e-07
nucleus	50.9%	44.7%	1.8e-07
organelle membrane	17.6%	13.4%	3.5e-07
membrane part	22.6%	18.1%	1.4e-06
cell	96.5%	94.3%	3.1e-05
cell part	96.5%	94.3%	3.1e-05
cytosol	16.0%	13.2%	0.00047
endomembrane system	10.4%	8.3%	0.0014
organelle envelope	10.5%	8.5%	0.0034
envelope	10.5%	8.5%	0.0034
intrinsic to membrane	13.0%	11.1%	0.0071
integral to membrane	13.0%	11.0%	0.0078
cellular component	97.7%	96.6%	0.0087

No enriched Cellular Component terms were found in human low-insularity proteins.

Table A.7: Enriched Cellular Component terms in human high-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
nucleoplasm	33.3%	18.8%	6.6e-71
nuclear part	42.0%	26.3%	3e-66
nuclear lumen	37.6%	23.3%	2.7e-59
intracellular organelle lumen	39.2%	26.0%	4.5e-48

membrane-enclosed lumen	39.6%	26.9%	7.1e-44
organelle lumen	39.4%	26.7%	9.8e-44
macromolecular complex	45.6%	32.4%	1.3e-42
protein complex	40.4%	28.3%	1.1e-38
nucleus	54.8%	42.3%	5.7e-35
nucleoplasm part	14.6%	7.9%	2e-31
intracellular organelle part	55.6%	44.6%	1.5e-27
organelle part	56.0%	45.2%	1.2e-26
intracellular membrane bounded organelle	67.0%	57.9%	6.5e-20
membrane-bounded organelle	67.1%	58.2%	6.9e-19
intracellular organelle	70.6%	64.2%	2.2e-11
organelle	70.7%	64.4%	7.4e-11
intracellular part	80.2%	76.6%	1.5e-05
cell	86.8%	83.7%	2.3e-05
cell part	86.8%	83.7%	2.3e-05
intracellular	80.2%	77.0%	8e-05
cellular component	88.6%	86.5%	0.001

Table A.8: Enriched Molecular Function terms in yeast low-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
catalytic activity	47.9%	38.6%	6.1e-15
transferase activity	16.9%	13.2%	5.4e-06

Table A.9: Enriched Molecular Function terms in yeast high-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
structural constituent of ribosome	12.6%	5.9%	2.6e-29
structural molecule activity	16.1%	9.4%	9.5e-20
RNA binding	10.5%	7.4%	7e-07
nucleic acid binding	18.8%	16.0%	0.0014
molecular function	97.7%	96.6%	0.0087

No enriched Molecular Function terms were found in human low-insularity proteins.

Table A.10: Enriched Molecular Function terms in human high-insularity proteins.

Term	Enrichment freq.	Background freq.	P-value
protein binding transcription factor activity	10.2%	6.2%	1e-15
nucleic acid binding	18.5%	13.5%	2.9e-12
DNA binding	12.1%	9.2%	7.9e-07
heterocyclic compound binding	21.0%	17.8%	2.2e-05
organic cyclic compound binding	21.1%	18.1%	7.5e-05

Bibliography

[1] Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nature Reviews Genetics* 2004, **5**:101–113.

[2] Uetz P, et al.: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae***. *Nature* 2000, **403**:623–627.

[3] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proceedings of the National Academy of Sciences* 2001, **98**(8):4569–4574.

[4] Fields S, Song OK: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**:245–246.

[5] Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Sraphin B: **The Tandem Affinity Purification (TAP) method: A general procedure of protein complex purification**. *Methods* 2001, **24**(3):218 – 229.

[6] Stark C, et al.: **The BioGRID Interaction Database: 2011 update**. *Nucleic Acids Research* 2011, **39**:D698–D704.

[7] Giancotti FG, Ruoslahti E: **Integrin signaling**. *Science* 1999, **285**(5430):1028–1033, [http://www.sciencemag.org/content/285/5430/1028.abstract].

[8] Barabási AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**(5439):509–512.

[9] Pandey J, Koyuturk M, Grama A: **Functional characterization and topological modularity of molecular interaction networks**. *BMC Bioinformatics* 2010, **11**(Suppl 1):S35.

[10] Freeman LC: **A set of measures of centrality based on betweenness**. *Sociometry* 1977, **40**:35–41.

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

[12] Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M: **The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics**. *PLoS Computational Biology* 2007.

[13] Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**(5765):1283–1287.

[14] Ohno S: *Evolution by gene duplication.* New York: Springer-Verlag 1970.

[15] Zhang J: **Evolution by gene duplication: an update**. *Trends in Ecology & Evolution* 2003, **18**(6):292–298.

[16] Hughes AL: **The evolution of functionally novel proteins after gene duplication**. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 1994, **256**(1346):119–124.

[17] Long M, Betrán E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old**. *Nature Reviews Genetics* 2003, **4**:265–875.

[18] Capra JA, Pollard KS, Singh M: **Novel genes exhibit distinct patterns of function acquisition and network integration**. *Genome Biology* 2010, **11**:R127.

[19] Qin H, Lu HHS, Wu WB, Li WH: **Evolution of the yeast protein interaction network**. *Proceedings of the National Academy of Sciences* 2003, **100**(22):12820–12824.

[20] Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network**. *Science* 2002, **296**(5568):750–752.

[21] Zotenko E, Mestre J, O'Leary DP, Przytycka TM: **Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality**. *PLoS Computational Biology* 2008, **4**(8):e1000140.

[22] Missiuro PV, et al.: **Information flow analysis of interactome networks**. *PLoS Computational Biology* 2009, **5**(4).

[23] Milo R, et al.: **Network motifs: simple building blocks of complex networks**. *Nature Genetics* 2002, **298**:824–827.

[24] Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**:C47–52.

[25] Wagner GP, Pavlicev M, Cheverud JM: **The road to modularity**. *Nature Reviews Genetics* 2007, **8**:921–931.

[26] Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks**. *Proceedings of the National Academy of Sciences* 2003, **100**(21):12123–12128.

[27] Rives AW, Galitski T: **Modular organization of cellular networks**. *Proceedings of the National Academy of Sciences* 2003, **100**(3):1128–1133.

[28] Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast**. *Nature Biotechnology* 2000, **18**:1257–1261.

[29] Song J, Singh M: **From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization**. *PLoS Computational Biology* 2013, **9**(2):e1002910.

[30] **Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: Large-scale organization and robustness**.

[31] Freeman LC, Borgatti SP, White DR: **Centrality in valued graphs: A measure of betweenness based on network flow**. *Social Networks* 1991, **13**:141–154.

[32] Das J, Yu H: **HINT: High-quality protein interactomes and their applications in understanding human disease**. *BMC Systems Biology* 2012, **6**(92).

[33] The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25–29.

[34] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED: **Saccharomyces Genome Database: the genomics resource of budding yeast**. *Nucleic Acids Research* 2012, **40**(D1):D700–D705.

[35] Teixeira MC, et al.: **The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae***. *Nucleic Acids Research* 2006, **34**:D446–D451.

[36] Monteiro PT, et al.: **YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in *Saccharomyces cerevisiae***. *Nucleic Acids Research* 2008, **36**:D132–D136.

[37] Heinicke S, et al.: **The Princeton Protein Orthology Database (P-POD): A comparative genomics analysis tool for biologists**. *PLoS ONE* 2007, **2**(8):e766.

[38] Li L, Christian J Soeckert J, Roos DS: **OrthoMCL: Identification of ortholog groups for eukaryotic genomes**. *Genome Research* 2003, **13**:2178–2189.

[39] Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL: **Automatic clustering of orthologs and inparalogs shared by multiple proteomes**. *Bioinformatics* 2006, **22**:e9–e15.

[40] Resnik P: **Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language**. *Journal of Artificial Intelligence Research* 1999, **11**:95–130.

[41] Jiang P, Singh M: **SPICi: A fast clustering algorithm for large biological networks**. *Bioinformatics* 2010, **4**(10).

[42] Newman MEJ, Girvan M: **Finding and evaluating community structure in networks**. *Physical Review E* 2004, **69**.

[43] van Dongen S: **Graph clustering by flow simulation**. *PhD thesis*, University of Utrecht 2000.

[44] Enright AJ, Dongen SV, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Research* 2002, **30**(7):1575–1584.

[45] Hagberg AA, Schult DA, Swart PJ: **Exploring network structure, dynamics, and function using NetworkX**. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Edited by Varoquaux G, Vaught T, Millman J, Pasadena, CA, USA 2008:11–15.

[46] Boyle EI, et al.: **GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**(18):3710–3715.

[47] Hunter JD: **Matplotlib: A 2D graphics environment**. *Computing In Science & Engineering* 2007, **9**(3):90–95.

[48] Myers C, Barrett D, Hibbs M, Huttenhower C, Troyanskaya O: **Finding function: Evaluation methods for functional genomic data**. *BMC Genomics* 2006, **7**(187).

[49] Song J, Singh M: **How and when should interactome-derived clusters be used to predict functional modules and protein function?** *Bioinformatics* 2009, **25**(23):3143–3150.

[50] Guimerà R, Amaral LAN: **Functional cartography of complex metabolic networks**. *Nature* 2005.

[51] Wagner A: **The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes**. *Molecular Biology and Evolution* 2001, **18**(7):1283–1292.

[52] Han JDJ, et al.: **Evidence for dynamically organized modularity in the yeast protein–protein interaction network**. *Nature* 2004.

[53] Komurov K, White M: **Revealing static and dynamic modular architecture of the eukaryotic protein interaction network**. *Molecular Systems Biology* 2007, **3**(110).

[54] Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W: **On the origin of new genes in *Drosophila***. *Genome Research* 2008, **18**(9):1446–1455.

[55] Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ: **Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression**. *Proceedings of the National Academy of Sciences* 2006, **103**(26):9935–9939.

[56] Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Mar Albà M: **Origin of primate orphan genes: A comparative genomics approach**. *Molecular Biology and Evolution* 2009, **26**(3):603–612.

[57] Cai J, Zhao R, Jiang H, Wang W: **De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae***. *Genetics* 2008, **179**:487–496.

[58] Katju V, Lynch M: **The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome**. *Genetics* 2003, **165**(4):1793–1803.

[59] Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution**. *Genome Biology* 2006, **7**(5):R43.

[60] Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**:459–473.

[61] Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW: **Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis***. *Molecular Biology and Evolution* 2006, **23**(2):469–478.

[62] Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW: **Adaptive evolution of young gene duplicates in mammals**. *Genome Research* 2009, **19**(5):859–867.

[63] Milinkovitch MC, Helaers R, Tzika AC: **Historical Constraints on Vertebrate Genome Evolution**. *Genome Biology and Evolution* 2010, **2**:13–18.

[64] Warnefors M, Eyre-Walker A: **The Accumulation of Gene Regulation Through Time**. *Genome Biology and Evolution* 2011, **3**:667–673, [`http://gbe.oxfordjournals.org/content/3/667.abstract`].

[65] Cai JJ, Borenstein E, Chen R, Petrov DA: **Similarly Strong Purifying Selection Acts on Human Disease Genes of All Evolutionary Ages**. *Genome Biology and Evolution* 2009, **1**:131–144.

[66] Finn RD, et al.: **The Pfam protein families database**. *Nucleic Acids Research* 2010, **38**:D211–D222.

[67] Heinicke S, et al.: **The Princeton Protein Orthology Database (P-POD): A comparative genomics analysis tool for biologists**. *PLoS ONE* 2007, **2**:e766.

[68] Le Quesne WJ: **The uniquely evolved character concept and its cladistic application**. *Systematic Biology* 1974, **23**(4):513–517.

[69] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Research* 2009, **37**(suppl 1):D5–D15.

[70] Hedges SB, Dudley J, Kumar S: **TimeTree: a public knowledge-base of divergence times among organisms**. *Bioinformatics* 2006, **22**(23):2971–2972.

[71] Capra J, Williams A, Pollard K: **ProteinHistorian: tools for the comparative analysis of eukaryote protein origin**. *PLoS Computational Biology* 2012, **8**:e1002567.

[72] Chen K, Durand D, Farach-Colton M: **NOTUNG: a program for dating gene duplications and optimizing gene family trees**. *Journal of Computational Biology* 2000, **7**:429–47.

[73] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proceedings of the National Academy of Sciences* 2004, **101**(16):6062–6067.

[74] Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**:207–210.

[75] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A: **NCBI GEO: archive for functional genomics data sets—update**. *Nucleic Acids Research* 2013, **41**(D1):D991–D995.

[76] Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes**. *Proceedings of the National Academy of Sciences* 2002, **99**(7):4465–4470.

[77] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Research* 2006, **34**(suppl 1):D535–D539.

[78] Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, ODonnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M: **The BioGRID interaction database: 2013 update**. *Nucleic Acids Research* 2013, **41**(D1):D816–D823.

[79] Gerstein MB, Kundaje A, Hariharan M, Landt SG, et al.: **Architecture of the human regulatory network derived from ENCODE data**. *Nature* 2012, **489**:91–100.

[80] Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, the Mouse Genome Database Group: **The Mouse Genome Database (MGD): mouse biology and model systems**. *Nucleic Acids Research* 2008, **36**(suppl 1):D724–D728.

[81] Dosztányi Z, et al.: **The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins**. *Journal of Molecular Biology* 2005, **347**:827–839.

[82] Dosztányi Z, et al.: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Bioinformatics* 2005, **21**:3433–3434.

[83] Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garca-Girn C, Gordon L, Hourlier T, Hunt S, Juettemann T, Khri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJP, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SMJ: **Ensembl 2013**. *Nucleic Acids Research* 2013, **41**(D1):D48–D55.

[84] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011, [`http://www.R-project.org/`]. [ISBN 3-900051-07-0].

[85] Farris JS: **Phylogenetic analysis under Dollo's Law**. *Systematic Biology* 1977, **26**:77–88.

[86] Gsponer J, Babu MM: **The rules of disorder or why disorder rules**. *Progress in Biophysics and Molecular Biology* 2009, **99**(2–3):94–103.

[87] Prachumwat A, Li WH: **Protein function, connectivity, and duplicability in yeast**. *Molecular Biology and Evolution* 2006, **23**:30–39.

[88] Pawson T: **Assembly of cell regulatory systems through protein interaction domains**. *Science***300**(5618):445–52.

[89] Prince VE, Pickett FB: **Splitting pairs: the diverging fates of duplicated genes**. *Nature Reviews Genetics* 2002, **3**:827–837.

[90] Conant GC, Wolfe KH: **Turning a hobby into a job: How duplicated genes find new functions**. *Nature Reviews Genetics* 2008, **9**:938–950.

[91] Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models**. *Nature Reviews Genetics* 2010, **11**:97–108.