Published in final edited form as:

J Phys Chem B. 2013 October 24; 117(42): . doi:10.1021/jp402944e.

Evolutionary Pressure on the Topology of Protein Interface Interaction Networks

Margaret E. Johnson and Gerhard Hummer

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, USA, Tel: 301-402-6290. Fax: 301-496-0825

Gerhard Hummer: Gerhard.Hummer@NIH.GOV

Abstract

The densely connected structure of protein-protein interaction (PPI) networks reflects on the functional need of proteins to co-operate in cellular processes. However, PPI networks do not adequately capture the competition in protein binding. By contrast, the interface interaction network (IIN) studied here resolves the modular character of protein-protein binding and distinguishes between simultaneous and exclusive interactions that underlie both co-operation and competition. We show that the topology of the IIN is under evolutionary pressure and we connect topological features of the IIN to specific biological functions. To reveal the forces shaping the network topology, we use a sequence-based computational model of interface binding along with network analysis. We find that the more fragmented structure of IINs, in contrast to the dense PPI networks, arises in large part from the competition between specific and nonspecific binding. The need to minimize nonspecific binding favors specific network motifs, including a minimal number of cliques (i.e., fully connected subgraphs) and many disconnected fragments. Validating the model, we find that these network characteristics are closely mirrored in the IIN of clathrin-mediated endocytosis. Features unexpected on the basis of our motif analysis are found to indicate either exceptional binding selectivity or important regulatory functions.

Keywords

Nonspecific binding; Protein Interaction Networks; Binding specificity; Clathrin-mediated endocytosis; Network analysis

Introduction

Protein-protein interaction (PPI) networks are designed to capture the cooperativity underlying biological functions at the molecular level by listing all binding partners of each protein in an organism. Network theory applied to PPIs has highlighted the collective character of complex cellular processes. The remarkable finding that the number of binding partners per protein is distributed in a power-law fashion established the presence of highly connected hub proteins in PPI networks across all organisms¹. Many of these hub proteins

Correspondence to: Gerhard Hummer, Gerhard.Hummer@NIH.GOV.

Supporting Information Available: The online supporting information contains extended details on the methods for optimizing the sequences for each network, for dealing with self-interactions, and for modifying the network connectivity. We include plots of the clustering coefficients and network motifs as a function of network topology, plots of the IIN interface degree distribution and the best fit, and a larger, labeled image of the endocytic IIN from Figure 2C, as well as examples of a randomized and optimized version of this network. An earlier version of the endocytic IIN and its network statistics are also included. This material is available free of charge via the Internet at <http://pubs.acs.org>.

appear to be essential, with important roles in connecting functionally distinct regions of the proteome²⁻⁵. Network clustering algorithms have been used to identify multi-protein complexes from PPIs and to identify functional modules^{3,6} in PPIs, providing a tool for predicting biological roles for proteins of unknown function.

However, competition as the second major biological force is only poorly captured in PPI networks. Their static nature makes it difficult to appreciate the spatial and temporal aspects of protein binding in the cell. In particular, PPI networks do not clarify whether a protein's binding partners compete with one another or can bind simultaneously (Figure 1).

By identifying the constituent binding interfaces on each protein⁷⁻⁹, we can create the corresponding interface interaction network (IIN)^{10,11}, illustrated in Figure 1, as a means to characterize the competition in protein-protein interactions². Overlaid on its parent PPI network, non-competitive interactions in the IIN point to possible multi-protein complexes. Moreover, the IIN serves as a tool to compare the structural and chemical similarities between proteins targeting specific interfaces¹⁰. Unlike the PPI, the IIN resolves the modular character of protein-protein binding in complex processes such as cell signaling¹², where specificity is conferred through multiple interactions. Every PPI network will have its own unique IIN, but we anticipate that, much like PPI networks, IINs will have common topological properties, and these properties will provide further insight into functional protein interactions.

The connectivity and topology of an IIN and its corresponding PPI can differ substantially because they are subject to different functional pressures. Protein interaction networks need to maintain a certain amount of global connectivity to ensure that disparate functional modules can communicate¹³. Because communication is maintained at the protein level, we propose that a major force shaping the interface interaction network is the requirement that interfaces bind with high specificity to their functional partners while simultaneously minimizing their affinity for any nonfunctional interface interactions. This requirement is implied by the principle of minimal frustration^{14,15}, where in its original application to protein folding, it leads to a first-order-like transition between native and random-coil states¹⁴, associated with a gap in the energy of the states. Applied to protein-protein binding, interfaces will have a gap in the binding energy between their specific binding partners and all other nonfunctional interactions, with specific interactions between proteins expected to compete favorably with non-specific interactions in the cell. As a result, the topology of the interaction networks is likely under evolutionary pressure^{11,16}. With this fundamental hypothesis, we can then study the characteristics of networks in which specific interactions dominate, just as one may determine the characteristics of foldable sequences in protein folding. In previous work¹¹ we established that the ability to evolve highly specific interfaces depends not only on the number of distinct interfaces but also on the topology of their interaction network. In particular, we identified that a hub network motif, where one interface binds multiple related partners, is compatible with high binding specificity. For example, a protein kinase domain may bind multiple partners via a phosphorylation site that is highly conserved among all partners. In contrast, we found that a chain of interacting interfaces is more difficult to optimize for high specificity. This motif would be observed if, for instance, two different SH3 domains both bound specifically to one proline-rich domain (PRD), but only the first of those SH3 domains bound specifically to a second PRD. The expected affinity of the second SH3 domain for the second PRD and the need to suppress that nonspecific interaction puts severe constraints on the evolution of highly specific interfaces.

Here we show that evolutionary pressures shape the topology of IINs by balancing functional requirements with optimization of binding specificity. We first identify properties

of IINs that confer high specificity to the binding interactions. We then test our predictions of important local motifs and global measures of connectivity against a detailed analysis of the IIN¹⁰ for clathrin-mediated endocytosis in yeast¹⁷, a system that exhibits a diverse array of protein-protein interaction types¹⁸. In this comparison, we find that much of the network conforms to our predictions. The agreement is not perfect with respect to the suppression of non-specific interactions, just as proteins may have non-ideal sequences to fold because in their folded state these proteins have to carry out specific functions. An example of such behavior is Rop dimer folding, where two basins with distinct but related topological structures dominate the energy landscape¹⁹, making the folding process susceptible to mutations. Importantly, we here find that the properties that would appear sub-optimal with respect to the need to suppress non-specific binding reflect on the requirement of proper biological function arising from protein interactions. As a result, our analysis not only provides us with an understanding of the forces shaping the network topology but also points to a connection between sub-optimal network features and biological function.

To identify network characteristics associated with high specificity, we generate families of networks that exhibit well-defined global and local properties. To quantify the specificity of these distinct network families we use a physics-based model of binding that calculates binding energies from the amino-acid sequences of the binding interfaces, and a Monte Carlo (MC) global optimization procedure that allows us to minimize nonspecific binding in each network. The overall specificity of the entire protein network is assessed by a minimum energy gap metric, E , that measures the separation between the most competitive nonspecific binding energy and the least competitive specific binding energy¹¹. Similar energy gap metrics have served as simple, yet powerful tools to quantify the binding energy landscape^{20,21} and the specificity of protein binding in the crowded environment of cells^{22,23}. We note that the different networks are both generated and optimized independently of the binding model and energy gap metric used to assess their global specificity. This allows us to rank welldefined network properties based on their specificity and translate these insights to real biological networks.

Our IINs are classified by three properties: (i) the number N of interfaces as a measure of their size, (ii) the degree distribution $p(k)$ as a measure of their global connectivity, and (iii) the clustering coefficient C as a measure of their local structure (illustrated by Figure 2). As the first feature varied in our analysis, the number of interfaces in an IIN almost always exceeds the number of proteins in the corresponding PPI, and does not exceed $2M$, where M is the number of edges representing protein interactions. The second feature we vary is the global connectivity, or the degree distribution of the IIN. Although PPI networks have power-law degree distributions¹, the number of interfaces per protein could potentially break the distribution up into a more Gaussian, or random distribution of partner numbers. Finally, we vary the local structure of the network, as represented by clustering coefficients and the distribution of 3- and 4-node subgraphs. We control the local structure or motifs²⁴ of the networks by using an optimization procedure that, inspired by our earlier results¹¹ on the binding specificity of hub versus chain motifs, acts to suppress interactions between partners of a node's partners. One of the ramifications is the elimination of cliques, or fully connected subgraphs, and a decrease in the clustering coefficient. Cliques may indicate notable regions of a PPI⁶ but, using the same argument applied to chain motifs, in IINs even the smallest cliques (triangles) would be difficult to evolve for high specificity while simultaneously suppressing all self interactions.

Methods

Binding specificity in interaction networks

To assess the specificity of an IIN, we calculate the minimum energy gap E that separates the least competitive specific interaction from the most competitive nonspecific interaction for any interface. To calculate E , we adapt a model¹¹ developed originally to evaluate the role of nonspecific binding as an evolutionary pressure acting to decrease the number of proteins in a cell and to shape the interaction network. The binding energy model was designed to retain relevant complexity of the physical protein interactions without becoming computationally intractable. Each protein binding interface is represented as a flat 25 residue surface, a typical size and shape of a protein-protein interface²⁵, that can interact with other binding surfaces by aligning the interfaces with partial or full overlap. Similar models have been used in the modeling of immune-system recognition²⁶ and protein evolution¹⁶. A large energy gap in units of the thermal energy, E/kT , indicates that the nonspecific interactions are much weaker than the specific interactions and therefore the network would exhibit high specificity.

IIN and topological properties

In our networks each node represents a separate interface, not a separate protein. Each edge represents an interaction between two interfaces and therefore the edges reflect the PPI network. Because the edges determine the functional interactions, their number M remains fixed in all of the networks we compare. Instead, we vary the number N of nodes (interfaces) and their connectivity to improve the overall specificity of the networks. Here we are interested in comparing a broad range of possible IIN structures to establish which network properties provide the highest degree of specificity to the interface interactions. Therefore we do not constrain the IINs to conform to any specific PPI network. Consequently, there are many possible PPIs that could give rise to the computationally generated IINs. However, in our analysis of the yeast endocytosis IIN, the corresponding PPI network is not changed.

Network generation

Model networks are grown to a specific number of nodes and edges and a specific degree distribution using the algorithm of Goh et al.²⁷ The parameter γ controls the resulting degree distribution of the network, with $\gamma=0$ creating random networks and larger values of γ developing increasingly longer tails. Each node i is assigned a weight $p_i = i^\gamma$. Then repeatedly two nodes are selected according to the product of normalized probabilities and an edge is created between them unless one already exists. Because this method creates orphaned nodes that have no partners, these nodes are rejoined into the network to produce a final network with the correct number of nodes and edges. Self-loop edges were excluded from the generated network families to simplify the analysis. The statistics on the clustering coefficients, node motifs and degree distributions of each class of network represent values averaged over 100 distinct networks for each class.

Optimizing networks

To create new networks that favor particular local network features while maintaining identical degree distributions we used a MC optimization procedure. The move sets were the same as those described in²⁸. The trial move was accepted according to a fitness function that was chosen to reduce chains and cliques in the IIN. The negative fitness f of a network of N interfaces is defined as

$$f = \sum_{i=1, k(i)>1}^N \exp \left(+\lambda \sum_{j=1}^{k(i)} k(n_i(j)) \right)$$

where $n_i(j)$ is the interface index of the j th partner of interface i , $k(n_i(j))$ is its number of partners, and λ is a positive free parameter set to 0.5. Singly connected interfaces (i.e., $k(i)=1$) are not included in f . We perform ~20,000 MC sweeps where a new network configuration is accepted or rejected based on a Boltzmann weight $\exp(-f/kT)$ with the effective temperature kT set to 1. The optimization procedure is largely insensitive to variations in λ ranging from 0.1–1 or kT ranging from 0.5–5. To generate randomized rather than optimized networks we set the temperature kT to infinity such that all moves will be accepted.

Clustering coefficients and motif probabilities

We evaluate clustering coefficients using both a local²⁹ and a global definition. The average local clustering coefficient is given by²⁹

$$C_{local} = \frac{1}{N} \sum_{i=1}^N \frac{2N_{closed}(i)}{k(i)[k(i)-1]}$$

where N is the number of interfaces i , $k(i)$ is the number of their partners, and $N_{closed}(i)$ counts how many pairs of interface i 's partners have an edge between them to form closed triangles with i . For nodes with a single partner the summand is set to zero. Any self-loop edges are ignored in this calculation and in the motif calculation below. Although scale-free networks typically have higher clustering coefficients, networks with a low average degree become disconnected and can have relatively low local clustering coefficients. We also use a global clustering coefficient C_{global} as the number of distinct closed triangles $N_{triangle}$ in the network divided by the total number of distinct triplets,

$$C_{global} = \frac{3N_{triangle}}{N_{open} + 3N_{triangle}}$$

with N_{open} the number of open triplets.

To enumerate the 4-node motifs present in our networks we identify all distinct sets of 4-node subgraphs that are connected by at least one path (each node in the subgraph can be reached by the others). There are six distinct 4-node subgraph architectures³⁰ and we note that they are all counted mutually exclusive to one another, i.e., a set of 4 nodes uniquely classifies as one of the six subgraphs. A single node may belong to more than one 4-node subgraph. Further details on methods is available in Supporting Information.

Results

1. IIN specificity as a function of network topology

To establish the properties of IINs that confer high binding specificity, we first examine the effects of varying the number N of interfaces given a fixed set of M protein interactions. We find that the specificity of interaction networks increases with their size, given a fixed

number of specific binding interactions (Figure 3A). As our measure of binding specificity, we use the minimum energy gap separating specific from nonspecific interactions for networks with a varying number N of interfaces and a fixed number of $M=150$ edges. The energy gap increases with N both for the original un-optimized networks and for networks that have been optimized, and irrespective of the parameter controlling the degree distribution. For optimized networks, the minimum energy gap saturates above $N \approx M$, where the number of distinct interfaces exceeds the number of binding interactions (edges). In a previous study¹¹ we showed that the specificity of networks with simplified topologies *decreases* as the number of interfaces, N , is increased. This decrease in specificity was due to the quadratic growth in nonspecific interactions as a function of N relative to a linear increase in specific interactions. Hence the increase in specificity observed here for larger networks may seem surprising because a larger number of nonspecific interactions must be suppressed. However, in the present study, as we increase N , we keep the number of edges fixed. As a result, the larger networks still have a quadratic growth in nonspecific interactions, but they are also sparser, such that the interfaces are relatively disconnected, with fewer of them having multiple competing partners. This simpler, less connected network structure in larger networks imposes fewer constraints on the evolution of highly specific interfaces.

Binding specificity also depends sensitively on network topology. The specificity decreases with increasing average degree, defined as the average number of binding partners $\langle k \rangle = 2M/N$ (Figure 3A). By contrast, the specificity depends in a more complex manner on the degree distribution. By varying the parameter γ , we created networks that favor short-tailed random distributions for $\gamma \approx 0$ and longer-tailed power-law distributions for $\gamma > 0$ (Figure 3B–C). For the densely connected networks ($N=90$), networks with random degree distributions ($\gamma = 0$) produce the highest specificity, suggesting that having a few very highly connected interfaces ($\gamma = 0.8$) makes achieving high specificity difficult. By contrast, for the more sparsely connected networks ($N=200$), a longer-tailed distribution ($\gamma = 0.6$) produces the highest specificity, implying that more highly connected interfaces are less problematic in sparse, disconnected networks. For even sparser networks the degree distributions will become less distinguishable, as ultimately for the maximum interface number of $N=2M=300$, each interface has one and only one partner.

The local network structure also affects the specificity of the IIN. In our network rewiring scheme, we altered the structure while keeping the degree distribution unchanged. We evolved the network to reduce the number of partners an interface's partners have, consequently suppressing cliques, such as closed triangles, and hub nodes that interact directly with other hub nodes. Figures 2A–B and S4 present these changes in local structure pictorially for representative networks. In Figure S1 we plot the change in clustering coefficients as a result of the optimization. Optimized networks have fewer cliques and higher specificity for all degree distributions and interface numbers. The improvement in specificity after optimization (Figure 3A) is largest for the longer tailed ($\gamma = 0.4$) distributions and this result is consistent with larger changes in the local connectivity and clustering for these networks.

We also analyzed the local structure of these networks by quantifying the prevalence of higher order 4-node subgraph architectures (illustrated in Table 1). We find that the optimized networks with higher specificity have more hub motifs and fewer chain motifs than the original networks (Figure S2–S3). We note, however, that 4-node local structural motifs are less important than the degree distribution in establishing specificity in dense (but not sparse) networks. For example, longer tailed distributions have a higher ratio of hubs to chains than random distributions but they have lower specificity for dense networks (see Figure S2). The 4-node flag motif contains a closed triangle and so its presence is markedly

reduced in the optimized networks, consistent with the reduced global clustering coefficient (Figure S3). The square motif is relatively uncommon in all networks and somewhat neutrally affected by the optimization procedure; however, this motif has biological implications that we discuss further below. The final two 4-node subgraphs are densely connected and contain triangles and, as expected, are nearly, if not fully absent from any of the optimized networks.

Summary of predicted IIN properties—On the basis of our model networks we conclude that biological IINs should optimally exhibit a sparse connectivity with a low average degree and a degree distribution that allows highly connected nodes and more closely resembles a scale-free than random distribution. The clustering coefficient should be very low with few closed triangles and a distribution of 4-node subgraphs that favors hub motifs over chain motifs. In the following, we will test each of these predictions against the actual network for the clathrin-mediated endocytosis in yeast.

2. Network properties of clathrin-mediated endocytosis in yeast

Yeast Endocytic IIN statistics—The PPI for CME has 56 proteins (Figure 4A) and has a corresponding IIN composed of 195 interfaces with 206 edges, 10 of which are self-loop edges (Figures 4B, 2C). The average number of interfaces per proteins is 3.5. This represents a highly disconnected network of average degree $\langle k \rangle = 2.06$ compared to 6.4 for the PPI (where self-loops have a degree of one), consistent with our predictions stated above. The network is highly fragmented¹⁰, distinguishing it from the PPI and most routinely studied network architectures^{29,31}. The clustering coefficient for this network is very low (zero), also in keeping with our predictions, with no triangles present in the network. For this sparse network we would expect highest specificity for a shorter tailed distribution than the one observed, which is best fit at $\gamma = 0.7$ (see Figures S5–S6).

In Table 1 we compare the local structural properties of the actual endocytosis network with a family of networks that have the same number of edges, interfaces, and the same degree distribution, but with edges reconnected either randomly, or under optimization. The observed network falls either in between the randomized and optimized networks or matches fully optimized networks (e.g., $C_{\text{global}} = 0$) for all properties except the square motifs. We also list p-values for the hypothesis that the local properties of the observed network are similar to the local properties of randomized networks. The low p-values confirm that the endocytosis IIN is distinct from randomized networks and unique in its local structure. None of the 1200 random networks we generated reproduced the 4-node motif structure of the endocytosis IIN (except the rarely occurring 5 or 6 edge ‘Other 4-node’ subgraphs), and an insignificant fraction produced the low clustering coefficients. The local structure of the endocytosis IIN not only deviates from randomized networks but its properties consistently approach those of our proposed ‘optimal’ network structure. Hence we can conclude that the local structure of the endocytosis IIN is not fully optimal as per our model predictions, but it is clearly modified compared to randomized networks and trends towards an optimized local structure (see Figure S7).

Of particular note is the fraction of square motifs observed in this biological network because it markedly exceeds the fraction observed in either the optimized or the randomized networks. This square motif is functionally quite reasonable, indicating related sequence domains that share related sequence partners, e.g., $A_1-B_1-A_2-B_2-A_1$, such that all A-type domains bind to all B-type domains but not to others of the same type. For example, if a pair of interacting interfaces (A-B) is duplicated and their specificity maintained, one would observe a square motif. Therefore even in our binding model this square motif should produce better specificity than a chain motif because the nonspecific interaction (no edge) is

replaced by a specific interaction (edge). However, our fitness function f (see Methods) does not explicitly favor their formation. In square subgraphs, an interface's partners have multiple partners and so our fitness function would have to be adapted to promote their presence in the network.

We also used our binding model to compare the specificity of this observed IIN with the randomized and optimized networks. The values for the minimum energy gaps are collected in Table 1 and show that the observed endocytic IIN has relatively high specificity that clearly exceeds the specificity of randomly reconnected networks. We include these values in Figure 3A. For the endocytosis network, the calculated minimum energy gaps after either optimization or randomization fall close to the curves obtained for computer-generated networks with ~ 0.7 , as inferred from the degree distribution (Figures S5–S6). Small deviations may reflect differences in the detailed structure of the endocytosis IIN and the family of generated networks, including a single, fixed degree distribution, and the presence of self-loops and of additional nodes and edges in the endocytosis network.

3. Functional relevance of suboptimal network motifs

Overall the endocytic IIN has characteristics consistent with a high specificity network. Nevertheless, there are a number of features that are not fully optimal in light of our model predictions. By analyzing these regions of the endocytic IIN with such suboptimal local structure compared to our network predictions, we gain insight into the forces driving the evolution of protein-protein interactions.

Chain motifs and sufficient binding variability—The endocytic IIN has a fairly high number of chain motifs relative to our optimized networks and this is unexpected because chain motifs are more difficult to optimize for high specificity. We first analyze the regions of the network that contain chain motifs mixed in with square motifs. As we noted above, square motifs, unlike chain motifs, are consistent with high specificity. A chain motif can be turned into a square by adding one edge and therefore this missing edge could simply indicate a missed interaction. However, the unexpected lack of binding could also indicate small differences between interaction partners that create a significant enough variance in binding energies to suppress what would otherwise have been a specific interaction. For example, the upper right corner of the IIN in Figure 2C contains a mixture of chain and square motifs between SH3 domains and their prolinerich domain (PRD) binding partners. A study of the specificity of SH3 domains towards PRDs demonstrated that there is sufficient difference between PRD sequences that not all SH3 domains prefer the same sets of partners³². Analysis of other square and chain motif regions of the IIN (e.g., involving kinases³³) suggests that small but significant variations in the binding domain sequences help maintain sufficient specificity, such that chain motifs alongside square motifs are tolerable.

Interestingly, although the 'missing' edge in a chain motif would be classified as nonspecific according to our rules, it is possible that the interaction retains functional relevance. For example, even related kinases target distinct phosphorylation sites, but may have weak cross-activity. Such weak binding between a set of non-partners, according to the network, may not be too detrimental to the cell. Hence optimizing interfaces for a large energy gap may not be necessary in all chain/square motif cases, and we have shown that even with chain motifs present in a network, a large positive energy gap can be achieved, indicative of sufficiently high network specificity.

Chain motifs and regulation of binding interactions—Ideally, the hub motif on the upper left of the IIN in Figure 2C, centered around the actin filament binding surface

(ACT1.2), would be disconnected from the rest of the IIN. Instead, some of ACT1.2's binding interfaces form unfavorable chains of interactions with distinct interfaces. While these chain features may again reflect tolerance of variation in binding sites, they also could point to spatial, temporal, posttranslational or allosteric regulation processes that keep competitive nonspecific interactions from forming. For example, ACT1.2's partner PAN1.3 branches off into two chains, one ending in SLA1.5. The intervening binding interfaces in this first chain, however, are regulated by phosphorylation and influenced by being on the same protein, PAN1. The ACT1.2 and SLA1.5 interfaces bind to PAN1 at different sites and only after phosphorylation do these two sites on PAN1 bind onto themselves. Hence the ability of the central region of PAN1 (PAN1.3) to bind to its partners is controlled by whether its N-terminal region (PAN1.4) is phosphorylated. Additionally, this PAN1 self-binding interaction may not be optimized for maximal specificity, since the constant co-localization of the two attached domains will effectively increase their likelihood of being bound to one another. Therefore the nonfunctional interactions generated by this chain motif are suppressed by regulating the availability and chemistry of the constituent binding interfaces.

Another unfavorable chain motif extending from the ACT1.2 hub interface also appears to indicate regulatory interactions. The ARC40 subunit of the ARP2/3 complex binds both to actin filaments and to the acidic motifs of LAS17 and others. The binding of LAS17 to the ARC40 subunit has been shown to interfere with the ability of the ARP2/3 complex to bind to actin filaments³⁴. Although there is some ambiguity in these interface assignments (to the point that a more precise resolution of the interface residues might remove the chain motif), the ability of LAS17 to inhibit binding suggests either direct or allosteric regulation via its acidic motif. In this case there is no obvious mechanism for suppressing the nonfunctional interactions generated by this chain motif. The regulatory function of the interfaces seems more critical than evolving the network topology for maximal specificity.

Discussion

We found that the observed properties of the IIN for proteins involved in clathrin-mediated endocytosis closely match the predictions of our binding model, both locally with respect to the prevalence of certain network motifs and globally with respect to the statistical properties. This good agreement underlines the strength of our initial hypothesis on the importance of minimizing nonspecific binding when defining protein interactions. Our model's emphasis on sequence specificity is supported by the fact that even proteins with the same domain structures do not always share the same binding partners³². The agreement is particularly notable given that our binding model does not capture the full complexity of interface binding, ignoring, for instance, differences in interface size and structure. Indeed, some seemingly sub-optimal local structures observed in the endocytic IIN could be explained by refined binding models that, for instance, vary the binding interface size. For a small binding interface, a smaller energy gap may be sufficient to prevent binding between pairs of proteins if the resulting affinities are weak and the concentrations are not too high. Size variation could therefore help reduce potential nonspecific binding, particularly for hub interfaces like those on actin (e.g. ACT1.2), where all interfaces have significant overlap but some partners utilize fewer residues to stabilize their interactions. Evidence suggests that hub interfaces with many partners have flatter surfaces and imperfect packing to accommodate binding partners of different sizes³⁵. Additionally, the fitness function we use to generate networks optimized for specific binding could be modified to accommodate square motifs that are consistent with high binding specificity, and would thereby further improve the agreement between our predicted optimized networks and real biological networks.

Ambiguity in interface assignment

In contrast to the PPI of the endocytic network, even simple visual inspection of the IIN not only helps identify competitive or co-operative protein interactions, but also reveals possible mistakes or ambiguities in the interface assignments. To illustrate this point we remark on the long chain in the center of the IIN (Figure 2C). The domains implicated in these interactions, such as on PAN1.1 and SCD5.0, stretch across hundreds of residues. Due to the ambiguity of the residues involved, these were marked as shared interfaces. However, it is quite possible that the residues used to mediate the interactions are in fact quite distinct and therefore the interfaces are not truly shared. If they were not shared then the chain would be broken and it would be easier to optimize the interfaces for higher specificity.

Robustness of network properties and extension to larger yeast IIN

The endocytic network is one subnetwork of the larger yeast PPI. How sensitive are our results to changes to the PPI? We have found our results to be quite robust to changes we have made over the course of the work to the PPI and therefore to the IIN. Expanding the starting set of edge interactions in our PPI network by adding 134 potential interactions from the BioGrid database³⁶, we were able to add 30 interfaces and 52 edges to our initial IIN. Despite this almost 20% increase, all of the trends in network properties were maintained, and only the denseness ($\langle k \rangle$) of the network increased (see Figure S8; Table S1). Binding specificity between protein interfaces is essential for proper functioning and we therefore anticipate that these same trends in clustering coefficients, 4-node motifs, and the degree distribution will be maintained in larger PPI networks.

We also expect that the IIN of a larger set of proteins could exhibit a degree distribution with a longer tail simply due to there being more potential binding partners per interface. However, we do not expect the average degree to grow when the network is expanded, because the proteins we have characterized thus far have all been part of the same functional module that is actually more densely connected than the full PPI. Also, when functionally unrelated proteins are introduced into the IIN we would anticipate more instances of temporal and spatial distribution of proteins acting to regulate accessibility of interfaces to one another. Variation in the relative abundance of individual proteins has also been shown to be an important mechanism for cells to evolve functional PPIs¹⁶.

Suboptimal network features point to function

The insights into protein function provided by our analysis of sub-optimal features of the endocytosis IIN can be directly translated to other IIN modules. The presence of chain motifs mixed in with squares can indicate tolerance of variability in binding partners due to, e.g., small changes in residue sequences and shape complementarity. It is also possible that the nonfunctional (missing) edge that distinguishes a chain from a square in these regions could become functional as a low-affinity rescue mechanism. The presence of chain motifs extending from hub motifs can indicate regions of regulation. In some cases regulation would be expected to shield interfaces, contributing to enhanced specificity, whilst in other cases it simply indicates that an interface's regulatory ability is more important than maximal specificity. By knowing the physical assumptions entering our network optimality measure, we can use any deviations between predicted and observed network motifs to help interpret how functional specificity is maintained in actual biological systems.

IIN motif analysis as a tool

The IIN is a resource for conveniently organizing more detailed information on protein interactions and formulating hypotheses on potentially missing or spurious interactions. IIN motif analysis can be used to identify protein interactions that may require tests whether

they are truly direct or only indirectly mediated through a third protein. This distinction helps clarify whether proteins are competing for binding partners, which in turn affects the dynamics and equilibrium of the process in which they are participating. IIN motif analysis can also be used to establish additional constraints in models of PPI network evolution³⁷. The evolution of new proteins occurs in part through domain duplication and recombination³⁸ and it is the structure of the IIN rather than the PPI that captures domain-level detail. The methods we have developed here could be used to create networks with disconnected, hub-heavy local structures and ultimately to help predict potential interface networks from PPIs. Network motif analysis should also prove useful in refining structural interaction networks determined through automated approaches^{2,7,39–42} by identifying regions with triangles or other suboptimal motifs as candidates for further inspection.

Conclusions

The cell is a crowded environment where proteins must navigate to form not only functional complexes with their specific binding partners, but avoid binding nonspecifically to all other surfaces they encounter. This physical pressure is common to all proteins and we showed here that it acts as a fundamental force shaping the topology of protein interface interaction networks. This perspective on the origins of network topology arising from the need to maximize specificity in binding interactions allows one to identify regions of the network where other forces may be equally important. In particular, wherever the network motifs differ from those predicted by our model, we found in the case of the CME network that the interfaces either exhibited high selectivity for individual protein partners, or access to the interfaces was regulated or contributed to regulating other interactions. In protein folding, non-optimal amino-acid interactions indicative of energetic frustration were similarly found to be concentrated near sites involved in protein function⁴³. The topological features of IINs derived from other protein interaction networks could also be used to identify spurious or indirect interactions between protein pairs. The IIN representation of protein-protein interactions thus provides unique insight into the evolution of protein interactions and provides a tool for curating and classifying experimentally observed protein interactions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Intramural Research Program of the NIDDK, NIH and by an NIH K99 award GM098371 to MEJ. The research used the Biowulf Linux cluster at the NIH.

References

1. Yook SH, Oltvai ZN, Barabasi AL. Functional and Topological Characterization of Protein Interaction Networks. *Proteomics*. 2004; 4:928–942. [PubMed: 15048975]
2. Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science*. 2006; 314:1938–1941. [PubMed: 17185604]
3. Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature*. 2005; 435:814–818. [PubMed: 15944704]
4. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, et al. Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network. *Nature*. 2004; 430:88–93. [PubMed: 15190252]
5. 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 Comput Biol*. 2008; 4:e1000140. [PubMed: 18670624]

6. Spirin V, Mirny LA. Protein Complexes and Functional Modules in Molecular Networks. *Proc Natl Acad Sci US A*. 2003; 100:12123–12128.
7. Clarke D, Bhardwaj N, Gerstein MB. Novel Insights through the Integration of Structural and Functional Genomics Data with Protein Networks. *J Struct Biol*. 2012; 179:320–326. [PubMed: 22343087]
8. Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R. Towards Inferring Time Dimensionality in Protein-Protein Interaction Networks by Integrating Structures: The P53 Example. *Mol BioSyst*. 2009; 5:1770–1778. [PubMed: 19585003]
9. Kiel C, Beltrao P, Serrano L. Analyzing Protein Interaction Networks Using Structural Information. *Annu Rev Biochem*. 2008; 77:415–441. [PubMed: 18304007]
10. Johnson ME, Hummer G. Interface-Resolved Network of Protein-Protein Interactions. *PLOS Comput Biol*. 2013 In Press.
11. Johnson ME, Hummer G. Nonspecific Binding Limits the Number of Proteins in a Cell and Shapes Their Interaction Networks. *Proc Natl Acad Sci US A*. 2011; 108:603–608.
12. Pawson T. Specificity in Signal Transduction: From Phosphotyrosine-Sh2 Domain Interactions to Complex Cellular Systems. *Cell*. 2004; 116:191–203. [PubMed: 14744431]
13. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al. Proteome Survey Reveals Modularity of the Yeast Cell Machinery. *Nature*. 2006; 440:631–636. [PubMed: 16429126]
14. Bryngelson JD, Wolynes PG. Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc Natl Acad Sci US A*. 1987; 84:7524–7528.
15. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins*. 1995; 21:167–195. [PubMed: 7784423]
16. Heo M, Maslov S, Shakhnovich E. Topology of Protein Interaction Network Shapes Protein Abundances and Strengths of Their Functional and Nonspecific Interactions. *Proc Natl Acad Sci US A*. 2011; 108:4258–4263.
17. Weinberg J, Drubin DG. Clathrin-Mediated Endocytosis in Budding Yeast. *Trends Cell Biol*. 2012; 22:1–13. [PubMed: 22018597]
18. Nooren IM, Thornton JM. Diversity of Protein-Protein Interactions. *EMBO J*. 2003; 22:3486–3492. [PubMed: 12853464]
19. Levy Y, Cho SS, Shen T, Onuchic JN, Wolynes PG. Symmetry and Frustration in Protein Energy Landscapes: A near Degeneracy Resolves the Rop Dimer-Folding Mystery. *Proc Natl Acad Sci US A*. 2005; 102:2373–2378.
20. Papoian GA, Wolynes PG. The Physics and Bioinformatics of Binding and Folding—an Energy Landscape Perspective. *Biopolymers*. 2003; 68:333–349. [PubMed: 12601793]
21. Wang J, Huang W, Lu H, Wang E. Downhill Kinetics of Biomolecular Interface Binding: Globally Connected Scenario. *Biophys J*. 2004; 87:2187–2194. [PubMed: 15454421]
22. Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI. Robust Protein Protein Interactions in Crowded Cellular Environments. *Proc Natl Acad Sci US A*. 2007; 104:14952–14957.
23. Zhang J, Maslov S, Shakhnovich EI. Constraints Imposed by Non-Functional Protein-Protein Interactions on Gene Expression and Proteome Size. *Mol Syst Biol*. 2008; 4:210. [PubMed: 18682700]
24. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. *Science*. 2002; 298:824–827. [PubMed: 12399590]
25. Jones S, Thornton JM. Principles of Protein-Protein Interactions. *Proc Natl Acad Sci US A*. 1996; 93:13–20.
26. Kosmrlj A, Read EL, Qi Y, Allen TM, Altfeld M, Deeks SG, Pereyra F, Carrington M, Walker BD, Chakraborty AK. Effects of Thymic Selection of the T-Cell Repertoire on Hla Class I-Associated Control of Hiv Infection. *Nature*. 2010; 465:350–354. [PubMed: 20445539]
27. Goh KI, Kahng B, Kim D. Universal Behavior of Load Distribution in Scale-Free Networks. *Phys Rev Lett*. 2001; 87:278207.
28. Maslov S, Sneppen K. Specificity and Stability in Topology of Proteins Networks. *Science*. 2002; 296:910–913. [PubMed: 11988575]

29. Watts DJ, Strogatz SH. Collective Dynamics of 'Small-World' Networks. *Nature*. 1998; 393:440–442. [PubMed: 9623998]
30. Itzkovitz S, Alon U. Subgraphs and Network Motifs in Geometric Networks. *Phys Rev E*. 2005;71.
31. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The Large-Scale Organization of Metabolic Networks. *Nature*. 2000; 407:651–654. [PubMed: 11034217]
32. Tonikian R, Xin XF, Toret CP, Gfeller D, Landgraf C, Panni S, Paoluzi S, Castagnoli L, Currell B, Seshagiri S, et al. Bayesian Modeling of the Yeast Sh3 Domain Interactome Predicts Spatiotemporal Dynamics of Endocytosis Proteins. *PLOS Biol*. 2009; 7:e1000218. [PubMed: 19841731]
33. Mok J, Kim PM, Lam HYK, Piccirillo S, Zhou XQ, Jeschke GR, Sheridan DL, Parker SA, Desai V, Jwa M, et al. Deciphering Protein Kinase Specificity through Large-Scale Analysis of Yeast Phosphorylation Site Motifs. *Sci Signal*. 2010; 3:ra12. [PubMed: 20159853]
34. Ti SC, Jurgenson CT, Nolen BJ, Pollard TD. Structural and Biochemical Characterization of Two Binding Sites for Nucleation-Promoting Factor Wasp-Vca on Arp2/3 Complex. *Proc Natl Acad Sci US A*. 2011; 108:E463–471.
35. Keskin O, Nussinov R. Similar Binding Sites and Different Partners: Implications to Shared Proteins in Cellular Pathways. *Structure*. 2007; 15:341–354. [PubMed: 17355869]
36. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al. The Biogrid Interaction Database: 2011 Update. *Nucleic Acids Res*. 2011; 39:D698–704. [PubMed: 21071413]
37. Evlampiev K, Isambert H. Conservation and Topology of Protein Interaction Networks under Duplication-Divergence Evolution. *Proc Natl Acad Sci US A*. 2008; 105:9863–9868.
38. Apic G, Gough J, Teichmann SA. Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes. *J Mol Biol*. 2001; 310:311–325. [PubMed: 11428892]
39. Nye TM, Berzuini C, Gilks WR, Babu MM, Teichmann SA. Statistical Analysis of Domains in Interacting Protein Pairs. *Bioinformatics*. 2005; 21:993–1001. [PubMed: 15509600]
40. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, et al. Structure-Based Prediction of Protein-Protein Interactions on a Genome-Wide Scale. *Nature*. 2012; 490:556–560. [PubMed: 23023127]
41. Shoemaker BA, Zhang DC, Tyagi M, Thangudu RR, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR. Ibis (Inferred Biomolecular Interaction Server) Reports, Predicts and Integrates Multiple Types of Conserved Interactions for Proteins. *Nucleic Acids Res*. 2012; 40:D834–D840. [PubMed: 22102591]
42. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-Resolution Protein Complexes from Integrating Genomic Information with Molecular Simulation. *Proc Natl Acad Sci US A*. 2009; 106:22124–22129.
43. Ferreira DU, Hegler JA, Komives EA, Wolynes PG. Localizing Frustration in Native Proteins and Protein Assemblies. *Proc Natl Acad Sci USA*. 2007; 104:19819–19824. [PubMed: 18077414]
44. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2. 8: New Features for Data Integration and Network Visualization. *Bioinformatics*. 2011; 27:431–432. [PubMed: 21149340]

Figure 1. Schematic illustrating (left) the binding of three proteins, where matching features indicate binding interfaces, (center) the corresponding interface-interaction network with orange and green indicating shared and non-shared binding interfaces, respectively, and (right) the protein-protein interaction network.

Figure 2.

Network evolution and validation. **A,B**) IINs with a fixed number of $M=150$ edges and $N=90$ (top) and 200 interfaces (bottom), degree distributions $\alpha=0$ (A) and 0.8 (B) before and after network evolution (left and right of arrows, respectively). Green interfaces have a single partner, orange interfaces are shared and have multiple partners, and purple indicates closed triangles in the network. The optimized networks have the same degree distribution but have been evolved to reduce the number of partners each node's partners have (see Methods). **C**) IIN of yeast clathrin-mediated endocytosis, closely resembling the optimized, sparse, $\alpha=0.8$ network in the bottom right corner of panel B). Larger, labeled version shown in Figure S7B. Network images created using Cytoscape⁴⁴.

Figure 3.

Properties of IINs. **A)** Specificity of the interface interaction networks with 150 edges as a function of interface numbers (N), degree distribution ($\langle k \rangle$), and local structure altered by the optimization scheme. Each data point represents an average over 10 networks with error bars indicating the standard deviation. The gray dashed lines correspond to the original, non-optimized networks, with degree distributions varying from $\alpha=0$ (light gray) to $\alpha=0.8$ (black). The same symbols are used for the optimized networks shown in solid colored lines. The degree distributions vary from $\alpha=0$ (light yellow) to $\alpha=0.8$ (red). The blue hexagram denotes the minimum energy gap (in units of the thermal energy kT) for the endocytic IIN, placed on the x-axis according to its average degree $\langle k \rangle = 2.06$ (upper x-axis). The black and red hexagrams denote the minimum energy gap for the randomized and optimized versions of the endocytic IIN, preserving its degree distribution including self-loops. **B)** Corresponding degree distributions for the sparse $N=200$ interface networks averaged over 50 networks. **C)** Corresponding degree distributions for the dense $N=90$ interface networks.

Figure 4.

Yeast clathrin-mediated endocytosis interaction sub-network. **A)** Protein-protein interaction network¹⁰ for the 56 proteins selected as part of the endocytic functional module¹⁷ showing all the 186 edges where interfaces were successfully assigned. **B)** The same protein interaction network as in A) with interfaces assigned to each protein interaction. The IIN has 206 edges because some of the protein interactions have two modes of binding to one another. Green interfaces have a single partner and orange interfaces are shared with multiple partners. The interface numbers range from one up to eleven (for LAS17). The interface interaction network of the endocytic network, shown separated from the proteins in Figure 1C, emphasizes the disconnected topology of the IIN.

Table 1

Local network properties for the endocytic IIN.

	Randomized	Observed	Optimized
C_{local}	0.0095 ± 0.006	0 ($p=0.008$)	0.0052 ± 0.0065
C_{global}	0.022 ± 0.01	0 ($p=0.008$)	0.0028 ± 0.0034
4-node hubs	0.44 ± 0.02	0.56 ($p<0.0008$)	0.86 ± 0.003
4-node chains	0.53 ± 0.02	0.37 ($p<0.0008$)	0.14 ± 0.004
4-node flags	0.025 ± 0.01	0 ($p=0.008$)	0.0012 ± 0.0015
4-node squares	0.0028 ± 0.001	0.061 ($p<0.0008$)	$8 \times 10^{-4} \pm 7 \times 10^{-4}$
Other 4-node	$0.0005 \pm 7 \times 10^{-4}$	0 ($p=0.29$)	0
Min Gap (Specificity)	2.95 ± 0.56	4.2	4.99 ± 0.13

Results are listed for randomized, actual, and computationally optimized networks. The clustering coefficients are defined in the SI Text. The numeric values are the fraction of all 4-node subgraphs observed in each of the networks. Randomized and optimized networks have the same degree distribution as the observed endocytic network (see Methods). Standard deviations are shown from averages over 100 network randomizations and 100 network optimizations. The last row is the minimum energy gap with averages taken over 10 randomized and optimized networks each. The p-values listed in parentheses are calculated based on the hypothesis that the endocytic IIN is related to randomized networks. 1200 randomized networks were generated. The low or minimum p-values consistently indicate the IIN is not similar to a randomized network but unique in its local structure. We note that the 'Other 4-node' subgraphs are rare in all networks because they are highly connected and the networks are relatively sparse.