Phys. Biol. **2** (2005) S24–S35 doi:10.1088/1478-3975/2/2/S03

Protein–protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach

Ozlem Keskin[1,2], Buyong Ma[2], Kristina Rogale[3], K Gunasekaran[2] and Ruth Nussinov[2,4]

[1] Koc University, Center for Computational Biology and Bioinformatics, and College of Engineering, Rumelifeneri Yolu, 34450 Sariyer Istanbul, Turkey
[2] Basic Research Program, SAIC-Frederick, Inc., Laboratory of Experimental and Computational Biology, NCI-Frederick, NCI-Frederick Building 469, Room 151, Frederick, MD 21702, USA
[3] Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08540, USA
[4] Department of Human Genetics and Molecular Medicine, Sackler Institute of Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

E-mail: ruthn@ncifcrf.gov

Received 9 February 2005
Accepted for publication 21 April 2005
Published 20 May 2005
Online at stacks.iop.org/PhysBio/2/S24

Abstract
Understanding and ultimately predicting protein associations is immensely important for functional genomics and drug design. Here, we propose that binding sites have preferred organizations. First, the hot spots cluster within densely packed 'hot regions'. Within these regions, they form networks of interactions. Thus, hot spots located within a hot region contribute *cooperatively* to the stability of the complex. However, the contributions of separate, independent hot regions are *additive*. Moreover, hot spots are often already pre-organized in the unbound (free) protein states. Describing a binding site through independent local hot regions has implications for binding site definition, design and parametrization for prediction. The compactness and cooperativity emphasize the similarity between binding and folding. This proposition is grounded in computation and experiment. It explains why summation of the interactions may over-estimate the stability of the complex. Furthermore, statistically, charge–charge coupling of the hot spots is disfavored. However, since within the highly packed regions the solvent is screened, the electrostatic contributions are strengthened. Thus, we propose a new description of protein binding sites: a site consists of (one or a few) self-contained cooperative regions. Since the residue hot spots are those conserved by evolution, proteins binding multiple partners at the same sites are expected to use all or some combination of these regions.

Introduction

Protein–protein interactions are critical for all cellular pathways, regulation and packaging [1, 2]. They are involved in all processes of a living organism. They are crucial to the understanding of all *in vivo* functions, cellular regulation, biosynthetic and degradation pathways, signal transduction, initiation of DNA replication, transcription and translation, multi-molecular associations, packaging, the immune response and oligomer formation. They relate to allosteric mechanisms, to turning genes on and off and to drug design. All biological processes are regulated through association and dissociation of protein molecules. These include hormone–receptor binding, protease inhibition, antigen–antibody recognition, signal transduction, enzyme–substrate binding, vesicle transport, RNA splicing and gene activation. Hence, solving the protein–protein interaction

1478-3975/05/020024+12$30.00 © 2005 IOP Publishing Ltd Printed in the UK

puzzle is at the center stage of protein science. The ability to predict the preferred method by which proteins interact [3–8] would facilitate assignment of protein function [9]. It would assist in the prediction of binding sites, in the construction of protein networks, in the prediction of multi-molecular assemblies, in mapping metabolic pathways and in drug design [10].

Yet, although much progress has been made, this problem is still far from being solved [3]. There are several reasons for this difficulty [11]. Proteins are flexible [12]. A given site may bind different ligands with different affinities. Protein–protein binding sites vary, with different relative contributions of the hydrophobic effect versus electrostatic interactions [13–17]. Moreover, depending on the protein function, the surface of a protein molecule is likely to contain a number of binding sites. For example, consider the proteins at the center of the organism's interaction map versus proteins at the map edges. Centrally located hub proteins have a few binding sites, with the sites binding a range of molecules, possibly under allosteric regulation. Examples of hub proteins may be signaling proteins, such as, for example, *Mdm2* or superantigens. In contrast, edge proteins may have a single binding site to interact with a specific ligand. Some binding sites are obligatory. Proteins experiencing such a binding mechanism are likely not to have a stable native structure in their unbound state [18–20]. Other proteins may have transient associations, depending on their functional state [20]. Obligatory protein–protein interactions are expected to be much more stable than the transient ones. The different binding states, different locations and functions and different stabilities explain why it has been difficult to extract general rules of protein–protein associations [14]. Moreover, binding sites are dynamic. Loop flexibility, a protruding flexible bridging β-sheet or a hinged α-helix, is often crucial for the binding, particularly in transient associations. Eventually, to understand the principles of protein–protein associations [21], the knowledge we obtain from analysis of the static crystal structures should be combined with the dynamic [12]. Figure 1 displays a ribbon diagram of a protein complex with its interface highlighted. Here, the yellow parts represent the two chains of the protein. The contacting residues are responsible for the interaction between the two chains and are colored magenta. The cyan residues display the residues nearby the contacting residues. Together, they form the scaffold for an interface between two proteins.

Here, we take a bottom-up approach. A bottom-up structure-based approach focuses on proteins. The goal is to predict which proteins interact and how the interactions will take place. Eventually, protein–protein interactions should be viewed within the context of Systems Biology. Within the Systems framework, predicting which proteins interact and how they interact is an extremely significant goal. It assists in assigning function, in obtaining information relating to their regulation, providing clues to the system dynamics, protein design and to competing pathways. It further provides essential information on the system robustness and drug design.

Below, we divide the review into two parts. We first address pairwise protein–protein interactions. Next we address the interactions within the Systems framework. In particular, we focus on the problem of 'hub' proteins with multiple interactions at the centers of the networks versus those at the edges.

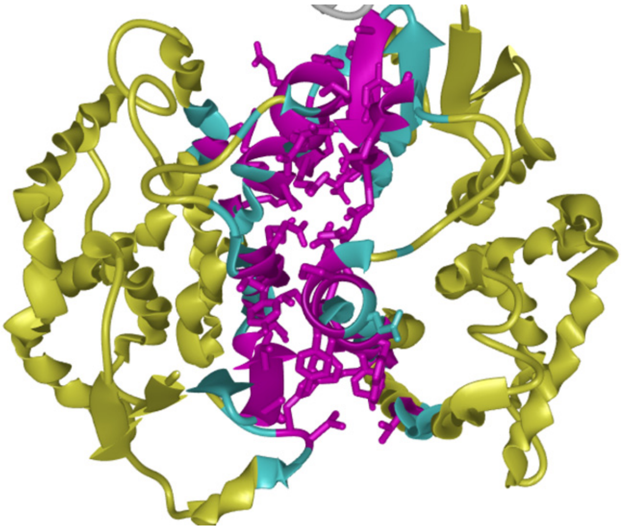

Figure 1. A ribbon diagram of a protein complex (glutathione S-transferase) with its interface highlighted (PDB code 1gwc). Here, the yellow parts represent the two chains of the protein (C and B chains). The contacting residues are responsible for the interaction between the two chains and are colored magenta. The cyan residues display the residues nearby the contacting residues. Together, they form the scaffold of an interface between two proteins.

1. Pairwise protein–protein interactions: cooperativity and organization

1.1. Cooperativity in protein folding and binding

Cooperativity is non-independence. Proteins are widely believed to fold cooperatively. Non-cooperative folding events would lead to an exhaustive search of the conformational space to reach the global minimum. Yet, an exhaustive conformational search would imply time scales not affordable in the biological world. Considerable literature has addressed the challenging question of the physical basis of cooperativity through which proteins would avoid an exhaustive search (e.g., [22, 23], and references therein). From the kinetic standpoint, cooperativity leads to preferred protein folding pathways. Cooperativity largely derives from the hydrophobic effect, the driving force of protein folding. Accounting for cooperativity has led to landmark experimental and computational investigations of the mechanisms and pathways of protein folding (reviewed in [22]), addressing the question of *how* the protein chain searches the immense number of possible nonlocal interactions to yield the hydrophobic core.

To understand cooperativity, we need to think of the system as a cohesive unit, where the behavior of the parts may depend on each other. That is, the overall behavior is the outcome of the properties of the entire system and not of the sum of the properties of its components. Below, we argue that the thermodynamic stability of the protein–protein complex is not a summation of the individual contributions of each of the residues independent of the other; rather, residues which are in direct spatial contact, or in close contact through a few tightly packed intermediate residues, impact the stability of the association in a non-additive manner. When a residue is in a tight physical (chemical) geometrical contact with others, its substitution would affect the structure and interactions of its neighboring residues. Thus, if this residue and its neighbors contribute significantly to the stability of the molecule or the complex, its mutation may affect the stability not only through the change of its own interactions, but in addition through the changes of its neighbors. This would affect the stability of the complex beyond the direct altered interactions of the mutated residue. Hence, if we were to simultaneously mutate two residues which are in close spatial contact in a densely packed environment, the change in the stability would not be the sum of the measured changes of each one separately. The measured change in the thermodynamic stability upon a mutation of a single residue already takes into account changes in the interactions of its closely packed neighboring residues. On the other hand, if the protein–protein interface can be separated into cohesive separate units, the impact of mutations in each of these is independent, i.e., non-cooperative.

1.2. Protein–protein binding sites consist of independent regions

Here, we propose that in protein–protein complexes, the binding sites consist of one or a few independent, tightly packed regions [24]. Residues which contribute significantly (more than 2 kcal mol^{-1} [25–27]) to the free energy of the protein–protein association are clustered within these regions. Within the tightly packed cluster, these so-called hot spot residues form a network of interactions, thus contributing *cooperatively* to the stability of the protein–protein complex. In contrast, the contribution of the independent regions is additive. We name these regions 'hot regions'. This type of tightly packed organization effectively screens the solvent from the charged groups, strengthening the charge–charge interactions. This rationalizes why the hot spots do not form more salt bridges and hydrogen bonds than other interacting residues at the protein–protein interface. Such a *hot region* organization is advantageous to the protein associations. Moreover, this organization of protein–protein binding highlights the similarity between protein folding and protein binding [28, 29]. In both packing plays a crucial role. Within the packed protein cores and the packed hot regions, there are residues that contribute significantly and cooperatively to the stability. As in cores, the hot spots are also highly conserved by evolution at the protein binding site [30–32]. The average conservation ratio of the neighboring residues in hot regions is 0.47 as compared to a 0.26 average conservation ratio for the rest of the interface residues. A residue is identified as a 'hot spot neighbor' if the distance between its Cα and a Cα of a residue is less than 6.5 Å.We also observe that the packing is higher around hot spots (on average 7.0 residues in the hot regions) and lower at the other regions (5.6 residues outside). For the packing calculations, residues whose Cα are closer than the cut-off distance are defined to be in contact, excluding the two bonded sequential neighbors. These numbers are obtained from a set of 44 interface clusters [24]. Combined, the hot region organization provides a new description of protein binding sites. It is useful since it explains why a summation of the hot spots contributions over-estimates (or, under-estimates) the binding free energy. This is expected to lead to better scoring schemes in the prediction of protein–protein associations. In addition, it suggests that a hot region should provide a good target for drug design. Figure 2 presents an example displaying the analogy of hot regions in protein interfaces with protein cores. The light yellow and cyan are the two chains of a protein, magenta atoms belong to the NADPH molecules, dark blue atoms represent the folding core of the protein. Dark yellow and red atoms indicate the hot region in the interface between the two chains. Both the folding core and the hot region have similar organization of highly packed clustered atoms.

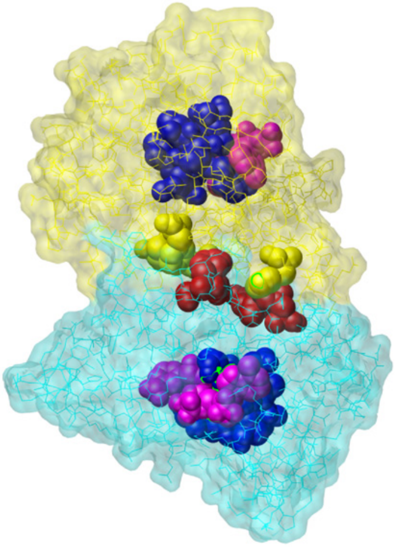

Figure 2. An example displaying the analogy of hot regions in protein interfaces with protein cores. The protein here is the carbonyl reductase complexed with NADPH and 2-propanol (PDB code 1cyd). A and B chains of the protein (light yellow and cyan colored regions) are displayed. The magenta atoms belong to the NADPH molecules dark blue atoms represent the folding core of the protein [101]. Dark yellow and red atoms indicate the hot regions in the interface between the two chains.

Below, we describe the steps and the results leading us to this new view of protein–protein binding sites. Overall, one may envision that different combinations of self-contained hot regions may be utilized when a protein binds different partners through the same binding site. It further suggests how one can conceivably modulate the binding, toward new protein partners. Finally, small molecule binding sites on the protein may consist of a single such region.

(*a*) *Not all interface residues contribute equally to the binding free energy; some residues were experimentally shown to be 'energy hot spots'*

The question of 'what makes a binding site a binding site' has already been posed a number of years ago [32–34].

While the definition is still unclear; conceptually, a binding site is usually described as a region that interacts with a region on a second protein. The residues which comprise the binding site are taken to be those that have atoms which are in contact with atoms belonging to the second protein. To estimate the binding free energy, one accounts for hydrogen bonds, electrostatic interactions, solvation, the hydrophobic effect and the vdW terms. The stability of the complexed proteins has been taken to be the sum of the interacting components.

A number of years ago, Jim Wells and his colleagues discovered the existence of 'energy hot spots', that is, residues that contribute significantly (>2 kcal mol^{-1}) to the binding free energy [35]. These residues have been identified through systematic substitutions of residues at the binding sites to alanine in a procedure commonly known as alanine scanning [25–27]. Subsequently, computational methods have been developed to predict these residues [35, 36]. Bogan and Thorn proposed that hot spots are surrounded by what they called 'O-rings' [26]. These are hydrophobic regions which may serve to exclude water from the hot spot residue. Thus, not all amino acids contribute equally. Some contribute marginally or not at all. On the other hand, a few others dominate the stability of the complex. As we discuss below, the structurally non-redundant dataset of protein–protein interfaces allowed a comparison between hot spots and structurally conserved residues [30, 31].

(*b*) *The derivation of a dataset of protein–protein interfaces: interfaces do not consist of residues arranged sequentially on the protein chain*

The first important step in the statistical exploration of protein–protein interactions is the availability of an appropriate dataset of protein–protein interfaces [37, 38]. The generation of such a dataset is not trivial. Most structural comparison algorithms are amino-acid sequence order dependent. Yet, an interface consists of bits and pieces of each of the chains, and some isolated residues. Consequently, a structural comparison algorithm which follows the chain order will be unable to generate a non-redundant structural dataset. Thus, to create the dataset, we used a structural comparison algorithm which is computer vision based and views protein structures as collections of (unconnected) points in three-dimensional space [39]. We applied the algorithm to all interfaces in the PDB [40]. Iterative clustering and loosening the criteria at subsequent levels [37, 38] led to 3799 clusters. Further filtering [41, 42], reduced the number of clusters to 103. Using MultiProt [43] we obtained the structurally conserved residues. MultiProt is also sequence order independent. It detects recurring motifs in an ensemble of proteins by simultaneously aligning multiple protein structures.

(*c*) *The diverse dataset allows a re-investigation of protein–protein binding sites and a new description*

The current description of a protein–protein binding site does not adequately account for all facts. It neither explains what makes some few residues hot spots, nor why the summation of the single residue contributions over- (or, under-) estimates the binding free energy. Furthermore, it does not explain why, even if the interfaces are very large, the maximal stability of the protein–protein interaction does not exceed a biological functional value. The non-redundant dataset allows a re-investigation, leading us to a new description: we propose that a binding site be viewed as consisting of unconnected hot regions. Within these regions, the interactions are optimized and cooperative. Between them, there is no optimization. Overall, the stability of the protein–protein complex is largely the sum of the interactions of the hot regions. This description is derived from several observations. First, we find that the hot spot residues are not homogeneously distributed across the interface; rather, they cluster in the 'hot regions'. Second, previously, we have shown that experimental hot spots correlate with structurally conserved residues. Now we find that conserved residues similarly cluster. Third, both the hot spots and the conserved residues are in locally highly packed regions. Within a 'hot region', the hot spots (and the computationally conserved residues) form a network of interactions among themselves and with other residues with high conservation ratios. Figure 3(*a*) displays a ribbon diagram of an example of the hot region clusters. The two encircled regions are the hot regions in the complex. The two chains composing the complex are in yellow and green. The hot spots, colored red and cyan, belong to the first and second chains of the protein, respectively. Figure 3(*b*) is an enlarged view of the upper hot region in a ball and stick representation. The coloring is the same as in figure 3(*a*). The dashed lines represent the pairs of atoms that are in close proximity (the distance between the pairs is less than 5 Å). Figure 4 illustrates the residue packing of Concanavalin lectin A. The protein is color coded according to the packing density of the residues, i.e., the highly packed regions are colored magenta. The least packed residues are colored green and cyan. The magenta-encircled residue clusters correspond either to the folding core or the hot regions of the complex. The picture that emerges is that protein binding sites consist of hot regions, with the residues within a hot region being tightly packed, and consequently likely to be highly conserved and to contribute significantly to the free energy of the association. This explains why such a hot region has a cluster of conserved residue hot spots. Since the conserved hot spot residues cluster and they interact with each other, their contribution is cooperative. Tight packing further contributes to screen the solvent, enhancing the strength of the electrostatic interactions. In addition, there are significant contributions of backbone hydrogen bonds in the hot spot and conserved residue interactions, further pointing to well packed regions. In contrast, between the hot regions, the packing is not optimal. On average, we do not observe hot spots, and similarly, no residue conservation. Hence, between the hot regions, there is no cooperativity in the residue contribution.

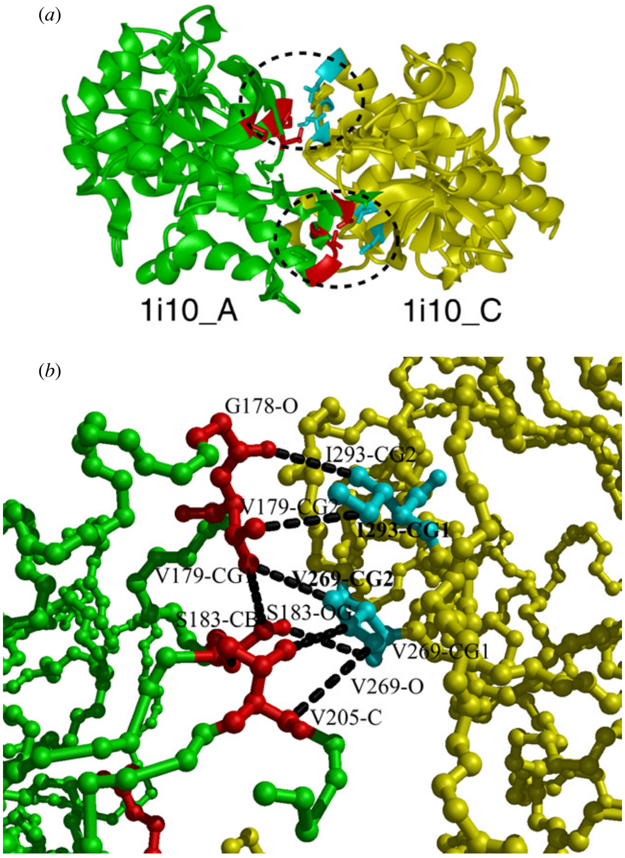

Figure 3. (*a*) The ribbon diagram of the protein human muscle l-lactate dehydrogenase (PDB code 1i10). This is an example for the hot region clusters. The two encircled regions are the hot regions of the complex between its A and C chains (yellow and green). The hot spots colored red and cyan belong to the first and second chains of the protein, respectively. (*b*) The enlarged view of the upper hot region in ball and stick representation. The coloring is the same as in part (*a*). The dashed lines represent the pairs of atoms that are in close proximity (the distances between the pairs are less than 5 Å).

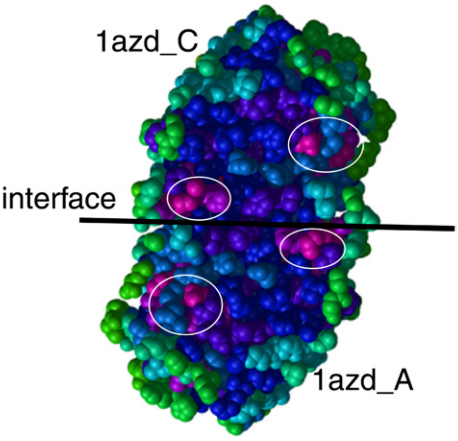

Figure 4. The residue packing of Concanavalin lectin A (PDB code 1azd). A and C chains are displayed. The protein is color coded according to the packing density of the residues; a magenta-to-cyan spectrum is used in the figure to represent the different levels of packing. The most highly packed regions are colored magenta. Dark blue, cyan and green residues represent the decreasing levels of packing, respectively. The horizontal line is the axis of symmetry of the complex assigned through the interface between the A and C chains. The magenta-encircled residue clusters correspond either to the folding core or the hot regions of the complex.

This view leads us to a new definition of a protein–protein binding interface: an interface is comprised of cooperative, locally densely packed 'hot regions'. This view of a cooperative hot region is attractive since it is consistent with current understanding of cores in protein folding. In agreement with the notion that protein folding and protein binding are similar processes with similar underlying mechanisms, a cooperative 'hot region' may resemble a core of a domain. The absence of cooperativity between hot regions may conceptually be viewed as two stable domains in a multi-domain protein. Hence, binding and folding are similar: cooperativity is observed in local tightly interacting regions and in the protein cores. Sheer counting under-represents the number of conserved charged residue couples in binding and in folding. In both interacting across-the-interface hot regions and in protein cores, electrostatics is enhanced through solvent screening.

Such an interface organization is entirely rational: when pre-folded proteins associate, one cannot expect optimization of the interaction across the interfaces. Local optimal interactions with less optimized regions in-between, allows for flexibility in the non-optimized regions and explains the observation that even presumably specific proteins, can bind a range of ligands, with different shapes, sizes and composition.

(*d*) *An experimental corroboration: protein–protein binding interfaces have modular architecture*

A nice experimental corroboration of the view presented here has recently been presented by Reichmann *et al* using the TEM1-beta-lactamase inhibitor protein (BLIP) system [44]. Through multiple-mutant analysis and x-ray crystallography, Schreiber and his colleagues have shown that the protein–protein interface consists of modules. A module is comprised of a number of closely interacting residues, with few interactions between the modules. The authors show that within a module, mutations cause complex energetic and structural consequences. On the other hand, the structural and energetic consequences of the removal of entire modules are small.

(*e*) *Hot spots (and structurally conserved residues) are coupled across the interface*

Further analysis of both the alanine scanning data and the non-redundant protein–protein interface dataset indicates that hot spots and conserved residues are coupled across the two chain interface [45]. Both are frequently found in complemented pockets [46]. Analysis of the across-the-interface conserved residue couples is particularly interesting: conserved charged residue couples are unfavored, despite the fact that using electrostatics one can predict the hot spots well [35, 36]. Furthermore, the flexible Gly is favorably coupled with aromatics, polar and small hydrophobic residues, again pointing to backbone H-bonds in the hot regions. Thorn and Bogan have collected the hot spot residues, making it a useful resource for studies of protein–protein interactions [47].

1.3. Interfaces have hydrophobic patches with the key residues pre-organized in the unbound state

Over the years, considerable efforts have been invested in studies of protein–protein interactions. The results were not surprising: hydrophobic residues are more frequent at the interfaces as compared to the rest of the protein surface. On the other hand, they are not as frequent as in the interior of the proteins. And, the complement also holds: polar residues are more frequent at the interfaces than in protein cores. However, they are not as frequent at the interface as compared to the remainder of the protein surface. Analyses have further been carried out on the frequencies of hydrogen bonds and salt bridges, on pairwise interactions and on interface sizes in different types of complexes. In particular, Janin and co-authors have shown that binding sites consist of patches of hydrophobic cores surrounded by more hydrophilic shells and that they may bury substantial amounts of surface area, reaching 3000–4000 Å^2 or more [13, 17, 34, 48].

Furthermore, recently, Rajamani *et al* [49] have noted that some side chains on the surface of the free, uncomplexed protein binding site are found in conformations similar to those observed in the bound complex. They further found that these 'ready-made' recognition motifs correspond to surface side chains that bury the largest solvent-accessible surface area after forming the complex ($\geqslant$100 Å^2). These side chains correspond to those of Li *et al* [46] which were detected based on the residue 'hot spot' conservation, found to frequently reside at the bottom of complemented pockets. Thus, both studies observe a pre-organization of these hot spot residues already in the unbound state.

1.4. Transient protein associations and disordered (versus ordered) protein complexes

Nooren and Thornton have characterized 'transient' protein–protein interactions [20, 50]. They have collected and analyzed two sets of complexes. The first contained 16 'weak' transient homodimers, with dissociation constants in the micromolar range. These are known to exist both as monomers and dimers at physiological concentrations. The second had 23 functionally validated transient heterodimers. The second set included more stable complexes, with nanomolar binding affinities. These complexes need a molecular trigger to form and break the interaction. Compared to the more stable homodimers, the weak homodimers have smaller, less hydrophobic and more planar contact areas at their interfaces. The physicochemical properties of these weak homodimers resemble those of non-obligate hetero-oligomeric complexes, whose composite monomers can exist on their own *in vivo*. On the other hand, the strong transient dimers often undergo large conformational changes upon binding. Overall, the molecular components of transient associations are stable in solution both as monomers and in their complexed form. Seraphin [51] has reviewed experimental methods for identification of transiently interacting proteins and of stable protein complexes.

We have analyzed structural characteristics of several types of complexes, such as natively unstructured (or disordered) proteins, ribosomal proteins, two-state and three-state complexes and crystal-packing dimers [18, 19]. The analysis revealed that the disordered proteins often have large intermolecular interfaces, the size of which is dictated by protein function (figure 5). Based on this observation, we proposed that disordered proteins provide a simple yet elegant solution to having large intermolecular interfaces, but with smaller protein, genome and cell sizes [18].

1.5. Different proteins may associate in similar ways

Remarkably, analysis of our interface clusters has indicated that while the interfaces are structurally similar, the proteins from which the interfaces are derived may be different, structurally and functionally [52]. The observation that different protein structures may associate in similar ways to yield preferred architectural motifs is again reminiscent of protein cores. There also, there are preferred motifs. Furthermore, in single chain proteins, similar motifs may be involved in different functions.

We further propose that similar principles and binding site description should hold for DNA/RNA binding. Indeed, there are indications for the validity of such a proposition. Experimentally, interaction networks have already been shown for protein–RNA interaction by Showalter and Hall [53].

2. Protein–protein interactions within the context of Systems Biology

2.1. Structural mapping protein–protein interactions

Above, we sought to understand the micro-organization of protein–protein interactions. We are spurred toward this goal not merely by the intellectual challenge; rather, figuring out the principles of protein–protein interactions will hopefully lead toward two practical goals. First, once the major components of the stabilizing interactions are understood, it may facilitate designing drugs to block the critical interactions in cases where the binding leads to disease [54]. And second, it should facilitate prediction of interactions. Mapping of protein–protein interactions assists in predicting protein function and in the construction of interaction maps. The function of newly discovered proteins such as those derived from the structural genomics initiative may be assigned through proteins binding to it. One would assume that if a protein is observed (or, predicted) to bind to proteins known to be components in a certain pathway, this new protein also plays a role in that pathway. Such an observation eventually leads to maps of the functional networks of the proteome and of all macromolecules in the cell, including protein-nucleic acids. This is essential to the understanding of how gene expression is controlled. Such a procedure may be viewed as a bottom-up strategy in Systems Biology. The availability of maps should not be looked at only as a mere enumeration of static interactions. Rather, a structural map of the macromolecular interaction

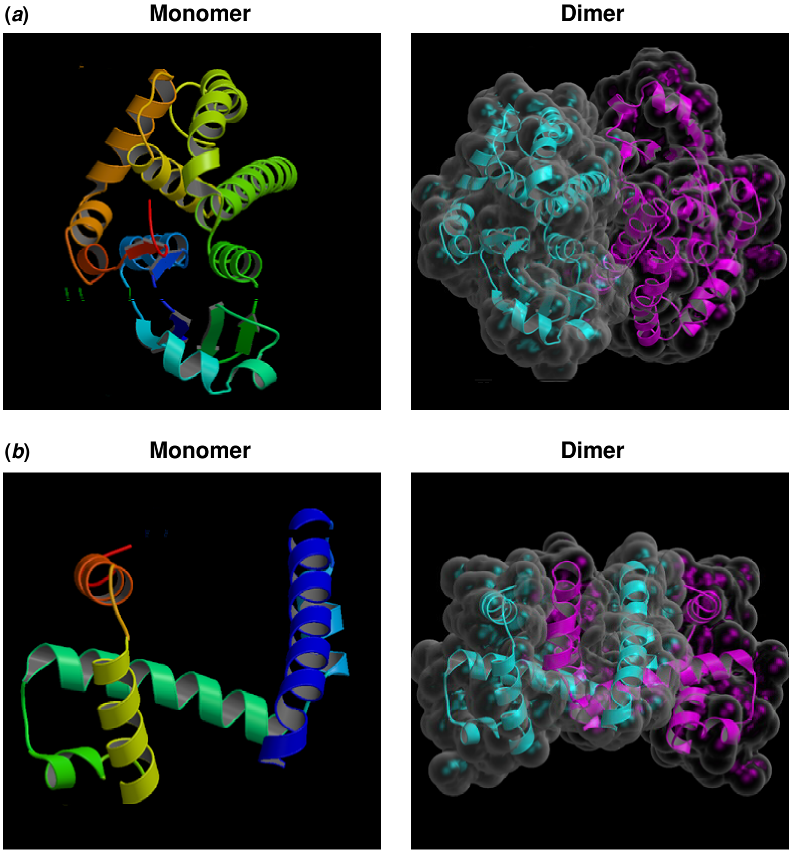

Figure 5. Ribbon representation of the monomer structure of the (*a*) ordered and (*b*) disordered proteins. The transparent solvent-accessible surfaces generated by GRASP for the dimers are shown on the right panels. (*a*) Glutathione S- transferase (1glq) which is stable as monomer or dimer (ordered case). (*b*) Beta-nerve growth factor (1bet) which is stable only as a dimer (disordered case). Disordered proteins (for example, two-state folders (*b*)) tend to have more extended shapes leading to a larger interface area compared to the globular and more compact ordered proteins (three-state folders (*a*)) [45, 46].

network may allow comprehension of the dynamics of the system. This is the essence of control mechanisms and of functional switches. Static maps of protein interactions tell us which proteins interact; however, they do not tell us under which conditions which paths dominate and how; and which intermolecular interactions overlap and which can co-exist. To understand the dynamics of the system on the molecular level we need to know not only *which* proteins interact, but *how* they interact. This implies that we need to have at our disposal the structures of the proteins and the structures of their associations.

The problem of protein–protein interactions within the structural context leads to a number of problems: on the technical side, obtaining the structures of large, multi-molecular assemblies at high-resolution is an extremely difficult task. Current high-resolution methodologies (x-ray and NMR) have difficulty obtaining such information. Cryo-electron microscopy (cryo-EM) is currently the method of choice [55–57], however, to date the resolution is not of atomic scale. From the computational standpoint while there are algorithms addressing this problem [58], they are still facing major hurdles toward this goal. It now appears that a promising strategy would involve a combination of low (cryo-EM) resolution of the assemblies, high (x-ray) resolution of the monomers (or in their absence, their modeled structures) and efficient algorithms to combinatorially put the monomers together and fit them against the EM maps [59]. The predictions of how the molecules interact imply knowledge of which association may—or may not—co-exist. If for a given protein X proteins Y and Z bind at the same site, these three proteins cannot form a complex simultaneously. Such information cannot be obtained from an interaction map enumerating protein–protein interactions. Which two proteins associate at a given time is a function of a conformational switch in protein X, Y or Z. This is the mechanics of the system. In addition, among the factors affecting which two proteins associate at any given time, temporal and spatial expression, compartmentalization and dynamics, mass action and competition clearly play crucial roles in the degradation of these processes.

Inspection of protein interaction maps or of databases of interactions (such as DIP [60] or BIND [61]) reveals that some proteins function as 'hubs'. These proteins can bind to a large number of partners. Even if we assume that these numbers contain a large error due to an experimental over-expression of the protein, nevertheless the experimental observations indicate that central hub proteins may have a large number of potential linkages. In contrast, some proteins have been observed to bind to a single or very few partners. This raises a few intriguing questions: what differentiates a hub protein from an edge one? Furthermore, if a protein functions as a hub and binds many different proteins, is its surface covered with binding sites or are the same binding sites utilized for binding to the various proteins? Since these proteins may have different sequences and different global folds, how can they be recognized by a given site on the host protein? Moreover, is there any characteristic property which distinguishes a protein, such that *a priori* it is earmarked to be a hub, or a binding site earmarked to be promiscuous toward a range of target proteins? Or, does it evolve toward this role, with optimization of its existing properties? Insight into these questions will be useful in addressing the crucial question of the Systems dynamics, and the network path choice under different sets of conditions.

In a recent insightful review, Beckett [62] described cases where a given site can bind to different proteins utilizing the same set of residues. This is not surprising. When analyzing the clustered protein–protein interface dataset that we have recently created [37], we have observed that whereas some clusters contain interfaces whose parent proteins are globally structurally similar and have the same function (type I clusters), this is not always the case. There are many clusters where the interfaces are similar; however the parent proteins have globally different structures and different functions (type II [52]). This suggests that regardless of function, there are favorable interface motifs, similar to preferred protein folds. Thus binding sites with given geometries and chemical properties can be utilized to bind a range of proteins. Furthermore, our studies indicate that there is no single property that differentiates between hub proteins and edge ones (Rogale *et al*, unpublished). This argues that hub proteins are likely not to be covered by binding sites; rather, the same sites will be employed for the many partner interactions.

If the same sites are utilized, what would be the micro-environment used in the different partner interactions? Since we observe that protein–protein interfaces may be described through *hot regions*, and since the hallmark of the hot regions are the tightly packed *hot spots* which are conserved throughout evolution, this argues that the hot regions are re-utilized by the different partners, although possibly with different combinations. Hence, evolution has not earmarked a protein or a site for multiple binding partners. Instead, as the proteome evolved and the systems developed to become more complex with webbed networks and multiple parallel routes, some proteins gradually acquired additional links. To obtain higher efficiency in their multi-binding capacity, some of their pre-existing properties were optimized. Since however the starting structures of these proteins differed from each other, the optimized configurations we currently observe involve a range of attributes. Thus, it appears that there is no prescribed set of properties which distinguishes a central protein or promiscuous binding site as compared to an edge protein.

2.2. *Centrally located proteins versus edge proteins*

As an example, here we focus on the yeast protein interaction network. To examine highly connected proteins from a structural viewpoint, we related the protein–protein interaction data with structural data from PDB. On the one hand, protein–protein interactions from several high-throughput experiments [63–66] with yeast two hybrid and tandem affinity purification techniques are available in DIP [67]; on the other hand, there are 346 yeast proteins with PDB structures as of 23 November 2004, although some of these are present only as fragments. The high-throughput experiments give rise to a protein–protein interaction map as in figure 6. Some ‘hub’ proteins clearly have high connectivity (i.e., a large number of proteins that interact with the given protein), while some others are ‘edge’ proteins and interact with very few proteins. The connectivity distribution follows a power law with the mean connectivity estimated at five [68].

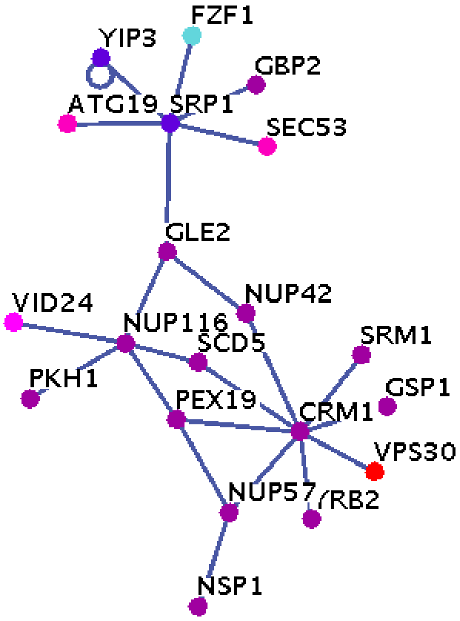

Figure 6. A portion of the yeast protein–protein interaction network [86] containing proteins involved in transport. Importin alpha (SRP1) and component of nuclear pore complex (NUP116) are ‘hubs’ or central proteins, while VID24 and VPS30 are examples of edge proteins.

One might ask what distinguishes the yeast hubs from the edge proteins. For example, (i) does the highly connected protein utilize more residues for interfacing? Figure 7(*a*) suggests this is not necessarily so. There are several classes of examples. First, metabolic proteins such as transketolase (connectivity 1) are usually edge proteins; but often function as homodimers with extensive homodimerization interfaces, 20–60 residues per interface. Second, large multi-protein complexes, such as RNA polymerase, nucleosome or DNA clamp–clamp loader complex, also have extensive interfaces involving 15–30 residues with other sub-units of the complex. While interface sizes within a complex may be comparable, connectivity among sub-units can vary greatly. In part, high connectivity of some components stems from interactions in the complex, but mostly it reflects the importance of the complex for the cell function; for example, histones which form the nucleosome interact with various chromatin remodeling and histone acetylation complexes and DNA repair

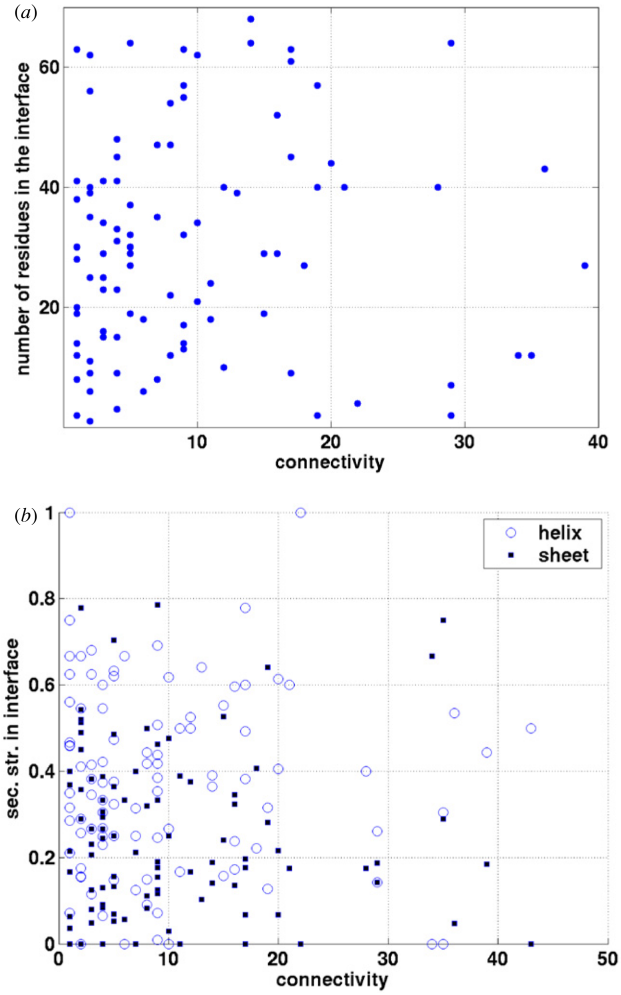

Figure 7. Two properties of interfaces of yeast proteins (determined from PDB structures) versus connectivity of yeast proteins (determined by large-scale experiments deposited in DIP). A residue is considered to be in the interface if its alpha carbon is less than 6.5 Å away from an alpha carbon from a different protein. All the interfaces found in the PDB are included (counting each residue once, even if it appears in multiple interfaces). Each dot is one protein. Outliers (extremely large connectivity or interface, such as largest sub-unit of RNA polymerase; less than five proteins) are not shown for clarity of the picture. (*a*) Number of residues in the interfaces of a yeast protein versus its connectivity. (*b*) The *y*-axis represents residues in secondary structure conformation (helical or beta sheet) as the percentage of the total number of residues in the interface of a protein. The scatter plot shows that no secondary structure is preferred by highly connected proteins.

proteins. Third, some proteins with low connectivity can have fairly large interfaces as heterodimers, such as the ubiquitin–hydrolase complex. Fourth, there are proteins with high connectivity whose biological unit is homo-oligomer and the interface sizes in the complex are comparable to those in hetero-oligomeric complexes, e.g., small nuclear ribonucleoprotein F. Fifth, both low and high connectivity proteins can have quite small interfaces, especially in transient complexes, such as signaling. (ii) Alternatively, does a highly interactive protein have more independently folding domains, each of which contributes to the connectivity of the protein? We inspected 907 yeast proteins which are covered by Pfam domains (a manually curated database of hidden Markov models obtained from high-quality alignments), but observed no correlation between the higher number of domains and the protein connectivity (data not shown). These results are consistent with the view that highly connected proteins adapt to this role mainly through the number of partners a single interface can accommodate, rather than through an increase in the number of interfaces or domains. (iii) Does the highly connected protein prefer a particular secondary structure in the interface, for example, do helices provide a better architecture for promiscuous binding sites? Figure 7(*b*) suggests again that this is not the case. Data from 104 proteins, where connectivity, complete PDB structure (as opposed to only a fragment) and complexation data with other proteins are simultaneously available, show no bias in the highly connected proteins toward any particular secondary structure in their interface(s). Of course, one potential caveat in our analysis is that PDB is not only sparse in the number of proteins (only 5% of the yeast proteome), but also in the number of structures containing two interacting proteins. For example, the maximum number of partners among yeast proteins in PDB is 7, while the highest connectivity exceeds 200. Moreover, in the yeast PDB dataset, homodimers and multi-protein complexes with many sub-units seem to be over-represented. Thus, figure 7 indicates a lack of correlation between the analyzed structural parameters and the protein connectivity.

So what can we finally say about distinguishing hub versus edge proteins? One conclusion is that protein function seems to be paramount, as observed for example in acquisition of partners during evolution. As an illustration, metabolic proteins are among the oldest proteins, found in all three kingdoms of life; yet they typically have very few interaction partners and are thus found at the edge of the interaction network. The highly connected proteins belong to the class of proteins that appeared in the eukaryotic radiation [69], especially in regulation. The fact that the interactivity of the protein depends on its function and not on its age is a further proof that the interactivity of proteins is optimized in evolution based on evolutionary pressures on the protein in the context of the entire system.

It is interesting to note that the most highly connected proteins are among those that perform the same function for many of their partners; for example, the second most highly connected protein in yeast is a kinase CDK1 that phosphorylates more than 200 proteins in the progression of cell cycle [70], and the third most highly connected protein is importin, which helps translocate proteins destined for the nucleus from the cytoplasm [71]. In such proteins, the interface with their partners is a single, highly promiscuous interface, representing the most economical way to perform a particular function in terms of the number of proteins needed for it. Incidentally, importin's interface also contains (at least) two hot regions, as exemplified in its partners' nuclear localization sequence (NLS). Some proteins destined for the nucleus contain a 'monopartite NLS', which overlaps both hot regions on the importin's interface, while others have 'bipartite NLS', two distinct basic sequences, separated by 10–12 residues, each of which fits in one of the two pockets.

2.3. Conclusions and outlook: protein interactions and Systems Biology

Here, we have addressed the micro- and the macro-scale protein interaction environment. We first examined the organization of protein–protein interfaces. Conserved hot spot residues cluster within the locally densely packed self-contained hot regions and form a network of interactions with each other and with other residues around them. This description of an interface implies that within a hot region the contribution of the residues to the stability of the protein–protein association is *cooperative*. On the other hand, between the hot regions, the complementarity across the interface of the interacting molecules is not as perfect. The packing density is not high and there are no clusters of networked hot spot residues allowing binding site flexibility. This view of the protein binding site highlights the analogy between binding and folding: in both there are regions of crucial cooperative interactions, whether in the densely packed protein cores or at the interface. The collection of independent hot regions in binding resembles the cores of individual domains in folding. This description of protein–protein binding may suggest how a given (e.g., signaling) protein may bind to different proteins. Optimization of different combinations of hot regions may suggest how a protein may efficiently participate in parallel (alternate) pathways, in multi-molecular associations and in cellular organization at different integrative levels. A hot region may further constitute a small molecule binding site and provide a target for drug design [54].

On the macro-scale, the challenging goal of Systems Biology is to integrate the different levels of information to explore the cellular complexity [72–82]. The aim is to combine available experimental and computational data to characterize the network of intermolecular interactions and their regulation. Eventually, all molecules and processes in the living organism are interconnected. Interconnectivity is reflected in the multiple symptoms of diseases, the changes that take place during disease progression and the side effects of drugs. Systems Biology probes how molecules interconnect as a network, and how the expression and the functional level are regulated within the network. As pointed out in the excellent reviews by Kitano [72–74] the properties of the system are key goals. A good example is the robustness of expression and regulation. Robustness implies that there are a myriad of ways for functional expression. This is why a given drug which blocks one pathway may not be an effective strategy, as observed in the fight against cancer.

Systems Biology is an extremely complex discipline with very noisy data. The data derive from a wide range of experimental proteomic tools, such as micro-array experiments, the yeast two hybrid screens, mass spectrometry, green fluorescent tagging and 2D polyacrylamide gel electrophoresis. The data are neither uniform nor clean: the problem of over-expression may lead to inaccurate results. Furthermore, studies are often performed *in vitro* and under different conditions. Such difficulties are inevitable when a problem is approached on this scale. Despite these problems, the aim is to fish out the relevant interactions between the molecules and to construct the global network recognizing that a certain error cannot be avoided. Experiments which can further be extremely useful entail gene knockouts, and analyses of which processes are affected. The goal is to combine large-scale data and specific information to model cellular processes by identifying the proteins which interact and the interactions between protein–DNA and RNA. Furthermore, it is essential to have data related to the dependence among the levels of expression and how the system corrects itself in the case of a malfunction of a specific protein. Combined, these data can be used to computationally simulate the cellular processes. In turn, predictions made through simulations can be tested by experiment. The volume of data is such that efficient computational schemes are crucial. Systems Biology is the next essential step in putting the molecules together within the framework of the cell. A Systems approach should allow the nature of the processes to be addressed, their regulation and the way they respond to a broad range of perturbations. It allows simulations of the dynamics of the cellular machinery, mimicking the innumerable ways in which the complex cellular machinery operates. This should facilitate design of an effective drug strategy, again based on a Systems approach.

To carry out this mission, an essential step in Systems Biology is a catalog of the interacting molecules and putting these together to create a map. An organized map is essential, as it provides the network of the cellular interactions [76]. As such, it serves as the basis for the understanding and the prediction of function. Nevertheless, such maps yield a global picture and do not tell us *how* the components are connected [83, 84]. Since what we have is the connectivity between the components, it can only provide static information. It further does not provide information regarding which intermolecular associations can co-exist simultaneously and which are exclusive of each other. Connectivity maps are insufficient to lead to an insight into the dynamics of the system, that is, how changes in one part affect the others.

A bottom-up structure-based approach focuses on proteins [87–97]. The goal is to predict which proteins interact and how the interactions will take place. Putting these together should create a structure-based map of interactions. Predictions of the structural associations, the pathways and the assemblies will provide information regarding which of the interacting proteins binds at the same site and which interactions can co-exist (i.e., do not overlap). This can conceivably be done in either of two ways: (1) by docking structures of proteins (known or unknown to interact from the connectivity map [33, 98–100]), or (2) using protein–protein interfaces derived from the PDB. Through structural comparisons of the binding sites of one side of the interface, and applying the transformation which is obtained by the structural superposition to ‘dock’ the complementary side of the interface, we may predict protein interactions and build (predicted) structural maps. The predicted interactions can be cross-checked with experimental databases of protein–protein interactions. There are advantages and disadvantages to each of these schemes. On balance, the second, structural comparison of interfaces approach appears more robust. Among the ingredients which are needed are non-redundant structure-based datasets of protein–protein interfaces; state-of-the-art pairwise and multiple structural comparisons and docking algorithms; and preliminary results which are verified in the experiment-based databases of protein–protein interactions [85].

Acknowledgments

We thank Drs C-J Tsai, Y Pan, D Zanuy, H-H (G) Tsai, members of the Nussinov–Wolfson group and Dr A Gursoy and S Aytuna from Koc University for their helpful discussions. The research of RN in Israel has been supported in part by the *Center of Excellence in Geometric Computing and its Applications* funded by the Israel Science Foundation (administered by the *Israel Academy of Sciences*). This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Glossary

Hub proteins versus edge proteins. Assume that we create a scatter plot of all proteins in the cell, where each protein is represented by a node. Next, we draw edges between nodes observed to interact, either *in vivo*, or in the test tube (*in vitro*) in some expression screens. Those nodes which are connected to many others are the *hub proteins*. On the other hand, proteins connected to one or very few proteins are the *edge proteins*.

Hot region organization. A *hot region* consists of residues spatially adjacent to each other, in a compact organization. These regions contain at least one ‘hot spot’ residue, i.e., a residue which has either been shown experimentally to contribute significantly (more than 2.0 kcal mol^{-1}) to the binding free energy or has been found to be conserved in a multiple structure (or, sequence) alignment. A hot spot residue is tightly packed and is in contact with other hot spot residues or other residues with high conservation ratios. Hot regions typically contain clusters of such residues. In contrast, the regions between the hot ones are not as optimally packed. The tight packing leads to the high conservation since it is difficult to accommodate mutations of these residues without either steric clashes or creation of ‘holes’.

Bottom-up strategy. A bottom-up strategy implies a strategy initiating from specific contacts between molecules and building the system up to create a map. In contrast, a top-down strategy initiates from the overall organization, trying to figure out the molecular components and the specific contacts which take place at each organizational level.

Systems Biology. A Systems Biology approach seeks to understand the entire system, rather than focuses on a specific molecule or a specific interaction. A Systems approach views the cellular machinery as a whole, and attempts to build the cellular interaction map, and in particular its dynamic regulation. Systems Biology studies are carried out with the explicit understanding that inevitably the data which are handled are noisy.

Structurally conserved residues. Structurally conserved residues are those residues that upon structural superposition of family members are observed to be conserved both structurally (i.e., occupy the same positions in space) and are sequentially in at least 50% of the compared molecules.

Cooperativity. Here, in the context of this paper, *cooperativity* implies non-independence in the contributions of the residues to the free energy of the protein–protein interactions. That is, the contributions of the residues are not additive. The sum of the contributions may either over- or under-estimate their actual contribution. This cooperative nature of their contribution arises since they are in a closely packed environment and interact with each other. Thus, mutation of one residue affects also the conformations and contacts of residues in its vicinity.

References

[1] Kleanthous C (ed) 2000 *Protein–Protein Recognition, Frontiers in Molecular Biology* (Oxford: Oxford University Press)

[2] Janin J and Wodak S 2003 Protein modules and protein–protein interactions *Adv. Protein Chem.* **61**

[3] Janin J, Henrick K, Moult J, Eyck L T, Sternberg M J, Vajda S, Vakser I and Wodak S J 2003 *Proteins* **52** 2–9

[4] Gallet X, Charloteaux B, Thomas A and Brasseur R 2000 *J. Mol. Biol.* **302** 917–26

[5] Ehrlich L, Reczko M, Bohr H and Wade R C 1998 *Protein Eng.* **11** 11–9

[6] Ofran Y and Rost B 2003 *FEBS Lett.* **544** 236–9

[7] Fariselli P, Pazos F, Valencia A and Casadio R 2002 *Eur. J. Biochem.* **269** 1356–61

[8] Lichtarge O and Sowa M E 2002 *Curr. Opin. Struct. Biol.* **12** 21–7

[9] Marcotte E M, Pellegrini M, Ng H L, Rice D W, Yeates T O and Eisenberg D 1999 *Science* **285** 751–3

[10] Eisenberg D, Marcotte E M, Xenarios I and Yeates T O 2000 *Nature* **405** 823–6

[11] Janin J 1995 *Biochimie* **77** 497–505

[12] Ma B, Shatsky M, Wolfson H J and Nussinov R 2002 *Protein Sci.* **11** 184–97

[13] LoConte L, Chothia C and Janin J 1999 *J. Mol. Biol.* **285** 2177–98

[14] Janin J and Chothia C 1990 *J. Biol. Chem.* **256** 16027–30

[15] Korn A P and Burnett R M 1991 *Proteins* **9** 37–55

[16] Young L, Jernigan R L and Covell D G 1994 *Protein Sci.* **3** 717–29

[17] Chakrabarti P and Janin J 2002 *Proteins* **47** 334–43

[18] Gunasekaran K, Tsai C J, Kumar S, Zanuy D and Nussinov R 2003 *Trends Biochem. Sci.* **28** 81–5

[19] Gunasekaran K, Tsai C J and Nussinov R 2004 *J. Mol. Biol.* **341** 1327–41

[20] Nooren I M and Thornton J M 2003 *J. Mol. Biol.* **325** 991–1018

[21] Chothia C and Janin J 1975 *Nature* **256** 705–8

[22] Dill K A, Fiebig K M and Chan H S 1993 *Proc. Natl Acad. Sci. USA* **90** 1942–6

[23] Kolinski A, Galazka W and Skolnick J 1996 *Proteins* **26** 271–87

[24] Keskin O, Ma B and Nussinov R 2005 *J. Mol. Biol.* **345** 1281–94

[25] Clackson T and Wells J A 1995 *Science* **267** 383–6

[26] Bogan A A and Thorn K S 1998 *J. Mol. Biol.* **280** 1–9

[27] DeLano W L 2002 *Curr. Opin. Struct. Biol.* **12** 14–20

[28] Tsai C J, Xu D and Nussinov R 1998 *Fold. Des.* **3** R71–80

[29] Haliloglu T, Keskin O, Ma B and Nussinov R 2005 *Biophys. J.* **88** 1552–9

[30] Hu Z, Ma B, Wolfson H and Nussinov R 2000 *Proteins* **39** 331–42

[31] Ma B, Elkayam T, Wolfson H and Nussinov R 2003 *Proc. Natl Acad. Sci. USA* **100** 5772–7

[32] Ringe D 1995 *Curr. Opin. Struct. Biol.* **5** 825–9

[33] Halperin I, Ma B, Wolfson H and Nussinov R 2002 *Proteins* **47** 409–43

[34] Bahadur R P, Chakrabarti P, Rodier F and Janin J 2004 *J. Mol. Biol.* **336** 943–55

[35] Kortemme T and Baker D 2002 *Proc. Natl Acad. Sci. USA* **99** 14116–21

[36] Guerois R, Nielsen J E and Serrano L 2002 *J. Mol. Biol.* **320** 369–87

[37] Keskin O, Tsai C-J, Wolfson H and Nussinov R 2004 *Protein Sci.* **13** 1043–55

[38] Tsai C J, Lin S L, Wolfson H J and Nussinov R 1996 *J. Mol. Biol.* **260** 604–20

[39] Nussinov R and Wolfson H J 1991 *Proc. Natl Acad. Sci. USA* **88** 10495–9

[40] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucl. Acids Res.* **28** 235–42

[41] Higgins D, Thompson J, Gibson T, Thompson J D, Higgins D G and Gibson T J 1994 *Nucl. Acids Res.* **22** 4673–80

[42] Henikoff S and Henikoff J G 1992 *Proc. Natl Acad. Sci. USA* **89** 10915–9

[43] Shatsky M, Wolfson H J and Nussinov R 2002 *Lecture Notes in Computer Science* vol 2452 (Berlin: Springer) pp 235–50

[44] Reichmann D, Rahat O, Albeck S, Meged R, Dym O and Schreiber G 2005 *Proc. Natl Acad. Sci. USA* **102** 57–62

[45] Halperin I, Wolfson H and Nussinov R 2004 *Structure* **12** 1027–038

[46] Li X, Keskin O, Ma B, Nussinov R and Liang J 2004 *J. Mol. Biol.* **344** 781–95

[47] Thorn K S and Bogan A A 2001 *Bioinformatics* **17** 284–5

[48] Bahadur R P, Chakrabarti P, Rodier F and Janin J 2003 *Proteins* **53** 708–19

[49] Rajamani D, Thiel S, Vajda S and Camacho C 2004 *Proc. Natl Acad. Sci. USA* **101** 11287–92

[50] Nooren I M and Thornton J M 2003 *EMBO J.* **22** 3486–92

[51] Seraphin B 2003 *Adv. Protein Chem.* **61** 99–118

[52] Keskin O and Nussinov R 2005 *Protein Eng. Des. Sel.* **18** 11–24

[53] Showalter S A and Hall K B 2002 *J. Mol. Biol.* **322** 533–42

[54] Arkin M R and Wells J A 2004 *Nat. Rev. Drug Dis.* **3** 301–17

[55] Rixon F and Chiu W 2003 *Adv. Protein Chem.* **64** 413–44

[56] Zhou Z H and Chiu W 2003 *Adv. Protein Chem.* **64** 93–130

[57] Ludtke S J, Chen D H, Song J L, Chuang D T and Chiu W 2004 *Structure (Camb.)* **12** 1129–36

[58] Inbar Y, Benyamini H, Nussinov R and Wolfson H 2003 *Proc. ISMB Bioinform. Suppl.* **1** I158–68

[59] Inbar Y, Benyamini H, Nussinov R and Wolfson H 2005 *J. Mol. Biol.* at press

[60] Salwinski L, Miller C S, Smith A J, Pettit F K, Bowie J U and Eisenberg D 2004 *Nucl. Acids Res.* **32** D449–51

[61] Bader G D, Betel D and Hogue C W 2003 *Nucl. Acids Res.* **31** 248–50

[62] Beckett D 2004 *Biochemistry* **43** 7983–91

[63] Gavin A C *et al* 2002 *Nature* **415** 141–7

[64] Ho Y *et al* 2002 *Nature* **415** 180–3

[65] Uetz P *et al* 2000 *Nature* **403** 623–7

[66] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y 2001 *Proc. Natl Acad. Sci. USA* **98** 4569–74

[67] Xenarios I, Salwinski L, Duan X J, Higney P, Kim S M and Eisenberg D 2002 *Nucl. Acids Res.* **30** 303–5

[68] Grigoriev A 2003 *Nucl. Acids. Res.* **31** 4157–61

[69] Kunin V, Pereira-Leal J B and Ouzounis C A 2004 *Mol. Biol. Evol.* **21** 1171–6

[70] Ubersax J A, Woodbury E L, Quang P N, Paraz M, Blethrow J D, Shah K, Shokat K M and Morgan D O 2003 *Nature* **425** 859–64

[71] Conti E, Uy M, Leighton L, Blobel G and Kuriyan J 1998 *Cell* **94** 193–204

[72] Kitano H 2002 *Science* **295** 1662–4

[73] Kitano H 2002 *Nature* **420** 206–10

[74] Kitano H 2004 *Nat. Rev. Cancer* **4** 227–35

[75] Davidov E J, Holland J M, Marple E W and Naylor S 2003 *Drug Discov. Today* **8** 175–83

[76] Kohn K W 1999 *Mol. Biol. Cell* **10** 2703–34

[77] Hood L and Galas D 2003 *Nature* **421** 444–8

[78] Kitano H 2002 *Curr. Genet.* **41** 1–10

[79] Huang Q, Raya A, Dejesus P, Chao S-H, Quon K C, Caldwell J S, Chanda S K, Izpisua-Belmonte J C and Schultz P G 2004 *Proc. Natl Acad. Sci. USA* **101** 3456–61

[80] LeBras M, Bensaad K and Soussi T 2003 *Oncogene* **22** 5082–90

[81] Haupt S, Berger M, Goldberg Z and Haupt Y 2003 *J. Cell Sci.* **116** 4077–85

[82] Tomit M 2001 *Trends Biotech.* **19** 205–10

[83] Bork P, Jensen L J, von Mering C, Ramani A K and Lee and Marcotte E M 2004 *Curr. Opin. Struct. Biol.* **14** 292–9

[84] Aloy P *et al* 2004 *Science* **203** 2026–9

[85] Aytuna A S, Gursoy A and Keskin O 2005 *Bioinformatics* at press

[86] Schwikowski B, Uetz P and Fields S 2000 *Nat. Biotechnol.* **18** 1257–61

[87] Wutchy S 2004 *Genome Res.* **14** 1310–4

[88] Arita M *Proc. Natl Acad. Sci. USA* **101** 1543–47

[89] Alberghina L, Chiaradonna F and Vanconi M 2004 *Chem. Bio. Chem.* **5** 1322–33

[90] Glaser P and Boone C 2004 *Curr. Opin. Struct. Biol.* **7** 489–91

[91] Hoffmann R and Valencia A 2003 *TiG* **19** 681–3

[92] Fraser H B, Hirsh A E, Wall D P and Eisen M B 2004 *Proc. Natl Acad. Sci. USA* **101** 9033–8

[93] Haugen A C *et al* 2004 *Genome Biol.* **5** R95

[94] Saucerman J J and McCulloch A D 2004 *Prog. Biophys. Mol. Biol.* **85** 261–78

[95] Jansen R *et al* 2003 *Science* **302** 449–53

[96] Xia Y *et al* 2004 *Annu. Rev. Biochem.* **73** 1051–87

[97] Baudot A, Jacq B and Brun C 2004 *Genome Biol.* **5** R76

[98] Fernandez-Recio J, Totrov M and Abagyan R 2003 *J. Mol. Biol.* **335** 843–65

[99] Russell R B, Alber F, Aloy P, Davis F P, Korkin D, Pichaud M, Topf M and Sali A 2004 *Curr. Opin. Struct. Biol.* **14** 313–24

[100] Janin J 2005 *Protein Sci.* **14** 278–83

[101] Townley H E, Sessions R B, Clarke A R, Dafforn A R and Griffiths W T 2001 *Proteins* **44** 329–35