doi:10.1016/j.jmb.2010.09.028 J. Mol. Biol. (2010) 403, 660–670

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution

Emmanuel D. Levy

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK
Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montreal, H3C 3J7, Canada

Received 24 June 2010;
received in revised form
19 August 2010;
accepted 13 September 2010
Available online
22 September 2010

Edited by M. Sternberg

Keywords:
protein;
interaction;
promiscuity;
interface;
evolution

Analysis of proteins commonly requires the partition of their structure into regions such as the surface, interior, or interface. Despite the frequent use of such categorization, no consensus definition seems to exist. This study thus aims at providing a definition that is general, is simple to implement, and yields new biological insights. This analysis relies on 397, 196, and 701 protein structures from *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens*, respectively, and the conclusions are consistent across all three species. A threshold of 25% relative accessible surface area best segregates amino acids at the interior and at the surface. This value is further used to extend the core–rim model of protein–protein interfaces and to introduce a third region called support. Interface core, rim, and support regions contain similar numbers of residues on average, but core residues contribute over two-thirds of the contact surface. The amino acid composition of each region remains similar across different organisms and interface types. The interface core composition is intermediate between the surface and the interior, but the compositions of the support and the rim are virtually identical with those of the interior and the surface, respectively. The support and rim could thus "preexist" in proteins, and evolving a new interaction could require mutations to form an interface core only. Using the interface regions defined, it is shown through simulations that only two substitutions are necessary to shift the average composition of a 1000-Å^2 surface patch involving ~28 residues to that of an equivalent interface. This analysis and conclusions will help understand the notion of promiscuity in protein–protein interaction networks.

Introduction

A central aim of biology is to understand the sequence–structure–function relationships of proteins and how protein evolution takes place in the space defined by these three elements. Relating the structure of proteins to their sequence, function, or evolution often involves defining structural regions. During the protein folding process notably, hydrophobic residues shield themselves from the solvent, yielding a hydrophobic protein interior and a hydrophilic surface.[1] This dichotomy between interior amino acids and surface amino acids has profound implications in stability and evolution. It was indeed observed early on[2] that amino acids at the surface of proteins are more free to mutate than those that are buried.[2–7] This is due to the more destabilizing nature of mutations in the protein interior, which are more likely to impair the structure and function of the protein[8] or even that of the cell.[9]

Amino acids involved in protein–protein interactions define an additional region in proteins. The structural properties of protein–protein interaction sites have been extensively studied, as witnessed by

E-mail address: emmanuel.levy@gmail.com.

Abbreviations used: ASA, accessible surface area; rASA, relative accessible surface area.

the many reviews[10–20] and books[21–24] published on the subject. A particular aspect of interest here is that the surface–interior dichotomy seen in protein folding can be applied to protein–protein interfaces, yielding the rim (~surface) and the core (~interior).[25,26] As for protein tertiary structure, amino acids at the interface core are more hydrophobic than those at the rim[14,25–28] and are more frequently hot spots.[29] This is also in agreement with evolutionary data showing that residues at the core are more conserved than those at the rim.[30–33]

Despite these clear similarities between protein surface and interface rim, and between protein interior and interface core, these regions have classically been studied separately. As a result, to the best of my knowledge, no existing definition of these regions allows their direct comparison. In this study, the first aim is therefore to define the interior and surface of proteins and to subsequently use that definition to decompose interfaces into structurally equivalent regions. This leads me to reexamine the current "core–rim" model and propose a third interface region called "support." Also, this novel way of partitioning protein structures provides new insights into both the organization and the evolution of protein–protein interactions sites. Additionally, the consistency of the results is validated across three distant species.

Results and Discussion

Defining surface and interior regions

A residue is usually considered buried if its relative accessible surface area (rASA) is below a cutoff value. Such a consideration may not seem to be biologically meaningful because the relative exposure of amino acids to the solvent is a continuous variable, not a discreet one. Nevertheless, two reasons motivate the definition of such discreet regions. The first reason is simplicity. It is indeed easier to compare a particular property (e.g., amino acid conservation) between two regions than according to a continuous variable.[6,8,30,34] The second reason that motivates the definition of discreet structural regions is given in the study of Tokuriki *et al.*,[8] where they conducted a large-scale *in silico* analysis of protein stability changes after the introduction of mutations. In all the proteins studied, the resulting distribution of $\Delta\Delta G$ values could be decomposed into a sum of two Gaussians, where one corresponds to surface residues and the second corresponds to buried ones.

This prompts the definition of two structural regions: the interior and the surface. Since it is known that amino acid composition differs greatly between these two regions, I aim at finding the rASA value that maximizes that difference, measured by the correlation of their amino acid frequency vectors. In order to ensure that the rASA value obtained is not strongly dependent on the data set used, this analysis is carried out on three data sets of proteins from *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens* containing 397, 196, and 701 proteins, respectively. Figure 1a shows that the composition between the surface and the interior is most different at a rASA value of ~25% for all three organisms. Interestingly, this value is the same as that described by Tokuriki *et al.*, which fits the bimodal distribution of the $\Delta\Delta G$ values that they observed.

Defining interface regions: Core, rim, and support

This idea of surface and interior can also be applied to protein–protein interfaces. Bogan and Thorn noted that among residues participating in the interface, those that contribute most to binding energy are shielded from the solvent.[29] They proposed a model where interfaces are made of a central desolvated region surrounded by energetically less important residues in contact with water. This was consistent with a model described in parallel—of interfaces made of a core of buried atoms surrounded by rings of accessible ones.[26] In line with these descriptions, the core–rim model was proposed,[25,28] where residues containing only solvent-accessible atoms form the rim and the remaining residues, with at least one buried atom, form the core.

This description can be extended using the rASA value of 25% determined above. Simply, if an interface residue (ΔASA>0; Fig. 1b) has more than 25% rASA in the complexed state, it is counted at the rim, which thus represents the intersection between the surface and the interface. Other residues can be assigned either to the core or to an additional category that I call support. The support includes residues with less than 25% rASA in the uncomplexed state and is thus the intersection between the interior and the interface. The remaining residues form the interface core. In other words, support residues are already largely buried in the monomer and become more buried in the complex, rim residues are largely exposed in the monomer and remain exposed in the complex, but interface core residues shift from being exposed in the monomer (>25% rASA) to being buried in the complex (<25% rASA).

Thus, in the same manner as proteins can be decomposed into interior and surface, interfaces can be decomposed into interior (support), surface (rim), and an intermediate category, the interface core. The size of these three categories changes with the rASA value used to separate the interior and the surface (Fig. 1c; Fig. S1). The larger is the rASA value, the more residues are at the interior and the larger is the support. Interestingly, at a rASA of 25%, the relative sizes of the three interface categories are comparable,

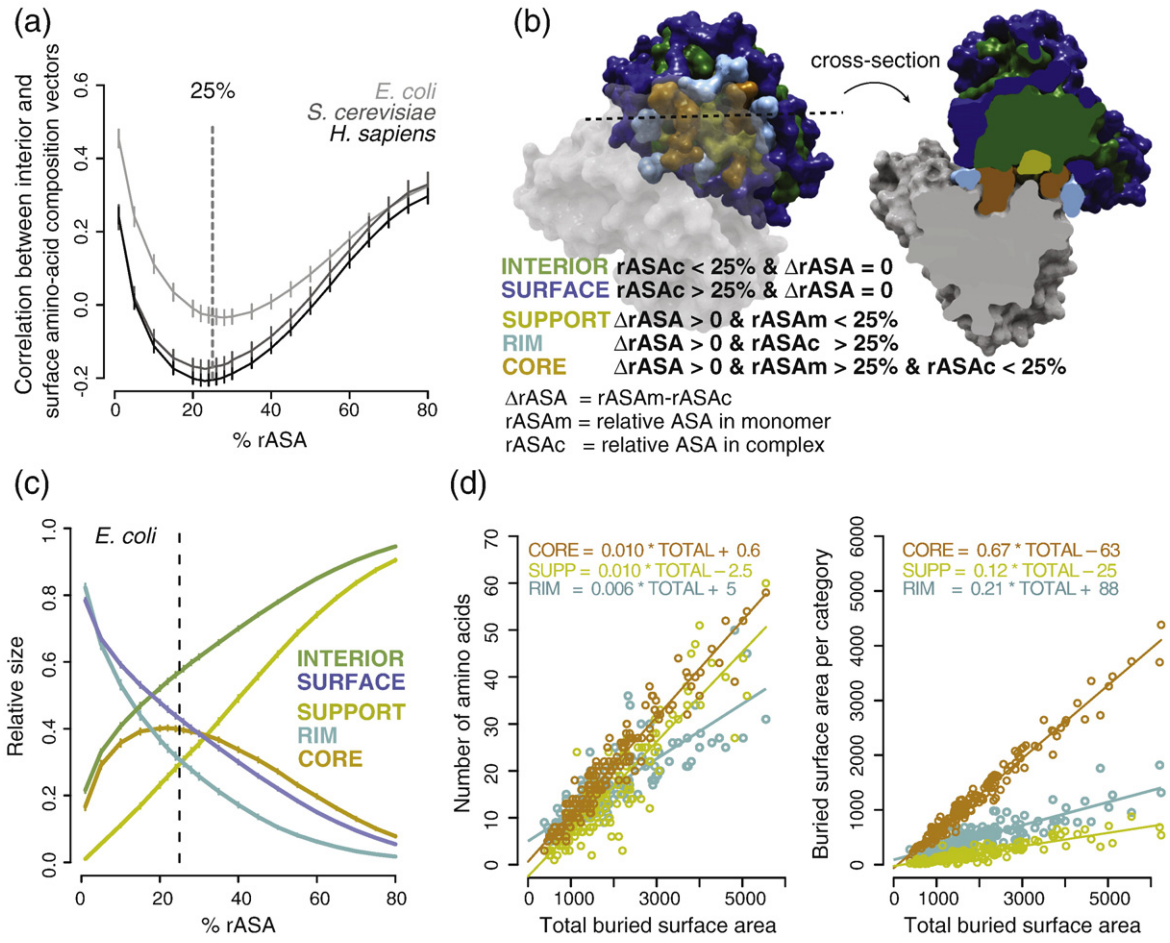

Fig. 1. Defining and characterizing the structural regions used in this study. (a) Amino acids are assigned to the surface or interior based on their rASA. For each rASA value, the frequencies of amino acids assigned at the surface and interior are calculated, and the correlation between their frequency vectors is plotted. The rASA value maximizing the difference between interior and surface is ~25% in the three species considered, and this value is used in the rest of the study. Bars indicate 95% confidence intervals. (b) Schematic depiction and definition of the structural regions proposed in this study. The definition of all regions depends on a single parameter (25% rASA) determined in (a). (c) Fraction of residues found in the different structural regions as a function of the rASA parameter among *E. coli* proteins. The relative contributions of the three interface categories sum to 1, as do the contributions of the interior and the surface. A 25% rASA yields comparable sizes for the three interface regions and maximizes the size of the core relative to the rim and the support. Bars indicate 95% confidence intervals (note that they are confounded with the line width). (d) The size of each interface region (in number of amino acids or buried surface area) as a function of interface size. Each protein is represented as three points, one for each region. The linear model that best fits the group of points corresponding to a particular category is given in the same color. Note that only dimers for these two plots are used.

and the size of the interface core is maximal. This is relevant to the prediction of protein–protein interaction sites, as interface core residues are expected to carry more "interface information" than the support or the rim. For example, amino acids at the support are expected to be hydrophobic, not necessarily because they contribute to the interface but simply because they are buried.

Below, I compare further the properties of the core, rim, and support defined at a rASA of 25%. Their contribution to the number of amino acids and buried accessible surface area (ASA) according to the total size of the interface is given in Fig. 1d. Importantly, all statistics are given per subunit. In terms of the number of amino acids, the core, rim, and support have comparable sizes. Looking at *E. coli* complexes, a linear regression shows an average of 1 core, 1 support, and 0.6 rim residue per 100 Å^2 of interface. These numbers are identical for *S. cerevisiae* complexes and remain close (i.e., 1, 0.8, and 0.7) for *H. sapiens* complexes (Fig. S2). When the slope intercept is taken into account, a typical 1000-Å^2 interface involves an average of ~28 amino acids, with ~10 amino acids forming the core, ~8 amino acids forming the support, and ~10 amino acids forming the rim.

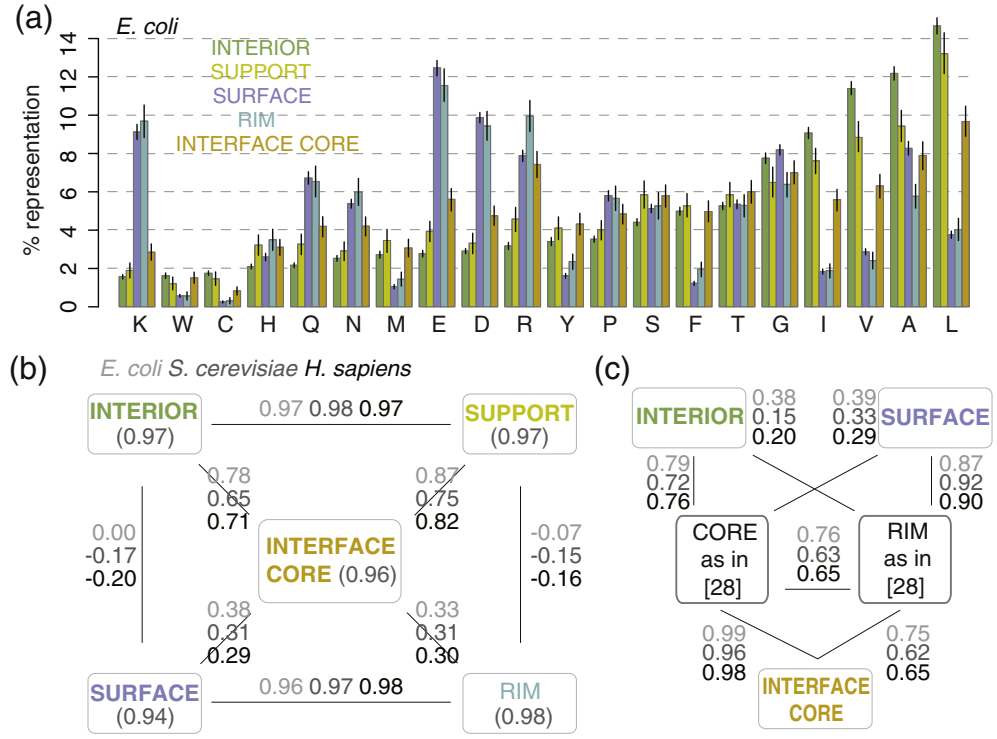

Fig. 2. Frequencies of amino acids among the five structural regions and their relationships to each other. (a) Bar plot of amino acid frequencies across the five structural regions. Bars indicate 95% confidence intervals. (b) Correlation between the amino acid frequency vectors of different regions. Values in parentheses indicate the correlation between the frequency vectors of the same region across different species (the lowest of the three correlation values is shown). The high values indicate that all regions have similar compositions across the different species. (c) Correlation between the amino acid frequency vectors of the regions defined in this study and the core–rim regions proposed in an earlier model.[25,28]

As noted previously,[25,28] the relative contributions are, however, quite different when considering the buried surface area, since the core alone buries twice more surface than the rim and the support together. An average core residue contributes 66–67 Å^2, against 11–14 Å^2 and 20–23 Å^2 for the support and the rim, respectively (the ranges reflect the differences among species, which are small; Fig. S2). Interestingly, however, the intercept of the core is negative, and that of the rim is positive for all three species. Although a negative intercept is not biologically meaningful, it reflects that the core contribution becomes weaker among smaller interfaces. If one assumes that amino acids at the core contribute most of the binding free energy,[29] this picture is consistent with the fact that the majority of transient interfaces bury less than ~800 Å^2 per subunit, a value around which the interface core contribution is moderate.

Amino acid compositions of the core, rim, and support

Although amino acid composition was used to determine the value of 25% rASA central to this study, the compositions that are obtained have not yet been described. In terms of interior and surface, Fig. 2a shows a classic picture where hydrophobic amino acids are preferred at the protein interior and charged amino acids are preferred at the surface. The interface core composition is intermediate between that of the interior and that of the surface, as described previously.[14,25–28] In contrast, the rim is nearly identical with the surface (light blue and dark blue; Fig. 2a), and the same applies to the support and the interior (light green and dark green) across all three species (Fig. S3). This is clearer in Fig. 2b, which shows the correlation coefficients between the amino acid frequency vectors of the different regions. The correlation values between the compositions of the rim and the surface are 0.96, 0.97, and 0.98 for *E. coli*, *S. cerevisiae*, and *H. sapiens*, respectively, and these are as high between the support and the interior (0.97, 0.98, and 0.97).

The interface core composition is closer to the interior ($R=0.78$, $R=0.65$, $R=0.71$) than it is to the surface ($R=0.38$, $R=0.31$, $R=0.29$). Therefore, the hydrophobic character of protein–protein interfaces comes mainly from the interface core. This is important to keep in mind in order to understand the existing differences in composition between obligate complexes and transient complexes. In

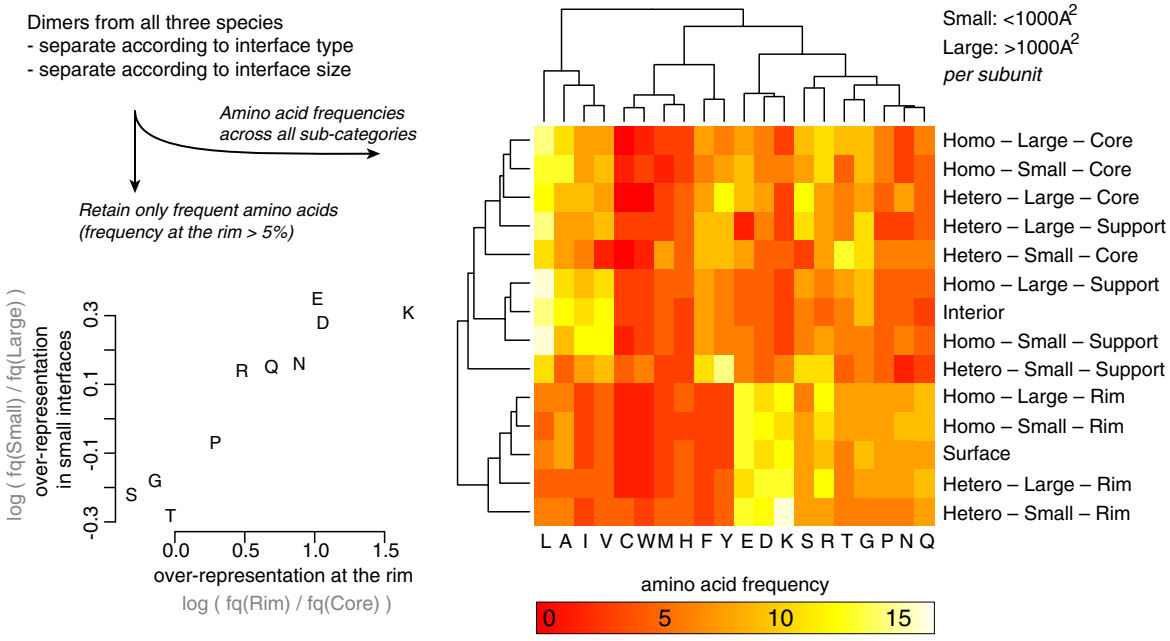

Fig. 3. Each structural region has a stable composition across different types of interfaces. Because the number of dimers—especially the number of heterodimers—is small, the data sets of the three species were combined together. Left: A look at whether the polar character of small interfaces is linked to the smaller contribution of the interface core among them. Propensities of amino acids are calculated from their rim and core frequencies (x-axis). On this scale, lysine, aspartic acid, and glutamic acid are the amino acids most overrepresented at the rim relative to the core. The same is then performed for small and large interfaces (y-axis), which shows that those same amino acids are also the most overrepresented among small interfaces. There is, in fact, a good correlation between overrepresentation at the rim and overrepresentation among small interfaces, suggesting that the polar character of small interfaces can be explained, at least in part, by their small size and necessarily smaller interface core. Right: A related picture in that the amino acid compositions of the different regions are relatively similar across different interface types.

particular, interfaces of transient complexes have been described as being more polar.[15,18,35–38] Looking back at Fig. 1d, the contribution of the core becomes smaller as interface size decreases. As a matter of fact, for interfaces below 1000 Å^2 (per subunit), the rim makes a larger contribution, on average. Since transient complexes bury, on average, less than 1000 Å^2 per subunit,[27,39] the contribution of the rim is more important among them, which is consistent with their interface being more polar as a whole. This is detailed in Fig. 3, which shows a correlation between amino acids overrepresented at the rim relative to the core and those overrepresented in small interfaces relative to large interfaces ($R=0.92$). Figure 3 also shows that the composition of the rim is similar across small and large interfaces, as well as across homointerfaces and heterointerfaces. In future studies, it will thus be interesting to assess the extent to which differing characteristics between interfaces types can be explained solely by differing contributions of the rim, core, and support regions.

In order to characterize further the properties of the regions defined here, they are compared to the core–rim model proposed previously by Chakrabarti and Janin, referred to as the CJ model.[28] In this model, a residue is assigned to the core if at least one atom is fully buried at the interface, or to the rim if all atoms are exposed to the solvent. These properties were computed and the resulting amino acid compositions were compared to those of the regions defined here (Fig. 2c). Interface cores from both models have similar amino acid compositions ($R=0.99$, $R=0.96$, $R=0.98$) and, accordingly, the relationships between the core and the interior, and between the rim and the surface are close to that described here. One can also note two differences. A first difference is that the rim and the surface are slightly more dissimilar in the CJ model ($R=0.87$, $R=0.92$, $R=0.90$), and a second difference is that the core and the rim appear closer in composition ($R_{CJ}=0.76$, $R_{CJ}=0.63$, $R_{CJ}=0.65$; $R_{\text{this study}}=0.33$, $R_{\text{this study}}=0.31$, $R_{\text{this study}}=0.30$). These differences can be explained by the fact that ~47% of residues at the rim in the CJ model are assigned at the core or the support here, and ~18% of the residues from the CJ core are at the rim here (Fig. S4). Therefore, globally, although both models are comparable, the way regions are defined in the model proposed here seems to yield more contrast between the interface rim and the core, and more similitude between the rim and the surface. Another important difference is that a third category is

introduced, the support, which is even more different from the rim than the core is ($R = -0.07$, $R = -0.15$, $R = -0.16$). Finally, note that the composition of structural regions can also be measured by the amino acid ASA contributions rather than by their frequency, as in Janin *et al.*,[14,27] Bahadur *et al.*,[25] Lo Conte *et al.*,[26] and Chakrabarti and Janin[28] for example. Yet interestingly, Fig. S5 shows that the conclusions remain similar when using this definition.

Two important implications follow from the near identity in composition between the support and the interior, and between the rim and the surface. The first is that different regions of the interface should be predicted independently, using different types of information. For example, amino acid composition information can be used to predict the interface core, but not the rim or the support. However, evolutionary conservation might be more suitable than amino acid composition for predicting the rim. The support, on the other hand, might be the most difficult region to predict, since both the composition and the evolutionary conservation are similar to those of structurally equivalent residues. Indeed, amino acids with less than 25% rASA are constrained not only by the interface but also by the structure.[8] Second, it implies that the rim and the support may, in principle, exist prior to any interaction, at least from a simple amino acid composition perspective. In other words, the evolution of a new interface might require the emergence of an interface core only. Following this idea, I examine the distance (in amino acid substitutions) that separates a protein interface from an equivalent surface patch.

Distance in amino acid substitutions between surface and interface compositions

The evolution of linear binding elements such as DNA binding sites or peptide motifs is simple to conceptualize because the information is two-dimensional. It is, however, more difficult to formalize how three-dimensional binding surfaces can appear and disappear during the course of evolution, although it is known that such events do occur both among heterooligomers[40–42] and among homooligomers.[43–47] A recent protein engineering experiment notably showed that new and stable homooligomeric protein interfaces could be induced even from single-point mutations.[48] Although homooligomeric interfaces are easier to form than heterooligomeric ones,[49–51] it is still difficult to imagine how a single-point mutation is enough to create an interaction that involves 20 residues or more.

An approach to this question is to analyze the extent to which a single-point mutation can shift the chemical properties of the surface towards that of an interface. The core–rim–support model can help in this analysis. Specifically, since the support and the rim have nearly the same composition as the interior and the surface, respectively, they may be considered as preexisting in the monomeric state. Following this idea, although an average of ~28 residues are involved in a typical 1000-Å^2 interface, only the core (corresponding to ~10 residues) differs in its composition with structurally equivalent noninteracting residues.

The distance between the surface and the interface core can be characterized using their amino acid log propensities (Fig. 4a). On this scale, a score of 0 means that the frequency in both environments is equal; a score of 1 means that the frequency at the interface core is twice that at the surface; and a score of −1 means the opposite. Given a set of 10 random amino acids, an interface propensity score can thus be calculated and corresponds to the sum of each amino acid propensity score (Fig. 4b). Ten thousand random sets of 10 surface amino acids were generated (where amino acid frequencies follow their frequencies at the surface) as were 10,000 interface cores. Figure 4c shows the density distribution of interface propensity scores for the 10,000 random surface sets (blue line) and for the 10,000 random interface core sets (orange line). This reveals a similar picture for all three species (Fig. S6), with two distinct but overlapping distributions. These distributions were compared to those measured on proteins [i.e., biological surface patches containing 10 surface residues (dotted blue line) or interface cores containing 10 residues (dotted orange line)]. As could be expected, the random and biological distributions of propensity scores are close, suggesting that the randomly generated surfaces and interface cores can be used to assess the minimal number of amino acid changes that separates them.

Figure 4c shows that the distance between the two means of the distributions is equal to ~4. This is small considering that a single substitution from a lysine (a propensity score of −1.7 in *E. coli*) to a phenylalanine (a propensity score of +2 in *E. coli*) introduces a change of 3.7. This is, of course, the most extreme case, as not all surface patches necessarily contain a lysine. In order to measure a more representative distance, the value corresponding to the residue with the lowest interface propensity score was substituted in each of the 10,000 surface patches, and the value of a randomly picked amino acid favored at interfaces (V, L, W, Y, M, I, C, F) was assigned instead. The resulting distribution of propensity scores for the 10,000 surface patches carrying one substitution is shown as the gray dotted line in Fig. 4c. The distribution is centered on zero, halfway between that of the surface and that of interface core. If the substitution process is iterated a second time, the distribution shifts further and overlaps almost perfectly with that of the interface core. This result is consistent across all three organisms (Fig. S6) and suggests that two amino

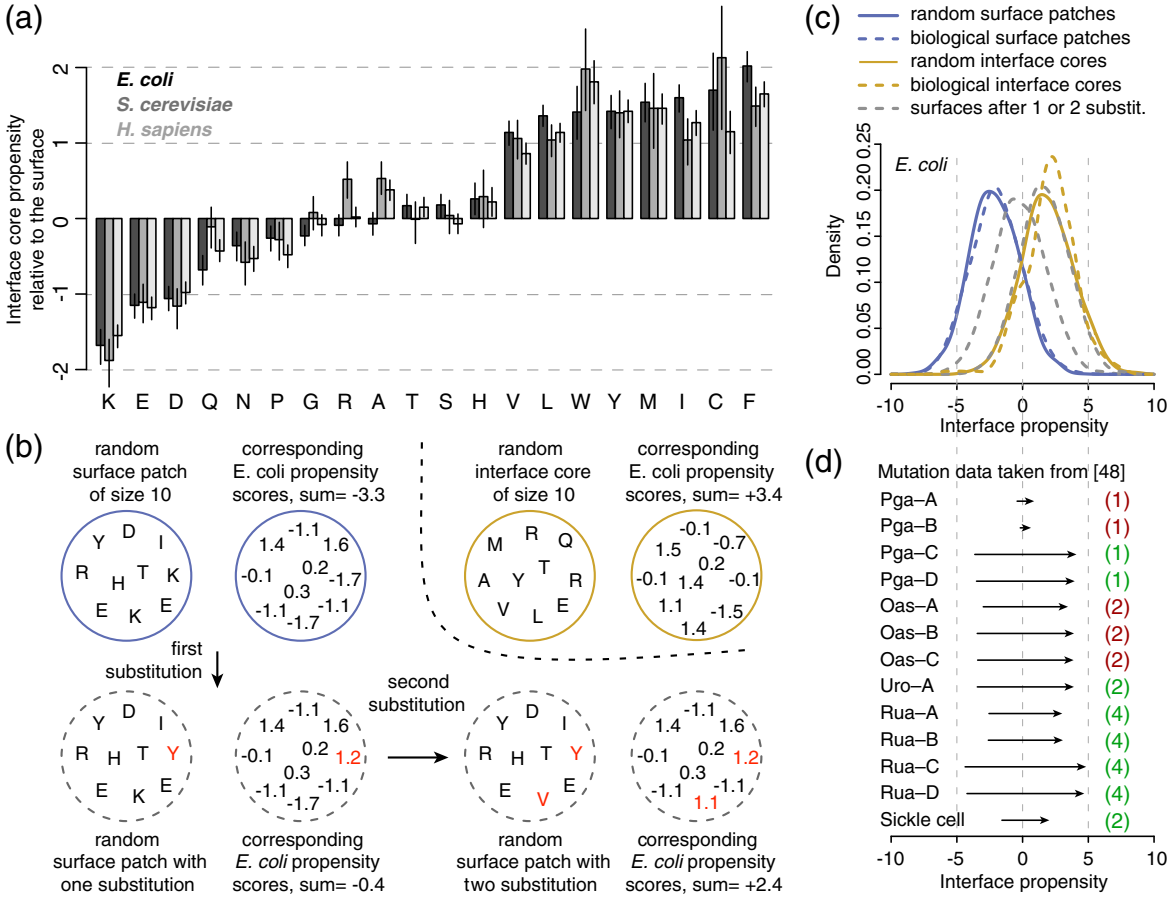

Fig. 4. Distance (in number of amino acid changes) that separates an average surface patch from an equivalent interface. (a) Propensities corresponding to the $\log_2$ ratio of interface–core frequency to surface amino acid frequency. Bars indicate 95% confidence intervals. (b and c) Illustration of the numeric simulation used to create the plot of (c). A random surface patch containing 10 amino acids (blue circle), where each amino acid sampling probability is equal to its frequency at the surface, is generated. Each amino acid is then converted into an interface propensity score, and those are summed. The distribution of the summed scores obtained from 10,000 surfaces patches is shown in (c) as the continuous blue line. Ten thousand interface core patches are generated in the same way, and their corresponding distribution of propensity scores is shown as a continuous orange line in (c). Within each surface patch, the amino acid with the lowest propensity score is then substituted and assigned the value of a randomly picked residue favored at interfaces (V, L, W, Y, M, I, C, F). The distribution of the propensity scores for patches with one substitution is shown as the gray dotted line. The second gray dotted line shows the result of a second iteration of the same process. (d) Interpretation of the effect of mutations on proteins' oligomeric state using the framework presented here. The mutation responsible for sickle cell anemia (E-to-V mutation) is also added. The names of the mutated proteins are given on the left, and the multiplicity of mutations is indicated on the right. For example, the A-to-F mutation in Rua-A is counted four times as Rua-A is a cyclic tetramer. Green indicates mutations resulting in the formation of a new interface (oligomer fraction >40%), while red indicates the absence of new interface formation (oligomer fraction <10%). For details on the mutations and oligomers, please refer to Table S1 in Grueninger *et al.*[48]

acid substitutions are enough, on average, to shift the composition of a 28-amino-acid surface patch to that of an interface of the same size.

Admittedly, amino acid composition is not sufficient to describe interface evolution, as additional factors should be taken into account (e.g., geometric complementarity). This is illustrated in Fig. 4d, which shows the interface propensity changes introduced by Grueninger *et al.* to engineer new protein–protein interfaces.[48] The changes introduced in Oas-A, Oas-B, and Oas-C are theoretically compatible with the formation of an interface but do not result in one, highlighting that composition alone is not necessarily sufficient. Interestingly, however, their other observations are compatible with the framework proposed here (i.e., the small propensity shifts introduced in Pga-A and Pga-B do not result in interface formation, while the more important changes introduced in Pga-C and Pga-D and in five other proteins do so).

Additionally, one should consider that thousands of different proteins coexist in cells[52] and, thus, a particular surface patch can associate with tens of thousands of potential surface patches on other proteins. If such a large number of possible associations balance out the geometric complementarity issue, promiscuous protein–protein interactions may frequently arise during the random course of evolution[53] and could even represent a burden that is selected against.[54,55] The idea that promiscuous interactions might be widespread is also supported by crystal contacts, which have not evolved to be complementary to each other but are nonetheless sometimes difficult to discriminate from biological interfaces (i.e., interfaces that can be observed *in vitro* or *in vivo*).[56]

Conclusions

I propose a definition for interior and surface, both within proteins and at protein–protein interfaces. The results obtained are consistent across three distant species, suggesting that they are general among globular proteins. The cutoff of 25% rASA best partitions the protein interior and the surface in that it maximizes the difference between their amino acid compositions. This cutoff is subsequently used to extend the core–rim model of protein–protein interfaces and to define an additional category, called support. With the proposed core–rim–support model, the amino acid compositions of the rim and the support are nearly identical with those of the surface and the interior, respectively. The interface core residues, on the contrary, are very different from equivalent surface residues. Interestingly, the compositions of the core, rim, and support remain similar across different types of protein complexes (i.e., among small or large interfaces, or among homooligomers or heterooligomers). However, the relative contribution of the interface rim decreases with interface size, which helps explain why smaller interfaces are generally more polar than larger interfaces.

A strength of the model proposed here is that all five structural categories depend on a single parameter (25% rASA), which is simple to compute. Moreover, from an interface prediction perspective, an advantage of this parameter is that it does not necessarily depend on the complexed state and can be applied on monomers when trying to predict protein–protein interfaces. More precisely, one can try to predict the support (<25% rASA in the monomeric state) and the interface core plus the rim (>25% rASA) using different types of information. These structural regions might be useful in reinterpreting some results on the properties of protein–protein interfaces. For example, Choi *et al.* found a correlation between interface size and conservation, but it might be interesting to control whether this trend can simply be explained by an increasing contribution of the interface core among larger interfaces.[57]

Here, these structural regions are used to formulate a minimal distance that separates the protein surface from "becoming" an interface. It is found that an average surface patch involving ~28 amino acids only requires two substitutions to adopt the composition of an equivalent interface, consistent with the results of a recent protein engineering experiment.[48] More generally, it is known that a single-point mutation of a hot-spot residue can destroy an interaction.[58,59] Symmetrically, this study suggests that single-point mutations could sometimes trigger the formation of new protein–protein interactions. This notion adds to established mechanisms such as domain swapping[60] and enabling or disabling loops[61,62] to help explain the evolution of protein–protein interactions. Moreover, combined with additional effects such as colocalization, which increases protein concentration by several orders of magnitude,[45] it suggests that promiscuous protein–protein interactions could be widespread[53] and might play an important role in shaping protein networks evolution.[63–65]

Methods

Data sets of protein structures

Species-specific structures were retrieved by sequence homology, that is, where the sequence from the SEQRES field is similar to that of proteins from *E. coli*, *S. cerevisiae*, or *H. sapiens* proteomes. A minimal sequence identity of 90% and a minimum overlap of 70% were imposed. Protein structures from the Protein Data Bank were used,[66] and the data set includes all structures present in the second release of 3D Complex.[67] All structures for which the biological state was manually annotated in the PiQSi database[68] as "error," "probable error," or "undefined," as well as all DNA binding proteins and membrane proteins, were discarded. Finally, only structures with a resolution below 3 Å were kept. A summary of the number of structures per organism and complex type is given in Table S1†.

Processing protein structures

ASA[69] was calculated with AREAIMOL from the CCP4 suite.[70] Relative solvent accessibilities were obtained by normalizing the absolute value with that of the same amino acid in a Gly-X-Gly peptide, as in Miller *et al.*[71] Figure 4c shows the calculations of the propensities of surface patches containing 10 amino acids. These were obtained as follows. For each residue r_i in each structure, all residues with >25% rASA contained in the 400-Å^2 patch

† All data are available at www.tinyurl.com/structuralregions

centered on r_i (i.e., with an α-carbon within a 11.28-Å radius of r_i). are retrieved. The propensity of r_i is then computed as the sum of the amino acid propensities contained in the surrounding patch. For Fig. 4c, I only considered surface patches containing 10 amino acids or interface core patches with over 5 amino acids. The score of the interface core patches was thus normalized by $10/n$, where n is the number of residues (this was necessary, as there were too few interface core patches of size 10).

Calculation of the distance between surface and interface

The interface propensity of an amino acid AA was obtained by taking the $\log_2$ ratio of its frequency at the interface core to its frequency at the surface: Interface propensity$_{AA} = \log_2(fq^{AA}_{\text{interface core}} / fq^{AA}_{\text{surface}})$. Numeric simulations and data analysis were carried out using R.[72]

Bootstrap procedure to calculate confidence intervals

The confidence intervals shown in Figs. 1a and c, 2a, and 4a were obtained by performing a bootstrap on the data set of each species. This consisted in resampling half of each data set one thousand times and extracting the interval containing 95% of the calculated values.

Acknowledgements

I am grateful to the MRC Laboratory of Molecular Biology and to the Human Frontier Science Project for financial support, and to Stephen Michnick and Université de Montréal for hosting part of this research. I thank Sarah Teichmann, Cyrus Chothia, Subhajyoti De, Joël Janin, and Georg Schulz for their constructive and helpful comments on the manuscript.

Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2010.09.028

References

1. Gething, M. J. & Sambrook, J. (1992). Protein folding in the cell. *Nature*, **355**, 33–45.
2. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 668–678.
3. Franzosa, E. A. & Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26**, 2387–2395.
4. Conant, G. C. & Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol. Biol. Evol.* **26**, 1155–1161.
5. Zhou, T., Drummond, D. A. & Wilke, C. O. (2008). Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Evol.* **66**, 395–404.
6. Sasidharan, R. & Chothia, C. (2007). The selection of acceptable protein mutations. *Proc. Natl Acad. Sci. USA*, **104**, 10080–10085.
7. Chothia, C. & Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 399–405.
8. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. (2007). The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* **369**, 1318–1332.
9. Drummond, D. A. & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**, 341–352.
10. Aloy, P. & Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7**, 188–197.
11. Chothia, C. & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, **256**, 705–708.
12. Deremble, C. & Lavery, R. (2005). Macromolecular recognition. *Curr. Opin. Struct. Biol.* **15**, 171–175.
13. Han, J. H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. (2007). The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319–330.
14. Janin, J., Bahadur, R. P. & Chakrabarti, P. (2008). Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.* **41**, 133–180.
15. Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
16. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. (2008). Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* **108**, 1225–1244.
17. Levy, E. D. & Pereira-Leal, J. B. (2008). Evolution and dynamics of protein interactions and networks. *Curr. Opin. Struct. Biol.* **18**, 349–357.
18. Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein–protein interactions. *EMBO J.* **22**, 3486–3492.
19. Ponstingl, H., Kabir, T., Gorse, D. & Thornton, J. M. (2005). Morphological aspects of oligomeric protein structures. *Prog. Biophys. Mol. Biol.* **89**, 9–35.
20. Reichmann, D., Rahat, O., Cohen, M., Neuvirth, H. & Schreiber, G. (2007). The molecular architecture of protein–protein binding sites. *Curr. Opin. Struct. Biol.* **17**, 67–76.
21. Nussinov, R. & Schreiber, G. (Eds.). (2009). *Computational Protein–Protein Interactions*. CRC Press, Boca Raton, FL.
22. Kleanthous, C. (Ed.). (2000). *Protein–Protein Recognition: Frontiers in Molecular Biology*. Oxford University Press, New York, NY.
23. Janin, J. & Wodak, S. (Eds.). (2003). *Protein Modules and Protein–Protein Interaction: Advances in Protein Chemistry*. Academic Press, San Diego, CA.
24. Golemis, E. & Adams, P. (Eds.). (2005). *Protein–Protein Interactions: A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
25. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

26. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
27. Janin, J., Rodier, F., Chakrabarti, P. & Bahadur, R. P. (2007). Macromolecular recognition in the Protein Data Bank. *Acta. Crystallogr. Sect. D*, **63**, 1–8.
28. Chakrabarti, P. & Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.
29. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
30. Eames, M. & Kortemme, T. (2007). Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure*, **15**, 1442–1451.
31. Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.
32. Guharoy, M. & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein–protein interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.
33. Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13**, 190–202.
34. Conant, G. C. (2009). Neutral evolution on mammalian protein surfaces. *Trends Genet.* **25**, 377–381.
35. Liu, B., Wang, X., Lin, L. & Dong, Q. (2009). Exploiting three kinds of interface propensities to identify protein binding sites. *Comput. Biol. Chem.* **33**, 303–311.
36. Ofran, Y. & Rost, B. (2003). Analysing six types of protein–protein interfaces. *J. Mol. Biol.* **325**, 377–387.
37. Nooren, I. M. & Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein–protein interactions. *J. Mol. Biol.* **325**, 991–1018.
38. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.
39. Mintseris, J. & Weng, Z. (2003). Atomic contact vectors in protein–protein recognition. *Proteins*, **53**, 629–639.
40. van Dam, T. J. & Snel, B. (2008). Protein complex evolution does not involve extensive network rewiring. *PLoS Comput. Biol.* **4**, e1000132.
41. Beltrao, P. & Serrano, L. (2007). Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput. Biol.* **3**, e25.
42. Mika, S. & Rost, B. (2006). Protein–protein interactions more conserved within species than across species. *PLoS Comput. Biol.* **2**, e79.
43. Dayhoff, J. E., Shoemaker, B. A., Bryant, S. H. & Panchenko, A. R. (2010). Evolution of protein binding modes in homooligomers. *J. Mol. Biol.* **395**, 860–870.
44. Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.
45. Kuriyan, J. & Eisenberg, D. (2007). The origin of protein interactions and allostery in colocalization. *Nature*, **450**, 983–990.
46. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332**, 989–998.
47. D'Alessio, G. (1999). The evolutionary transition from monomeric to oligomeric proteins: tools, the environment, hypotheses. *Prog. Biophys. Mol. Biol.* **72**, 271–298.
48. Grueninger, D., Treiber, N., Ziegler, M. O., Koetter, J. W., Schulze, M. S. & Schulz, G. E. (2008). Designed protein–protein association. *Science*, **319**, 206–209.
49. Schulz, G. E. (2010). The dominance of symmetry in the evolution of homo-oligomeric proteins. *J. Mol. Biol.* **395**, 834–843.
50. Andre, I., Strauss, C. E., Kaplan, D. B., Bradley, P. & Baker, D. (2008). Emergence of symmetry in homo-oligomeric biological assemblies. *Proc. Natl Acad. Sci. USA*, **105**, 16148–16152.
51. Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J. & Shakhnovich, E. I. (2007). Structural similarity enhances interaction propensity of proteins. *J. Mol. Biol.* **365**, 1596–1606.
52. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. & O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
53. Levy, E. D., Landry, C. R. & Michnick, S. W. (2009). How perfect can protein interactomes be? *Sci. Signaling*, **2**, pe11.
54. Pechmann, S., Levy, E. D., Tartaglia, G. G. & Vendruscolo, M. (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proc. Natl Acad. Sci. USA*, **106**, 10159–10164.
55. Zhang, J., Maslov, S. & Shakhnovich, E. I. (2008). Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Mol. Syst. Biol.* **4**, 210.
56. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein–protein interfaces. *J. Mol. Biol.* **336**, 943–955.
57. Choi, Y. S., Yang, J. S., Choi, Y., Ryu, S. H. & Kim, S. (2009). Evolutionary conservation in multiple faces of protein interaction. *Proteins Struct. Funct. Bioinf.* **77**, 14–25.
58. Keskin, O., Ma, B. & Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* **345**, 1281–1294.
59. Cunningham, B. C. & Wells, J. A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, **244**, 1081–1085.
60. Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299.
61. Hashimoto, K., Madej, T., Bryant, S. H. & Panchenko, A. R. (2010). Functional states of homooligomers: insights from the evolution of glycosyltransferases. *J. Mol. Biol.* **399**, 196–206.
62. Akiva, E., Itzhaki, Z. & Margalit, H. (2008). Built-in loops allow versatility in domain–domain interactions: lessons from self-interacting domains. *Proc. Natl Acad. Sci. USA*, **105**, 13292–13297.
63. Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nat. Rev. Genet.* **9**, 965–974.

64. Nobeli, I., Favia, A. D. & Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167.
65. Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* **8**, 803–813.
66. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K. *et al.* (2002). The Protein Data Bank. *Acta Crystallogr. Sect. D*, **58**, 899–907.
67. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. (2006). 3D Complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2**, e155.
68. Levy, E. D. (2007). PiQSi: Protein Quaternary Structure Investigation. *Structure*, **15**, 4.
69. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
70. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. Sect. D*, **50**. (1994)., 760–763.
71. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.
72. R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.