Shelling the Voronoi interface of protein-protein complexes predicts residue activity and conservation

Benjamin Bouvier, Raik Gruenberg, Michael Nilges, Frédéric Cazals

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Shelling the Voronoi interface of protein-protein complexes predicts residue activity and conservation

Benjamin Bouvier — Raik Grünberg — Michael Nilges — Frédéric Cazals

N° 1

Janvier 2008

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--1--FR+ENG

Shelling the Voronoi interface of protein-protein complexes predicts residue activity and conservation

Benjamin Bouvier [*] , Raik Grünberg [†] , Michael Nilges [‡] , Frédéric Cazals [§]

Thème BIO — Systèmes biologiques
Projet ABS

Rapport de recherche n° 1 — Janvier 2008 — 39 pages

Abstract: The accurate description of protein-protein interfaces remains a challenging task. Traditional criteria, based on atomic contacts or changes in solvent accessibility, tend to over or underpredict the interface itself and cannot discriminate active from less relevant parts. A recent molecular dynamics simulation study by Mihalek and co-authors (2007, JMB 369, 584-95) concluded that active residues tend to be 'dry', that is, insulated from water fluctuations. We show that patterns of 'dry' residues can, to a large extent, be predicted by a fast, parameter-free and purely geometric analysis of protein interfaces. We introduce the shelling order of Voronoi facets as a straightforward quantitative measure of an atom's depth inside an interface. We analyze the correlation between Voronoi shelling order, dryness, and conservation on a set of 54 protein-protein complexes. Residues with high shelling order tend to be dry; evolutionary conservation also correlates with dryness and shelling order but, perhaps not surprisingly, is a much less accurate predictor of either property. Voronoi shelling order thus seems a meaningful and efficient descriptor of protein interfaces. Moreover, the strong correlation with dryness suggests that water dynamics within protein interfaces may, in first approximation, be described by simple diffusion models.

Key-words: Protein-protein complex, interface activity, hotspots, conservation, Voronoi models.

[*] INRIASophia-Antipolis;ABSandInst.PasteurParis;Benjamin.Bouvier@sophia.inria.fr
[†] EMBL-CRG Systems Biology, Barcelona, Spain; rg@raiks.de
[‡] Unité de Bioinformatique Structurale, Institute Pasteur Paris, France; nilges@pasteur.fr
[§] INRIASophia-Antipolis;ABS;Frederic.Cazals@sophia.inria.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

L'épluchage des interfaces protéine-protéine permet de prédire la conservation des résidus et leur activité

Résumé : La description précise des interfaces protéine-protéine demeure une tâche délicate. Des critères tels que la distance relative des partenaires ou bien ou le changement de l'accessibilité au solvant ne permettent en effet ni d'identifier précisément les atomes à l'interface, ni de mettre en évidence ses parties critiques. En se basant sur des résultats de dynamique moléculaire, Mihalek et al (2007, JMB 369, 584-95) ont conclu récemment que les résidus actifs à l'interface ont une propension marquée à être *secs*, i.e. isolés des fluctuations du solvant. Nous montrons que dans une large mesure, ces résidus secs sont identifiés par une analyse géométrique sans paramètre de l'interface. Cette analyse fait appel à la profondeur d'un atome (et par extension d'un résidu) à l'interface de Voronoi du complexe. Sur un jeu de 54 complexes protéine-protéine, nous analysons en outre la corrélation entre cette profondeur, le caractère *sec* d'un résidu, et sa conservation. Les résidus profonds on tendance à être secs; la conservation des résidus corrèle également avec la profondeur et le caractère sec, mais dans une moindre mesure. La profondeur d'un résidu mesurée à partir de l'interface de Voronoi permet donc de quantifier un paramètre physiquement fondé. De plus, cette profondeur suggère que la dynamique du solvant peut, en première approximation, être décrite par un simple modèle de diffusion.

Mots-clés : Complexe protéine-protéine, activité, résidus critiques, conservation, modèles de Voronoi.

1 INTRODUCTION

Specific recognition between proteins plays a crucial role in almost all cellular processes and most proteins are embedded in highly connected (and dynamically changing) networks of interaction partners [1]. Despite much progress [2], identifying the exact interface between two proteins remains difficult. On the one hand, exact predictions are hindered by the complex and dynamic nature of proteins [3, 4]; on the other hand, the descriptors we employ to study the interface may be flawed or ill-chosen.

A protein-protein interface is traditionally defined by the 'geometric footprint', which refers to all atoms within a given distance of the interaction partner. Somewhat more precise definitions rely on the loss of solvent accessibility (SA) upon binding [5]. Yet, as much as half of this footprint can seemingly be irrelevant to binding [6]. As contributions to specificity and affinity appeared very unevenly distributed, substantial effort has been spent on the identification of areas or residue patches that are actively involved in molecular recognition [7, 8, 9, 10]. This lead to the definition of 'hotspot' residues [11, 12]. Hotspots refer to the usually very small number [12] of 'key' residues in a protein-protein interface, the mutation of which causes large changes in the binding free energy. Contrary to this focus on isolated residues, more recent studies have revealed strong non-additive, collective effects [13] which point to a modular organization of interfaces into interaction clusters [14].

Also the evolutionary record seems of limited use for distinguishing relevant from irrelevant. The sequence conservation of protein-protein interfaces is hardly statistically significant and depends heavily on surface-patch selection techniques [15]. A commonly adopted view states that, unlike catalytic sites that are highly unlikely to transform in a series of discrete steps without complete loss of activity [16], the assembly of proteins involves a continuous scale of binding modes, from transient to stable, leaving more freedom for evolution to proceed in incremental steps [17, 18, 19]. Interestingly, conservation signals become more convincing if one turns away from individual– and towards patches [20] or clusters of residues [21].

Water forms an essential part of protein-protein interfaces [9, 22]. The occlusion of bulk solvent is a common denominator not only of classical hotspots [23], but also of the more recently identified interaction modules [14], which are delimited by structural water. In fact, the removal of water from partially solvated backbone hydrogen bonds has been argued to be a driving force of binding [24, 25].

Recently, Mihalek and coworkers [26] went one step further and classified interface residues by the dynamics of surrounding water molecules. They asserted that the important residues are the ones whose interactions are not disturbed by water fluxes. These 'dry' residues (some of which may actually be in contact with immobile, structural water molecules) were found to correlate better with conservation than the overall geometric footprint and to feature some characteristic properties of classical hotspots. The dryness results collated by these authors on a variety of systems thus represent valuable information as a measure of residue importance; we will constantly refer to them during this work.

However, the method suffers from some drawbacks. It relies on molecular dynamics simulations which are computationally expensive and sensitive to setup and parameteriza-

tion. Furthermore, it cannot itself distinguish between interface and noninterface residues. Mihalek and coworkers addressed this problem by discarding residues that are also dry in the isolated partners, hereby further increasing computational costs and neglecting the possibility of conformational transitions upon binding.

All in all, the combination of the large size of protein-protein interfaces, the relatively small areas that appear actually important and the lack of unambiguous ways to identify them, amounts to a difficult problem for which novel approaches are highly desirable. We present a method based on the shelling of the Voronoi interface of protein-protein complexes. The method quantifies the depth of any given atom inside the interface, in a manner accounting for both the geometry and the topology of the interface. The method is simultaneously accurate, computationally inexpensive, and elegant in that it does not require parameterization. Voronoi shelling order features an excellent correlation with the water shielding observed by Mihalek et al., without the need for simulations or geometric footprinting. We analyze the relationship between three quantities of interest (Voronoi shelling order, dryness and conservation) on the same set of protein complexes. We illustrate the advantages as well as potential improvements of the geometric measure with detailed examples and elaborate on the more complex correlation with evolutionary information.

2 THEORY

2.1 Voronoi description of protein-protein interfaces

In this section, we briefly summarize the Voronoi model of protein-protein interfaces, which is described in more detail in [27], together with a comprehensive bibliography. Given a collection of sample points equipped with the Euclidean distance, the Voronoi diagram is the space partition which assigns to every sample the convex polyhedron containing all points in space closer to it than to any other sample. In 3D space, these Voronoi regions are bounded by Voronoi facets (resp. edges, vertices) which consist of points equidistant from two (resp. three, four) samples.

The Euclidean Voronoi diagram of atom centers in a molecule, first employed by Richards [28] to investigate packing properties in proteins, is unable to account for the fact that different atoms have different radii. A convenient generalization thereof, which overcomes this limitation while retaining non-curved bissectors, is the power diagram[29]. It replaces the Euclidean distance with the 'power distance' of a point to a sphere centered at $\mathbf{a}$ and of radius r: $p(\mathbf{x}) = |\mathbf{a} - \mathbf{x}|^2 - r^2$. The power diagram is an extension of the Voronoi diagram (to which it reverts for atoms of equal radii); hence, we continue to refer to it as such in the text. Throughout the study, we compute it for atomic spheres whose radii are the so-called group radii [30], expanded by the radius of a probe water molecule r_w =1.4 Å. This effectively models the solvent-accessible surface (SAS) of the protein, as defined by Lee and Richards [31]. An example Voronoi diagram for a hypothetical two-dimensional molecule is shown on Figure 1.

The Voronoi diagram has a dual (an associated and strictly equivalent structure) called the Delaunay triangulation; in practice, Voronoi diagrams are calculated via their Delaunay triangulation rather than directly. The Delaunay triangulation consists of edges (resp. triangles, tetrahedra) that connect the centers of two (resp. three, four) adjacent spheres whose corresponding Voronoi regions share a facet (resp. an edge, a vertex).

When modeling molecules, a drawback of the Voronoi diagram is that atoms located on the convex hull have unbounded Voronoi regions (all but the region of atom a_2, on Figure 1). An elegant way of solving this problem is to use a restriction of the Delaunay triangulation called the α-complex [32]. For a fixed value of α, each ball of center $\mathbf{a_i}$ and radius r_i is replaced by a ball of center $\mathbf{a_i}$ and radius $\sqrt{r_i^2 + \alpha}$. Given these expanded balls, the construction of the α-complex mimics that of the Delaunay triangulation, to the extent that one focuses on the intersection of the restriction of each expanded ball to its Voronoi region rather than the Voronoi region itself; see Figure 1 for an illustration. Varying the value of α allows for the investigation of properties at different scales. In particular, for very large values of α the α-complex is identical to the Delaunay triangulation. In rare occurrences of desolvated models, an additional filtering step may be necessary to discard all instances of unphysically large facets at the rim of the interface [27]; we do not discuss this issue further since this study involves solvated models only.

We now apply this methodology to model the interface between two proteins A and B. Following [27], the AB interface consists of the Delaunay edges found in the 0-complex –

the α-complex for $\alpha = 0$, and whose endpoints belong to A and B. Because of the duality between the Delaunay and Voronoi representations, the interface can also be described using the Voronoi facets dual to the aforementioned edges. The interface model can be extended to accommodate interface water molecules W, defined as sharing at least one edge with each partner in the 0-complex. This allows for the definition of the following interfaces: AB between the protein partners; AW (resp. BW) between partner A (resp. B) and interface water; $AW - BW$ as the union of the interfaces AW and BW; ABW as the union of the interfaces AB and $AW - BW$. Like other methods mentioned above, our model correctly identifies any atom losing solvent accessibility as an interface atom. Unlike these methods however, it also detects interface atoms that do not lose solvent accessibility – essentially buried backbone atoms, these represent a non-negligible 13% of the interface [27].

2.2 Shelling the ABW interface

The next step of the algorithm attributes a Voronoi shelling order (VSO) to each facet of the ABW interface. This represents the number of 'jumps' between adjacent facets that needs to be performed, from the currently considered location, to reach the rim of the interface (Figures 2a and 3a). The Voronoi interface is thus partitioned into concentric shells of increasing selling order.

The calculation of VSO values for all interface facets requires two passes. During the first pass, boundary Voronoi facets located at the rim of the interface are enumerated and given a VSO of one. Voronoi facets are bounded by Voronoi edges, each of which is incident to exactly three Voronoi facets in the Voronoi diagram; however, some of these facets may not belong to the interface (their dual Delaunay edges are not in the 0-complex). This allows us to detect rim Voronoi facets as the ones featuring at least one Voronoi edge that is incident to one interface Voronoi facet only. The second pass explores the interface breadth-first starting from the previously identified rim facets. Given an interface Delaunay edge (of shelling order n), the algorithm checks all incident Delaunay triangles, as each such triangle contributes zero, one or two additional interface edges. If these have not already been shelled, they are given a VSO of $n+1$. To speed up the search operations, a temporary map storing edges of VSO $n-1$, n and $n+1$ is used, since these are the only ones that can be encountered at level n; the contents of this map are copied over to a permanent structure each time n increases.

The outcome of this process is the association of an integer VSO value to each Delaunay edge (or equivalently, Voronoi facet) of the ABW interface. However, our ultimate goal is to quantify the depth of any given atom inside the interface. This is done by tagging the atom with the minimum value among the shelling orders of the Delaunay edges to which the atom contributes (Figures 2b and 3b). The maximum or average values have also been considered as candidates, but their variation throughout the interface were found to closely mimic that of the minimum. Finally, the shelling order of a residue, defined as the average VSO value over its constituent atoms contributing to the Voronoi interface, is employed when comparing to residue-based measures such as conservation or dryness.

3 RESULTS

3.1 Voronoi shelling order, conservation and water dynamics

A recent simulation study examined the rate at which residues in protein-protein interfaces exchange surrounding water molecules [26]. Residues that were mostly shielded from mobile water molecules, defined as "dry" by Mihalek et al., turned out to be more conserved and were thus interpreted as the active part of the interface. Our initial goal is to assess how well shelling order is able to predict dryness on the set of homo- and heterodimer complexes studied by Mihalek et al. [26]. As a yardstick, we compare to the previously established correlation between conservation and dryness. Conservation is determined from pFam [33] hidden Markov models [34] using a relative entropy scheme [35]. In order to characterize all possible relationships, we also examine, further down in the text, how good a predictor of shelling order conservation is. We generate three ROC plots for each complex, describing the performance of shelling order as predictor of dryness, of conservation as predictor of dryness and of conservation as predictor of shelling order, respectively. A representative example set of ROC curves is shown in Figure 4. The area between each ROC curve and the diagonal quantifies the predictive power of a score (i.e. VSO, conservation) in terms of sensitivity and specificity. An area of 0.5 corresponds to a perfect prediction, which in the example of shelling order predicting dryness means that the n dry residues in the interface perfectly match the n residues with highest shelling order without any over-prediction. By contrast, a ROC area of 0 corresponds to the performance of a pure random classifier. See Section 5.4 for details.

The results are compiled in Tables 1 and 2 for heterodimers and homodimers, respectively, and summarized in Figure 5. Evidently, Voronoi shelling order is a very good predictor of dryness and outperforms conservation for 35 of the 36 homodimers and 17 of the 18 heterodimers. VSO always performs better than a purely random classifier, whereas conservation fails to do so in seven cases (five homodimers and two heterodimers). The third columns of Tables 1 and 2 quantify the ability of sequence conservation to predict Voronoi shelling order. We define the n_{core} residues with highest VSO as 'core' and the remainder as 'rim' and test the ability of conservation to discriminate between the two. We adjust n_{core} for each complex so as to exactly match the number of residues classified as dry. We thus tie ourselves to a threshold chosen by Mihalek et al. [26] rather than optimizing our own. Nevertheless, the connection from conservation to Voronoi shelling order appears as good as it is to dryness. While the results differ in detail, the average ROC area is 0.15 for heterodimers and 0.12 for homodimers, which compares well with the respective figures of 0.14 and 0.13 for the prediction of water shielding. However, both conservation-based predictions are outperformed by the much closer correlation between shelling order and dryness, reflected by average ROC areas of 0.31 and 0.34. This notable discrepancy indicates a more direct link between the two latter properties, both of which are structure-based.

3.2 Spatial distribution of conserved residues

The analysis of the ROC curves provides insight into the location of highly conserved residues across the interface shells: conservation becomes a mediocre predictor for Voronoi shelling order when highly conserved residues are found at low VSO (such residues are expected to be wet) and/or when poorly conserved residues are found at high VSO (such residues are expected to be dry). However, this simplified focus on extreme values can not fully capture the spatial distribution of conservation. We therefore now address two complementary points, namely (i) the average residue conservation as a function of VSO, and (ii) the cumulated conservation score over consecutive shells.

(i) Guharoy and Chakrabarti showed that residues at the interface core are, on average, more conserved than those on the rim [36]. Their binary interface model defined the rim as all residues that are not fully buried inside the complex. Our more quantitative description helps to refine the prior conclusion. We normalize conservation scores and Voronoi shelling order so that both span the range 0 to 1 for each interface. We then compute the average conservation score as a function of VSO using a large moving window comprising 1/4 of all interface residues. Figures 6 and 7 show this running average for all complexes. The relation between residue conservation on the one hand, and depth within the interface on the other, is evidently not a simple one. The non-averaged original values (gray lines) highlight the scattering of conservation across shells: highly conserved residues are found even at the very rim. Only the extensive averaging reveals a clear correlation between increases in shelling order and residue conservation. This observation is not sensitive to the actual averaging window and the curves remain very similar for window sizes between 1/8 and 1/2 of the interface (data not shown).

The overall correlation between shelling order and conservation can be quantified in a single number by double integration over the running average. We denote $c(x)$ the average conservation score at $VSO = x$ and reset the baseline of this function to 0 by substracting the minimum value m: $\overline{c}(x) = c(x) - m$. We now define $A = \int_0^1 \overline{c}(t)dt$ to be the area under this running average and we normalize $\overline{c}(x)$ to cover an area of 1: $f(x) = (c(x) - m)/A$. Function $f(x)$ can be seen as a probability density function, with associated cumulated distribution function $F(x) = \int_0^x f(t)dt$ (dash-dotted line in figures 6 and 7). One always has $F(1) = 1$, but the speed at which F reaches 1 depends on whether conserved residues are picked up early (in the outer shells) or late (inner shells). F thus encodes the cumulative conservation score up to shelling order x. To provide a concise measure of this property, we report $g(x) = \int_0^x F(t)dt$ (dotted line in figures 6 and 7). The total area under F depends on the overall distribution of conservation across shells. Lower values of $g(1)$ thus indicate that conserved residues tend to cluster towards the *core* of the interface; values above 0.5 (the double integral over a flat line) denote clustering near the *rim*. The deviation $\Delta = g(1) - 0.5$ is reported in the lower right corner of each plot in figures 6 and 7. $g(1)$ falls below a value of 0.5 for 15 out of 18 heterodimers and 28 out of 36 homodimers. Conservation thus generally increases towards the interface core. Nevertheless, apart from the few obvious exceptions, closer inspection also reveals some interesting systematic deviations: (i) Conservation density

often reaches its maximum before the innermost shell – the interface center thus appears under less constraint than a surrounding outer core; (ii) contrary to the overall trend, a pronounced secondary peak of conservation is sometimes apparent at the very edge of the interface.

(ii) While the previous analysis focuses on the spatial distribution of conservation per se, it is also worthwhile to compare the spatial distribution of conservation for two sets of residues: the interface residues and the dry residues. The detailed analysis is described in section 8.1 of the supplemental material. Non-interface residues account for a proportion of the total conservation score (over the whole protein) in the range 60% to 84% in heterodimers (average 76%), and 36% to 97% for homodimers (average 73%) –see the second column of Tables 3 and 4 in the supplemental material. These results alone show that the effect of the majority of conserved residues on the interface is at best an indirect one –for example, through the imposition of a protein fold which in turn dictates interface structure. Moreover, the comparison of the area under the cumulated distribution function for interfacial and dry residues performed in Section 8.1 confirms that the rim amino-acids account for a non-negligible part of the conservation. The good agreement with the scattered conservation signals and conserved interface rims observed in figures 6 and 7 allows us to rule out a purely statistical effect where a large number of moderately conserved rim residues might end up having more weight than a small number of highly conserved core amino-acids: highly conserved residues do occur on a non fortuitous basis at the rim of protein-protein interfaces.

The in-depth examination of average and cumulated conservation thus confirms the general trend of higher conservation towards core shells but also hints at a more complex fine structure. The very center of an interface often appears more amenable to change than its immediate surroundings; furthermore, numerous interfaces seem to bear substantial evolutionary pressure on their outer rims. From the inspection of examples, we speculate this latter signal to be a signature of electrostatic steering [37] but the issue deserves further scrutiny.

3.3 Case-studies: best and worst case scenarios for shelling order

To identify in more detail the incentives and shortcomings of using shelling order for the description of interfaces and as a predictor of water dynamics, we focus on three extreme cases of application, which are presented in Figure 8.

The ideal case. The interface of the homodimer complex 1E2D (left) features a compact and planar core composed of a single patch of atoms with high shelling orders (large panel), which the MD simulations of Mihalek and coworkers also identify as dry (lower left-hand panel). Such compact interfaces with disk-like topologies and no holes represent best case scenarios for the predictive power of our model. Also conservation performs well for this complex. However, in contrast to shelling order, the conservation score delimitates a patch which extends far beyond the dry residues, resulting in a good sensitivity but a poor selectivity. In fact, the most highly conserved residues are catalytic in nature, and located at the entrance of a finger-like cavity which extends, from the other side of the protein,

in the direction of the interface (not visible in the figure). The co-crystallized thymidine monophosphate and adenosine diphosphate substrates [38] allowed Mihalek and coworkers to identify these residues as catalytic and as such to exclude them from their analysis. However, the detection of catalytic residues is not always as straightforward and the influence of this and a variety of other factors hamper the use of conservation measures for specific predictions.

Stacks of water molecules. The interface of the homodimer 1L5W is quite extensive and highly non planar, consisting of two 'prongs' separated by a cleft. Two high-VSO patches are found on either of the prongs. The *ABW* interface is discontinuous in the region of the cleft, due to the presence of more than one layer of solvent molecules sandwiched between the partners (Figure 9); this resets the shelling order to low values in that area. On the other hand, MD simulations find a much smaller patch of dry residues that extends inside the cleft, which means that some of the aforementioned solvent molecules are in fact structural in nature, and do not move during the simulation. A remarkable example of this occurs for tryptophane 203 (located inside the cleft), which is classified as dry by Mihalek and coworkers but is surrounded by numerous water molecules on Figure 9. Here we are confronted with the main advantage of MD simulations over our model: they are able to discriminate structural water on the basis of residence times, whereas our static model relies on the fact that buried interfacial water does not usually form multiple layers. However, it is clear from Tables 1 and 2 that situations featuring water molecules structured along more than one layer rarely occur; we discuss this issue further in section 4. Within the interface, conservation fares better since one of the prongs and the cleft region are fairly well conserved. However, the most conserved regions lie at the protein core (not visible on the figure) and, to a lesser extent, elsewhere on the protein surface.

Discontinuities of the interface. Figure 8 shows a graphical representation of shelling, conservation and dryness for complex 1A59. 1A59 has an intricate topology, consisting of two monomers of predominantly globular nature linked by long 'tails' wrapped around the partner. Dry residues appear both on the globular part and on the first segment of the tail (Figure 8). Voronoi shelling order very accurately predicts the latter patch of dry residues, but over-predicts the entire tail as being dry or active, too. More interestingly, it also misses the lower part of the dry patch on the globular side of the protein. A careful inspection of the interface reveals two holes in the AB interface which reset the shelling order there, preventing the shelling order from peaking in this region (Figure 10). The fact that such holes are visible in the AB interface hints at a sizable packing issue: minute defects do not usually result in such discontinuities of the AB interface[27]. Indeed, the gaps between the atoms of the two monomers [1] span the range 5.2-6.2 Å and 5.9-6.3 Å, respectively, and could accommodate a water molecule each. Since the crystal structure does not contain structural water, we cannot ascertain whether this is the case and our fast solvation procedure proved unable to fill the holes – even though it did successfully place isolated water molecules in three other locations. By comparison, conservation correlates with dryness on the globular

[1] Hole 1: residues 209 to 213 (chain A) and 583 to 587 (chain B); hole 2: residues 206 to 210 (chain A) and 586 to 590 (chain B).

part of the interface, but also features widespread conserved patches covering most of the protein surface.

4 DISCUSSION AND CONCLUSION

4.1 A quantitative interface definition

Among the various definitions of what exactly constitutes a protein-protein interface, the planar facets obtained from a Voronoi tessellation [39, 40] arguably present the closest ties to the literal meaning of the term 'interface'. Indeed, such facets stem from pairs of directly interacting atoms, and the definition of the interaction area is simpler than that required by analytical interface models [41]. The Voronoi model shows excellent correlation with classically defined curvature and solvent accessible area but captures the interface more fully than methods based on solvent accessibility [27] —see also [42] for a review on the use of Voronoi diagrams in protein structure and interface analysis. By contrast, the widely used geometric footprint (based on residue contacts) yields an ambiguous interaction layer biased towards large residues and subject to an arbitrary distance cut-off [3].

Here, we go beyond the binary classification of whether or not a given atom is part of the interface and furthermore quantify how many facets separate it from the edge of the interface. The idea is related to the concept of residue or atom depth [43, 44] which shows some correlation with thermodynamic properties [43] and residue conservation [45] in globular proteins. Previous studies have defined atomic depth as the simple Euclidean distance to the closest solvent molecule. By contrast, Voronoi shelling order partitions the interface into concentric shells, accounting for both the geometry and topology of the interface and appears closer to physical reality. Yet other previous studies have dissected protein interfaces into "inner" and "outer" or "core" and "rim" residues (for example, [46, 47, 48, 36]). Although a number of general trends emerge, conclusions from these works are hindered by distinct definitions of the interface combined with different classifications for core and rim. Voronoi shelling order provides a more quantitative, parameter-free and unambiguous alternative to the ad-hoc classifications previously employed.

4.2 Shelling order and water dynamics

The shelling of the Voronoi interface yields an accurate quantification for the concept of burial depth. Shelling order quantifies the number of atomic shells a water molecule must pass on the shortest path to a given position (facet) in the interface. This description is particularly valuable for highly curved interfaces (1A59, 1L5W...) which the Euclidean distance cannot correctly measure. We have here revealed a clear correlation between Voronoi shelling order and the 'dryness' of a residue, that is, its shielding from itinerant bulk solvent molecules. While one could expect some ties between the two measures, the extent of the agreement over a representative set of complexes is intriguing. After all, dryness was derived from exhaustive molecular dynamics simulations which consider hundreds of additional parameters and details that are totally ignored by our model. On the contrary, Voronoi shelling order is a purely geometric property, calculated from a static set of atomic positions without any further parameter. In particular, we do *not* consider: electrostatic charges, polarity, hydrogen bonds, or any kind of fluctuations – all of which are expected

to influence water dynamics. This suggests that the seemingly complex dynamic exchange of bulk solvent with interfacial water primarily depends on a simple path length and could tentatively be approximated by an analytical model of diffusion along a gradient.

4.3 Complementarity of conservation and Voronoi shelling order

Evolutionary conservation alone cannot usually be employed to predict the active part of an interface, let alone the interface itself. Hence the necessity to cross-correlate it with some other measure (like geometric footprint or change in solvent accessibility) before using it for such purposes. By comparison, Voronoi shelling order simultaneously offers an unambiguous definition of the protein-protein interface and a more fine-grained classification within this interface.

Furthermore, the quantification of evolutionary signals is not trivial. pFam sequence alignments are considered high quality but are not guaranteed to be homogeneously distributed between protein families, hereby introducing bias. Moreover, some protein stretches cannot be aligned at all, and needed to be excluded from our analysis of conservation. We quantify conservation with an entropy-based measure that has been shown to outperform other conservation scores [35]; alternative means can be employed but the actual method of choice seems to have limited effect on the correlation with dryness[26].

Bearing in mind the interference from many other factors, sequence conservation can, nevertheless, provide independent testimony of an area's importance. It confirms the notion of water shielding as an indicator of binding activity and it supports the functional relevance of shelling order. In fact, conservation and VSO are best used in conjunction rather than as competitors. We find a general correlation between shelling order and conservation but, in contrast to a simple classification into rim and core, our continuous measure also resolves interesting deviations from this trend. Such deviations hint at catalytic sites, defects in solvation and packing, but may also indicate binding contributions that do not directly rely on water shielding.

4.4 Methodological improvements

As previously discussed, discrepancies between dryness and shelling order arise for cases where structural (slow moving) water molecules form more than one layer inside a cavity. This is due to the fact that in our current model, interfacial water molecules must make simultaneous contact with both protein partners; any additional layer of water molecules not fulfilling this criterion will be considered as bulk and lead to the splitting of the ABW interface. However, 'trapped' water molecules are known to stabilize turns and bends through hydrogen bonding with main-chain atoms in otherwise unstructured regions [49], and cannot be ignored. Their behavior is so different from that of bulk water that it is debatable whether they should be considered as delimiters for the interface, even when stacked in more than one layer – dryness results from MD simulations tend to show that they shouldn't.

The most straightforward approach to alleviate discrepancies between dryness and shelling order in these difficult cases would be to optimize the threshold separating 'dry' from 'wet',

instead of using Mihalek's choice [26]. Our model could also be extended so as to declare as interface water all solvent molecules W_i found on a path $AW_1 \dots W_kB$ joining both partners. Using $k = 2$ or $k = 3$ could allow to infer similar properties for water molecules organized in layers, as in complex 1L5W. Nevertheless, the current interface model, despite using $k = 1$, demonstrates that it is legitimate to infer dryness/activity from a purely geometric perspective. This effectively replaces a costly MD simulation by a very fast computation on a structure taken directly from the PDB.

Another worthwhile methodological improvement would address rare cases where discontinuities in the interface appear due to packing or solvation defects. An example thereof is the previously discussed 1A59 interface (Figure 10). Regardless of the quality of the structure or the equilibration procedure, such cases could be accommodated by using a water probe radius larger than 1.4 Å, or by devising an adaptive scheme for the value of α ($\alpha > 0$) employed to construct the α-complex. In any case, these extensions should be investigated in conjunction with the threshold used to define dryness.

4.5 Conclusion

In this paper, we present a novel method to explore protein-protein interfaces. The interface is defined using the Voronoi diagram of interacting atom pairs; unlike geometric footprinting methods, all atoms involved in the interface are identified with little to no over-prediction and without resorting to a distance threshold. We have shelled the Voronoi interface from the rim to the core, thus associating an interface depth to each atom. This Voronoi shelling order (VSO) correlates very well with the protection of residues from itinerant water fluxes, as computed by Mihalek and coworkers [26] which, in turn, can be considered a measure of residue activity. The calculation of shelling orders, however, is about five orders of magnitude faster than a typical MD simulation. Moreover, the rather accurate prediction from a simplistic and purely geometric model hints at the possibility to approximate the complex dynamics of interfacial water by simple analytic diffusion models. Comparison with evolutionary signals confirms the functional relevance of 'dry' residues and, likewise, reveals a general increase of conservation towards inner interface shells. Systematic deviations from this trend may inform about distinct binding mechanisms, catalytic activities but also modeling errors. Our accurate and continuous scale of burial depths could also be used to delimitate patches on an interface. Hence, it appears as a worthy candidate for the theoretical study of collective effects in protein-protein interfaces [13], which are progressively replacing the traditional 'hotspot' view.

5 METHODS

5.1 Complex preparation

The coordinates for the homo- and heterodimer complexes listed in Tables 1 and 2 originate from the PDB database. Crystallographic water molecules were removed in order to exclude bias from different structure qualities. Missing atoms, including polar hydrogens, were added and briefly minimized. The structure was surrounded by a 9 Å layer of water molecules from an equilibrated TIP3P box. The water was briefly minimized by 3 rounds of conjugate-gradient optimization of 40 steps each with, initially (round 1), frozen and later (rounds 2 and 3) harmonically restrained protein coordinates. Keeping this restraint, the water was then further relaxed by 100 2-fs steps of molecular dynamics at 100 K, followed by 40 steps conjugate gradient minimization. Optimizations and simulations were performed using the CHARMM19 force field [50] and an electrostatic cutoff of 12 Å with force shifting [51] inside the X-PLOR package. This structure preparation protocol is automated by the pdb2xplor.py program which is part of the open source Biskit package [52]. The final structure was stripped of its hydrogen atoms and used as input for the Voronoi interface calculations (see below).

To test the legitimacy of this economical solvation procedure, a more thorough approach was employed on complex 1M0S. After an initial re-optimization of the crystal structure (retaining crystal water), the complex was placed inside a triclinic box, solvated with SPC water molecules from an equilibrated box and neutralized by 8 Na^+ ions. The solvent molecules were then relaxed around the fixed solute by a steepest-descent optimization followed by 100 ps of molecular dynamics (MD) simulation with position restraints on the solute. The entire system was then simulated for 5 ns without restraints, with a 300 K Maxwellian distribution of initial velocities. MD simulations employed the particle-mesh Ewald treatment of long-range electrostatics and periodic boundary conditions, as well as couplings to heat (300 K, 1 ps) and pressure (1 bar, 1 ps) baths; they were performed with GROMACS 3.3.2 [53] using the OPLS all-atom force field [54]. The final equilibrated box had dimensions 76x92x69 Å and comprised 13460 water molecules. Convergence of the protein structure was reached after 2 ns of simulation, at a mean RMSD of 1.90 Å from the crystal structure.

Section 8.2 of Supplemental Material compares the Voronoi interfaces of complex 1M0S using these two equilibration procedures. The very similar results, both in terms of interface topology and the identification of interfacial water, justify the economical solvation method and indicate the robustness of our model against minor changes both in protein conformation and hydration patterns.

5.2 Calculation of shelling orders

The program Intervor, responsible for the actual computation and shelling of the Voronoi interface, is based on the CGAL computational geometry library [55]; an online version of Intervor is available [56]. On an Intel Pentium IV 3 GHz CPU, an Intervor run for

a typical complex takes less than 5 seconds. We also provide a wrapper (Biskit.Intervor) for integrating the stand-alone program in Biskit workflows. Residue shelling orders were calculated by averaging over a residue's interface atoms.

5.3 Dryness and conservation

Dryness results were those discussed in [26] and were kindly provided to us by O. Lichtarge and coworkers.

Multiple sequence alignments were obtained from the pFam database [33] of HMMER profiles [34] using the HMMER software version 2.3.1. Protein family profiles matching a given sequence were identified with hmmpfam using a conservative E-value and bit score cutoff of 1e-8 and 60, respectively. The sequence was then aligned to the matching profile with the hmmalign program. Following [35], the conservation of each alignment position was quantified by the Kullback-Leibler divergence (relative entropy) between the HMM emission probabilities p and the background distribution of amino acids in SwissProt q:

$$s = \sum_{i=1}^{20} p_i \log \frac{p_i}{q_i}.$$

The complete procedure is automated in the Hmmer.py module of Biskit. Before further analysis, residues outside the interface (average $VSO = 0$) or lacking conservation scores were removed and conservation scores were independently normalized to the maximum of each monomer face.

5.4 ROC curves

Receiver Operating Characteristics (ROC) curves[57] are an efficient way of representing the accuracy of a binary classifier. A binary classifier maps instances of an object into two categories, positive or negative, based on each instance's position relative to a threshold. The quality of the classifier is then assessed by how well the prediction relates to the actual value of the instance. Four cases are possible: true positive (both the outcome from a prediction and the actual value are positive), false positive (the prediction is positive while the actual value is negative), true negative (prediction and value are both negative) and false negative (prediction is negative while value is positive). From this contingency table, the notions of selectivity and sensitivity can be defined as

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

and

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}.$$

A ROC curve is the 2D plot of sensitivity versus specificity, where each point corresponds to a different threshold value. A perfect predictor, which features neither false positive nor

false negative occurrences, should pass through the point (1,1) for the optimal threshold value. Therefore, the closer the ROC plot is to the upper right corner, the higher the overall accuracy of the test [58]. A purely random classifier, with equal chances of making correct or erroneous predictions, has a linear ROC curve connecting points (0,1) and (1,0) – the first diagonal. How much better than random a predictor is can hence be quantified by calculating the area between its ROC curve and the diagonal, which varies from -0.5 (worst-case classifier) to 0.5 (perfect classifier) through 0 (pure random classifier). ROC curve and ROC area calculations were performed with the Biskit.ROCalyzer module.

By way of example, figure 4 shows typical ROC curves for shelling order and conservation as predictors for dryness, in the specific case of the 1HE1 complex. For this system, shelling order is systematically better than conservation at predicting dryness, regardless of the threshold chosen to discriminate between positive and negative predictions in each case. This translates into a larger area between the diagonal (representing a random prediction) and the shelling order ROC plot, than between the diagonal and the conservation ROC plot.

5.5 Miscellaneous

The Biskit python package [52] was also used for various other scripting tasks and the collation of results. All parts of Biskit are open source and available at http://biskit.sf.net. Pymol [59], Ipe [60] and CGAL-Ipelets [61] were employed for the rendering of figures.

Acknowledgments. *We would like to express our gratitude to Olivier Lichtarge and Tuan Anh Tran for providing us with their detailed dryness results. The automatic generation of conservation profiles was implemented by Johan Leckner. B. Bouvier acknowledges funding from the INRIA Cooperative project ReflexP. R. Grünberg is supported by the Human Frontiers Science Program.*

6 TABLES

Table 1 Heterodimers. Performance of shelling order (VSO) as a predictor for dryness, of conservation as a predictor for dryness, and of conservation as a predictor for shelling order, for each of the considered heterodimer complexes.

PDB Id.	VSO → dry	Conservation → dry	Conservation → VSO
1HE1	0.42	0.28	0.02
1CXZ	0.39	0.24	0.19
1CEE	0.39	0.12	0.11
1C1Y	0.36	0.17	0.05
1RRP	0.34	0.22	0.21
1FIN	0.34	0.10	0.18
1E96	0.34	-0.02	0.15
1ZBD	0.33	0.09	0.19
1FOE	0.33	0.19	0.27
1A0O	0.32	0.23	0.12
2TRC	0.32	-0.08	0.11
1GOT	0.32	0.13	0.23
1WQ1	0.31	0.19	0.08
1IBR	0.30	0.01	-0.14
1A2K	0.26	0.15	0.28
1LFD	0.25	0.26	0.15
1AGR	0.19	0.10	0.25
1YCS	0.16	0.16	0.29
avg.	0.31	0.14	0.15

Table 2 Homodimers. Performance of shelling order (VSO) as a predictor for dryness, of conservation as a predictor for dryness, and of conservation as a predictor for shelling order, for each of the considered homodimer complexes.

PDB Id.	VSO → dry	Conservation → dry	Conservation → VSO
2BIF	0.45	0.09	0.02
1E5Q	0.45	0.15	0.31
1E2D	0.45	0.37	0.38
1H7T	0.45	0.12	0.17
1TB5	0.43	0.14	0.02
2DOR	0.42	0.19	0.13
1QIN	0.42	0.14	0.14
1E98	0.42	0.40	0.45
1J79	0.40	-0.09	-0.08
1NYW	0.40	-0.09	0.04
1BTO	0.38	0.27	0.12
1Y6R	0.38	0.17	0.03
1KER	0.37	0.14	0.08
1EK4	0.37	0.15	0.21
1LBX	0.37	0.21	0.11
1L9W	0.36	0.29	0.27
1AI2	0.36	0.18	-0.05
1W1U	0.35	0.07	-0.03
1DQX	0.33	0.10	-0.09
1E7Y	0.32	0.24	-0.06
1HKV	0.32	0.09	0.04
1M0S	0.32	0.07	0.34
1KC3	0.32	0.35	0.32
1M4N	0.31	0.17	0.14
1A59	0.31	0.15	0.19
1DQR	0.31	0.09	0.08
1AN9	0.30	0.11	0.06
1M7P	0.29	0.01	0.08
1TC2	0.29	-0.01	0.17
1AD3	0.28	-0.03	0.16
1ALN	0.27	0.14	0.04
1H16	0.27	-0.06	-0.02
1M9N	0.26	0.09	0.20
1L5W	0.24	0.18	0.25
1CG0	0.22	0.12	0.05
1LXY	0.21	0.10	0.11
avg.	0.34	0.13	0.12

7 FIGURES

Figure 1 Voronoi diagram (light solid lines) for a hypothetical molecule consisting of four atoms (a_1 to a_4), and restriction of the balls to their Voronoi regions. The α-complex ($\alpha = 0$) consists of the four vertices a_1 to a_4, of the three edges a_1a_2, a_1a_3, a_2a_3, and of the triangle $a_1a_2a_3$ formed between them.

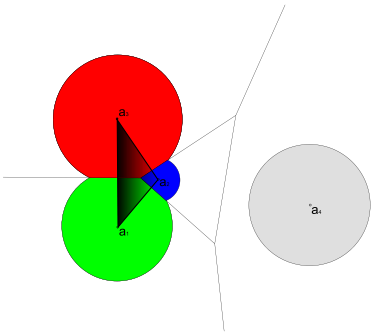


Figure 2 (a) Shelling of the Voronoi interface of a dimer complex, seen from the top. Solid dots represent protein atoms' centers, hollow dots water atoms' centers; for clarity, all atomic radii have been taken equal and the corresponding spheres omitted. The Voronoi facets composing the protein-protein interface are colored according to their shelling order: one (light gray, at the rim), two (middle gray), three (dark gray). (b) Two-dimensional illustration of the Voronoi interface shelling of a dimer complex. Red and blue circles represent the atoms of each partner, the green circle a water molecule. Interface Delaunay edges, which connect atoms on different partners, are shown as solid black (AB interface) or green ($AW - BW$ interface) lines; the Voronoi facets are shown as dashes. Black numerals denote the shelling order of each Delaunay edge/Voronoi facet, from which the atomic shelling orders (red, blue and green numerals) can be derived (refer to text for details). On this simple illustration, the high curvature of the $AW - BW$ interface due to the water molecule accounts for the high shelling order of the blue atoms.

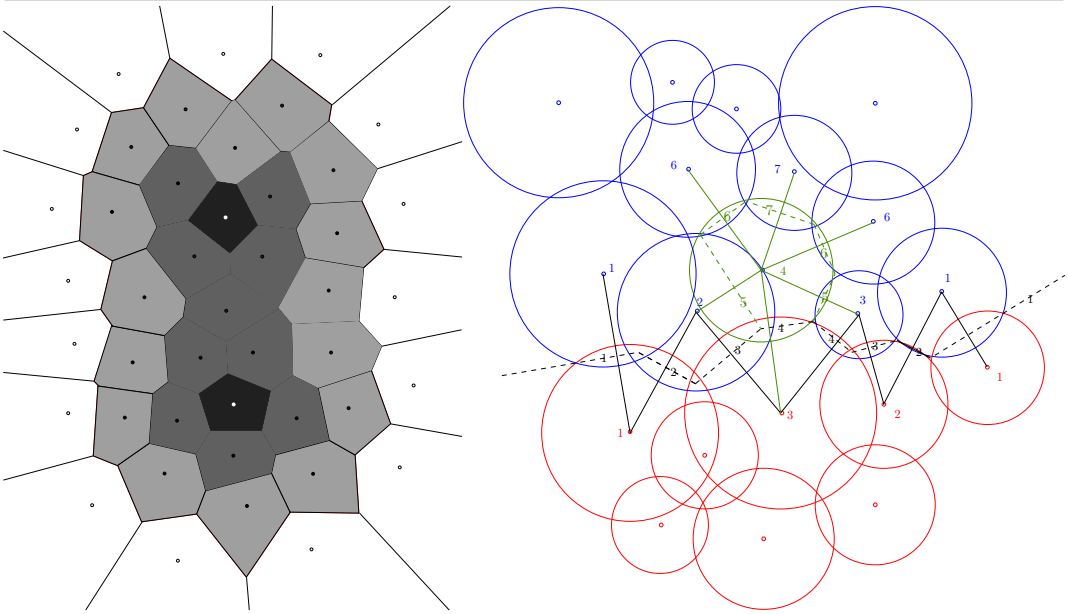


Figure 3 (a) Voronoi interface of the 2DOR homodimer complex, superimposed on the solvent accessible surface representation of one of the monomers (gray); for clarity, the second monomer is not shown. The facet shelling order varies from 1 (blue) to 6 (red). (b) Solvent accessible surface of one monomer of the 2DOR complex, showing the shelling order of interface atoms (color-coded as in panel b).

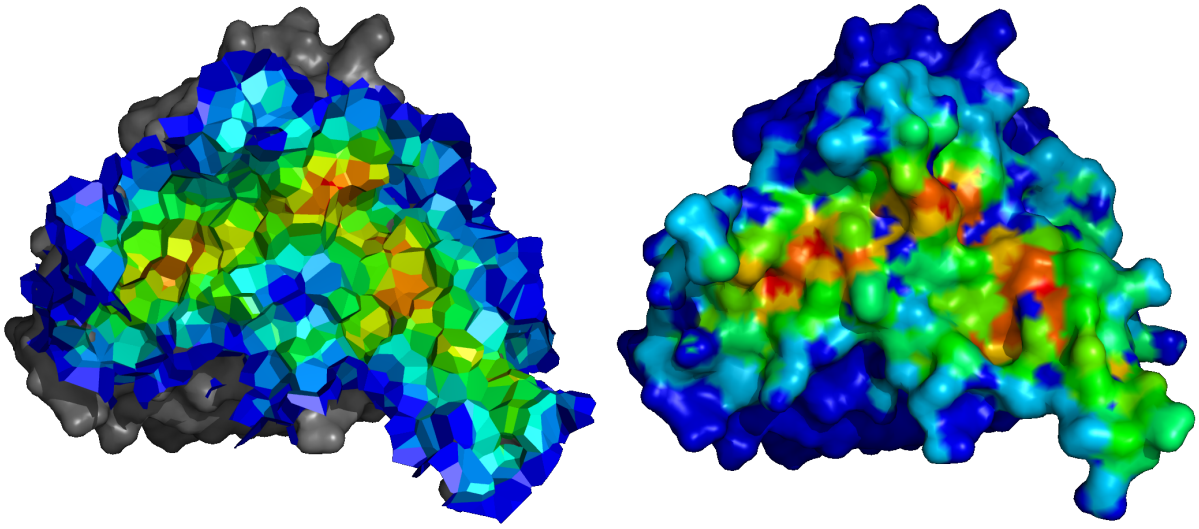


Figure 4 ROC plots evaluating shelling order (solid line) and conservation (dashed line) as predictors for dryness. Each point on a ROC plot corresponds to a different threshold value for the prediction. The plot for a perfect predictor should pass through $(1, 1)$; that of a random predictor (on average) is the diagonal (dotted line). The area between the ROC curve and the diagonal measure the performance of the predictor compared to random.

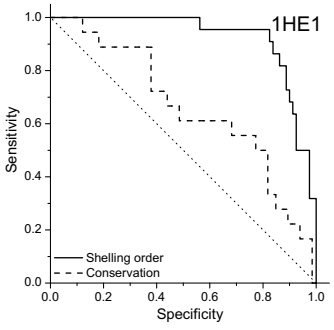


Figure 5 Performance of shelling order (circles, solid line) and conservation (squares, dashed line) as predictors of dryness, for all studied heterodimer (left panel) and homodimer (right panel) complexes. Scores are measured as the area between the corresponding ROC curve and the diagonal; complexes are sorted by decreasing shelling order score. Negative values (hatched area) denote a performance that is no better (on average) than that of a purely random classifier.

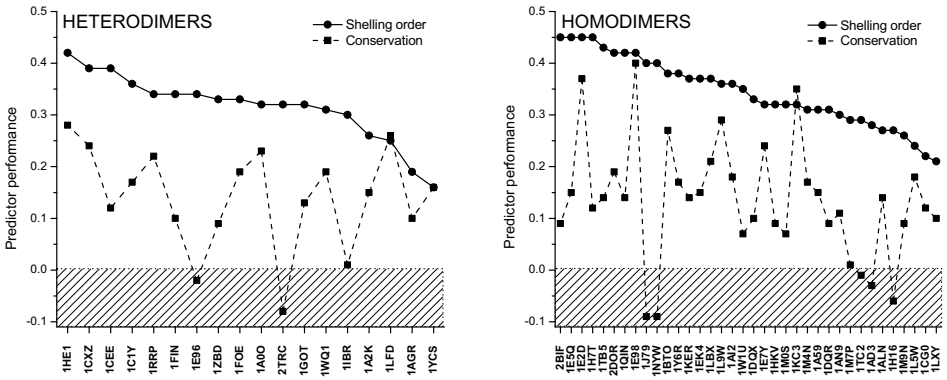


Figure 6 Spatial distribution of conservation across heterodimer interfaces. The conservation score for each interface residue, normalized to the maximum score, is plotted against its normalized shelling order. Black –: running average with a large window size (1/4 of all interface residues); Gray –: all data points; – · –: Integral over running average; · · · : Double integral over running average; Δ: deviation of the double integral from 0.5 – values below zero indicate conservation bias towards high shelling order (the core).

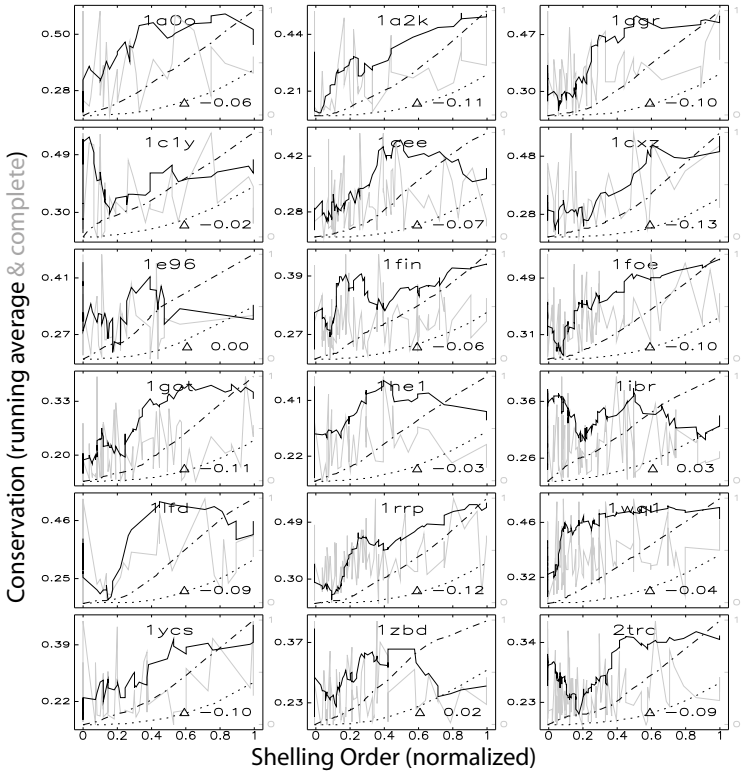


Figure 7 Spatial distribution of conservation across homodimer interfaces. See figure 6 and text for a detailed description.

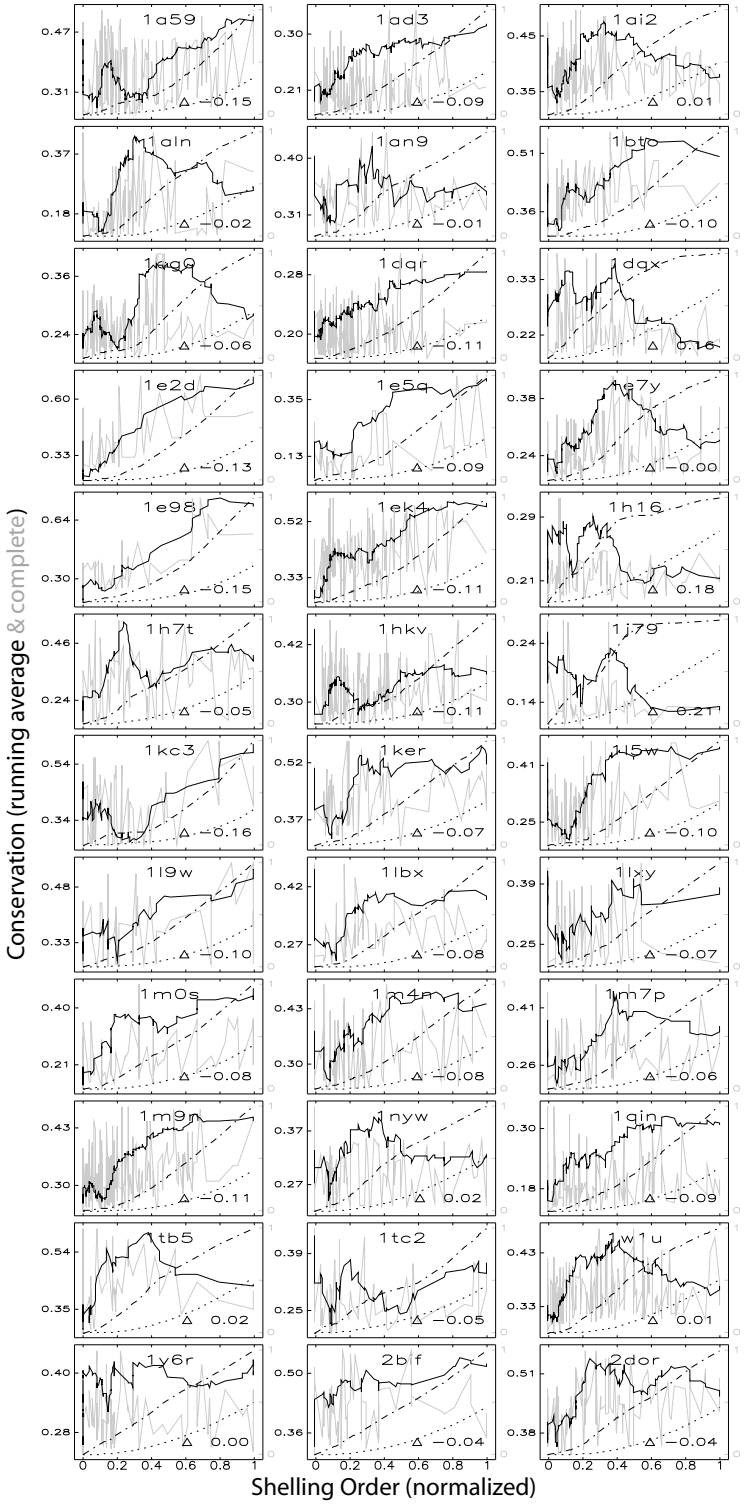


Figure 8 Projection of shelling order (large panels), dryness (lower left-hand panel) and conservation (lower right-hand panel) on the molecular surface of homocomplexes 1E2D (left), 1L5W (center) and 1A59 (right); one of the monomers was removed for clarity. Cold (resp. hot) colors represent low (resp. high) values; gray areas denote residues for which conservation information was unavailable.

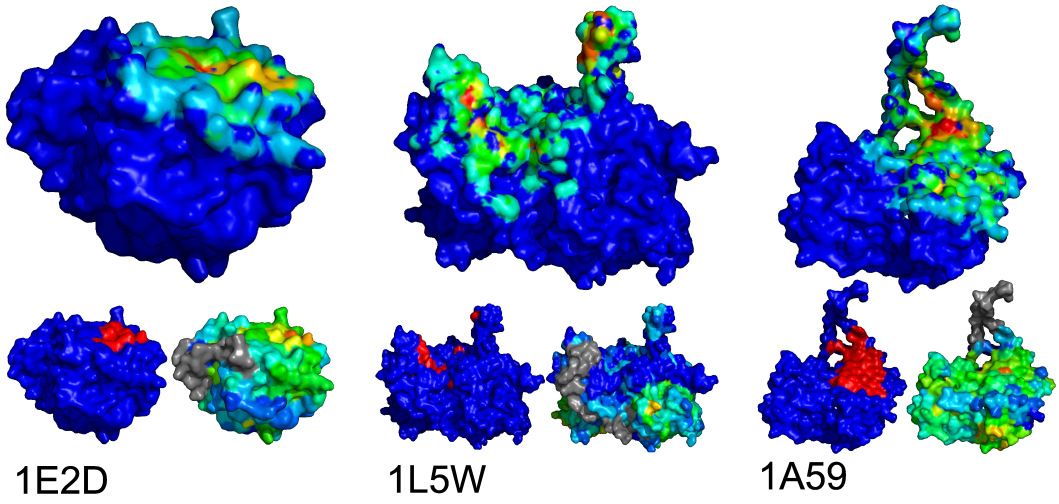


Figure 9 View of the cleft region of the 1L5W interface, showing the two protein partners as solid and mesh surfaces, respectively. Colors code for shelling order, which is low inside the cleft due to the presence of numerous water molecules which fragment the interface.

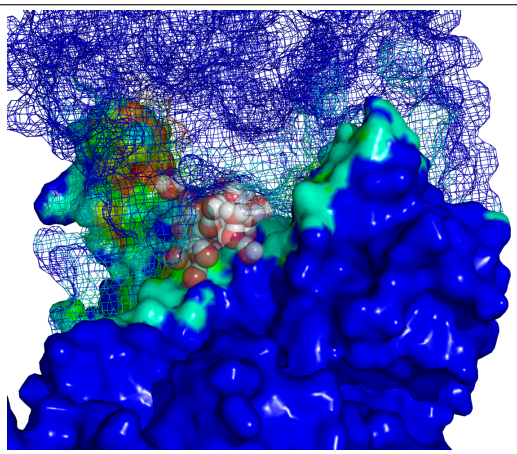
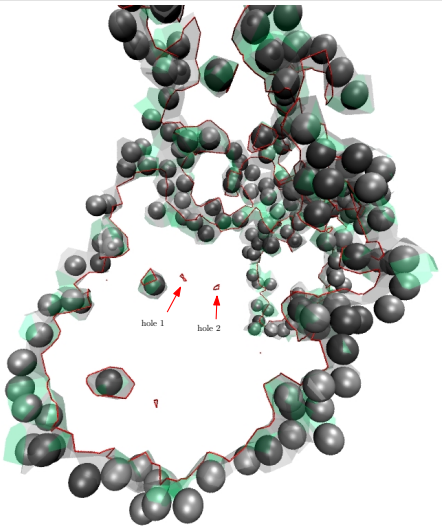


Figure 10 Boundary of the AB interface of complex 1A59 (red line), interfacial water (gray spheres), and $AW - BW$ interface (grey and green Voronoi polygons). The holes pointed out by arrows prevent the shelling order from peaking in the middle of the interface patch –compare to the bottom left panel of complex 1A59 on Fig. 8.



References

[1] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. DÃ¼mpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery., Nature 440 (7084) (2006) 631–636.
URL `http://dx.doi.org/10.1038/nature04532`

[2] J. J. Gray, High-resolution protein-protein docking., Curr Opin Struct Biol 16 (2) (2006) 183–193.
URL `http://dx.doi.org/10.1016/j.sbi.2006.03.003`

[3] R. Grünberg, J. Leckner, M. Nilges, Complementarity of structure ensembles in protein-protein binding., Structure 12 (12) (2004) 2125–2136.
URL `http://dx.doi.org/10.1016/j.str.2004.09.014`

[4] R. Grünberg, M. Nilges, J. Leckner, Flexibility and conformational entropy in protein-protein binding., Structure 14 (4) (2006) 683–693.
URL http://dx.doi.org/10.1016/j.str.2006.01.014

[5] C. Chotia, J. Janin, Principles of protein-protein recognition, Nature 256 (1975) 705–708.

[6] J.-L. K. Kouadio, J. R. Horn, G. Pal, A. A. Kossiakoff, Shotgun alanine scanning shows that growth hormone can bind productively to its receptor through a drastically minimized interface., J Biol Chem 280 (27) (2005) 25524–25532.
URL http://dx.doi.org/10.1074/jbc.M502167200

[7] S. Jones, J. Thornton, Analysis of protein-protein interaction sites using surface patches, J. Mol. Biol. 272.

[8] B. Ma, T. Elkayam, H. Wolfson, R. Nussinov, Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces, Proceedings of the National Academy of Sciences 100 (10) (2003) 5772–5777.
URL http://www.pnas.org/cgi/content/abstract/100/10/5772

[9] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein-protein recognition sites, Journal of Molecular Biology 285 (1999) 2177–2198.
URL http://www.sciencedirect.com/science/article/B6WK7-45R884M-RY/2/6f4a9866e2495a34273695de046893dc

[10] S. A. Teichmann, Principles of protein-protein interactions, Bioinformatics 18 (suppl. 2) (2002) S249–.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/suppl_2/S249

[11] T. Clackson, J. Wells, A hot spot of binding energy in a hormone-receptor interface, Science 267 (5196) (1995) 383–386.
URL http://www.sciencemag.org/cgi/content/abstract/267/5196/383

[12] I. S. Moreira, P. A. Fernandes, M. J. Ramos, Hot spots – a review of the protein-protein interface determinant amino-acid residues, Proteins: Structure, Function, and Bioinformatics 68 (4).
URL http://dx.doi.org/10.1002/prot.21396

[13] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, G. Schreiber, The modular architecture of protein-protein binding interfaces, PNAS 102 (1) (2005) 57–62.
URL http://www.pnas.org/cgi/content/abstract/102/1/57

[14] D. Reichmann, O. Rahat, M. Cohen, H. Neuvirth, G. Schreiber, The molecular architecture of protein-protein binding sites, Current Opinion in Structural Biology 17 (2007) 67–76.

URL http://www.sciencedirect.com/science/article/B6VS6-4MVDVF3-3/2/60d484544e9900423f219f319b0936b5

[15] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, E. S. Huang, Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?, Protein Sci 13 (1) (2004) 190–202.
URL http://www.proteinscience.org/cgi/content/abstract/13/1/190

[16] B. E. Shakhnovich, N. V. Dokholyan, C. DeLisi, E. I. Shakhnovich, Functional fingerprints of folds: Evidence for correlated structure-function evolution, Journal of Molecular Biology 326 (1) (2003) 1–9.
URL http://www.sciencedirect.com/science/article/B6WK7-47RC32Y-3/2/780a30e8b2d84bdefba0ea80b0a6b6b2

[17] A. Valencia, Automatic annotation of protein function, Current Opinion in Structural Biology 15 (3) (2005) 267–274.
URL http://www.sciencedirect.com/science/article/B6VS6-4G7X9HS-3/2/90d887674957c53860854c3254a94748

[18] P. Aloy, H. Ceulemans, A. Stark, R. B. Russell, The relationship between sequence and interaction divergence in proteins, Journal of Molecular Biology 332 (5) (2003) 989–998.
URL http://www.sciencedirect.com/science/article/B6WK7-49HMCGT-4/2/6d13750d3d40cc10b07ef096ef54961b

[19] R. P. Bahadur, P. Chakrabarti, F. Rodier, J. Janin, A dissection of specific and non-specific protein-protein interfaces, Journal of Molecular Biology 336 (4) (2004) 943–955.
URL http://www.sciencedirect.com/science/article/B6WK7-4BRTKR9-C/2/ce716dee1132d2605f8e1c96d1154253

[20] O. Lichtarge, H. R. Bourne, F. E. Cohen, An evolutionary trace method defines binding surfaces common to protein families, Journal of Molecular Biology 257 (2) (1996) 342–358.
URL http://www.sciencedirect.com/science/article/B6WK7-45PV59P-1T/2/08c0824641e25a3fe8dba3e81635a653

[21] O. Rahat, A. Yitzhaky, G. Schreiber, Cluster conservation as a novel tool for studying protein-protein interactions evolution., Proteins.
URL http://dx.doi.org/10.1002/prot.21749

[22] F. Rodier, R. Bahadur, P. Chakrabarti, J. Janin, Hydration of protein - protein interfaces, Proteins 60 (1) (2005) 36–45.

[23] A. A. Bogan, K. S. Thorn, Anatomy of hot spots in protein interfaces, Journal of Molecular Biology 280 (1998) 1–9.

URL http://www.sciencedirect.com/science/article/B6WK7-45S49GB-9C/2/b3d9c6f299c1eec3933d2774dffaf67d

[24] A. Fernandez, R. S. Berry, Extent of Hydrogen-Bond Protection in Folded Proteins: A Constraint on Packing Architectures, Biophys. J. 83 (5) (2002) 2475–2481.
URL http://www.biophysj.org/cgi/content/abstract/83/5/2475

[25] A. Fernandez, H. A. Scheraga, Insufficiently dehydrated hydrogen bonds as determinants of protein interactions, Proceedings of the National Academy of Sciences 100 (1) (2003) 113–118.
URL http://www.pnas.org/cgi/content/abstract/100/1/113

[26] I. Mihalek, I. Res, O. Lichtarge, On itinerant water molecules and detectability of protein-protein interfaces through comparative analysis of homologues, Journal of Molecular Biology 369 (2) (2007) 584–595.

[27] F. Cazals, F. Proust, R. P. Bahadur, J. Janin, Revisiting the Voronoi description of protein-protein interfaces, Protein Sci 15 (9) (2006) 2082–2092.
URL http://www.proteinscience.org/cgi/content/abstract/15/9/2082

[28] F. M. Richards, The interpretation of protein structures: Total volume, group volume distributions and packing density, Journal of Molecular Biology 82 (1974) 1–14.

[29] F. Aurenhammer, Power diagrams: properties, algorithms and applications, SIAM J. Comput. 16 (1987) 78–96.

[30] C. Chotia, The nature of accessible and buried surfaces in proteins, J. Mol. Bio. 105 (1976) 1–12.

[31] B. Lee, F. M. Richards, The interpretation of protein structures: Estimation of static accessibility, Journal of Molecular Biology 55 (3) (1971) 379–380.
URL http://www.sciencedirect.com/science/article/B6WK7-4DNGV9F-34/2/659a5209b127663bc3133c0b801a29a5

[32] H. Edelsbrunner, E. P. Mücke, Three-dimensional alpha shapes, ACM Trans. Graph. 13 (1) (1994) 43–72.

[33] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucl. Acids Res. 34 (suppl. 1) (2006) D247–251.
URL http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_1/D247

[34] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998, Ch. The theory behind profile HMMs.

[35] K. Wang, R. Samudrala, Incorporating background frequency improves entropy-based residue conservation measures., BMC Bioinformatics 7 (2006) 385.
URL http://dx.doi.org/10.1186/1471-2105-7-385

[36] M. Guharoy, P. Chakrabarti, Conservation and relative importance of residues across protein-protein interfaces., Proc Natl Acad Sci U S A 102 (43) (2005) 15447–15452.
URL http://dx.doi.org/10.1073/pnas.0505425102

[37] T. Selzer, G. Schreiber, Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction., J Mol Biol 287 (2) (1999) 409–419.
URL http://dx.doi.org/10.1006/jmbi.1999.2615

[38] N. Ostermann, I. Schlichting, R. Brundiers, M. Konrad, J. Reinstein, T. Veit, R. S. Goody, A. Lavie, Insights into the phosphoryltransfer mechanism of human thymidylate kinase gained from crystal structures of enzyme complexes along the reaction coordinate, Structure 8 (6) 629–642.
URL http://www.sciencedirect.com/science/article/B6VSR-40H578S-B/2/72acba30cd1e54cc5111cf865a8374b2

[39] A. Varshney, F. P. Brooks, D. C. Richardson, W. V. Wright, D. Manocha, Defining, computing, and visualizing molecular interfaces, in: IEEE Visualization, Atlanta, USA, 1995, pp. 36–43.

[40] Y.-E. A. Ban, , H. Edelsbrunner, J. Rudolph, Interface surfaces for protein-protein complexes, in: RECOMB, San Diego, 2004, pp. 205–212.

[41] R. R. Gabdoulline, R. C. Wade, D. Walther, Molsurfer: A macromolecular interface navigator., Nucleic Acids Res 31 (13) (2003) 3349–3351.

[42] A. Poupon, Voronoi and voronoi-related tessellations in studies of protein structure and interaction., Curr Opin Struct Biol 14 (2) (2004) 233–241.
URL http://dx.doi.org/10.1016/j.sbi.2004.03.010

[43] S. Chakravarty, R. Varadarajan, Residue depth: a novel parameter for the analysis of protein structure and stability., Structure 7 (7) (1999) 723–732.

[44] A. Pintar, O. Carugo, S. Pongor, Atom depth in protein structure and function., Trends Biochem Sci 28 (11) (2003) 593–597.

[45] A. Pintar, O. Carugo, S. Pongor, Atom depth as a descriptor of the protein interior., Biophys J 84 (4) (2003) 2553–2561.

[46] P. Chakrabarti, J. Janin, Dissecting protein-protein recognition sites, Proteins 47.

[47] I. M. A. Nooren, J. M. Thornton, Structural characterisation and functional significance of transient protein-protein interactions., J Mol Biol 325 (5) (2003) 991–1018.

[48] A. Bordner, R. Abagyan, Statistical analysis and prediction of protein-protein interfaces, Proteins 60 (3) (2005) 353–66.

[49] J. G. S. Sheldon Park, Statistical and molecular dynamics studies of buried waters in globular proteins, Proteins: Structure, Function, and Bioinformatics 60 (2005) 450–463.
URL http://dx.doi.org/10.1002/prot.20511

[50] B. Brooks, R. Bruccoleri, Olafson B.D., D. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization and dynamics calculations., J Comp Chem 4 (1983) 187–217.

[51] P. Steinbach, R. Loncharich, B. Brooks, The effects of environment and hydration on protein dynamics: A simulation study of myoglobin., Chem Phys 158 (1991) 383–94.

[52] R. Grünberg, M. Nilges, J. Leckner, Biskit–A software platform for structural bioinformatics, Bioinformatics 23 (6) (2007) 769–770.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/6/769

[53] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, Gromacs: fast, flexible, and free., J Comput Chem 26 (16) (2005) 1701–1718.
URL http://dx.doi.org/10.1002/jcc.20291

[54] W. Damm, A. Frontera, J. Tirado-Rives, W. L. Jorgensen, Opls all-atom force field for carbohydrates, Journal of Computational Chemistry 18 (1997) 1955–1970.
URL http://dx.doi.org/10.1002/(SICI)1096-987X(199712)18:16<1955::AID-JCC1>3.0.CO;2-L

[55] CGAL, Computational Geometry Algorithms Library, http://www.cgal.org.

[56] http://cgal.inria.fr/Intervor.

[57] D. M. Green, J. M. Swets, Signal detection theory and psychophysics, John Wiley and Sons Inc., New York, 1966.

[58] M. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine [published erratum appears in Clin Chem 1993 Aug;39(8):1589], Clin Chem 39 (4) (1993) 561–577.
URL http://www.clinchem.org/cgi/content/abstract/39/4/561

[59] W. DeLano, The Pymol molecular graphics system, http://www.pymol.org (2002).

[60] O. Cheong, The Ipe extensible drawing editor, http://tclab.kaist.ac.kr/ipe/ (1993-2007).

[61] http://cgal-ipelets.gforge.inria.fr/.

8 Supplemental Material

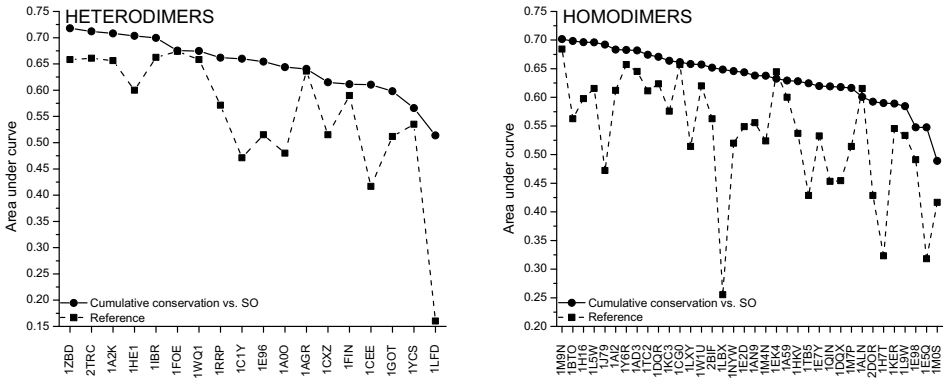
8.1 Distribution of conserved residues: interface residues versus dry residues

As outlined in section 3.2, we compare the spatial distribution of conservation in the entire set of interface residues with that of the dry residues.

We first consider all interface residues. To study the cumulated conservation score over consecutive shells, we compute the proportion of the interface conservation score which is contained in the subset of residues whose average VSO is lower than some value. Normalizing over shelling orders and varying the threshold yields a curve that rises from $(0, 0)$ (no residues selected, zero cumulative conservation) to $(1, 1)$ (all residues selected, 100% cumulative conservation). The area under this curve provides information about the variation of conservation with shelling order, since numerous highly conserved residues with low (high) shelling order will cause the curve to rise early (late) and result in large (small) areas.

Next, we focus on the dry residues and construct references with which to compare the previously computed areas, that quantify the relevance of rim residue conservation in each case. Denoting n_{dry} the number of dry residues of a given complex as reported in [26], we sort the interface residues by decreasing shelling order and assume the first n_{dry} only to be conserved –those with highest shelling orders. Let m and M be the minimum and maximum shelling orders in this subset, respectively (note that M is also the highest VSO found in the entire complex), and let $x = m/M$. The step function which is null from 0 to x, and equal to 1 from x to 1, maximizes the area $1 - x$ under the curve relative to the conservation of the subset of n_{dry} residues.

Figure 11 Area under the normalized cumulative conservation vs. shelling order curve (circles, solid line) and reference area (squares, dashed line), for all studied heterodimer (left panel) and homodimer (right panel) complexes – see text for details. Areas larger than the reference denote complexes for which rim residues are significantly conserved.



As seen from Figure 11, the rim residues account for a non-negligible part of the conservation: the area under the corresponding curve was found to be greater than the reference in all but two homodimer complexes, for which both measures were roughly equal. This could, in part, be due to a purely statistical effect: a large number of moderately conserved rim residues might end up having more weight than a small number of highly conserved core amino-acids. However, the peak in average conservation observed at the rim of many complexes (Section 3.2 (i)) proves that highly conserved residues occur on a non fortuitous basis at the rim of protein-protein interfaces – most likely as anchors for important electrostatic interactions that dictate complex formation and activity.

Table 3 Relationship of shelling order and conservation for the heterodimer set: proportion of total conservation provided by noninterface residues, area under the normalized cumulative conservation vs. VSO curve (see text), area under the corresponding 'reference' curve (see text).

PDB Id.	Proportion of conservation score for noninterface residues	Area under curve, interface residues	Reference
1YCS	0.76	0.57	0.53
1RRP	0.61	0.66	0.57
1E96	0.83	0.65	0.52
1CXZ	0.78	0.61	0.52
1LFD	0.80	0.51	0.16
1WQ1	0.64	0.67	0.66
1FOE	0.77	0.68	0.67
1AGR	0.77	0.64	0.64
1IBR	0.77	0.70	0.66
1FIN	0.75	0.61	0.59
1HE1	0.61	0.70	0.60
1A2K	0.70	0.71	0.66
1A0O	0.71	0.64	0.48
1ZBD	0.79	0.72	0.66
1GOT	0.83	0.60	0.51
2TRC	0.71	0.71	0.66
1CEE	0.62	0.61	0.42
1C1Y	0.77	0.66	0.47

Table 4 Relationship of shelling order and conservation for the homodimer set: proportion of total conservation provided by noninterface residues, area under the normalized cumulative conservation vs. VSO curve (see text), area under the corresponding 'reference' curve (see text).

PDB Id.	Proportion of conservation score for noninterface residues	Area under curve, interface residues	Reference
1A59	0.72	0.63	0.60
1H16	0.89	0.70	0.60
1M0S	0.73	0.49	0.42
1E5Q	0.97	0.55	0.32
1H7T	0.83	0.59	0.32
1E7Y	0.86	0.62	0.53
1ALN	0.64	0.60	0.62
1CG0	0.71	0.66	0.66
1E2D	0.81	0.64	0.55
1W1U	0.84	0.66	0.62
1KER	0.86	0.59	0.55
1EK4	0.74	0.63	0.64
1BTO	0.74	0.70	0.56
1QIN	0.36	0.62	0.45
1TB5	0.84	0.62	0.43
1M4N	0.76	0.64	0.52
2BIF	0.86	0.65	0.56
1M9N	0.57	0.70	0.68
1M7P	0.74	0.62	0.51
1E98	0.83	0.55	0.49
1L5W	0.95	0.70	0.62
1AD3	0.74	0.68	0.65
1J79	0.85	0.69	0.47
1AI2	0.62	0.68	0.61
1L9W	0.90	0.58	0.53
1LXY	0.87	0.66	0.51
1NYW	0.64	0.65	0.52
1KC3	0.87	0.66	0.58
1Y6R	0.72	0.68	0.66
1LBX	0.76	0.65	0.26
2DOR	0.72	0.59	0.43
1DQR	0.64	0.67	0.62
1AN9	0.85	0.64	0.56
1TC2	0.79	0.67	0.61
1HKV	0.72	0.63	0.54
1DQX	0.57	0.62	0.45

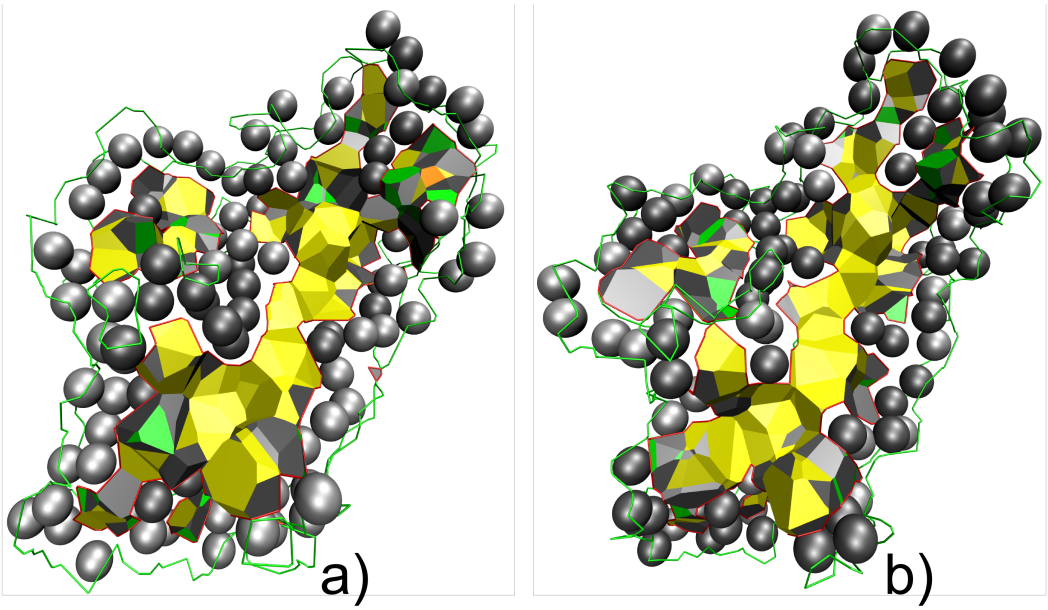
8.2 Validation of the sample preparation procedure

The procedure employed for the rehydration and equilibration of each of the complexes (Section 5) has deliberately been kept short, and can be run in minutes on a desktop computer. In this paragraph, we ascertain whether the placement and equilibration of the water molecules added using this fast protocol are of sufficient quality for the current application. Of particular interest are the interfacial water molecules. When in simultaneous contact with both protein partners, they form the $AW - BW$ interface (Figure 2b and 10); but several layers of water inside a larger pocket will create holes in the interface, possibly splitting it into several connected components. The implications for shelling orders are crucial: in the first case, the water molecules will not affect the SO, while in the second scenario a boundary is created and the SO consequently reset to 1.

The complex 1M0S, which features a large pocket filled with crystal water molecules, was used for the test. A rigorous equilibration procedure, retaining the crystal water molecules and involving a 5 ns molecular dynamics simulation with state-of-the-art algorithms and parameters (Section 5), provided us with a reference structure. Both this structure and the one from the fast procedure were used as input to Intervor. Figure 12 shows the tessellation of the AB interface and the interfacial water molecules for both cases. Due to minor conformational transitions that have occurred during the 5 ns MD simulation, the two interfaces are not superposable. However, they retain the same shape and number of connected components. In both cases, the central cavity is filled with interfacial water that participates to the ABW interface. Both interfaces feature boundaries of comparable lengths and topologies.

This difficult test case provides justification for our sample preparation methodology. It also represents a tribute to the robustness of our model, which delivers stable results upon variation of the solvation of the complex within a reasonable range.

Figure 12 The AB interface (colored Voronoi facets) and the interfacial water molecules W (grey spheres) for two distinct rehydration and equilibration procedures – a fast (a) and a more exhaustive one (b); see text for details. Boundaries of the AB and $AW - BW$ interfaces are shown as red and green sticks, respectively.



Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399