PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Defining and characterizing protein surface using alpha shapes

Laurent-Philippe Albou,[1†] Benjamin Schwarz,[1,2†] Olivier Poch,[1*] Jean Marie Wurtz,[1] and Dino Moras[1]

[1] Department of Biology and Structural Genomics, IGBMC, CNRS, INSERM, ULP, Ilkirch, France
[2] LSIIT UMR 7005 CNRS, Université de Strasbourg, Strasbourg, France

ABSTRACT

The alpha shape of a molecule is a geometrical representation that provides a unique surface decomposition and a means to filter atomic contacts. We used it to revisit and unify the definition and computation of surface residues, contiguous patches, and curvature. These descriptors are evaluated and compared with former approaches on 85 proteins for which both bound and unbound forms are available. Based on the local density of interactions, the detection of surface residues shows a sensibility of 98%, whereas preserving a well-formed protein core. A novel conception of surface patch is defined by traveling along the surface from a central residue or atom. By construction, all surface patches are contiguous and, therefore, allows to cope with common problems of wrong and nonselection of neighbors. In the case of protein-binding site prediction, this new definition has improved the signal-to-noise ratio by 2.6 times compared with a widely used approach. With most common approaches, the computation of surface curvature can be locally biased by the presence of subsurface cavities and local variations of atomic densities. A novel notion of surface curvature is specifically developed to avoid such bias and is parametrizable to emphasize either local or global features. It defines a molecular landscape composed on average of 38% knobs and 62% clefts where interacting residues (IR) are 30% more frequent in knobs. A statistical analysis shows that residues in knobs are more charged, less hydrophobic and less aromatic than residues in clefts. IR in knobs are, however, much more hydrophobic and aromatic and less charged than noninteracting residues (non-IR) in knobs. Furthermore, IR are shown to be more accessible than non-IR both in clefts and knobs. The use of the alpha shape as a unifying framework allows for formal definitions, and fast and robust computations desirable in large-scale projects. This swiftness is not achieved to the detriment of quality, as proven by valid improvements compared with former approaches. In addition, our approach is general enough to be applied on nucleic acids and any other biomolecules.

Proteins 2009; 76:1–12.

Key words: surface; alpha shape; patch; curvature; binding site; interaction; knob; cleft; structural bioinformatics; computational biology.

INTRODUCTION

The biological function of a protein essentially relies on its interactions with solvent and other biomolecules. Chemical and structural diversity observed at molecular surfaces allow for the wide variety of interactions necessary for cellular life. To decipher biological processes, it is thus crucial to accurately define the nature and shape of these surfaces. The determination of the surface in terms of atoms, residues, and surface patches has already allowed to conduct numerous studies in the protein–protein interaction fields[1–3] as well as to develop several prediction algorithms for the detection of binding sites and the modeling of complexes.[4–7] More detailed characterization of the surface in terms of clefts and knobs (respectively concavities and convexities) was also used for the study of interface complementarity and docking of molecules.[8–11]

Amongst the methodologies used for the description of the surface of a molecule, the alpha shape theory[12] is probably one of the most promising. The alpha shape model of a molecule is a polyhedral representation that uniquely decomposes the space occupied by its atoms and retains interesting characteristics such as the shape of the molecule and a notion of interatom neighborhood. Despite the relative complexity of the theory, alpha shapes have been used to address a wide variety of problems in structural biology, such as the computation of protein surface and volume[13] as well as their derivatives,[14] the detection of pockets in known structures,[15–17] the construction of molecular surface meshes,[18,19] the validation of structures,[20,21] or the study of interfaces.[22,23]

Additional Supporting Information may be found in the online version of this article.

Grant sponsors: The Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM), Structural Proteomics in Europe (SPINE2-Complexes, CEE FP7 LSHG-CT-2006-031220), the Décrypthon program initiated by the Association Française contre les Myopathies (AFM, CAMI 12727), IBM, the Ligue Nationale contre le Cancer, comité du Haut-Rhin and the Université Louis Pasteur de Strasbourg (ULP).

[†] Laurent-Philippe Albou and Benjamin Schwarz contributed equally to this work.

*Correspondence to: Olivier Poch, 1 rue Laurent Fries, BP 10142, 67404 Illkirch CEDEX, France. E-mail: poch@igbmc.fr

Received 26 June 2008; Revised 30 September 2008; Accepted 1 October 2008
Published online 21 October 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22301

In this article, the alpha shape theory is used as a unifying framework to compute various properties depicting the surface of a biomolecule. The definition of s*urface atoms* is straightforwardly provided by the alpha shape model. For the definition of *surface residues* a novel notion is introduced, the *valence Vr* representing the density of surface interactions around an accessible residue. By radiating on the surface around a surface residue, we give a novel and intuitive definition of *contiguous surface patches*. Curvature computations based on solid angle approaches[8,24] usually do not differentiate the empty space due to subsurface cavities from the empty space above the surface; as a result they detect more protrusions than they should. To tackle this problem, we define the *exposure* of an atom as a modified solid angle computed locally above the alpha shape surface. This exposure is then smoothed in a surrounding region to define its local surface curvature. Based on this latest notion, clefts and knobs are detected on the surface and characterized in terms of accessibility and composition.

The biological relevance of these definitions is validated on a dataset of 85 proteins involved in transient heterodimeric interactions for which both the bound and unbound forms are available. We consider a protein chain to be in an unbound form if it participates only to crystal packing contacts.[25] As some conformational changes can occur during an assembly formation, it has been proposed to predict protein-binding sites using only these unbound forms.[5] In our dataset where small conformational changes are observed, we verified that most of the interacting residues (IR) seen in bound forms are also found on the surface in their respective unbound forms. In the context of protein-binding site analysis, we show that our surface patches have a better overlap with known binding sites and a better signal-to-noise ratio than the commonly used approach of Jones and Thornton.[26] In addition, our conception of local surface curvature correlates well with visual inspection and is compared with a former approach.[24] It allows a fast detection of clefts and knobs, dividing the surface in 38% knobs, the remaining being clefts. Knob residues are found to interact 30% more with partner proteins than cleft residues. IR are also shown to be, respectively, 15 and 18% more accessible than noninteracting residues (non-IR) in knobs and clefts. A more detailed analysis of these accessibilities reveal that hydrophobic and aromatic IR have 54% more accessibility in clefts than in knobs with respect to non-IR. The IR in knobs are indeed found to be more charged and less hydrophobic and less aromatic than those in clefts.

Our implementations benefit from fast and robust algorithms recently developed in Computational Geometry and provided by the CGAL library.[27] The geometric representation of alpha shapes combined with our geometric descriptors is generic and applicable to any molecular structure such as proteins, nucleic acids, or lipids. Furthermore, these tools are fast enough for use in large-scale projects such as interactomics.

METHODS

Alpha shapes

As depicted in Figure 1(A), molecules are generally described as a union of balls representing either van der Waals (VdW) or solvent accessible (SA) models.[28] Another common model is provided by the molecular surface,[29] defined as the limit of space around the molecule that a rolling probe sphere can actually touch.

Molecules can also be modeled with their *Delaunay complex*[30] [Fig. 1(B)] which is a unique partition of the three-dimensional (3D) space in nonoverlapping tetrahedra whose vertices are atom centers. This construction bears information on the atom neighborhood: a Delaunay edge links the two nearest atoms in the direction of that edge. Such edges can be arbitrarily long, covering for instance a surface cavity [Fig. 1(C)], segment [ab]). By trimming the "largest" Delaunay edges, triangles, and tetrahedra, it is possible to distinguish between the voids surrounding the molecule and the actual molecular object [Fig. 1(C), gray colored triangles]. This task is achieved through the *alpha complex*,[12] a filtration of Delaunay edges, facets, and tetrahedra based on the growth of soft balls virtually placed on all atom centers of the molecule. The size of these so called α-balls increases with alpha, and the alpha complex registers contacts between alpha balls: when two (respectively three or four) alpha balls touch each other, the corresponding Delaunay edge (respectively facet or tetrahedron) belongs to the alpha complex for this specific alpha value. In the present study we restrict our use to an alpha parameter of 0 [Fig. 1(C)]. This particular construction (also referred as *the dual complex* in the literature) corresponds to the case, where the radius of a ball modeling an atom measure the van der Waals radius of this atom raised by a probe sphere radius (generally 1.4 Å corresponding to a water molecule). In the following descriptions and discussions, we will assume a value of 0 for every occurrence of alpha. As demonstrated by Edelsbrunner[31] this construction is a unique and precise representation of the molecule.

The *alpha shape* (with alpha equal to 0, also known as *dual shape*) of a molecule is the border of its alpha complex [Fig. 1(C), bold segments]. It is a polyhedron with triangular facets that precisely depicts the surface of a molecule (see Fig. 2). A facet in the alpha shape links a triplet of surface atoms blocking a probe sphere, whereas an edge links two atoms that allow the same probe sphere to roll from one blocking position to another. The vertices of the alpha shape are exactly the atoms with a strictly positive accessible surface area (ASA > 0 Å).

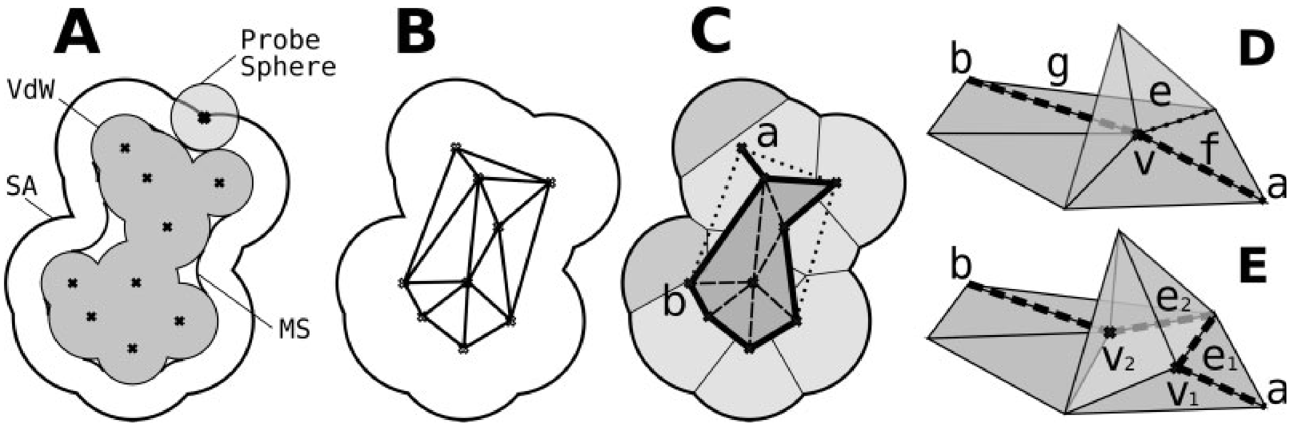

Figure 1

Molecular surface representations. (**A**) 2D example of van der Waals (VdW) representation of the molecule modeled as a union of balls with VdW radii. By rolling a probe sphere on the VdW surface one defines the surface accessible (SA) and Molecular Surface (MS) models. (**B**) In 2D, the Delaunay complex of a molecule (here represented in its SA model) is composed of straight lines between atom centers and the triangles they delineate. In 3D this "triangulation" contains also tetrahedra. (**C**) The molecule's 0-balls (matching the atoms in their SA representation) have been represented and their contacts emphasized by thin straight lines separating them. The alpha complex ($\alpha = 0$) comprises all Delaunay edges except the three dotted ones (for instance, the edge between vertices a and b is stripped because the corresponding grayed balls do not touch each other). For similar reasons three Delaunay triangles are stripped and only the seven grayed triangles belong to the alpha complex. The alpha shape is the border of the alpha complex, it is pictured here with bold edges. (**D**) A case of 3D surface ambiguity, for clarity the upper facets are depicted transparent. Traveling on the surface from vertex a to b with edges f and g allows one to cross the upper facets through vertex v. (**E**) To forbid such surface crossings, v is split in two, as well as the transparent facets and its two basis edges. Traveling from a to b now necessitates to visit edges e1 and e2.

The surface of the alpha shape may present ambiguities (nonmanifoldness) in cases where an atom (vertex) is shared by two sides of the surface [Fig. 1(D)]. Ambiguous vertices, edges and triangles are virtually split to prohibit surface crossing of facets [Fig. 1(E)]. The resulting surface is stored in a half-edge data structure. Ulti-

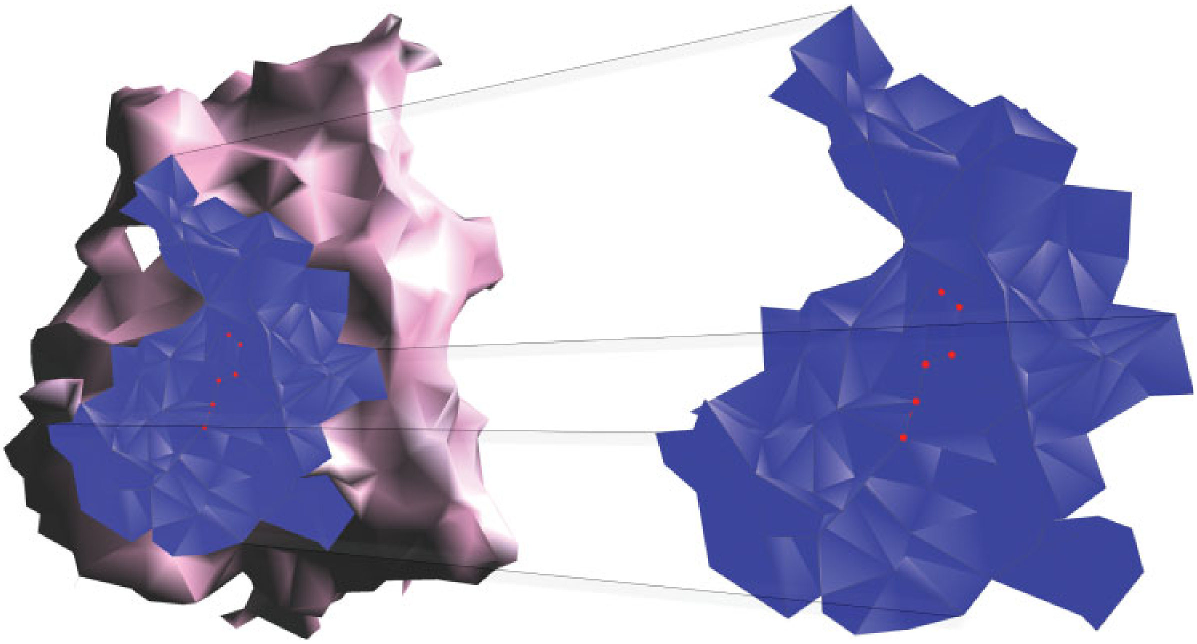

Figure 2

Alpha shape and a surface patch. Alpha shape of the protein RAR α (1dkfB) in purple. In blue, a contiguous surface patch generated by our approach. For better visualization, the right part of the figure presents a focus on the surface patch.

mately, the vertices and edges of this modified alpha shape provide a *graph depicting the neighborhood of atoms* on the protein surface. Only atoms that share a surface intersection in the SA model will be connected by an edge in this graph. A *graph of surface residue* neighbors is also constructed by connecting residues that share at least one edge in their atom neighborhood graph.

To compute alpha shapes, we rely on the library CGAL,[24] and use as parameters an alpha value of 0, a probe sphere radius of 1.4 Å, and common van der Waals radius for atoms. For a deeper insight into the alpha shape theory and for its relationship with molecular models please refer to other articles from Edelsbrunner and Mucke[12] and Edelsbrunner.[31] An introduction to these models is also provided by A. Poupon.[32]

Dataset of bound and unbound protein structures

To evaluate our surface descriptors, we have built a dataset of 85 proteins for which the structures of both their bound and unbound forms are available. Bound forms correspond to the structure of the protein extracted from the structure of an assembly, whereas the unbound forms correspond to the structure of the protein that participates only to crystal contacts.[25] Each of these proteins is involved in transient heterodimers (following the definition of Nooren and Thornton[33]), and therefore both the bound and unbound forms have a biological meaning.

As a first step, structures of protein assemblies with resolution better than 3 Å are extracted both from the Protein Data Bank (PDB)[34] and already published datasets.[3] Then, a non-redundant dataset of 225 transient heterodimers is built with a maximum sequence identity of 30%. Antigen-Antibody structures and assemblies with fragmented proteins are also removed. The Average length of protein chains is 240 amino acids and no protein chains have less than 50 amino acids or more than 576 amino acids.

The transient state is inferred *in silico* by checking in the PDB if known IR detected in a structure assembly are found to interact with at least one different partner in another assembly. However, these contacts may result from crystal packing and therefore do not necessarily occur *in vivo*. For this reason, all our assemblies and their transient aspects have been manually verified by consulting experiments described in the literature.

The unbound forms are then retrieved using the following approaches:

(a) Protein structures that are described as monomers in PISA[35] are retrieved.
(b) Each protein chain is compared with the monomers of PISA, using BLAST.[36] To lower the occurrence of conformational changes due to key mutations and emphasize only those due to the assembly formation, only structural candidates with at least 95% residue identity over 95% of the sequence length are retrieved.
(c). If several unbound structures remain, the one with the best resolution and b-factor is selected.

As a result of this process, 85 proteins are obtained for which both their bound and unbound forms are available (Supp. Info. Table I).

IR are then detected on bound forms by a change of accessibility of at least 1 Å^2 during the assembly formation.[3] The ASA values of atoms and residues are computed using the Naccess program[37] with default parameters. IR are then mapped on the corresponding unbound forms, using a pairwise alignment.

Surface residues

A *surface atom* is defined as an *accessible atom* (ASA > 0 Å^2). As previously stated, accessible atoms correspond exactly to the vertices of the alpha shape. The *valence Va of an accessible atom* is defined as the number of its accessible atom neighbors (the number of its edge connected atoms in the alpha shape). The *valence Vr of an accessible residue* (ASA > 0 Å^2) is defined as the number of edges connecting atoms from that residue to atoms of other accessible residues. An accessible residue is then considered as a surface residue by combining its number of surface atoms Nr, and its valence Vr (see "Results and Discussion" section).

In the approach of Miller *et al.*,[38] surface residues are defined as those having an observed ASA of at least 5% of their reference ASA. The reference ASA of a residue X is the ASA of the residue in a polypeptide extended-state Gly-X-Gly.

More recently Chakravarty *et al.*,[39] have proposed a novel way of defining surface residues by computing a notion of depth for every atom and residue of a protein structure. To compare this approach with ours, we implemented this notion of depth using the surface atoms as reference, as proposed by Pintar *et al.*[40]

To optimize our definition of surface residues, a measure of sensitivity is assessed by considering the fraction of known IR that are described as being part of the surface in bound forms. Residues that are not described as surface residues are considered part of the protein core. A measure of specificity is then evaluated by keeping trace of the fraction of amino acids that constitutes the protein core.

We perform an optimization of our surface residue detection by varying simultaneously Nr and Vr. During this optimization, the best definition of surface residues is attained when almost all IR are detected as being surface residues, whereas the protein core contains the biggest fraction of amino acids. For comparisons with the Miller *et al.*[38] and Pintar *et al.*[40] approaches, we vary respectively the percentage of accessibility and the threshold of depth used to detect surface residues and evaluate the fraction of core residues.

Surface patches

Two kinds of surface patches are generally considered in the literature: surface patches of variable size that correspond to subregions of an interface assembly[3] and surface patches of a given size that are generated evenly over the surface of a single protein.[26] Although the first approach is aimed at better characterizing a known interface, the second approach is commonly used to average properties on a specific region to predict a biologically relevant fact such as protein or nucleic acid binding sites. We focus on this second definition.

To construct an atom surface patch around a surface atom, we gather the nearest surface atoms that are reachable over a continuous surface from that center (see Fig. 2). This is achieved by computing minimal distances from the central atom to every other atom in the graph of surface atoms introduced in the Alpha shapes section. This computation relies on the Dijsktra shortest distance path algorithm[41] where edges linking two surface atoms are weighted according to the euclidian distance separating them. Essentially, the distance over the surface computed for any two atoms is the sum of the edge lengths forming the shortest path between these two atoms.

Residue surface patches are computed in the same way and the weight of an edge linking two accessible residues correspond to the minimal euclidian distance between any of their atoms. This is achieved by assigning a distance of 0 to each atom of the central residue in the Dijsktra algorithm.

By construction, our surface patches are edge connected. This means that any atom of the patch is reachable from any other atom of the patch through a list of atomic intersections over the surface. In the following evaluation, this property has been chosen for the study of *surface patch contiguity*.

In the commonly used definition of surface patches of Jones and Thornton in 1997,[26] surface residues are characterized by their C_α and a solvent vector pointing toward the solvent. To select only surface residues that are on the same side of the surface, surface residues are added to the patch if the angle between their solvent vectors and the solvent vector of the central residue is less than 110°.

Every surface patch of 20 residues[5,7] was generated over the surface of every protein of our dataset of bound forms with both approaches. These surface patches are mapped into sub-graphs, taking as reference our graph of surface atoms. Then, several measures are analyzed:

1. For each protein chain, the maximum overlap (in terms of residues) between the known binding site and any of the surface patches.
2. The number of contiguous subregions that compose a surface patch, that is the number of connex composants of the surface patch graph. By construction, our surface patches always define a unique region.

To further understand the differences observed between these approaches, we evaluated the number of surface atoms and residues, that are contiguous to a central surface residue in our method and that are not present in the corresponding patch obtained by the former approach.

For the prediction of protein-binding sites, where the interacting potential of a residue is determined by the analysis of properties in the surrounding region, it is necessary to generate automatically the surface patch that will best overlap with a binding site while having the lowest number of non-IR. This problem consists in finding the patch size N that will optimize the *signal-to-noise* ratio Q [Eq. (1)].

$$Q_{(\mathrm{P},N)} = \frac{\mathrm{O}_{(\mathrm{P},N)}}{\mathrm{NIR}_{(\mathrm{P},N)}} \tag{1}$$

$\mathrm{O}_{(\mathrm{P},N)}$ is the best observed percentage of overlap between a surface patch of size N (expressed in terms of number of atoms, number of residues or distance) and the known binding site of the protein P. $\mathrm{NIR}_{(\mathrm{P},N)}$ is the percentage of non-IR inside that patch.

Surface curvature

The *relative exposure* Ω of a surface atom *a* is defined as the fraction of a tiny sphere centered on *a*, that lies outside the alpha shape. In the present two-dimensional (2D) example [Fig. 3(A)], this value corresponds to the sum of normalized angles ω_1, ω_2, ω_3 of « empty » Delaunay triangles at atom *a*. In 3D this generalizes to a sum of tetrahedra solid angles [Fig. 3(B)]. To better differentiate the values corresponding to clefts and knobs, Ω was normalized to define values ranging from −1 (cleft) to 1 (knob), with 0 defining a flat region. The *relative exposure of a residue* is defined as the mean of its atom exposures and as such, follows the same rules of normalization.

In some "degenerate" cases, the surface of an atom might be scattered in disconnected atomic components as illustrated in the 2D example [Fig. 3(C)]. The initial split of vertices in the alpha shape (presented in "Material and Methods" section) allows us to maintain a distinct value for each atomic component. A per atom value is obtained by summing up all component of the atom with the exception of cavities [Fig. 3(C), b2].

To define the *local surface curvature* for an atom, relative exposures are smoothed on a surrounding concentric region. Starting from a central atom, a *smoothing region* is determined by considering all surface atoms

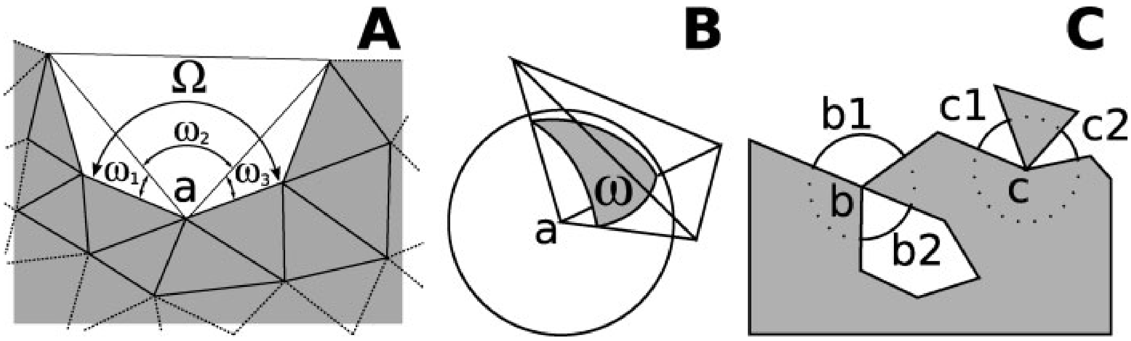

Figure 3

Relative exposure. (**A**) In 2D the relative exposure Ω of vertex a corresponds to an "empty angle" around this vertex. It is computed as a sum $\omega_1 + \omega_2 + \omega_3$ of solid angles. (**B**) The solid angle ω at vertex a of a tetrahedron is the portion of surface area lying inside the tetrahedron of a tiny sphere centred on a. (**C**) A 2D example where a vertex may have more than one component. The atom b has two components, the component b1 is on the surface of the protein and the component b2 is in a cavity. Atom c has two components, both on the surface of the protein.

accessible through a maximum of *s* edges, where *s* is a size parameter.

To emphasize local features, more importance is given to atoms near the patch center than to remote ones:

$$C(a) = \sum_{i \in \text{surfacepatch}} \frac{\Omega(i)}{d(a,i)} \qquad (2)$$

where C(*a*) is the local surface curvature of atom *a*, *d*(*a*,*i*) is the distance over the surface (in the graph) between atom *a* and atom *i*, and Ω(*i*) is the relative exposure of atom *i*. The local surface curvature of a residue is computed as the mean of its atom values.

To validate our definition of local surface curvature, we compared our values with those computed by the CX program,[27] an approach similar to common solid angle approaches.[8–11] To assess the curvature, CX approximates the amount of space filled by atoms within a sphere of 10 Å. We further used our definition of local surface curvature to define clefts and knobs over the surface and characterized them in terms of accessibility and composition.

RESULTS AND DISCUSSION

Defining the protein surface

Surface residues

Following the definition of sensitivity and specificity proposed in "Material and Methods" section, the best definition of a surface residue is found to be an accessible residue that either possesses five surface atoms or has a valence Vr higher than 10. With these parameters, 98% of the IR were detected as being part of the surface, whereas 20% of the residues were assigned to the protein core (see Fig. 4).

Similar results are observed for the approach of Miller *et al.*[38] with the default parameter of 5% accessibility, leading to the detection of 96% of IR as being part of the surface, while keeping a protein core formed on average of 24% of the residues.

The best result that could be achieved with the residue depth approach[40] was obtained with a depth of 2.3 Å. Although 98% of IR are detected as part of the surface, the protein core is less well defined with only 15% of the residues. Other depth-based approaches that use water molecules as surface referents might perform better, but are far more time consuming due to the placement of these referents either by Molecular Dynamics or Monte Carlo approaches.

We confirm the good distinction between protein surface and core by analyzing several physicochemical properties known to differentiate these two structurally different regions: the hydrophobicity and flexibility are computed with the amino acid scales of Argos (1982) and Creamer (2000), and the amino acid conservation is computed as a Shannon Entropy[42] derived from a multiple alignment generated by PipeAlign.[43] These three approaches that divide residues into surface and protein core emphasize the same differences in physicochemical and evolutive properties: the protein surface is on average (1) 40% less hydrophobic, (2) 65% more flexible, and (3) 75% less conserved than the protein core (Supp. Info. Table II).

Our approach to define surface residues is comparable with the one of Miller *et al.*,[38] which is widely used to differentiate protein surface and core. This suggests the importance of a novel parameter introduced in this study, the valence Vr, which represents the density of interactions at the surface. Interestingly, IR are detected as being part of the surface equivalently in both bound and unbound forms (data not shown). This remark further supports the possibility to predict protein-binding

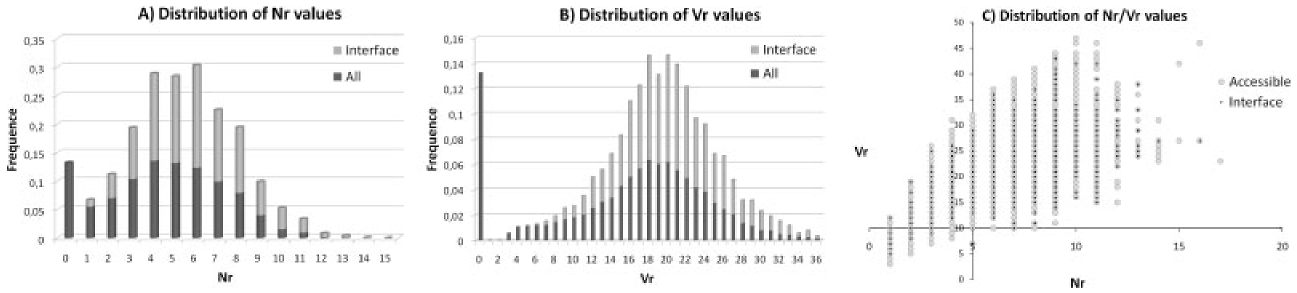

Figure 4

Selection of parameters to define surface residues. The distribution of Nr values for both Interface and All residues is shown. To detect most of the interface as being part of the surface, it seems reasonable to select an Nr superior or equals to 2 if the parameter is used alone. For Vr distribution, a threshold of 10 or 11 can be chosen to detect by itself most of the interface. Finally, the plot of Nr/Vr leads to select Nr = 5 and Vr = 10 to detect more than 98% of the interface as being part of the protein surface.

sites using unbound forms where only small or moderate conformational changes occurred during the molecular assembly formation.

Surface patches

All surface patches were generated on the dataset of 85 bound forms, with a patch size similar to previous studies (N = 20 residues).[4–7] Binding sites in our dataset are composed of 27 amino acids on average. For each protein chain, our best overlapping surface patch contains on average 15–16 IR, corresponding to an average overlap of 62.3% with known binding sites, while retaining only four to five (22%) non-IR. The signal-to-noise-ratio Q was thus improved by 63.6%, ranging from 2.2 for the Jones and Thornton approach, to 3.6 for our method (Table I).

Generating residue surface patches of 20 residues on each surface atom rather than on each surface residue further increases the signal-to-noise ratio to 4.9. For such a definition, the best overlap with a known binding site is 65.9% on average, while retaining only 17.4% of non-IR, thus improving the signal quality by more than two times compared with Jones and Thornton.

Finally, to obtain the best signal-to-noise ratio with our approach, an optimization was performed by varying the size of the patch. Experiments revealed a peak for 15 residues, with respectively Q = 5.8 for bound forms and Q = 5.1 for unbound forms. With this patch size, the average best overlap of a surface patch with a binding site is respectively 55.4% in bound forms and 55.8% in unbound forms, while the percentage of non-IR inside the patch represents no more than 9.3% in bound forms and 11.8% in unbound forms. Compared with the approach of Jones and Thornton, the use of residue surface patches generated on a per atom basis combined with a patch size of N = 15 residues thus improved the signal-to-noise ratio by a factor of 2.6.

Although the per atom approach generates about 10 times more patches than the per residue approach, it is fast enough to be applied in large-scale projects thanks to the combination of two fast computational tools : alpha shapes and Dijkstra graph travel.

As a further refinement, our surface patches can be extended to define core and rim patches. Further experiments will be conducted to explore a potential correlation between these definitions and the notions of interface core and interface rim.[2,3]

Characterizing the protein surface

Our definition of relative exposure shares similitudes with the ASA. This statement was verified on our dataset

Table I

Comparisons of the Best Overlapping Patches with Binding Sites

	Alpha-shape[a]	Alpha-shape[b]	Jones and Thornton
Maximum overlap ($O_{(P,M)}$)	62.3%	65.9%	56%
Fraction of non interacting residues ($NIR_{(P,M)}$)	22%	17.4%	25.5%
Signal-to-noise ($Q_{(P,M)}$)	3.6	4.9	2.2
Number of missed contiguous atoms$^+$	—	—	18
Number of missed contiguous residues$^+$	—	—	3.9
Number of subregions selected per patch$^+$	—	—	1.5

Surface patches are generated using a size parameter of 20 residues on the dataset of bound forms. Values of missed atoms and residues as well as the number of subregions are not applicable to our approach.

Values for parameters noted with a ($^+$) have not been computed for a and b as alpha shape is taken as reference.

[a] Residue surface patches generated for each surface residues.

[b] Residue surface patches generated for all surface atoms.

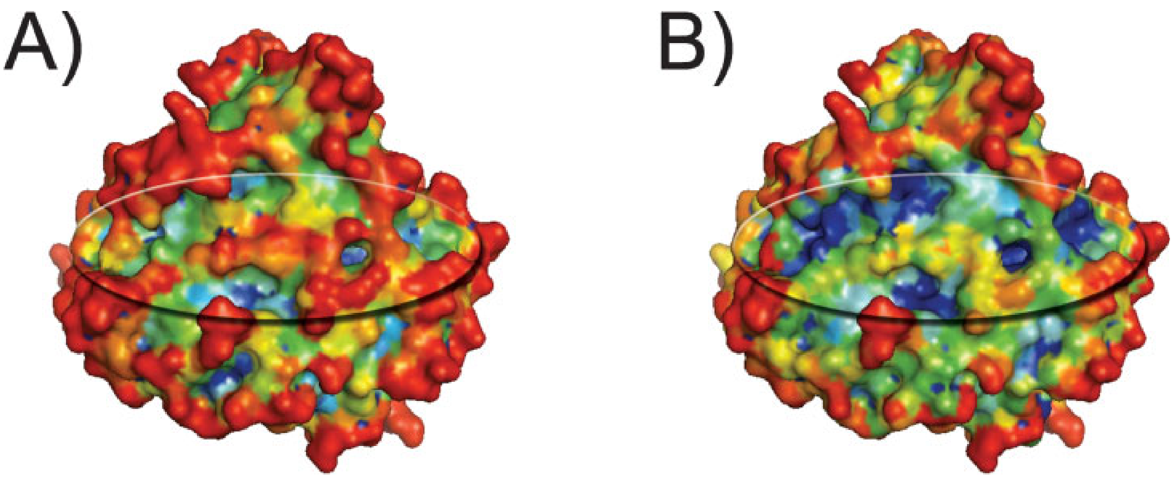

Figure 5

Protein structure of Cytochrome P450 (1eup:A) visualized with PyMol.[44] CX values are represented in (**A**), our local surface curvature index in (**B**). To allow comparisons between (A) and (B), values have been normalized using 90% of all atomic values and using a threshold of 1 to differentiate between knobs and clefts for CX values and 0 for our solid angle approach. Blue indicates clefts (values near −1), green indicates flat surfaces (values near 0) and red indicates knobs (values near 1). The cavity of 855 Å^3 (SA model) below the ellipsoid region increases the global amount of empty space and bias the curvature of CX toward the detection of flat and knob regions.

by a strong Pearson product-moment correlation coefficient between these two notions (respectively 0.86 for atoms and 0.89 for residues).

Like the ASA values, the relative exposure values are subject to great variations in the neighborhood of an atom. This property is explicitly depicted by the lacunary nature of the alpha shape surface. To reflect a local trend of the surface around an atom, we introduced the notion of local surface curvature by smoothing the relative exposures in the atom neighborhood. The size of the smoothing region is a critical parameter used to define the level of details to be observed on the surface. When this smoothing region is small, the accent is placed on local details, whereas when it is larger global features are emphasized. A visual inspection was performed and a reasonable balance between local and global surface features was achieved for a size of smoothing region $s = 2$ [Fig. 5(B)].

Several approaches have already been proposed to address the determination of protein surface curvature. Because of the intuitiveness of this notion, no quantitative evaluation exists to differentiate poor and good approaches. Nevertheless, we compared our results with CX,[27] an existing method similar to common solid angle approaches.[8–11] In both cases, curvature values were computed for all surface atoms and only a moderate correlation (0.64) was observed. For a more detailed understanding of this low correlation, we proceeded to a quantitative comparison to verify if both approaches were able to detect the same protruding regions. Considering the 10% of atoms with higher values, an overlap of 66% is observed between both approaches. The main differences were observed for regions detected either as flat (near 0) or cleft (near −1) by our local surface curvature score. This low overlap can be explained by the difference in methodology behind the two approaches. CX being based on local atomic densities can be biased by the presence of subsurface cavities, a problem common to most solid angle approaches.[8–11] In the extreme case, when influenced by local variations of densities or by the presence of cavities, even clefts can be detected as protruding [Fig. 5(A)]. In contrast, our method allows to distinguish between the empty space above and below the surface and only considers the empty space above the surface to reflect the real local curvature.

To study the protein surface topography, we defined *knob residues* as surface residues with a local surface curvature greater than 0, and *cleft residues* as those with a local surface curvature smaller than 0. Following this definition the surface is composed on average of 62% of clefts and 38% of knobs. Furthermore, knob IR contributes to 6.7% of the surface of the protein whereas cleft IR contribute to 8.6% of this surface. Therefore, the bayesian probability of having an interacting residue in a knob is 0.174 and 0.139 in a cleft. IR are thus 30% more frequent in knobs than in clefts, a result that correlates with the fact that IR are known to be, on average, relatively accessible (Table II).

To further understand the differences between cleft and knob regions, we proceeded to a detailed analysis of

Table II
Amino Acid Accessibility and Composition of Protein Surfaces and Interfaces

	Accessibility[a]								Area[b]						Propensities[c]			
	Surface			Interface			Interface/ surface[d]		Surface			Interface			Ln (Knob/cleft) (1)		Interface/surface (2)	
Residue	All	Knob	Cleft	All	Knob	Cleft	Knob	Cleft	All	Knob	Cleft	All	Knob	Cleft	Surface	Interface	Ln (Knob/knob)	Ln (Cleft/cleft)
PHE	41	67	29	76	108	40	61	38	1.8	1.2	3.4	4.9	4.8	4.9	−1.01	−0.00	1.36	0.36
TYR	58	94	38	80	113	49	20	29	3.3	2.5	5.2	7.4	7.0	8.3	−0.73	−0.17	1.03	0.47
TRP	51	79	37	71	93	59	18	59	1.1	0.8	2.0	2.1	1.3	4.3	−0.95	−1.22	0.47	0.74
ALA	44	56	28	47	58	30	4	7	4.6	4.7	4.6	3.9	3.9	3.8	0.02	0.04	−0.18	−0.20
VAL	45	66	30	52	73	35	11	17	3.5	2.8	5.3	4.0	3.7	4.9	−0.65	−0.30	0.27	−0.08
LEU	47	77	33	66	88	43	14	30	4.8	3.5	8.4	6.5	5.9	7.9	−0.89	−0.30	0.53	−0.06
ILE	45	74	33	63	97	41	31	24	2.4	1.6	4.5	5.3	4.1	8.3	−1.03	−0.71	0.92	0.61
MET	59	93	33	85	110	51	18	55	1.4	1.3	1.7	2.8	2.9	2.4	−0.23	0.19	0.79	0.37
ASP	66	80	42	70	87	39	9	−7	7.7	8.1	6.5	4.8	5.2	3.9	0.21	0.29	−0.44	−0.52
GLU	83	99	49	84	107	50	8	2	10.8	12.1	7.2	7.1	7.3	6.7	0.52	0.09	−0.51	−0.08
LYS	99	116	62	110	125	65	8	5	12.9	14.3	9.2	7.7	8.9	4.8	0.45	0.62	−0.48	−0.66
ARG	94	120	58	114	143	69	19	19	8.6	8.8	7.9	10.9	11.6	9.1	0.11	0.25	0.27	0.13
SER	53	66	31	58	74	31	12	0	6.1	6.6	4.8	5.8	6.1	5.2	0.30	0.17	−0.07	0.07
THR	54	69	36	63	79	40	14	11	5.5	5.3	6.3	5.4	5.0	6.1	−0.17	−0.19	−0.05	−0.03
ASN	68	86	43	72	90	45	5	5	6.5	6.6	6.2	5.0	5.0	5.1	0.07	−0.03	−0.29	−0.19
GLN	79	97	50	82	108	50	11	0	6.1	6.3	5.4	4.1	3.9	4.7	0.16	−0.19	−0.50	−0.15
CYS	25	40	19	30	38	26	−5	37	0.5	0.3	1.0	0.7	0.4	1.4	−1.20	−1.39	0.17	0.37
HIS	65	88	36	72	103	34	17	−6	2.5	2.6	2.3	3.2	3.5	2.4	0.15	0.39	0.29	0.04
PRO	63	80	34	74	87	36	9	6	5.8	6.3	4.3	4.3	5.3	1.9	0.40	1.01	−0.17	−0.79
GLY	35	45	21	42	51	28	13	33	4.1	4.2	3.9	4.3	4.3	4.2	0.08	0.03	0.02	0.07
AVG	59A^2	80A^2	37A^2	71A^2	92A^2	43A^2	15%	18%										

Accessibility and composition have been computed on the bound dataset of 85 protein structures. Surface residues are those detected by our method and show a 0.97 correlation with previously published amino acid scale. As for interface residues, our amino acid scale has a 0.9 correlation with the amino acid scale of Chakrabarti.[3]

[a] Accessibility of amino acids for either the surface or the interface and decomposed following knob and cleft regions.

[b] Amino acid contributions to the surface area or the interface area.

[c] Propensity for a residue (1) to be part of a knob rather than a cleft or (2) to be part of an interacting knob/cleft rather than a noninteracting knob/cleft.

[d] Indicates the raise of residue accessibility (in percentage) between interacting and noninteracting residues.

amino acid accessibility and composition for both non-IR and IR. First, our scale of amino acid contribution to the surface area (Table IIb: Surface: All) shows a strong correlation of 0.97 with a previously published amino acid scale[3]; a strong correlation of 0.9 for the composition of IR was also observed. Then this amino acid contribution to the surface area was decomposed into knob surface area and cleft surface area. Phe, Tyr, and Trp constitute the aromatic cluster, Ala, Val, Leu, Ile, Met constitute the hydrophobic cluster, Asp and Glu the anionic cluster and Lys, Arg the cationic cluster. Whereas knob surfaces are composed of only 4.5% aromatics and 13.9% hydrophobic residues, cleft surfaces are composed on average of twice as many aromatics (10.7%) and hydrophobic residues (24.5%). For interacting surfaces, aromatics represent 13.2% of knob regions (compared with 4.5% for noninteracting surfaces) and 17.5% of cleft regions (compared with 10.6%). These interacting surfaces show also more hydrophobic residues than noninteracting surfaces, in particular for knobs (20.5% of the surface area versus 13.9% for noninteracting surfaces) and somewhat less for clefts that are already very hydrophobic (27.3% against 24.5%). Furthermore, cationic and anionic residues are shown to be more present in knob than in cleft regions for both IR and non-IR, although knob IR are less anionic (12.5%) than knob non-IR (20.2%). Cleft non-IR are also shown to be more charged (30.8% for both anionic and cationic residues) than cleft IR (24.3%).

To conclude, knob regions are shown to be more charged, less hydrophobic and less aromatic than cleft regions. Furthermore, greater differences of surface contribution are seen between knob IR and knob non-IR than for cleft IR and cleft non-IR. Knob IR resemble cleft non-IR and are shown to be less charged, more hydrophobic and more aromatic than knob non-IR. It is, therefore, easier to distinguish knob IR from knob non-IR than to distinguish cleft IR from cleft non-IR.

To sum up these conclusions, euclidian metrics were computed (see Fig. 6), as proposed by Chakrabarti and Janin,[3] by computing the distance Δf between two compositions f_i and f'_i:

$$(\Delta f)^2 = 1/19 \sum_{i=1}^{20} (f_i - f'_i)^2$$

Finally, this finer description of the surface should help to improve the prediction of protein-binding sites by allowing to compare separately knob and cleft regions.

Nucleic acids and other biomolecules

With the exception of the definition of surface residues, our approaches can be directly used to analyze and characterize the surface of other biomolecules such as DNA, RNA, or lipids. As defined for protein structures, the surface atoms of these biomolecules still correspond to the accessible atoms extracted from the alpha shape. By construction, our local surface curvature depends only on a continuous surface and is influenced neither by variations of density nor by the presence of cavities. Therefore, this surface descriptor can also be applied to these biomolecules [Fig. 7(A)]. Finally, our approach also allows for the generation of surface patches along nucleic acid surfaces [Fig. 7(B)].

By unifying and facilitating the analysis and comparison of molecular surfaces, we hope that this geometrical approach will benefit the emergent structural studies both on current and newly characterized biomolecules.

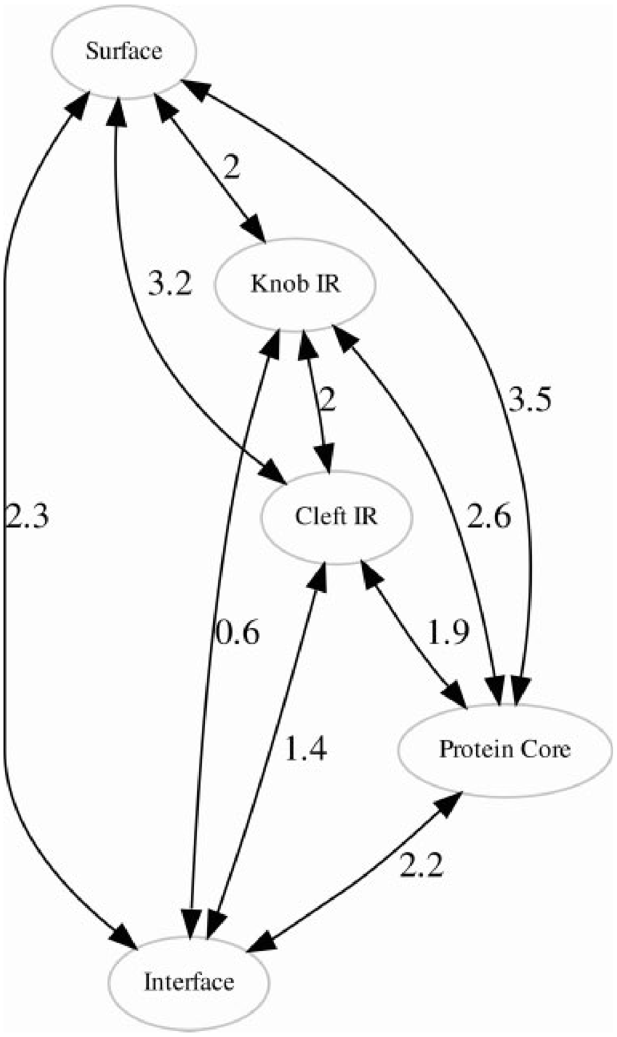

Figure 6

Distances between amino acid compositions. Distances are those defined in the text as Δf and are expressed as percentages. The area-based composition of each region (except for the protein core) is listed in Table II. Area-based composition for the protein core is taken from Lo Conte *et al.*[2] The noninteracting surface is more distant from Cleft IR than Knob IR, and Cleft IR are more similar to the protein core than the noninteracting surface.

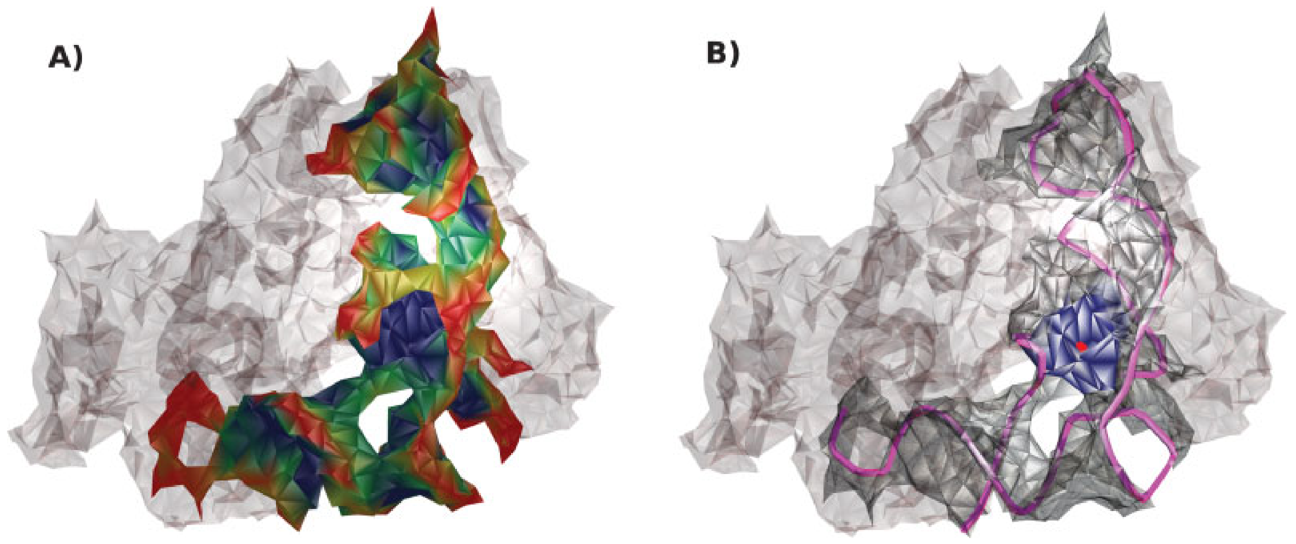

Figure 7

Structure of the Arginyl-tRNA synthetase complexed with the tRNA(Arg) (1f7u). The Arginyl-tRNA synthetase is transparent in the background to precise the orientation of the tRNA. (**A**) tRNA mapped with the local surface curvature following the same color code as in Figure 5. (**B**) The backbone of the tRNA is in mauve, the alpha shape of the tRNA in transparent gray and the edges connecting surface atoms are represented with black lines. A surface patch around O5052 (red point) of U908 is delineated and pictured blue.

CONCLUSIONS

In this study, alpha shapes have been used to model and study properties of protein surfaces that are relevant to the description of the surface and the analysis of molecular interactions. Using this framework, we were able to define surface atoms and residues, as well as to generate contiguous surface patches. Using the field of protein-binding site prediction to evaluate the relevance of our definitions, we achieved a significant improvement in the determination of surface patches, where the signal-to-noise ratio in the definition of the interacting potential of a residue is increased by 2.6 times with respect to a previous approach.

The alpha shape framework was further used to define a conception of surface curvature that is biased neither by the variation of atomic density nor by the presence of cavities below the surface. In the characterization of the molecular surface topography, this conception revealed a landscape composed on average of 38% knobs (the remaining being clefts), where IR are 30% more frequent in knobs than in clefts. This distinction is important for IR as demonstrated by the differences in accessibility and composition between these two regions. These results remain true when considering unbound forms, where only small conformational changes occurred during the assembly formation.

The robust geometric framework of alpha shapes has allowed us to unify the computation of several properties relevant to the analysis and comparison of any molecular surfaces, with a proven improvement compared with former approaches. Our algorithms are fast enough to be used in large-scale projects such as interactomics, and will be applied in the future for the analysis of protein and nucleic acid interactions.

ACKNOWLEDGMENTS

The authors thank Mme Dominique Bechmann (IGG-LSIIT) for cosupervision of Benjamin Schwarz as well as Professor Joël Janin (IBBMC) for his invaluable comments and encouragement.

REFERENCES

1. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
2. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites J Mol Biol 1999;285:2177–2198.
3. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47:334–343.
4. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
5. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338:181–199.
6. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 2006;34:3698–3707.
7. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007;66:630–645.
8. Connolly ML. Measurement of protein surface shape by solid angles. J Mol Graph 1986;4:3–6.
9. Cazals F, Chazal F, Lewiner T. Molecular shape analysis based upon the morse-smale complex and the connolly function. In: Proceed-

ings of the 19th Annual Symposium on Computational Geometry (SCG) 2003. pp 351–360.

10. Norel R, Wolfson HJ, Nussinov R. Small molecule recognition: solid angles surface representation and molecular shape complementarity. Comb Chem High Throughput Screen 1999;2:177–191.
11. Norel R, Lin SL, Wolfson HJ, Nussinov R. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 1995;252:263–273.
12. Edelsbrunner H, Mucke EP. Three-dimensional alpha shapes. ACM Trans Graph 1994;13:43–72.
13. Liang J, Edelsbrunner H, Fu P, Sudharkar PV, Subramaniam S. Analytic shape computation of macromolecules I: molecular area and volume through alpha shape. Proteins 1998;33:1–17.
14. Edelsbrunner H, Koehl P. The weighted-volume derivative of a space-filling diagram. Proc Natl Acad Sci USA 2003;100:2203–2208.
15. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
16. Liang J, Edelsbrunner H, Fu P, Sudharkar PV, Subramaniam S. Analytic shape computation of macromolecules II: inaccessible cavities in proteins. Proteins 1998;33:18–29.
17. Edelsbrunner H, Facello MA, Liang J. On the definition and the construction of pockets in macromolecules. Discrete Appl Math 1998;88:83–102.
18. Edelsbrunner H. Deformable smooth surface design. Discrete Comput Geom 1999;21:87–115.
19. Bajaj CL, Lee HY, Merkert R, Pascucci V. NURBS based B-rep models for macromolecules and their properties. In: Proceedings of symposium on Solid Modeling and Applications. 1997. pp 217–228.
20. Zomorodian A, Guibas L, Koehl P. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. Comput Aided Geomet Des 2006;23:531–544.
21. Li X, Hu C, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins 2003;53:792–805.
22. Cazals F, Proust F, Bahadur RP, Janin J. Revisiting the Voronoi description of protein-protein interfaces, Protein Sci 2006;15:2082–2092.
23. Ban YA, Edelsbrunner H, Rudolph J. Interface surfaces for protein-protein complexes. J ACM 2006;53:361–378.
24. Pintar A, Carugo O, Pongor S. CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics 2002;18:980–984.
25. Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Biol 1997;12:973–974.
26. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
27. CGAL, Computational Geometry Algorithms Library, http://www.cgal.org.
28. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
29. Connolly ML. Analytical molecular surface calculation. J Appl Crystallogr 1983;16:548–558.
30. Delaunay B. Sur la sphère vide. Izvestia Akademii Nauk SSSR Otdelenie Matematicheskii i Estestvennyka Nauk 1934;7:793–800.
31. Edelsbrunner H. The union of balls and its dual shape. Discrete Comput Geom 1995;13:415–440.
32. Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol 2004;14:233–241.
33. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
35. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–797.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
37. Hubbard SJ, Thornton JM. NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College London 1992.
38. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
39. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 1999;7:723–732.
40. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J 2003;84:2553–2561.
41. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math 1959;1:269–271.
42. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci USA 2005;102:15447–15452.
43. Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson JD, Wicker N, Poch O. PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res 2003;31:3829–3832.
44. DeLano WL. The PyMOL Molecular Graphics System. San Carlos, CA, USA: DeLano Scientific; 2002.