REVISITING THE VORONOI DESCRIPTION OF PROTEIN-PROTEIN INTERFACES

Frédéric Cazals[1], Flavien Proust[1], Ranjit P. Bahadur[2], Joël Janin[2]

1. INRIA Sophia-Antipolis, Project Geometrica, BP 93, F-06902 Sophia-Antipolis, France
2. IBBMC UMR 8619 CNRS Université Paris-Sud Bât. 430, F-91405 Orsay, France

Corresponding author:
J. Janin
IBBMC Université Paris-Sud Bât. 430, F-91405 Orsay, France
Joel.Janin@ibbmc.u-psud.fr
tel +33 1 69 15 7966

Running title: Voronoi description of protein-protein interfaces
Manuscipt: 21 pages (texte), 4 tables, 6 figures

Keywords

protein-protein interaction / algorithmic geometry / alpha-complex / interface connectivity

Abstract

We develop a model of macromolecular interfaces based on the Voronoi diagram and the related alpha-complex, and we test its properties on a set of 96 protein-protein complexes taken from the Protein Data Bank. The Voronoi model provides a natural definition of the interfaces, and it yields values of the number of interface atoms and of the interface area that have excellent correlation coefficients with those of the classical model based on solvent accessibility. Nevertheless, some atoms that do not lose solvent accessibility are part of the interface defined by the Voronoi model. The Voronoi model provides robust definitions of the curvature and of the connectivity of the interfaces, and leads to estimates of these features that generally agree with other approaches. Our implementation of the model allows an analysis of protein-water contacts that highlights the role of structural water molecules at protein-protein interfaces.

Abbreviations:

ASA, accessible surface area; BSA, buried surface area; VIA, Voronoi interface area; cc, connected component; scc, significant connected component.

Proteins make non-covalent interactions that are essential elements of their biological function. The study of such interactions relies in part on modeling the geometry and physical chemistry of the interfaces built by interacting proteins. When atomic coordinates are available, the Voronoi description of proteins is a useful geometric tool that has been applied in a variety of settings. The pioneering work of Richards (1974) used the Euclidean Voronoi diagram to analyze the atomic packing inside macromolecules, followed by many other authors (Harpaz et al. 1994; Gerstein et al. 1995; Pontius et al. 1996; Nadassy et al. 2001; Tsai & Gerstein 2002; McConkey et al. 2002). The Voronoi diagram associates to each atom its Voronoi cell, a convex polyhedron that contains all points of space closer to that atom than to any other atom. More recently, it has been used to define contacts in macromolecules without applying a distance cut-off: two atoms are in contact if and only if their Voronoi cells have a facet in common. Similarly, Voronoi cells can be drawn around amino acid residues to define residue-residue contacts (Singh et al. 1996; Munson & Singh 1997; Soyer et al. 2000; Dupuis et al. 2005). Given this definition of a contact, the set of facets shared by atoms of two macromolecules forming a complex represents their interface. There is however a major difficulty: atoms on the molecular surface have unbounded, or at least poorly defined, Voronoi cells. This may be circumvented by surrounding the protein with solvent molecules (Soyer et al. 2000), but their position must be fixed, which is not physically meaningful. An alternative is to use the alpha-complex, an extension of the Voronoi diagram proposed by Edelsbrunner & Mucke (1994). Applications of the alpha-complex to macromolecules, reviewed by Poupon (2004), include the computation of molecular surfaces (Akkiraju & Edelsbrunner 1996), and that of interfaces in an implementation where the unbounded facets that extend out of the molecular surface are removed by an iterative process called retraction (Ban et al. 2004). Cazals & Proust (2006) recently offered a simpler, and possibly more natural, way to define the interface between molecules by removing facets based on purely geometric criteria. Here, we apply their procedure to a set 96 protein-protein complexes taken from the Protein Data Bank (PDB, Berman et al. 2002), and compare the results to those of the classical approach where interfaces are defined by changes in solvent accessibility (Chothia & Janin 1975; Janin & Chothia 1990; Jones & Thornton 1995, 1996; Lo Conte et al. 1999; Chakrabarti & Janin 2002).

METHODS AND RESULTS

A Voronoi model of macromolecular interfaces based on the alpha-complex

The Euclidean Voronoi diagram assumes that all atoms have the same radii, and its application to molecules must make approximations to fit atoms of different sizes (Richards 1974; Harpaz et al. 1994, Pontius et al. 1996). A closely related geometric construction provides a mathematically correct way to accommodate different radii: the *power diagram*, where the Euclidean distance is replaced by the power distance with respect to a sphere (Gellatly & Finney 1982; Aurenhammer 1987). The power of a point **x** relative to a sphere of radius r centered at point **a** is:

$$p(\mathbf{x}) = |\mathbf{a}\text{-}\mathbf{x}|^2 - r^2 \qquad (1)$$

Points **x** that have the same power relative to two spheres belong to the radical plane of the spheres, which contains their intersection if it exists. Given two atoms A_1 and A_2 represented by two balls (hard spheres) of radii r_1 and r_2 centered at $\mathbf{a}_1$ and $\mathbf{a}_2$, we may also define the power of A_1 relative to A_2, or A_2 relative to A_1:

$$p(A_1, A_2) = |\mathbf{a}_1\text{-}\mathbf{a}_2|^2 - r_1^2 - r_2^2 \qquad (2)$$

When $p(A_1, A_2) = 0$, the two balls are orthogonal (they intersect at a right angle).

The power diagram reduces to the Voronoi diagram when all the balls have the same radius. Thus, we shall call it also a Voronoi diagram and associate it with its dual, the *Delaunay triangulation*. This is built by drawing edges spanning pairs of atoms that have a Voronoi facet in common, triangles spanning triplets that have a common Voronoi edge, and tetrahedra spanning quartets that have a common Voronoi vertex. In Fig. 1A, the Delaunay triangulation includes four vertices placed at the centers of the four atoms, six edges linking these atoms, and three triangles. The Voronoi facets shared by the four atoms are drawn (in two dimensions) as lines orthogonal to the Delaunay edges; three of them extend outside the molecular surface and are unbounded.

Noting that the power diagram is invariant if the same quantity α is added to all the square radii, Edelsbrunner & Mucke (1994) introduced the *alpha-complex*. For a given α, the alpha-complex is built as the Delaunay triangulation, except that one restricts each Voronoi cell to its associated ball and seeks intersections between these restricted regions. Thus, a Delaunay edge between two atoms is drawn if and only if the common facet lies inside the associated balls. This condition, which we shall call *condition alpha*, is satisfied in Fig. 1A by the facets drawn in full line between atoms A_1 and A_2 or A_3. These facets are inside the balls representing the atoms, and the Delaunay edges spanning these atoms are part of the alpha-complex. In contrast, the facet between A_1 and A_4, drawn in dashes is entirely outside the balls, and the a_1a_4 edge is not part of the alpha-complex for this particular value of α. If we increase α, all the square radii increase and more facets satisfy condition alpha, until the alpha-complex reduces to the standard Delaunay triangulation at large values of α.

In our implementation, the ball radii are atomic or group radii augmented of the water probe radius, and α=0. Under these conditions, the surface of the union of the balls, represented in two dimensions by arcs of circle drawn in full lines in Fig. 1A, is the solvent accessible surface as defined by Lee & Richards (1971). In a complex between two molecules, we color their atoms in red and blue respectively, and represent the interface by the set of bicolor Voronoi facets associated with the Delaunay edges linking atoms of different colors in the alpha-complex. In Fig.1A, the interface comprises the two facets orthogonal to the a_1a_2 and a_1a_3 edges, but not the facet between A_1 or A_4 due to condition alpha. These facets are drawn in green, whereas those in blue are internal to the blue molecule.

With large molecules such as proteins, condition alpha imposes a stringent selection that removes from the interface nearly all of the facets that stick out of the molecular surface. Nevertheless, some unbounded or excessively large facets remain, such as the one between A_1 and A_2 in Fig. 1A. These are discarded based on *condition beta:*

$$m/r > M \qquad (3)$$

where r is the radius of the smaller of the two balls, and m the radius of the largest ball orthogonal to the balls representing the two atoms. M is a threshold value that we set to M=5 after checking that the number of discarded facets is very small (0.16%) and that similar results are obtained for M in the range 2.4 to 7. Condition beta is illustrated by Fig. 1B; there,

r is the radius of A_1, m, that of the ball drawn in dashes. This ball is centered at the Voronoi vertex x defined by A_1, A_3 and A_4, it is orthogonal to the three balls representing these atoms, and it is the largest ball orthogonal to A_1 and A_4. Condition alpha rejects the facet between A_1 and A_3, condition beta accepts or rejects the facet between A_1 and A_4 depending on the value of M.

Computing the Delaunay triangulation of a collection of balls, and subsequently the alpha-complex, is demanding in terms of efficiency and numerical issues. Our implementation is based upon the Alpha_shape_3 package of the CGAL library *www.cgal.org*, and it is accessible at *http://bombyx.inria.fr/Intervor/intervor.html*.

The sample of protein-protein interfaces

The sample used in calculations here comprises protein-protein interfaces in 96 entries of the Protein Data Bank listed in Table 1. The calculation deals either with the proteins alone (AB model), or with the proteins and the structural water reported into the entry (ABW model). In the latter case, the sample is restricted to 30 entries reporting crystal structures at 2 Å resolution or better (2 Å set), as the water structure is likely to be less reliable in lower resolution studies. We call AW-BW the protein-water interface of the ABW model. The sample is split in five classes: PI complexes between proteases and protein inhibitors, ESI complexes between enzymes other than proteases and protein substrates or inhibitors, AA antigen-antibody complexes, ST complexes involved in signal transduction or the cell cycle, MI miscellaneous complexes. All are non-obligate or transient assemblies in the sense of Noreen & Thornton (2003): A and B are proteins that fold separately and remain independent entities until they associate.

All protein atoms are tagged as A or B and included in the calculations. Water molecules, ignored in the AB model, appear in the ABW model if their crystallographic temperature factor is less than 80 $Å^2$, and are considered as part of the interface if they make at least one contact with atoms of both A and B. Other non-protein atoms (HETATM in PDB entries) are included and tagged as U (unknown). Group radii are taken from Chothia (1976): C atoms 1.87 Å aliphatic, 1.76 Å trigonal; N atoms 1.65 Å neutral, 1.50 Å charged; O atoms 1.4 Å; S atoms 1.85 Å; all other atoms 2.0 Å. These radii are augmented of the probe radius (1.4 Å) for the Voronoi construction.

Size of the interfaces

In the AB model, interface atoms are all atoms of protein A (resp B) that share a Delaunay edge with an atom of protein B (resp A) in the alpha-complex for α=0. The set of the bicolor facets dual of such edges constitutes the interface. Thus, the size of an interface can be evaluated in at least three ways: by counting interface atoms (N_{Vor}), counting facets (N_{facet}), or computing the Voronoi interface area VIA as the sum of the individual facet areas. In the classical approach, interfaces are sets of atoms that lose solvent accessibility when a complex forms. Then, the interface size is commonly evaluated as a buried surface area (BSA), which is the difference between the solvent accessible surface area (ASA) of the protein atoms in isolated A and B and in the complex (Chothia & Janin, 1975). The solvent accessibility model has no equivalent to N_{facet}, but the number of atoms that lose accessibility should correspond to N_{Vor}, and the buried surface area to the VIA. Data of Lo Conte et al. (1999) will be used for comparison with ours.

Counting atoms

Table 2 shows that the AB interfaces in our sample comprise an average of 239 atoms, but the range of N_{Vor} is wide (117-581) and the standard deviation large. AA interfaces, which are the most regular in size, have an average of 208 atoms with a small standard deviation. All but one of the 28 AA interfaces have N_{Vor} in the range 160-260. Antigen-antibody interfaces are described as 'standard-size' in Lo Conte et al. (1999). The range N_{Vor}=160-260, which corresponds to that standard size. also comprises a great majority (22 out of 29) of the PI interfaces, and 70% of the 96 interfaces in the set. ST interfaces tend to be larger and more heterogeneous in size than in the other classes.

Fig. 2A shows that N_{Vor} is linearly correlated to the number N_{at} of atoms that lose ASA. For the 72 complexes common to our sample and that of Lo Conte et al. (1999), the correlation is excellent (R^2=0.992), but N_{Vor} exceeds N_{at} by about 13%. This excess is present in similar proportion in all complexes. Thus, some atoms that share facets with atoms of the other protein do not lose solvent accessibility. An examination of individual interfaces indicates that two-thirds of the atoms that contribute to N_{Vor} but not N_{at} have zero or nearly zero ASA in the isolated A or B components. Most belong to the protein main chain and are largely buried by their covalent environment. Fig. 3 shows an example of that situation: the red ball is an atom of A that, when its neighbors in A are removed, is seen to intersect the

blue ball figuring an atom of B; when the neighbors are present, the red ball is completely screened and has no solvent accessible surface. There are also cases of solvent accessible atoms that have bicolor facets, yet do not lose ASA in the complex. In addition, 0.12% of the atoms that lose ASA are not counted in N_{Vor} because they contribute only to facets that do not pass condition beta.

N_{Vor} increases in the ABW model, due to atoms that share a facet with an interface water molecule but not with atoms of the other protein component. These atoms are part of the protein-water interface, but not the AB interface. On average in the 30 entries of the 2 Å set, N_{Vor} is 45% larger in the ABW than the AB model, the ABW interface comprising 330 protein atoms and 34 water molecules.

To test whether A and B may make equal contributions N_A and N_B to N_{Vor}, we evaluated the ratio:

$$r_{AB} = \max (N_B/N_A, N_A/N_B) \quad (4)$$

r_{AB} measures the asymmetry of the contributions. Its average value, 1.22 in our sample, is larger (1.47) in PI interfaces. As often noted, protease active sites tend to have a concave shape, the inhibitors a complementary convex shape, and the concave surface contributes more atoms to the interface than the convex one. In the extreme case of the kallikrein-pancreatic trypsin inhibitor complex (2kai), the protease contributes twice as many atoms as the inhibitor. In contrast, r_{AB} has a low value (1.11) in the AA class, compatible with the observation that antibodies raised against protein antigens tend to have flat combining sites (Mariuzza et al. 1987; McCallum et al. 1996). Table 2 indicates that ST interfaces resemble AA interfaces from this point of view.

Counting facets

AB interfaces contain N_{facet}= 423 bicolor facets on average, that is, 1.77 facet per interface atom. As each facet implicates two atoms, the average interface atom has twice as many neighbors across the interface: n_{neigh} has an average of 3.53, a small standard deviation, and similar values in the five classes of complexes (Table 2). Thus, the linear correlation between the numbers of facets N_{facet} and of interface atoms N_{Vor} is excellent (R^2=0.984). In the ABW model, the average number of facets between protein atoms is essentially the same as in the AB model, but many new facets appear between protein atoms and water: of the 769

facets reported in Table 2 for the average interface in the ABW model, 53% are with water molecules.

The facets vary widely in size. The facet area averages 3.0 $Å^2$, but the median is only 1.65 $Å^2$ and small facets with an area below 1 $Å^2$ form 38% of the sample. Condition beta removes excessively large facets, yet 5% of the facets retained have areas above 10 $Å^2$ and up to 113 $Å^2$.

Interface area

The Voronoi interface area (VIA) of the 96 interfaces ranges averages 1263 $Å^2$ with a broad range (733-2960 $Å^2$) and a large standard deviation (Table 2). VIA is linearly related to N_{Vor} (R^2 =0.964), to N_{facet} (R^2 =0.926) in spite of the variability of the facet size, and also to BSA, the interface area defined by solvent accessibility. The correlation to the values of BSA reported by Lo Conte et al. (1999) is very good (R^2=0.982). Noting that two atoms that are in contact at an interface contribute twice to BSA, but only once to VIA, the Voronoi model yields interface areas that are approximately 31% larger than BSA/2 (Fig. 2B). In the ABW model, VIA increases by 30% as new protein atoms and water molecules become part of the interface.

Topology and shape

Connectivity

The Voronoi model provides a simple definition of *connected components* (cc) within an AB interface: a cc is a set of facets that have edges in common. On average, the 96 interfaces contain 1.90 cc. Some connected components were very small, and we removed those that contributed less than 7.5% of the VIA. Calling the remainder *significant connected components* (scc), we observe that the interfaces in our sample contain 1.21 scc on average. A large majority, 81 out of 96, have only one scc; 9 have two, and 6 have three. All but 2 of the 29 PI interfaces, and all but 2 of the 28 AA interfaces have a single scc. In contrast, multi-component ST interfaces are common: 7 out of 19.

We compared the scc to the patches of interface atoms defined by the geometric clustering procedure of Chakrabarti & Janin (2002) with a distance cutoff of 15 Å. Of 70 complexes analyzed by these authors, 50 have an interface that has a single patch and also a single scc. In two cases, a single patch interface is split in two scc (Table 3), but the smaller

of the two is only just above the 7.5% VIA cutoff. On the other hand, 8 interfaces that form a single scc are split by the clustering algorithm. When both procedures split the interface, they do it in very similar ways: the fraction N_{com}/N_{at} of the atoms that belong both to the same patch and the same scc is at least 0.74. The very large interface of the *E. coli* EF-Tu/Ts complex (1efu) is split into three scc and four patches (Fig. 4). The blue and green patches coincide with two of the scc, the other two forming a single large scc. In the ribonuclease-ribonuclease inhibitor complex (1dfj), the interface comprises three patches and three scc; one of the patches coincides with a scc, but the remainder of the interface is split in two different ways, so that the N_{com}/N_{at} fraction is only 0.74. In total, the two procedures yield identical results on 53 of the 70 interfaces; they disagree on the number of fragments in 11 cases including 1efu; in the remaining 6 interfaces, some of the patches do not coincide with a scc.

Water and connectivity

In the ABW model, connected components may be identified separately in the protein-protein interface AB and the protein-water interfaces AW-BW. In the 2 Å set, the average number of cc in an AB interface is 2.7, that of scc, 1.37, taking the same 7.5% VIA cut-off as above to define a scc (Table 2). Nine interfaces have two scc and one three. The AW-BW interface is much more fragmented and has an average of 6.6 cc. In a second step, we merge the connected components of the AB and the AW-BW interfaces that share a common edge. This reduces the number of cc, and the number of scc becomes 1 in all 30 interfaces of the 2 Å set. In other terms, interface water molecules connect the scc in all these interfaces.

Fig. 5 illustrates the merging process in the chymotrypsin–eglin (1acb) and the transducin G_{α}-$G_{\beta\gamma}$ (1got) complexes. The chymotrypsin-eglin interface is a standard-size PP interface. In the AB model, it forms a single scc with holes that contain water (Fig. 5A). In the ABW model, water in the larger hole splits the interface into two scc that the merging procedure fuses into one. In transducin, a ST archetype, the interface between $G_{\alpha}$and $G_{\beta\gamma}$ is larger than in 1acb and comprises two well-defined scc lined with water molecules (Fig. 5B). Some of these waters connect the scc and cause them to fuse during the merging procedure. In both examples, comparing the connectivities of the AB and ABW interfaces yields information on packing defects filled by water molecules.

Curvature

The curvature carried by a Voronoi edge ε may be defined as:

$$h(\varepsilon) = \beta(\varepsilon)\, l(\varepsilon) \qquad (5)$$

where $\beta(\varepsilon)$ is the dihedral angle between the two bicolor facets sharing that edge, and $l(\varepsilon)$ is the length of the edge (Cohen-Steiner & Morvan 2003). In Fig. 6A, the two facets are shared by the Voronoi cell of an atom of A centered in **a**, and the cells of two atoms of B centered in $\mathbf{b_1}$and $\mathbf{b_2}$. Alternatively, the facets may belong to an atom of B and two of A. By convention, β is positive in the first case, negative in the other. In the $\mathbf{b_1ab_2}$ Delaunay triangle, the $\mathbf{ab_1}$ and $\mathbf{ab_2}$ edges represent non-covalent contacts atom A makes with B_1 and B_2. The $\mathbf{b_1b_2}$ edge may be a covalent bond or a Van der Waals contact. Its length is near 1.5 Å in the first case, above 3.5 Å in the other case. Thus, the absolute value of β, equal to the $\angle b_1ab_2$ angle, is likely to be smaller when B_1 and B_2 are covalently bonded. This is observed in the distribution of $|\beta|$ (Fig. 6B), which is bimodal. The curvature is the range 12-24 degrees when B_1 and B_2 are covalently bonded, 20-80 degrees when the bond is non-covalent.

To get a global view of the shape of an interface, we may calculate a mean curvature angle by averaging $h(\varepsilon)$ over all interior edges and normalizing by the total length of the edges:

$$s_H = \left[\Sigma_{edges}\ \beta(\varepsilon)\, l(\varepsilon)\right] \ / \ \left[\Sigma_{edges}\ l(\varepsilon)\right] \qquad (6)$$

The average value of s_H is 5.2 degrees, but the range is wide: 0-17 degrees. The smaller values are for AA and ST interfaces. PI interfaces have larger mean curvatures, the largest value being for the kallikrein-pancreatic trypsin inhibitor complex (2kai, Fig. 6C) as for the asymmetry ratio r_{AB}. In that interface, most pairs of facets are concave towards the inhibitor, and the local curvatures tend to add up. In a flat AA or ST interface, the two orientations are equally frequent, and local curvatures of opposite sign cancel. Thus, the shape information derived from the mean curvature is similar to that obtained above from the r_{AB} ratio.

Chemical composition, accessibility and interactions

Chemical groups

The chemical composition of the facets that form the AB interfaces is given in Table 4: 58% of the facets involve a non polar (carbon containing) chemical groups, 30%, a neutral polar (O, N, S containing) group, and 12%, a charged group from a Asp, Glu, Lys or Arg side chain. The non-polar fraction is similar in the 96 interfaces, but charged groups are highly

variable. The three types of chemical groups contribute respectively 56, 29, and 15% of the BSA in the sample analyzed by Lo Conte et al. (1999). Thus, the composition based on surface areas is similar to that obtained by counting Voronoi facets. Nevertheless, the composition of the set of atoms that contribute the facets is different: 65% non-polar, 27% neutral polar and 8% charged, which implies that the average polar or charged group contributes more facets than a non polar group. In addition, we noted above that about 13% of the atoms that contribute to N_{Vor} do not lose ASA. This set of atoms is significantly enriched in non-polar groups (73% vs. 65%) and lacks charged groups (2% vs. 8%), in line with the observation that a majority have zero ASA to start with, and the protein main chain contributes 58% of the set.

Even though some interface atoms are already buried in free A or B, most remain accessible to solvent even in the complex. The fraction of the N_{Vor} interface atoms that have zero ASA in the complex is 35% on average, with a standard deviation of 7% and a wide range (13-58%). This buried fraction is the same as in the solvent accessibility model (Lo Conte et al. 1999) in spite of the fact that there are 13% more interface atoms in the Voronoi model. When water is taken into account, many more interface atoms are buried, and the buried fraction increases sharply from 38% to 62% in the 2 Å set.

Interactions

A bicolor Voronoi facet indicates an interaction between an atom of A and one of B. The average number of interactions per interface atom is the same, n_{neigh}=3.52, as the average number of neighbors. In Table 4, we distribute facets into three types that represent different types of interactions: non polar/non polar interactions between two carbon-containing groups; polar/polar interactions between two O, N, or S containing groups; and non polar/polar interactions. On average, 44% of the facets are of the non polar/non polar type, 12%, polar/polar and 44%, non polar/polar. In these statistics, charged groups count as polar, and only 1% of the facets represent a positive/negative charge interaction (salt bridge). These fractions are close to those expected for random pairing given the atomic composition of the interfaces. Statistics based on the contributions to the VIA rather than the number of facets give the same non polar/polar fraction (44%), a slightly lower non polar/non polar fraction (39%) and a larger (17%) polar/polar fraction that includes 2.9% of charge-charge interactions. The composition of the Voronoi facets reproduces the known atomic preferences

for interfaces (Tsai et al. 1997; Lo Conte et al. 1999), but contact preferences at the atomic level are much less obvious (Robert & Janin 1998; Mintseris & Weng 2003), and their detection requires a more detailed statistical analysis.

In the ABW model, facets involving water molecules indicate the interaction of a protein atom with interface water, which we label water/polar or water/non polar depending on the type of protein atom. Like Rodier et al. (2005), we find water-mediated interactions to be at least as abundant at interfaces as direct protein-protein interactions. The average number of bicolor facets in the 2 Å data set increases from 405 in AB to 769 in the ABW model. The additional interactions are 64% water/non polar, 36% water/polar, the same proportions as for non polar and polar protein atoms in N_{Vor}.

DISCUSSION

The Voronoi construction has been extensively used to measure atomic volumes and describe the atomic packing inside proteins (Richards 1974; Harpaz et al. 1994; Gerstein et al. 1995; Pontius et al. 1996). Its first application to protein-protein interfaces was to show that they pack as densely as the protein interior by comparing the Voronoi volumes of interface atoms to those of atoms buried inside proteins (Janin & Chothia 1976; Lo Conte et al. 1999). Later applications include atomic and residue contacts (Munson & Singh 1997; McConkey et al. 2002). We use here an updated and enhanced implementation of that construction to define interfaces and examine their properties. Like Ban et al. (2004), we define a protein-protein interface by the set of facets shared by atoms of the two proteins after discarding excessively large facets that extend out of the protein surface. However, the way we treat the large facets is more direct, and it leads to significantly different results when applied to a set of protein-protein complexes taken from the PDB. In addition, our construction defines accessibility to solvent and handles water molecules, which were not considered by Ban et al. (2004).

Geometric and chemical features of protein-protein interfaces that have been examined in studies based on solvent accessibility are easily retrieved on our model. The Voronoi and solvent accessibility models are in good agreement concerning the size of the interfaces, expressed either as the number of atoms or a surface area. The observed correlation between the numbers N_{Vor} and N_{at} of interface atoms is very high, the correlation between the areas

VIA and BSA also. Ban et al. (2004) cite values of a surface area similar to the VIA for 70 complexes analyzed by Chakrabarti & Janin (2002) and included in this study. They report a correlation with BSA values of 0.85, whereas we obtain 0.982. As both constructions apply the alpha-complex to atomic protein models, the better fit to the solvent accessibility model must be attributed to the different way we handle the large facets on the protein surface.

Although the solvent accessibility model and our implementation of the Voronoi model agree on the size of the interfaces, they differ in their definition of interface atoms. Both models find the same fraction of the interface atoms to be buried in the complex, and all atoms that lose solvent accessibility are part of the Voronoi interface. However, the converse is not true: a remarkable result of our study is the presence at interfaces of atoms that are already buried in the component subunits. In the complex, these atoms share Voronoi facets with one or several atoms of the other component, yet removing that component does not make them accessible to a water probe. They do not contribute to the hydrophobic effect and, being mostly non polar, may form few polar interactions. But they contribute to van der Waals interactions and to the close-packing of the interface. A solvent accessible atom may also fail to lose accessibility because the additional contacts it makes in the complex concern a region of its surface that is buried in the component subunit. We find that the solvent accessibility criterion misses about 13% of the interface atoms for that reason. Main chain atoms, which account for 19% of the BSA (Lo Conte et al. 1999) represent 39% of N_{Vor} and are a majority among the interface atoms that do not lose accessibility. Thus, the Voronoi model suggests that the protein main chain plays a role in protein-protein interaction that is even more important than suggested by previous studies.

The Voronoi model also gives a quantitative basis to features that are not easily estimated otherwise. For instance, the connectivity of an interface has a simple definition: connected components are sets of bicolor facets that have edges in common. By that criterion, a majority of the interfaces in Table 1 are singly connected, a single scc including all or nearly all of the facets. The larger interfaces may contain two or three scc of comparable size. Interfaces have been split in various ways in the past, for instance by considering segments of the protein sequence (Jones & Thornton 1997), or by clustering interface atoms based on a distance criterion (Chakrabarti & Janin 2002; Reichmann et al. 2005). The geometric clustering procedure of Chakrabarti & Janin (2002) distributes interface atoms into patches that are essentially identical to a scc in three-quarters of the complexes of Table 3, and in

most other cases, it splits a scc into two patches as in Fig. 4. Thus, the two approaches yield very similar results, but the Voronoi definition does not depend on a cutoff distance like the clustering procedure.

The curvature of interface is another parameter that can be defined in the Voronoi model. The quantity h(ε) measured at a Voronoi edge (Eq. 5) is an extension to a polyhedral surface, of the mean curvature of a smooth surface (Cohen-Steiner & Morvan 2003). Its sign indicates whether the interface is locally convex towards the A or the B component of the complex. When h(ε) is averaged over the whole interface to yield the s_H angle (Eq. 6), the large value obtained for some PI complexes reflects the complementary concave/convex surfaces of the protease and the inhibitor. In AA and ST complexes, the interaction involves mostly flat patches on the protein surfaces, and s_H is small. The curvature defined by h(ε) is distinct from the angle deficiency of Ban et al. (2004), which is estimated at the vertexes of Voronoi polyhedra, not at their edges. It also differs from the planarity estimated by fitting a least-square plane through the interface atoms (Argos 1988; Jones & Thornton, 1996), yet the same qualitative conclusions can be drawn concerning the shapes of different classes of interfaces.

Unlike the solvent accessibility model, which identifies the interface atoms (albeit not all of them), but says nothing about their partners in the other subunit, the Voronoi model identifies the pairs in contact in a natural way without requiring a distance cut-off. This property has been used to analyze contacts and generate empirical potentials between protein atoms (Munson & Singh 1997; McConkey et al. 2002). We show here that the Voronoi model also handles protein-water interactions, which are abundant at protein-protein interfaces (Janin 1999; Rodier et al. 2005). Our data highlight the role of structural water, which fills packing defects, and links together the components of interfaces that are split into several scc when only protein atoms are taken into account.

As a conclusion, we believe that this study introduces a new tool to analyze interactions between biological macromolecules, and give a geometric, topological and chemical description of their interfaces starting at the atomic level.

Acknowledgements

We are grateful to Dr. J. Bernauer (Orsay) and F. Rodier (Gif-sur-Yvette) for discussion. JJ

acknowledges support of the EIDPP program of Action Concertée Incitative IMPBio (Ministère de la Recherche).

References

Akkiraju, N. & Edelsbrunner, H. (1996). Triangulating the surface of a molecule. *Discrete Appl. Mathematics* **71**:5-22.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Engin.* **2**:101-113.

Aurenhammer, F. (1987). Power diagrams: properties, algorithms and applications. *SIAM J. Computing* **16**:78-96.

Ban, Y. E.A., Edelsbrunner, H. & Rudolph, J. (2004). Interface surfaces for protein-protein complexes. *RECOMB*. 205-212.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr.* **58**:899-907.

Cazals, F. & Proust, F. (2006). Revisiting the description of Protein-Protein interfaces. Algorithms. Submitted.

Chakrabarti, P. & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins* **47**:334-343.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins *J. Mol Biol.* **105**:1-12.

Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature* **256**:705-708.

Cohen-Steiner, D. & Morvan, J. M. (2003). Restricted Delaunay triangulations and normal cycle. *Symposium on Computational Geometry*. 312-321.

Dupuis, F., Sadoc, J., Jullien, R., Angelov, B. & Mornon, J. P. (2005). Voro3D: 3D Voronoi tesselation applied to protein structures. *Bioinformatics* **21**:1715-1716.

Edelsbrunner, H. & Mucke, E. P. (1994). Three-dimensional Alpha-Shapes. *ACM Trans. Graphics* **13**:43-72.

Gellatly, B. J. & Finney, J. L. (1982). Calculation of protein volumes: an alternative to the Voronoi procedure. *J Mol Biol* **161**:305-322.

Gerstein, M., Tsai, J. & Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* **249**:955-966

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure* **2**:641-649.

Janin, J. & Chothia, C. (1976). Stability and specificity of protein-protein interactions: the case of the trypsin-trypsin inhibitor complexes. *J Mol Biol.* **100**:197-211.

Janin, J. & Chothia, C. (1990) The structure of protein-protein recognition sites. *J Biol Chem.* **265**:16027-16030.

Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure Fold Des.* **7**:R277-279.

Jones, S. & Thornton, J. M. (1995) Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol.* **63**:31-65.

Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci USA*. **93**:13-20.

Jones, S. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **272**:121-132.

Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**:379-400.

Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol.* **285**:2177-98.

MacCallum, R. M., Martin, A. C. & Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J Mol Biol.* **262**:732-45.

Mariuzza, R. A., Phillips, S. E. & Poljak, R. J. (1987). The structural basis of antigen-antibody recognition. *Annu. Rev. Biophys Biophys Chem.* **16**:139-59

McConkey, B. J., Sobolev, V. & Edelman, M. (2002). Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics* **18**:1365-1373.

Mintseris, J. & Weng, Z. (2003). Atomic Contact Vectors in protein-protein recognition. *Proteins* **52**:629-639.

Munson, P. & Singh, R. (1997). Statistical significance of hierarchical multi-body potentials based on the Delaunay tesselation and their application in sequence-structure alignment. *Protein Sci.* **6**:1467-1481.

Nadassy, K., Tomas-Oliveira, I., Alberts, I., Janin, J. & Wodak, S. J. (2001). Standard atomic volumes in double-stranded DNA and packing of protein-DNA interfaces. *Nuc. Ac. Res.* **29**:3362-3376.

Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions. *EMBO J.* **22**:3486-3492.

Pontius, J., Richelle, J. & Wodak, S. J. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **264**:121-136.

Poupon, A. (2004). Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.* **14**:233-41

Reichmann, D., Rahat, O., Albeck, S., Meged, R., Dym, O. & Schreiber, G. (2005). The modular architecture of protein-protein binding interfaces. *Proc. Natl. Acad. Sci. USA* **102**:57-62.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**:1-14.

Robert, C. H. & Janin, J. (1998). A soft, mean-field potential derived from crystal contacts for predicting protein-protein interaction. *J. Mol. Biol.* **283**:1037-1047.

Rodier, F., Bahadur, R. P., Chakrabarti, P. & Janin, J. (2005). Hydration of protein-protein interfaces. *Proteins* **60**:36-45.

Singh, R., Tropsha, A. & Vaisman, I. (1996). Delaunay tesselation of proteins: four body nearest-neighbour propensities of amino acid residues. *J. Comput. Biol.* **3**:213-221.

Soyer, A., Chomilier, J., Mornon, J. P., Jullien, R. & Sadoc, J. (2000). Voronoi tesselation reveals the condensed matter character of folded proteins. *Phys. Rev. Lett.*. **85**:3532-3535.

Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**:53-64.

Tsai, J. & Gerstein, M. (2002) Calculation of protein volumes: sensitivity analysis and parameter database. *Bioinformatics* **18**:985-995.

Legend of figures

Fig. 1: Delaunay and Voronoi descriptions of a protein-protein interface. Atoms are drawn as balls centered in a_1 for the red molecule, a_2, a_3, a_4 for the blue molecule. The ball radii are group radii augmented of the water probe radius. (A) Condition alpha: Delaunay edges are drawn in blue between the blue atoms and in green between blue and red atoms; the green edge between A_1 and A_4 is dashed to indicate that it is not part of the alpha-complex for α=0. Thus, the interface between the red and blue molecules comprises only the two Voronoi facets drawn in thick green line. (B) Condition beta: The dashed circle represents a ball orthogonal to the balls representing atoms A_1, A_3 and A_4. Its radius being m, the facet between A_1 and A_4 (thick green line) will be accepted or discarded depending on the ratio m/r, where r is the radius of A_1.

Fig. 2: Comparison of the Voronoi and solvent accessibility models of protein-protein interfaces. Data for the solvent accessibility model are taken from Lo Conte et al. (1999). (A) Number of interface atoms in the Voronoi AB model (N_{Vor}) vs. atoms that lose ASA (N_{at}). The slope of the line is 1.13, the correlation coefficient, 0.992. (B) Voronoi interface area VIA vs. half of the buried surface area. The slope of the line is 1.31, the correlation coefficient, 0.982.

Fig. 3: A solvent inaccessible atom that is part of the Voronoi interface. Balls representing atoms have radii equal to the group radius plus the water probe radius (1.4 Å). The red ball is an atom of A, the blue ball an atom of B. On top, the two balls are seen to intersect, and both atoms are part of the AB interface. On bottom, balls representing other atoms of A occlude the red ball; these atoms were omitted in the top part of the figure.

Fig. 4: Interface connectivity: patches and scc. The *E. coli* EF-Tu/Ts interface is split by the geometric clustering procedure of Chakrabarti & Janin (2002) into four patches, but it contains only three scc. (A) The patches are in different colors on the molecular surface of EF-Tu. (B) Heavy lines mark the edges of the three scc. EF-Ts is drawn as a red ribbon in both panels.

Fig. 5: Water and the connectivity of protein-protein interfaces. (A) Chymotrypsin-eglin (1acb). The facets belong to the AB interface and form a single scc. In the ABW model, the interface is split into two scc marked by the heavy lines. Water is located around the interface

and in the gap separating the two scc. (B) The transducin G_{α}-$G_{\beta\gamma}$ interface (1got). The interface is in two parts, each lined by water molecules. They form a single scc in the AB model, and two in the ABW model. In both 1acb and 1got, the two scc of the ABW model merge into one due to connecting waters.

Fig. 6: Curvature. (A) An atom of protein A centered in **a** has common facets with two atoms of protein B centered in $\mathbf{b_1}$ and $\mathbf{b_2}$, and these facets share an edge ε. The three atoms form a Delaunay triangle (heavy line). The discrete curvature at edge ε is the product of the length of the edge by the dihedral angle β, which is equal to the angle in **a** of the Delaunay triangle. (B) Distribution of the values of |β| in the 1udi interface. The peak near 15° represents triangles where the atoms centered in $\mathbf{b_1}$ and $\mathbf{b_2}$ are covalently linked. (C) The kallikrein-pancreatic trypsin inhibitor (2kai) interface, which has the largest mean curvature s_H=17 degrees in our sample, is concave towards the inhibitor drawn as a ribbon. The concavity is particularly marked around Lys 15 (drawn in van der Waals spheres), which occupies a well-defined pocket on the protease surface.

Table 1: Protein-protein complexes

PI: Protease-inhibitor complexes (29)

1acb*	1avw*	*1brc*	1bth	1cbw	*1cgi*	1cho*	1cse*	1dan*	1fle*
1hia	1mct*	*1ppe**	1ppf*	*1spb**	1stf	*1tab*	1tbq	1tgs*	1toc
2kai	2ptc*	2sic*	2sni	*2tec**	3sgb*	3tpi*	4cpa	4htc	

ESI: Other enzyme-substrate or inhibitor complexes (12)

1brs*	1dfj	1dhk*	1fss	1gla	*1kkl*	*1mah*	1udi	*1ugh**	1ydr
2mta	2pcc								

AA: Antigen-antibody complexes (28)

1ahw	1ao7	*1bql*	*1bvk*	*1dqj**	1dvf*	1eo8	1fbi	1iai	1jhl
1kb5	*1kxq**	*1kxt**	*1kxv**	1mel	1mlc	1nca	1nfd	1nmb	1nsn
1osp*	1qfu	1vfb*	*1wej**	2jel	*2vir*	3hfl	3hfm		

ST: Signal transduction, cell cycle (19)

1a0o	1a2k	1agr	1aip	*1avz*	1ebp	1efn	1efu	1fin	*1fq1*
1gg2	1got*	1gua*	1hwg	1tx4*	*1wq1*	1ycs	2trc	*3hhr*	

MI: Miscellaneous (8)

1ak4	1atn	1dkg	1fc2	1igc	*1l0y*	1seb	2btf

PDB entries for the 96 protein-protein complexes; 72 entries were taken from Lo Conte et al. (1999); the other 24 are in italics; asterisks mark the 30 entries of the 2 Å set used in the ABW model.

Table 2: Geometric properties of the interfaces

Type of interface	PI		ESI		AA		ST		MI		All		2Å set		ABW model	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
Number of interfaces	29		12		28		19		8		96		30		30	
Interface atoms																
number N_{Vor}	238	80	239	74	208	34	311	136	187	46	239	89	227	62	330	88
asymmetry ratio r_{AB}	1.47	0.24	1.14	0.12	1.11	0.08	1.08	0.05	1.16	0.16	1.22	0.23	1.30	0.26	1.25	0.22
buried fraction f_{bur}	0.42	0.06	0.30	0.07	0.31	0.06	0.34	0.04	0.35	0.06	0.35	0.07	0.38	0.08	0.62	0.08
Number of facets N_{facet}	434	141	452	147	362	69	522	250	322	86	423	162	405	109	769	217
Number of neighbours n_{neigh}	3.7	0.1	3.4	0.2	3.5	0.2	3.5	0.1	3.4	0.2	3.53	0.18	3.57	0.15	3.63	0.13
Interface area VIA ($Å^2$)	1219	486	1270	330	1100	203	1691	701	1019	272	1263	488	1168	368	1526	434
Connected components scc	1.1	0.4	1.3	0.6	1.1	0.3	1.5	0.8	1.4	0.7	1.21	0.54	1.13	0.43	1.37	0.6
Mean curvature s_H (degrees)	10.2	3.2	4.4	2.7	4.5	3.0	2.3	1.8	3.9	3.1	5.2	4.2	6.6	4.5		

Mean value and standard deviation of geometric parameters in the set of 96 protein-protein interfaces and in subsets corresponding either to the five classes of Table 1 or to the 30 high-resolution X-ray structures (2 Å set). The ABW model applies only to the 2 Å set.

Table 3: Patches vs. connected components[a]

1. One patch - one scc (50)											
1a0o	1a2k	1acb	1agr	1ak4	1atn	1avw	1brs	1bth	1cbw	1cho	1cse
1dhk	1dvf	1efn	1fbi	1fc2	1fle	1gla	1gua	1hia	1iai	1igc	1jhl
1kb5	1mct	1mel	1mlc	1nca	1nfd	1nmb	1nsn	1osp	1ppf	1qfu	1seb
1stf	1tgs	1udi	1vfb	2jel	2kai	2ptc	2sic	2sni	3hfl	3hfm	3sgb
3tpi	4cpa										
2. One patch - two scc (2)											
	1ao7	2pcc									
3. Two patches - one scc (8)											
1fin	1fss	1toc	1tx4	1ydr	2btf	2trc	4htc				
4. Two patches - two scc (5) [b]											
1gg2 (0.99)	1got (1.0)	1hwg (0.85)	1tbq (0.95)	1ycs (0.90)							
5. Three or four patches - one to three scc (5) [b]											
1aip (0.99)	1dan (0.85)	1dfj (0.74)	1dkg (0.82)	1efu (0.99)							

(a) Patches are taken from Chakrabarti & Janin (2002), except for the antigen-antibody complexes, which were reported as having two patches (or three patches in the case of 1kb5) due to the clustering algorithm being run separately on the H and L chains of the antibody. The same clustering algorithm finds a single patch when the search is done in one step on the two chains.

(b) Numbers in parentheses are the fraction N_{com}/ N_{at} where N_{at} is the total number of atoms that lose ASA, and N_{com} the number of atoms belonging to patches that are entirely contained in a scc.

Table 4: Chemical properties of the interfaces

Composition[a]	All		2Å set		ABW model	
	mean	SD	mean	SD	mean	SD
Atom types						
non polar (C)	58	2	58	1		
neutral polar (N, O, S)	30	3	31	3		
charged (N, O)	12	3	11	3		
main chain	39	3				
Pairwise interactions[b]						
non polar/non polar	44	4	44	4	47	4
polar/polar	12	2	12	2	11	2
non polar/ polar[a]	44	3	44	3	43	3
water/polar[c]					36	2
water/non polar[c]					64	2

(a) percent of N_{facet}

(b) 'polar' includes charged groups.

(c) percent of AW-BW facets

Fig.1A

Fig.1B

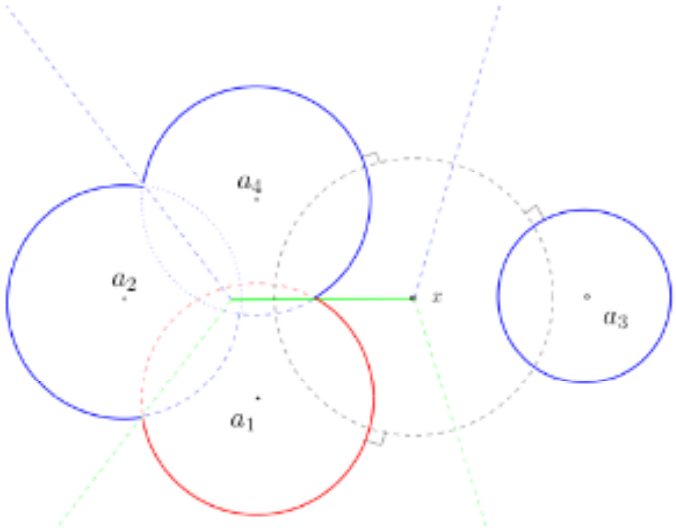

Fig.2A

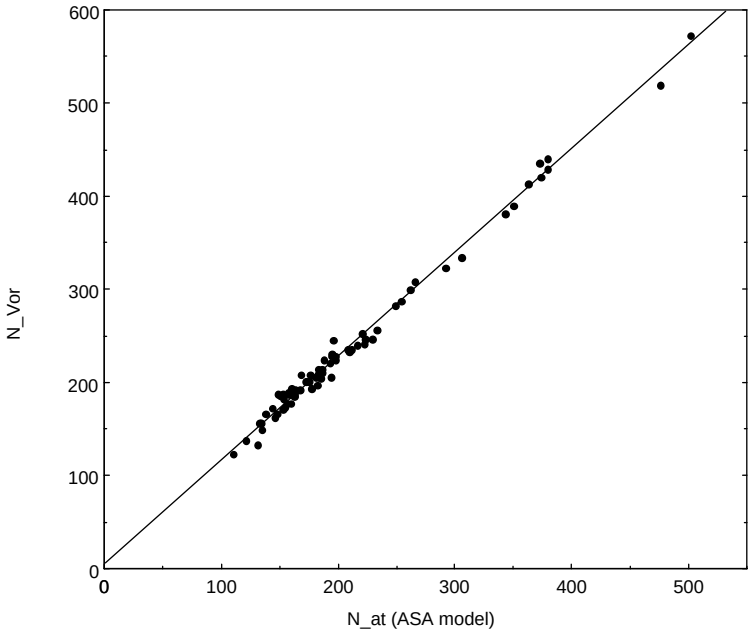
600
500
400
300
200
100
0
N_Vor
0
100
200
300
400
500
N_at (ASA model)

Fig.2B

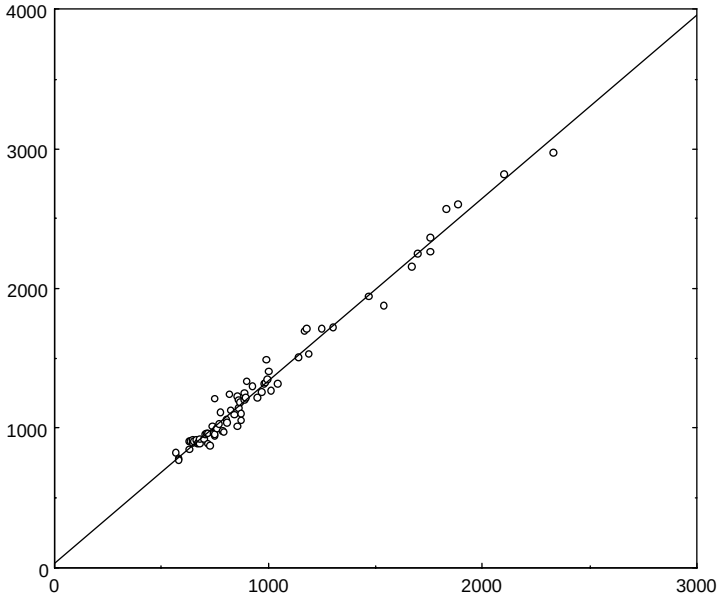
4000
3000
2000
1000
0
0
1000
2000
3000

Fig.3

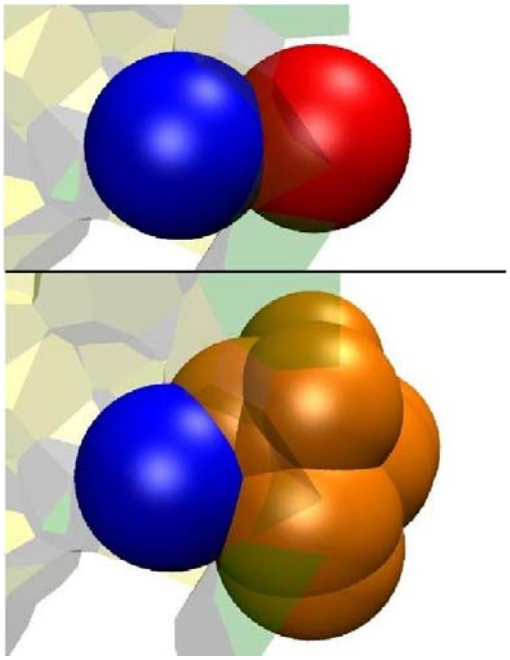

Fig.4A

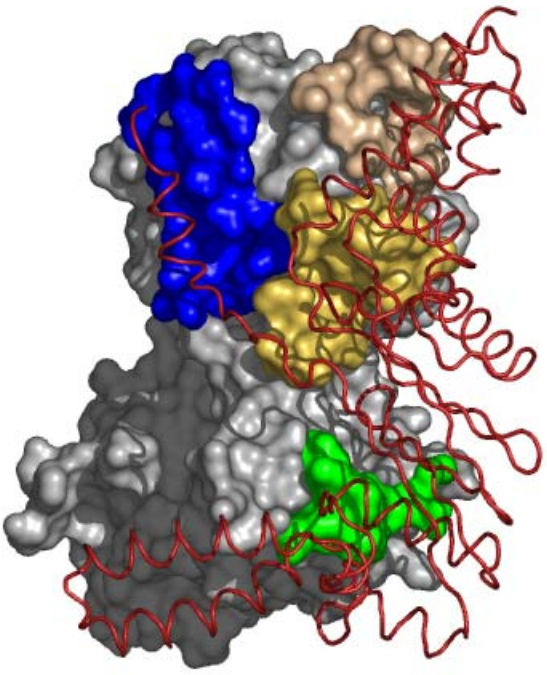

Fig.4B

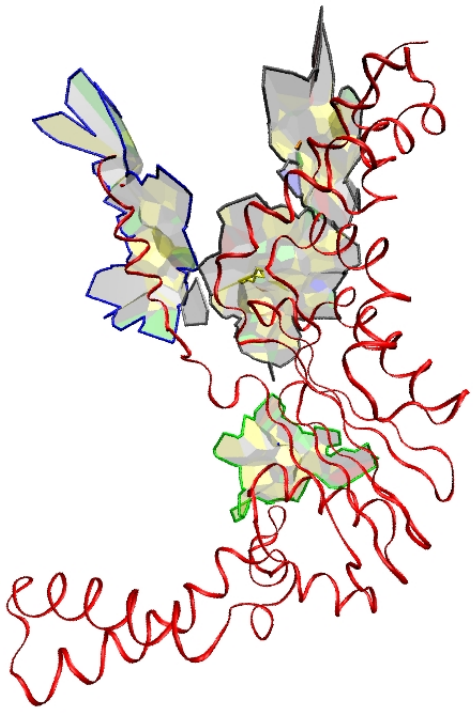

Fig.5A

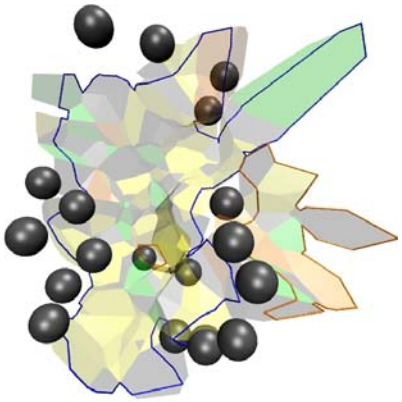

Fig.5B

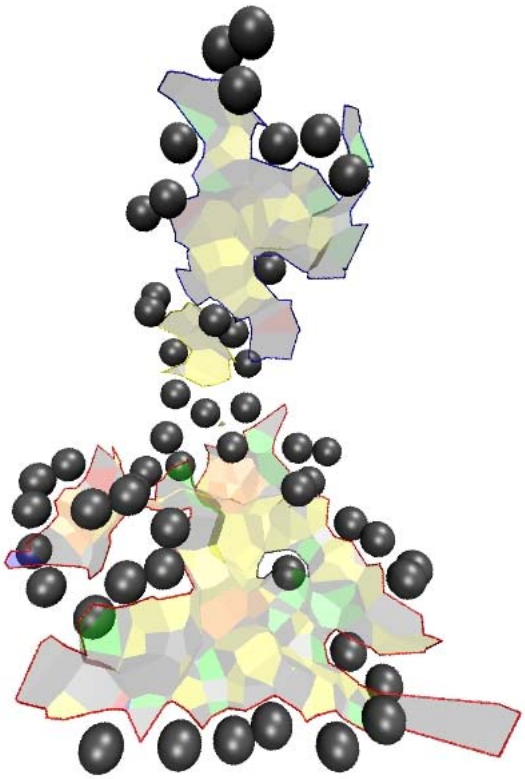

Fig.6A

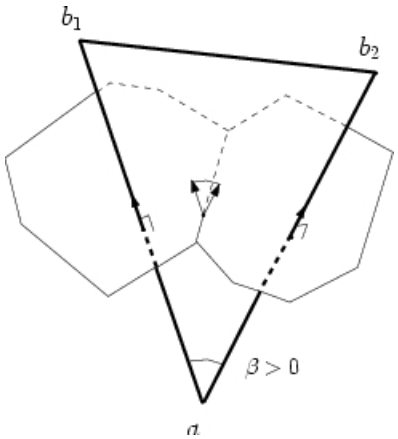

Fig.6B

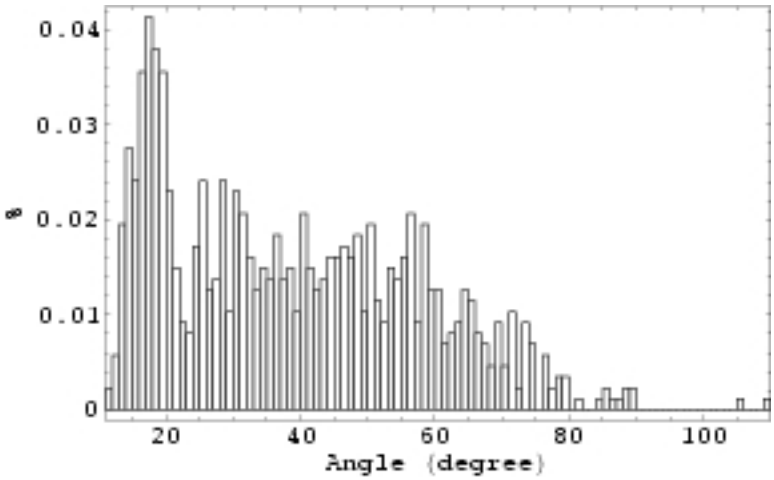

Fig.6C