BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 792–800
doi:10.1093/bioinformatics/btt644

Structural bioinformatics

Advance Access publication November 7, 2013

Computing cavities, channels, pores and pockets in proteins from non-spherical ligands models

Lydia Benkaidali[1,2], François André[3], Boubekeur Maouche[4], Pridi Siregar[5], Mohamed Benyettou[6], François Maurel[1] and Michel Petitjean[7,*]

[1]ITODYS, UMR 7086, CNRS, Université Denis Diderot, Paris 7, [2]Programme Doctoral International, Université Pierre & Marie Curie, Paris 6, [3]CEA/DSV/iBiTec-S/SB2SM, CNRS, UMR 8221, Saclay, France, [4]LPCTCI, Université des Sciences et de la Technologie Houari Boumédiène, Algiers, Algeria, [5]IBC, Integrative BioComputing SARL, Place du Granier, 35135 Rennes-Chantepie, France, [6]LAMOSI, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, Oran, Algeria and [7]MTi, UMR-S 973, INSERM, University Denis Diderot, Paris 7, France

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Identifying protein cavities, channels and pockets accessible to ligands is a major step to predict potential protein–ligands complexes. It is also essential for preparation of protein–ligand docking experiments in the context of enzymatic activity mechanism and structure-based drug design.

Results: We introduce a new method, implemented in a program named CCCPP, which computes the void parts of the proteins, i.e. cavities, channels and pockets. The present approach is a variant of the alpha shapes method, with the advantage of taking into account the size and the shape of the ligand. We show that the widely used spherical model of ligands is most of the time inadequate and that cylindrical shapes are more realistic. The analysis of the void parts of the protein is done via a network of channels depending on the ligand. The performance of CCCPP is tested with known substrates of cytochromes P450 (CYP) 1A2 and 3A4 involved in xenobiotics metabolism. The test results indicate that CCCPP is able to find pathways to the buried heminic P450 active site even for high molecular weight CYP 3A4 substrates such as two ketoconazoles together, an experimentally observed situation.

Availability: Free binaries are available through a software repository at http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html

Contact: michel.petitjean@univ-paris-diderot.fr

Received on July 24, 2013; revised on October 30, 2013; accepted on November 3, 2013

1 INTRODUCTION

Protein concavities play an important role for biological functions at the molecular level such as ligand binding. Knowledge of protein–ligand binding sites and their access routes contributes to optimize virtual screening of chemical libraries and then to propose new potential drugs. The concavity containing the active site may be located at various places in the protein. When it is at the surface of the protein, the concavity is called a pocket, and it is called a cavity when the active site is buried inside the protein, in which case the route to the active site is called a channel. In any case, defining the shape of the protein needs to locate the boundary between the protein and its void parts. Many algorithms were proposed in the literature to solve this non-trivial problem. In a recent review (Pérot *et al.*, 2010), 20 geometric algorithms were reported, plus 10 algorithms based either on genomic information or on energy calculations. These latter being devoted to protein–ligand complexes energy calculation by docking (see Sperandio *et al.*, 2006 for a review), they are outside the scope of this article. We report in Table 1 a non-exhaustive list of 35 geometric algorithms and their variants.

There are several ways to classify them. The main criteria are: (i) the computation is dependent or not on probes spheres assumed to represent ligands, (ii) the computation is based either on a rectangular grid or on a Delaunay triangulation or on none of them. It has to be noticed that building the Delaunay triangulation of N input vertices in the 3D space consists practically in partitioning the smallest convex polyhedron enclosing the N input vertices into non-overlapping adjacent tetrahedra with endpoints edges at input vertices, such that none interior of the spheres circumscribed to these tetrahedra contains any of the input vertices (Edelsbrunner, 1987). The dual structure of the Delaunay triangulation is the Voronoi diagram (Edelsbrunner, 1987), which, roughly speaking, is the union of the N polyhedral cells defined from the set of the planes mid-perpendicular to the sides of these tetrahedra. To each of the input vertices K, $1 \leq K \leq N$, is associated its polyhedral cell (possibly unbounded) which is the domain containing all points of the space closest to K than to any of the $N - 1$ other vertices. This important property explains why Voronoi tessellations were used to visualize the shape of proteins, such as in Voro3D (Dupuis, 2003; Dupuis *et al.*, 2005).

Grid methods offer several drawbacks. The orientation of the 3D grid is arbitrary, and most of the time an increase of the precision by 10 needs to increase the number of nodes by 10^3, thus increasing computing times by 1000. It is the case, when the measures of cavity volumes are estimated from the number of nodes or from the number of voxels (i.e. the 3D analogs of pixels). When the grid nodes are in a graph in which paths are seeked to identify channels, the number of paths to be enumerated increases exponentially with the number of nodes. To overcome this problem, the user can define an arbitrary upper bound

*To whom correspondence should be addressed.

Table 1. Main geometric algorithms for pockets or channels detection

Name	References	Method
Cavity search	Ho and Marshall (1990)	Gridded slices and voxels
POCKET	Levitt and Banaszak (1992)	Grid (marching cubes); probe sphere
VOIDOO	Kleywegt and Jones (1994)	Grid and voxels; probe sphere
SURFNET	Laskowski (1995)	Grid; fitting spheres into the spaces between atoms
Masuya and Doi	Masuya and Doi (1995)	Grid; probe sphere; cavity boundaries are spherical surfaces that are parts of the probe sphere
LIGSITE	Hendlich *et al.* (1997)	Grid; variant of POCKET; visualization from VRML files
Variant in LIGSITE	Stahl *et al.* (2000)	Grid; clusters of cavity points; pocket atoms were defined as being the protein grid points closest to any surface point
LigandFit	Venkatachalam *et al.*, 2003	Grid; nodes of cavities boundaries are got with a probe cube (orientation fixed by the grid)
Travel Depth	Coleman ans Sharp (2006)	Grid: channels are defined from shortest path in grid to convex hull boundary
CAVER	Petřek *et al.* (2006)	Grid; minimal cost paths channels; the contribution to the cost function at each node is arbitrarily set to the inverse of a geometrically interpretable quantity in order to avoid a meaningless maximal cost path search
PocketPicker	Weisel *et al.* (2007)	Grid; pockets are clusters of grid points; an integer buriedness index is defined from the presence or absence of a protein atom at fixed distance along 30 search directions
PocketDepth	Kalidas and Chandra (2008)	Grid; shortest Euclidean travelling paths in grid to given protein atoms are identified; pockets are clusters of site points
PoreWalker	Pellegrini-Calace *et al.* (2009)	Grid; computation of an initial pore axis from the protein secondary structure (variable diameter from test spheres at various locations), then the axis is relocated and the process is iterated; the algorithm is restricted to some classes of transmembrane channels proteins
McVol	Till and Ullmann (2010)	Grid; filling probe spheres; volumes are Monte-Carlo measured (counting inner points)
POCASA	Yu *et al.* (2010)	Gridded slices; pockets are sets of free points between a probe sphere and protein surface
VICE	Tripathi and Kellogg (2010)	Grid (with integer arithmetic); pockets are lists of protein atoms; paths are lists of voxels
T-RRT	Jaillet *et al.* (2010), Cortés *et al.* (2011)	Grid; Monte-Carlo path planning through voxels
PROPORES	Lee and Helms (2012)	Grid and voxels; hybrid variant of POCKET and SURFNET; shortest paths in voxels with arbitrary cost functions
David	David (1988)	Voronoi tessellation is used to estimate void volumes; these latter are divided in two three size classes; clefts are continuous sets in the class of the intermediate size
FindSurf	Lewis (1989)	Voronoi tessellation altered with dummy tiles; the method is used in complement of David's one; clefts are lists of atoms
CAST, CASTp	Edelsbrunner and Mücke (1992, 1994), Edelsbrunner *et al.* (1998)	Original alpha shapes, based on deletions in the Delaunay triangulation of the protein atoms with the help a probe sphere of user fixed radius alpha (hence the name 'alpha shape')
APROPOS	Peters *et al.* (1996)	Alpha shapes, list of atoms in clusters by difference between two alpha shapes
MOLE	Petřek *et al.* (2007)	Variant of CAVER based on Delaunay triangulation; minimal cost paths channels; the cost function is a variant of the one in CAVER, and is still the inverse of a geometrically interpretable quantity
GeometricPotential	Xie and Bourne (2007)	Delaunay triangulation of C_α atoms, removal of terahedra based on arbitrary cut-off values
MolAxis	Yaffe *et al.* (2008a, b)	Channels are alpha shapes based minimal cost pathways; the name comes from the use of a medial axis (here, it is a collection of points having more than one closest point on the van der Waals surface of the protein)
SplitPocket	Tseng and Li (2009), Tseng *et al.* (2009)	Alpha shape based on weighted Delaunay triangulation; pockets are sets of empty Delaunay triangles with at least one acute triangle
Fpocket	Le Guilloux *et al.* (2009), Schmidtke *et al.* (2010)	Pockets are lists of atoms deduced from Delaunay triangulation followed by filling spheres clustering

(continued)

Table 1. Continued

Name	References	Method
UnionBall	Mach and Koehl (2011)	Enhanced version of AlphaVol (part of the original alpha shape package); the calculation of volumes of spheres unions are based on the Voronoi diagram (Edelsbrunner, 1995)
CAVER 3.0	Chovancova *et al.* (2012)	Variant of CAVER implementing new algorithms for the calculation and clustering of pathways
MOLE2	Berka *et al.* (2012)	Variant of MOLE; needs a user defined starting point; a vertex of the Voronoi diagram is removed if a sphere with interior threshold radius cannot pass through any of the tetrahedron sides; channels are defined from shortest paths detection
HOLE	Smart *et al.* (1993, 1996, 1997)	Monte-Carlo simulated annealing exploration of the space available to a sphere; at each step the maximal sphere radius is computed
PASS	Brady and Stouten (2000)	Layers of non-overlapping filling spheres are generated in cavities
SCREEN	Nayal and Honig (2006)	An envelope surface is computed with the help of a 5 Å radius probe sphere, allowing to assign depth values to vertices on the protein surface; the clusters of vertices below a depth threshold define the surfaces of the cavities
PHECOM	Kawabata and Go (2007)	Two probe spheres radii are used; small probes not overlapping large ones define pockets
Binding Response	Zhong and MacKerell (2007)	Overlapping filling spheres are clustered (two steps); clusters define pockets

Grid means 3D rectangular or cubic mesh of nodes. Voxels are 3D analogs of pixels. Alpha shapes are substructures of the Delaunay triangulation

of the shortest paths, but in this case the major part of the network of channels is lost.

At the opposite of a grid of virtual nodes, the Delaunay triangulation of the N atomic centers is more natural, and efficient algorithms exist, the main one being based on a 4D convex hull (Preparata and Shamos, 1985; Edelsbrunner, 1987). It is why the seminal concept of alpha shape appears in most modern pockets and channels calculation algorithms, as seen in Table 1. This concept was first introduced in the plane by Edelsbrunner *et al.* (1983), and then was generalized to higher dimensional spaces and proposed for molecular applications by Edelsbrunner and Mücke (1992). Alpha shapes are a generalization of the convex hull of the N points, this latter being the smallest convex polyhedron enclosing the N points. A spherical probe sphere of radius alpha erases from the polyhedral hull any point at distance less than alpha to the sphere center for all positions where the sphere does not enclose any of the N points. The resulting object is called an alpha hull. Substituting straight edges for the circular ones and triangles for the spherical caps, the final object is the alpha shape of the N points set. When alpha tends to zero, the alpha shape reduces to the N points. When alpha is large enough, the alpha shape is the convex polyhedral hull of the points. For intermediate alpha values, the alpha shape is not convex and can be used as a model of the protein shape. The original variant of the alpha shape concept that we introduce is described in the next section.

2 METHODS

2.1 Approach

Our approach differs from the existing ones based on alpha shapes variants at three levels: (i) we are interested in the structure of the network of channels, i.e. the complement of the protein alpha shape to its convex hull, (ii) we considered several geometric models for the ligand, not only spheres and (iii) the solid representing a ligand is used to flag the triangular faces as 'can be passed' or 'cannot be passed', rather than erasing edges, triangles and tetrahedra.

The first level is of crucial importance when it is necessary to identify potential access channels to a buried active site, as in cytochromes P450 (Guengerich, 2005): the shape of the channels must be analyzed, not the shape of the protein. The second level is detailed in Section 2.2, in which it is shown that the spherical model of ligands is most of the time inadequate. The third level can be understood as follows: to travel inside the polyhedral protein hull, the ligand must pass through some sequence of triangular faces, each face separating two adjacent tetrahedra, just as triangular doors can be open or closed in a labyrinth of tetrahedral rooms: see Section 2.3. The final goal is to exhibit the full network of paths through the labyrinth, or from the exterior of the labyrinth to the chamber neighboring the hidden active site.

2.2 Ligands shapes analysis

In order to analyze the shapes of the ligands, we report in Table 2 the statistics for 21 geometric parameters on the dataset of 70 substrates of the cytochrome P450 (CYP) 3A4 used by Meslamani *et al.* (2009), containing from 16 to 186 atoms, hydrogens included. The CYP 3A4 substrates are known to offer a wide structural diversity (Rendic, 2002), so it is why we considered that the statistics on these 70 substrates are meaningful.

There are several ways to see in Table 2 that the spherical model of ligands is irrealistic. The first way is to look at the sphericity coefficient G, which is equal to 1 for a sphere and is less than 1 for other solids. It is the case of our dataset, for which G is always <0.5, but G reflects the shape of the van der Waals solid and thus does not include internal void cavity volumes although the need to compute them was pointed out (Sonavane and Chakrabarti, 2008). Also, G depends on the atomic radii values and there is no agreement about these values (see, for example, some suggested set of values in Gavezzotti, 1983; Richards, 1985; Scott and

Table 2. Statistics on 21 geometric parameters (physical units: Å, Å^2, Å^3)

Parameter	Minimum	Maximum	Mean	Standard deviation
S_W (van der Waals surface[a])	147.4	1324.6	387.8	185.9
V_W (van der Waals volume[a])	115.5	1162.5	334.5	166.8
D (diameter[b])	6.9	28.9	13.0	4.1
R (radius[c])	3.7	14.5	6.6	2.0
V_S [d]	201.3	12673.8	1582.6	1985.7
V_{ch} (convex hull volume)	1.10^{-2}	1282.9	216.1	208.8
H_h (half height of the MHEC[e])	4.10^{-4}	5.5	2.3	0.9
R_h (radius of the MHEC[e])	3.6	14.4	6.5	2.0
V_H (volume of the MHEC[e])	4.10^{-2}	3340.8	732.4	677.1
H_r (half height of the MREC[f])	2.2	7.2	3.7	1.0
R_r (radius of the MREC[f])	3.3	14.2	6.2	2.1
V_R (volume of the MREC[f])	191.0	4829.4	1091.3	1006.4
G (sphericity coefficient[g])	0.066	0.471	0.246	0.077
I_v (volumic shape coefficient[h])	4.10^{-5}	0.404	0.166	0.080
I_g (geometric shape index[i])	0.816	1.000	0.972	0.045
H_h/R	8.10^{-4}	0.649	0.364	0.129
R_h/R	0.876	1.000	0.985	0.018
H_r/R	0.257	0.830	0.592	0.141
R_r/R	0.786	1.000	0.947	0.045
V_{ch}/V_H	0.134	0.589	0.304	0.077
V_{ch}/V_R	5.10^{-5}	0.418	0.206	0.081

All computations were performed with the ASV and RADI freewares, available on the same software repository than CCCPP. [a]S_W and V_W were calculated with the analytical method (Petitjean, 1994, 2013), implemented in ASV. [b]Length of the largest atom pair. [c]Radius of the smallest enclosing sphere; $2R \geq D \geq 2R\sqrt{(6/3)}$ (Petitjean, 1992). [d]Volume of the smallest enclosing sphere. [e]MHEC, minimal height enclosing cylinder, calculated with the algorithm of Brandenberg and Theobald (2006), implemented in RADI. [f]MREC, minimal radius enclosing cylinder, calculated with the algorithm of Petitjean (2012), implemented in RADI. [g]$G = 36\pi V_W^2/S_W^3$; G takes values in [0; 1]. [h]Ratio V_{ch}/V_S of the volume of the convex hull to the volume of its smallest enclosing sphere. [i]$I_g = (D - R)/R$; $I_g \leq 1$ (Petitjean, 1992).

Scheraga, 1966; Zefirov and Zorkii, 1989). In fact these values depend on how they are defined (see Bondi, 1964, for a discussion). Thus we neglect the atomic spheres and we look at other parameters. The geometric shape coefficient $I_g = (D - R)/R$ (Petitjean, 1992) relates the diameter of the hull to its radius, but it varies few in this context. Assuming that the shape is convex, the best shape model is the convex hull of the atomic centers, i.e. the smallest convex polyhedron containing these centers: by definition, any other enclosing convex solid will contain this polyhedron. The observed distribution of the ratio I_v of volume of the convex hull to the volume of the smallest enclosing sphere exhibit a mean value <0.17, which is quite low, thus proving the inadequation of the sphere model. However it is much more difficult to test if a convex polyhedron can pass through pathways without colliding a set of punctual or spherical obstacles rather than testing if a sphere can pass these obstaches without colliding.

It is why the cylindrical shape was devised for molecules (Petitjean, 2012). A cylinder has a convex shape defined by only two parameters: radius and height, although a sphere has only one, a rectangular 3D box has three ones and a polyhedron has many ones. However cylinders are easier to handle than rectangular boxes for collision tests (see Section 2.3). Since there are two parameters, there is an infinity of ways to combine them to define the minimality criterion for a cylindrical solid (e.g. minimize its volume or its surface, or else), but we retained only two of them: the minimal height cylinder which is relevant for rather flat ligands, and the minimal radius cylinder which is relevant for rather elongated ligands. Intuitively, the former reflects atomic centers distributions for which the smallest eigenvalue of the inertia matrix (i.e. N times the covariance matrix) is small compared to the two other eigenvalues, although the latter reflects atomic centers distributions for which two eigenvalues are small compared to the largest one. We see in Table 2 that the radius of the minimal height cylinder and the height of the minimal radius cylinder remain close to 1, although the respective minimized height and minimized radius of these cylinders are significantly <1. Then the distributions of the hull volume to the cylinder volume offer values significantly <1, and of the same magnitude than the sphericity coefficient G and than the volumic shape coefficient I_v, thus confirming the pertinence of the cylindrical model over the spherical model. It is of major importance to understand that, although a ligand conformer has a unique shape, it does not preclude that this unique shape can be modeled through several ways, such as the minimal height enclosing cylinder, the minimal radius enclosing cylinder, the smallest enclosing sphere (even if that latter encloses a large void volume) and so on. Furthermore, it is recalled that most ligands are flexible, and that is a second reason to insist that several realistic shape models can be simultaneously useful. However, the data of Table 2 show that cylindrical shapes are more realistic than spherical shapes, as already mentioned in that section and illustrated in Figure 1.

2.3 Geometric tests of passage

In the basic alpha-shape approach, a tetrahedral cell with its circumscribed sphere of radius alpha can contain a spherical probe ligand of radius alpha (protein atomic spheres are neglected). Assuming that the circumcenter is in the tetrahedra (it is most of the time the case), the spherical ligand cannot exit from the tetrahedra because each radius of the four circles circumscribed to the triangular faces is smaller than alpha. Thus the knowledge of the standard protein alpha shape does not suffice to conclude that the ligand can travel in a channel: a ligand is ensured to travel in a channel if and only if it passes through its associated sequence of triangular faces, and that is a major originality of our approach.

To implement this approach we consider the graph of the Delaunay triangulation, in which the N nodes of the graph are the tetrahedra and

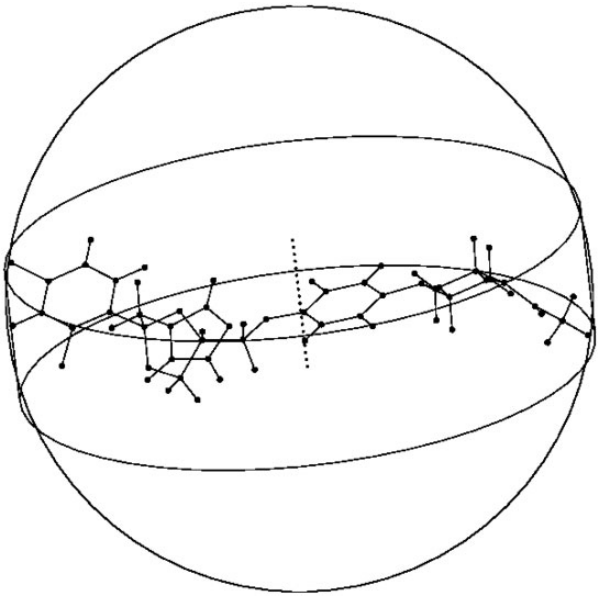

Fig. 1. The ketoconazole, a known CYP 3A4 substrate, modeled by its minimal enclosing sphere ($R = 10.1$ Å) and by its minimal height enclosing cylinder (half height: $H_h = 2.4$ Å; $R_h = 10.0$ Å)

the edges of the graph are the triangular faces, i.e. two adjacent tetrahedra are two nodes joined by a non-directed edge. This graph, which underlies the Voronoi diagram structure, is called here the 'facial graph' and must not be confused with the one where the nodes are the atoms and the edges are the sides of the tetrahedra. Each node of the facial graph has at most four neighbors. Optionnally, a dummy node representing the exterior of the convex hull can be added to the graph, in which case each of the N nodes has exactly four neighbors. Then it suffices to perform a loop on the edges (i.e. the triangles) of the facial graph and flag the ones which cannot be passed by the ligand as deleted edges. The resulting network of channels is a subgraph of the facial graph, not necessarily connected.

The most difficult task is to decide whether or not an object (the ligand) can pass through a triangle. It is hard in the case of a non-convex object, and even for convex objects such as a polyhedral hull or a straight box it is not trivial to do. In fact the problem is to define the criterion for passage rather than to compute if yes or no the passage is possible once its criterion is defined. Thus we followed a simplified approach: we considered punctual atoms and then we defined a passage test through a triangle for the two cylindrical models under the assumption that we can neglect collisions with atoms not at the triangle vertices.

For cylinders with a ratio radius/height sufficiently close to zero, the smallest planar projected section of the cylinder is its circular section because its 2D image can be contained in any of its other planar projected sections. The axis of the cylinder is optimally set orthogonal to the plane of the triangle and we are left to a 2D problem. Then, we flag the passage test as being successful in two cases (a) or (b): (a) the triangle is acute and the radius of the circular section (i.e. the one of the cylinder) is smaller than the circumradius of the triangle, or (b) the triangle is obtuse and the circular section of the cylinder does not contain any of the triangle vertices when the center of this circular section is constrained to lie in the triangle. In the case (a), the largest possible cylinder radius is the circumradius, and in case (b) it is the distance from the vertex at the opposite of the largest side to the point at the intersection of the largest side and of the mid-perpendicular of the second largest side. In both cases (a) and (b), the center of the circular section must pass inside the triangle, not outside (the passage of a punctual solid through a 3D triangle is always defined).

It should be noticed that the tests (a) and (b) are rigorous in the case of the largest section of a spherical ligand, for which the condition of a lack of collision with a fourth vertex can be relaxed.

For cylinders with a ratio height/radius sufficiently close to zero, the axis of the cylinder is optimally set parallel to the plane of the triangle and we are left to a 2D problem. The cylinder height is also the distance between the two closest parallel planes enclosing the ligand. Here, still under the assumption that we can neglect collisions with atoms not at the triangle vertices, we flag the passage test as being successful when the cylinder height is smaller than the largest height of the triangle (or than the second largest side of the triangle when it is obtuse). This passage test can be applied with the same limitations as above to any ligand shape provided that we know the distance between the two closest slabs enclosing the ligand, this latter distance being equal to the height of the smallest height cylinder defined in Section 2.2.

3 RESULTS AND DISCUSSION

To exemplify how works CCCPP, we first considered the CYP 1A2, which is known to accept rather flat substrates (Sridhar *et al.*, 2011; Zaretzki *et al.*, 2012). The data were extracted from the crystallized complex of the 1A2 with α-naphtoflavone, PDB code 2HI4. The flatness, i.e. the thickness of the ligand, is defined as $2H_h$ (see Table 2). Figure 2 shows that ligands of thickness 4.5 Å and higher can access to surface pockets although no pathway is found to the buried heminic site. Lowering this critical value from 4.5 Å to 4.45 Å, the network of channels available to the ligand appears in front of the distal face of the heme, and CCCPP indicates that the network connects the exterior of the enzyme to the active site. When visualized in superimposition of the channel computed for a 4.45 Å critical value, the flat ligand of the complex (0.32 Å) appears indeed in this predicted channel, in front of the distal face of the heme. A remarkable result is that all parts of the central area of the access channel become unavailable quasi simultaneously for any ligand of thickness slightly increasing from 0.05 Å above the critical value of 4.5 Å although the surface pockets are still available for ligands of thickness 4.5 Å. Another result is that the network of channels does not reduce to a small number of chains of nodes: there is NOT a small number of channels, and counting them is meaningless, at the opposite to what seems intuitively suggested by some other channels modeling approaches (see Table 1). CCCPP provides relevant information on this complex network rather than outputting a misleading small number of paths, and the full network structure can be stored both in a text file and in a molecular file to be reused or displayed by any molecular viewer.

The CYP 3A4 is known to accept a much wider diversity of substrates than the 1A2 (Rendic, 2002), some of them being very large. The case of the ketoconazole is of special interest. Its radius, as defined in Table 2, is 10.1, and its thickness is 4.8 Å. A conformational study showed that only a limited number of conformers suffices to handle flexibility (Benkaidali *et al.*, 2012), and the effect on the radius and on the thickness can be neglected to compute the network of channels. A crystallized complex containing two ketoconazoles is available, PDB code 2V0M. There are four chains A, B, C, D in the PDB file. Only surface residues are missing in chains A and D, so that no reconstruction is needed for these chains in our context. The 3D similarity between chains A and D was checked with the SDM algorithm (Petitjean, 1998) after 3D alignment of the two backbones. Only the C_α of

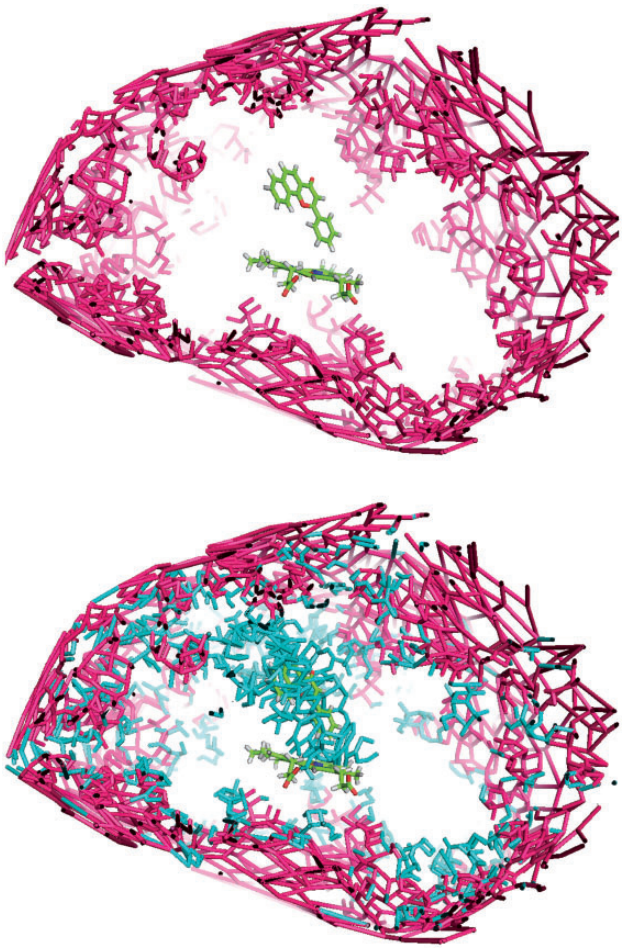

Fig. 2. The facial graph of the CYP 1A2 channels network available for ligands of, respectively, 4.50 Å (top) and 4.45 Å (bottom) thickness, visualized with PyMOL. For clarity, the channels are superimposed with the ligand (α-naphtoflavone) of the CYP 1A2 complex, pdb code 2HI4. Edges indicate the pathways in the network, except for the ligand and for the heme, where edges indicate chemical bonds

ARG 212 was outside the maximal common 3D motif, thus we neglected this difference and we considered only the chain A. Using a passage test based on a spherical model (it is an option in CCCPP), CCCPP did not find any access to the heme. However CCCPP indeed found an access for the cylindrical model. Figure 3 shows that the two ketoconazoles are indeed in the predicted channels, one in front of the distal face of the heme, and the other one just above in a dead end part of the network.

CCCPP outputs many surface pockets. It is of interest to observe that they can be used. Figure 4 shows such a situation for the complex of the CYP 3A4 with progesterone, PDB code 1W0F.

CCCPP can predict channels from the 3D structure of the enzyme without the help of a complex. We used two 3A4 structures crystallized without ligand, 1TQN and 1W0E. Due to missing residues, we reconstruct the mid of the sequence with HYDRO_PDB (Azuara *et al.*, 2006). Figure 5 shows the results of CCCPP for two 3A4 structures found in the PDB, codes

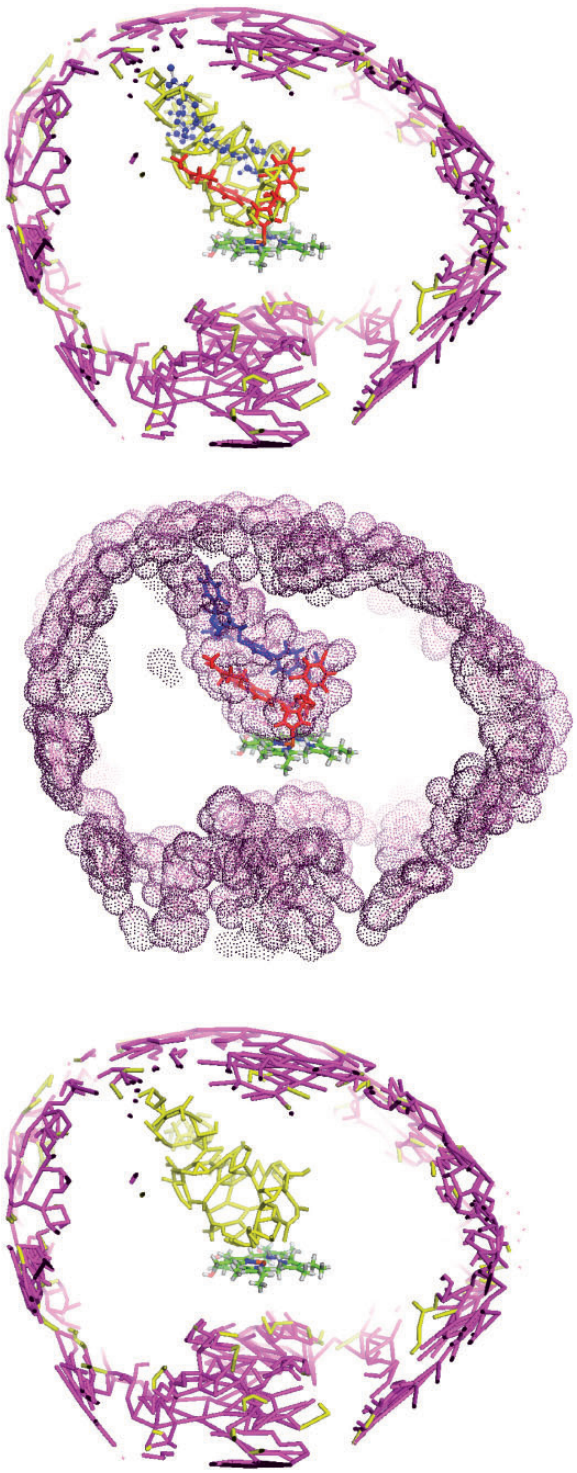

Fig. 3. The facial graph of the CYP 3A4 channels network available for the ketoconazole, retrieved from PDB data, code 2V0M, visualized with PyMOL with the two ketoconazoles (top). Other display mode with virtual spheres centered on the facial graph nodes (middle). The facial graph without the ketoconazoles (main channel part in yellow) (bottom)

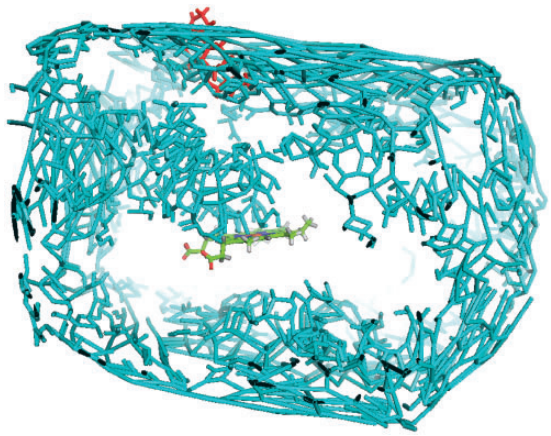

Fig. 4. The facial graph of the CYP 3A4 channels network available for the progesterone, retrieved from PDB data, code 1W0F, visualized with PyMOL. The progesterone lies on a pocket surface on the top

1TQN and 1W0E. Both the initial spatial structures and the reconstructed ones slightly differ, but the central part of the channels network is visible in front of the distal face of the heme. The critical thickness above which the access to the heme disappears was computed at 6.25 Å (1TQN) and at 5.25 Å (1W0E). The same values were got without reconstruction.

The critical thickness has been computed for other 3A4 complexes, such as with metyrapone and ritonavir, PDB codes 1W0G and 3NXU, respectively. For 3NXU, there are two chains A and B. For these letters, the SDM algorithm did not found significant differences between the backbones: all the 457 C_α were in the common 3D motif. The critical thickness was 5.0 Å for 1W0G and 6.0 Å for 3NXU. All critical thickness values we computed for the 3A4 ranged in the interval 5–6 Å. The differences are interpretable in terms of malleability of the enzyme: some CYPs have been observed to adopt different open and closed conformations when bound to different ligands (Yu *et al.*, 2013).

No objective numerical criterion exists to compare channels computed from different algorithms. That fact is mainly due to the deep differences in the nature of the output produced by the softwares, thus introducing a major difficulty to compare channels. For instance, let us compare the channels of the CYP 3A4 computed by Cojocaru *et al.* (2007) with CAVER, cited in Table 1. The main differences are:

- CCCPP produced channels depending on the shape and size of the ligand, i.e. the routes sterically available to a small ligand were more numerous that the routes available to a large ligand, although CAVER produced 'universal' channels, i.e. independent of the ligand.
- CCCPP concluded about the sterical ability of a given ligand to pass or not through the channels, although CAVER did not.
- Owing to the examples presented in this section for the CYP 3A4, CCCPP found a complex topology of the network of channels, all of them being in intersection, although three 3A4 channels are mentioned by Cojocaru *et al.* (2007).

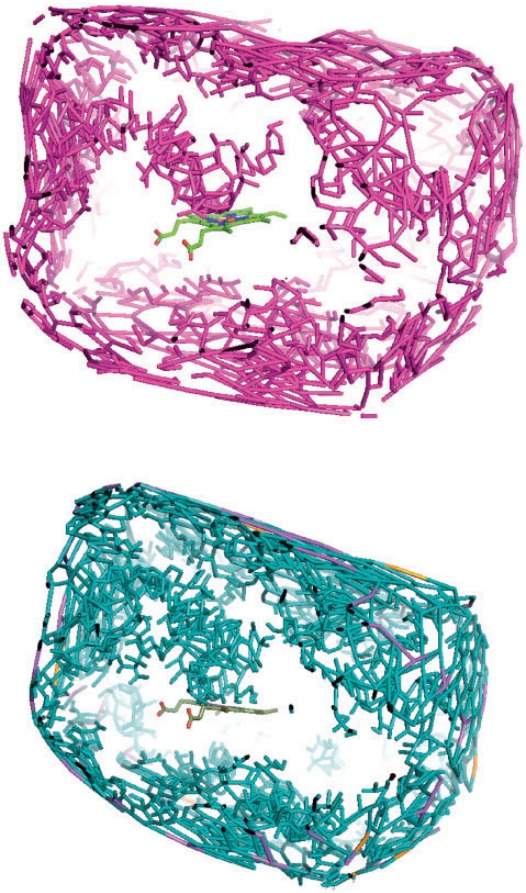

Fig. 5. The facial graph of the CYP 3A4 channels network retrieved from PDB data, codes 1TQN (top) and 1W0E (bottom), visualized with PyMOL

Since these CAVER channels (visually, sequences of intersecting spheres) were described from their closeness (not numerical) to specific loops and helices of the CYP 3A4, it is impossible to compute some degree of similarity they offer with our single network of channels got for a specific ligand. Nevertheless, on the basis of a visual checking, the egress of the CAVER channels were found to be seemingly correct by CCCPP, although the network we found was not indicated by CAVER. CYP 3A4 channels were briefly evoked by Berka *et al.* (2012) as an example of use of a MOLE2, a variant of MOLE, this latter being considered as returning similar results to CAVER (Petřek *et al.*, 2007). The authors of MOLE2 found four channels but did not perform channels comparisons with the three CAVER channels mentioned above, although CAVER and MOLE are algorithms having much in common. Anyway, we insist that the user must restrict channels comparisons to results produced by a common software, or at least produced by softwares based on similar algorithms. CCCPP computes only voids, not binding sites. Numerical comparisons of void calculations need at least that there exist computer representations of the voids as numerical objects. Even though these computerized objects are unambiguously defined, most of the times they have deeply different meanings because they are issued from different methods.

Comparing objects of deeply different nature is not always possible because it needs to define a numerical criterion to compare these objects. It is stressed that the objects differ in nature, not by the values of their measured parameters: there is no list of parameters common to the objects. In the case of CCCPP, the computerized objects are non-connected graphs, in which each node of the graph is associated to a tetrahedra (but not to an atom) of the Delaunay triangulation of the protein. Actually none of the methods cited in Table 1 offers a computer representation compatible with the one of CCCPP (thus CCCPP offers indeed a new insight on the voids), even with those based on Delaunay triangulations (e.g. CASTp, MOLE), thus precluding extensive numerical comparisons with the help of data banks. This is a major difference with the computation of binding sites, because these latter can be checked from experimental values stored in data banks.

The examples given in this section are for illustration of our methodology: a deeper investigation of the CYP450 channels is outside the scope of this article. Reviews on specific members of this enzyme's family are available in Wade *et al.* (2004) and in Cojocaru *et al.* (2007). The computing times were around 2 min on linux and macosx workstations. Nearly a half of this time is devoted to compute the Delaunay triangulation, and most of the rest of the time is used to generate the output molecular file containing the network for display. Once the triangulation done, finding pathways for a ligand from the exterior of the protein to an end point in the center of the enzyme is done in less than 1 s. Also, we point out that CCCPP did not predict binding pockets. Computing pockets can be a preliminary step to the computation of binding pockets, but the latter problem is out of the scope of that paper [see Chen *et al.* (2011) for a recent review].

4 CONCLUSION

We presented a new algorithm to compute cavities, channels and pockets in proteins, and we developed a software called CCCPP which implements this algorithm and outputs the full network of cavities. The void parts are bounded by the convex hull of the atoms of the protein. The structure of the network is stored in a text file and is written in a molecular file to be displayed with any molecular viewer. It was proved that cylindrical ligands shapes are much more realistic than spherical ligands shapes. On the basis of this unusual cylindrical model, CCCPP was successful in explaining how large molecules can be admitted in front of the buried heminic active site of CYP450s. For these latter, CCCPP indicated that there is a complex network of channels rather than a small finite number of channels. It also indicated that the final part of the network in front of the heme lies at its distal face, which is in agreement with physico chemical data. We do not claim that our methodology is superior to the tenths of available other ones. We just point out that our methodology gave rise to an alternative tool providing new information about pockets and channels in proteins. It can be used in a virtual screening context, eventually in addition to energy-based methods.

Funding: Programme Doctoral International 'Modélisation des Systèmes Complexes', IRD (Institut de recherche pour le développement) Paris 6, France (to L.B.) and UPMC (Université Pierre & Marie Curie), Paris 6, France (to L.B.); Integrative BioComputing (IBC company), Rennes, France (to L.B.).

Conflict of Interest: none declared.

REFERENCES

Azuara,C. *et al.* (2006) PDB_Hydro: incorporating dipolar solvents with variable density in the Poisson-Boltzmann treatment of macromolecule electrostatics. *Nucleic Acids Res.*, **34**, W38–W42.

Benkaidali,L. *et al.* (2012) How well is conformational space covered? In: Putz,M. (ed.) *Chemical Information and Computational Challenges in the 21st Century*. NOVA Science Publishers, New York, pp. 299–313.

Berka,K. *et al.* (2012) MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels. *Nucleic Acids Res.*, **40**, W222–W227.

Bondi,A. (1964) Van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441–451.

Brady,G.P. and Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, **14**, 383–401.

Brandenberg,R. and Theobald,T. (2006) Radii minimal projections of polytopes and constrained optimization of symmetric polynomials. *Adv. Geom.*, **6**, 71–83.

Chen,K. *et al.* (2011) A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds. *Structure*, **19**, 613–621.

Chovancova,E. *et al.* (2012) CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.*, **8**, e1002708.

Cojocaru,V. *et al.* (2007) The ins and outs of cytochrome P450s. *Biochim. Biophys. Acta*, **1770**, 390–401.

Coleman,R.G. and Sharp,K.A. (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J. Mol. Biol.*, **362**, 441–458.

Cortés,J. *et al.* (2011) Encoding molecular motions in voxel maps. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **8**, 557–563.

David,C.W. (1988) Vornoi polyhedra as structure probes in large molecular systems–VII. Channel identification. *Comput. Chem.*, **12**, 207–208.

Dupuis,F. (2003) Tessellations de Voronoï appliquées aux structures protéiques. PhD Thesis, Université Denis Diderot, Paris 7, France.

Dupuis,F. *et al.* (2005) Voro3D: 3D Voronoi tessellations applied to protein structures. *Bioinformatics*, **21**, 1715–1716.

Edelsbrunner,H. (1987) Voronoi diagrams. In: Brauer,W., Rozenberg,G. and Salomaa,A. (eds) *Algorithms in Combinatorial Geometry*. Springer-Verlag, Berlin, pp. 293–334.

Edelsbrunner,H. (1995) The union of balls and its dual shape. *Discrete Comput. Geom.*, **13**, 415–440.

Edelsbrunner,H. and Mücke,E.P. (1992) Three-dimensional alpha shapes. In: *Proceedings of the 1992 Boston Workshop on Volume Visualization*. ACM, pp. 75–82.

Edelsbrunner,H. and Mücke,E.P. (1994) Three-dimensional alpha shapes. *ACM Trans. Graphics*, **13**, 43–72.

Edelsbrunner,H. *et al.* (1983) On the shape of a set of points in the plane. *IEEE Trans. Info. Theory*, **29**, 551–559.

Edelsbrunner,H. *et al.* (1998) On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.*, **88**, 83–102.

Gavezzotti,A. (1983) The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity. *J. Am. Chem. Soc.*, **105**, 5220–5225.

Guengerich,F.P. (2005) Human cytochrome P450 enzymes. In: Ortiz de Montellano,P.R. (ed.) *Cytochrome P450, Structure, Mechanism, and Biochemistry*. 3rd edn. Kluwer/Plenum, New York, pp. 377–530.

Hendlich,M. *et al.* (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, **15**, 359–363.

Ho,C.M. and Marshall,G.R. (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput. Aided Mol. Des.*, **4**, 337–354.

Jaillet,L. *et al.* (2010) Sampling-based path planning on configuration-space costmaps. *IEEE Trans. Robotics*, **26**, 635–646.

Kalidas,Y. and Chandra,N. (2008) PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.*, **161**, 31–42.

Kawabata,T. and Go,N. (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins*, **68**, 516–529.

Kleywegt,G.J. and Jones,T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D*, **50**, 178–185.
Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.
Lee,P.H. and Helms,V. (2012) Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. *Proteins*, **80**, 421–432.
Le Guilloux *et al.* (2009) Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
Levitt,D.G. and Banaszak,L.J. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.*, **10**, 229–234.
Lewis,R.A. (1989) Determination of clefts in receptor structures. *J. Comput. Aided Mol. Des.*, **3**, 133–147.
Mach,P. and Koehl,P. (2011) Geometric measures of large biomolecules: surface, volume, and pockets. *J. Comput. Chem.*, **32**, 3023–3038.
Masuya,M. and Doi,J. (1995) Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *J Mol. Graph.*, **13**, 331–336.
Meslamani *et al.* (2009) Assessing the geometric diversity of cytochrome P450 ligand conformers by hierarchical clustering with a stop criterion. *J. Chem. Inf. Model.*, **49**, 330–337.
Nayal,M. and Honig,B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, **63**, 892–906.
Pellegrini-Calace *et al.* (2009) PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PloS. Comput. Biol.*, **5**, e1000440.
Pérot,S. *et al.* (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, **15**, 656–667.
Peters,K.P. *et al.* (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, **256**, 201–213.
Petitjean,M. (1992) Applications of the Radius-Diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **32**, 331–337.
Petitjean,M. (1994) On the analytical calculation of van der Waals surfaces and volumes: some numerical aspects. *J. Comput. Chem.*, **15**, 507–523.
Petitjean,M. (1998) Interactive maximal common 3D substructure searching with the combined SDM/RMS algorithm. *Comp. Chem.*, **22**, 463–465.
Petitjean,M. (2012) About the algebraic solutions of smallest enclosing cylinders problems, *Appl. Alg. Eng. Comm. Comp.*, **23**, 151–164.
Petitjean,M. (2013) Spheres unions and antersections and some of their applications in molecular modeling. In: Mucherino,A. *et al.* (ed.) *Distance Geometry: Theory, Methods, and Applications*. Springer, New York, pp. 61–83.
Petřek,M. *et al.* (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, **7**, 316.
Petřek,M. *et al.* (2007) MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, **15**, 1357–1363.
Preparata,F.P. and Shamos,M.I. (1985) Convex hulls: basic algorithms. In: *Computational geometry*. Springer-Verlag, Berlin, pp. 95–149.
Rendic,S. (2002) Summary of information on human CYP enzymes: human P450 metabolism data. *Drug. Metab. Reviews*, **34**, 83–448.
Richards,F.M. (1985) Calculation of molecular volumes and areas for structures of known geometry. *Meth. Enzymol.*, **115**, 440–464.
Schmidtke,P. *et al.* (2010) Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.*, **38**, W582–W589.
Scott,R.A. and Scheraga,H.A. (1966) Conformational analysis of macromolecules. III. Helical structures of polyglycine and poly-L-alanine. *J. Chem. Phys.*, **45**, 2091–2101.
Smart,O.S. *et al.* (1993) The pore dimensions of gramicidin A. *Biophys J.*, **65**, 2455–2460.
Smart,O.S. *et al.* (1996) HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graphics*, **14**, 354–360.
Smart,O.S. *et al.* (1997) A novel method for structure-based prediction of ion channel conductance properties. *Biophys J.*, **72**, 1109–1126.
Sridhar,J. *et al.* (2011) QSAR models of cytochrome P450 enzyme 1A2 inhibitors using CoMFA, CoMSIA and HQSAR. *SAR QSAR Environ. Res.*, **22**, 681–697.
Sonavane,S. and Chakrabarti,P. (2008) Cavities and atomic packing in protein structures and interfaces. *PLoS Comput. Biol.*, **4**, e1000188.
Sperandio,O. *et al.* (2006) Receptor-based computational screening of compound databases: the main docking-scoring engines. *Curr. Prot. Peptide Sci.*, **7**, 369–393.
Stahl,M. *et al.* (2000) Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.*, **13**, 83–88.
Till,M.S. and Ullmann,G.M. (2010) McVol - A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.*, **16**, 419–429.
Tripathi,A. and Kellogg,G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins*, **78**, 825–842.
Tseng,Y.Y. and Li,W.-H. (2009) Identification of protein functional surfaces by the concept of a split pocket. *Proteins*, **76**, 959–976.
Tseng,Y.Y. *et al.* (2009) SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.*, **37**, W384–389.
Venkatachalam,C.M. *et al.* (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.*, **21**, 289–307.
Wade,R.C. *et al.* (2004) A survey of active site access channels in cytochromes P450. *J. Inorg. Biochem.*, **98**, 1175–1182.
Weisel,M. *et al.* (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Central J.*, **1**, 7.
Yaffe,E. *et al.* (2008a) MolAxis: efficient and accurate identification of channels in macromolecules. *Proteins*, **73**, 72–86.
Yaffe,E. *et al.* (2008b) MolAxis: a server for identification of channels in macromolecules. *Nucleic Acids Res.*, **36**, W210–W215.
Xie,L. and Bourne,P.E. (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics*, **8(Suppl. 4)**, S9.
Yu,J. *et al.* (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, **26**, 46–52.
Yu,X. *et al.* (2013) Conformational diversity and ligand tunnels of mammalian cytochrome P450s. *Biotechnol. Appl. Biochem.*, **60**, 134–145.
Zaretzki,J. *et al.* (2012) RS-Predictor models augmented with SMARTCyp reactivities: Robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model.*, **52**, 1637–1659.
Zefirov,Y.V. and Zorkii,P.M. (1989) Van der Waals radii and their application in chemistry. *Russ. Chem. Rev.*, **58**, 421–440.
Zhong,S. and MacKerell,A.D. (2007) Binding response: a descriptor for selecting ligand binding site on protein surfaces. *J. Chem. Inf. Model.*, **47**, 2303–2315.