

# MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories

Peter Schmidtke<sup>1,\*</sup>, Axel Bidon-Chanal<sup>2</sup>, F. Javier Luque<sup>1</sup> and Xavier Barril<sup>1,3</sup>

<sup>1</sup>Departament de Fisicoquímica and Institut de Biomedicina (IBUB), Facultat de Farmàcia, Universitat de Barcelona, 08028 Barcelona, Spain, <sup>2</sup>Equipe de Dynamique des Assemblages Membranaires, UMR No. 7565, Centre National de la Recherche Scientifique-Université Henri Poincaré, Nancy, France and <sup>3</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA)

Associate Editor: Anna Tramontano

## ABSTRACT

**Motivation:** A variety of pocket detection algorithms are now freely or commercially available to the scientific community for the analysis of static protein structures. However, since proteins are dynamic entities, enhancing the capabilities of these programs for the straightforward detection and characterization of cavities taking into account protein conformational ensembles should be valuable for capturing the plasticity of pockets, and therefore allow gaining insight into structure–function relationships.

**Results:** This article describes a new method, called MDpocket, providing a fast, free and open-source tool for tracking small molecule binding sites and gas migration pathways on molecular dynamics (MDs) trajectories or other conformational ensembles. MDpocket is based on the fpocket cavity detection algorithm and a valuable contribution to existing analysis tools. The capabilities of MDpocket are illustrated for three relevant cases: (i) the detection of transient subpockets using an ensemble of crystal structures of HSP90; (ii) the detection of known xenon binding sites and migration pathways in myoglobin; and (iii) the identification of suitable pockets for molecular docking in P38 Map kinase.

**Availability:** MDpocket is free and open-source software and can be downloaded at <http://fpocket.sourceforge.net>.

**Contact:** pschmidtke@ub.edu

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on July 6, 2011; revised on August 28, 2011; accepted on September 27, 2011

## 1 INTRODUCTION

Over the past two decades, a variety of algorithms have been proposed with the aim to identify binding pockets for small molecules in biomolecular targets (An *et al.*, 2005; Brady and Stouten, 2000; Hendlich *et al.*, 1997; Kim *et al.*, 2008; Kleywegt and Jones, 1994; Laskowski, 1995; Laurie and Jackson, 2005; Le Guilloux *et al.*, 2009; Peters *et al.*, 1996; Weisel *et al.*, 2007). These algorithms can be classified into three broad classes depending on the general method used for cavity detection: (i) geometry-based, (ii) energy-based or (iii) sequence-based methods. The first class relies on geometrical features of pockets with no or few

considerations for the interaction energy between a putative ligand and the pocket. Energy-based algorithms estimate the suitability of a pocket to bind a molecule using probe–pocket interaction energy calculations. The latter methods are usually computationally more expensive and require specific atom typing and the use of underlying force fields. Sequence-based methods exploit the propensity of conserved residues in the binding site. Last, various hybrid methods, combining at least two of the previous approaches, also exist (Halgren, 2007; Huang, 2009; Liang *et al.*, 1998). The reader is directed to recent reviews for a detailed discussion of different cavity detection algorithms (Henrich *et al.*, 2010; Leis *et al.*, 2010; Pérot *et al.*, 2010).

The vast majority of cavity detection algorithms have been developed to treat static structures, like crystal structures of proteins available in the protein data bank (PDB). However, this represents a serious limitation to account for the intrinsic plasticity of the binding pocket. Protein dynamics act on a multitude of aspects in protein function. For instance, side chain flipping or domain motions can obstruct or free internal cavities or channels that allow migration of ligands and reshape the binding sites (Bidon-Chanal *et al.*, 2006; Carrillo and Orozco, 2008; Spyraakis *et al.*, 2011). In turn, these findings raise challenging questions about the impact of protein flexibility on the topological features of cavities and their binding properties.

Few works have attempted to account for the dynamical behaviour of proteins in the identification of binding cavities and tunnels. The interplay between protein dynamics and ligand migration pathways can be characterized by tools that rely on prior molecular dynamics (MDs) simulations and further post-processing of the trajectory. VOIDOO (Kleywegt and Jones, 1994) allows internal cavity and volume calculations, but it is rather time consuming and its use is not straightforward. CAVER (Beneš *et al.*, 2010) is a PyMOL plugin that allows internal channel detection on MD trajectories. CAVER was recently improved to a software called MOLE using computational geometry principles instead of grid-based calculations (Petrek *et al.*, 2007). More recently, Wolfson and coworkers have proposed a method called MolAxis for detection of channels from the interior of the protein to the bulk solvent (Yaffe *et al.*, 2008). These methods are designed to detect channels on static structures or conformational ensembles of the protein.

In order to examine the suitability of internal pathways for ligand migration, more time consuming techniques like implicit ligand sampling have been proposed (Cohen *et al.*, 2008). Glazer *et al.*

\*To whom correspondence should be addressed.

(2009) have used MDs for the identification of calcium binding sites and have shown that the inclusion of dynamic behaviour can improve function prediction. Recently, a MATLAB-based approach called Dynamic Map Ensemble (DyME) (Lin and Song, 2011) has been proposed for the analysis of putative internal channels from MD trajectories using Voronoi tessellation and clustering techniques. Regarding the binding of ligands in cavities, protein dynamics has been accounted for by using a set of privileged static structures, which are chosen as representative conformational states of the pocket based on experimental X-ray structures or from MD simulations of the target (Barril and Morley, 2005; Kua *et al.*, 2002; Novoa *et al.*, 2010). More recently, Eyrisch and Helms established a protocol for the detection of transient cavities in protein–protein interfaces by using MD simulations and the cavity detection algorithm PASS (Brady and Stouten, 2000; Eyrisch and Helms, 2007). Other approaches combine conformational sampling and/or selection from MD simulations to allow pocket adaptation to a given ligand positioned in the binding site (Sherman *et al.*, 2006). Finally, a distinct strategy has been adopted in PELE (Borrelli *et al.*, 2005), as internal pockets and channels are identified based on localized perturbations that combine Monte Carlo sampling and energy minimization calculations to track and predict ligand migration pathways.

In this work, a new generic pocket detection program called MDpocket is presented. The aim of MDpocket is to identify and characterize binding sites and channels that might be transiently formed in the protein from the analysis of conformational ensembles generated by MDs or other sources. The core of this new program is the recently published open-source platform Fpocket (Le Guilloux *et al.*, 2009), which is a very fast geometry-based cavity detection algorithm. The platform relies on three programs: (i) *fpocket*, which identifies cavities in a protein; (ii) *dpocket*, which extracts descriptors of the pocket; and finally (iii) *tpocket*, which allows assessment of pocket scoring functions. One main advantage of *fpocket* is its adaptability to a given problem: initially developed for the discovery of small molecule binding sites, it can detect different types of cavities, including very small pockets, ligand binding sites or even tunnels via a proper choice of parameters available through command line.

MDpocket is fast and well suited for the study of processes where tracking of cavities is of interest. In particular, the capabilities of MDpocket are illustrated by examining (i) the plasticity of the HSP90 binding site using a set of X-ray crystallographic structures, (ii) the analysis of xenon binding sites and migration pathways in myoglobin (Mb) and (iii) the selection of suitable docking sites for P38 (asp-phe-gly) DFG-in binders.

The software is freely available as part of the Fpocket software package and can be downloaded from <http://fpocket.sourceforge.net>.

## 2 METHODS

### 2.1 MDpocket input

The general input format of MDpocket is a text file listing filenames to all pdb files to be considered for the analysis. This choice is motivated by the fact that MD trajectories are stored in different file formats depending on the specifications defined in programs such as Amber (Case *et al.*, 2005), Charmm (Brooks *et al.*, 2009; MacKerel *et al.*, 1998), Gromacs (Hess *et al.*, 2008) or NAMD (Phillips *et al.*, 2005). Due to the lack of a common format, we have decided to transform the trajectory to a set of pdb files corresponding

to snapshots taken along the simulation. Moreover, when those files are ordered by time, MDpocket permits the analysis of time-dependent events. In addition, the use of pdb files also facilitates the analysis of X-ray structures taken from the PDB. Finally, the PDB files do not need to be identical (no generic topology required), which makes MDpocket easily adaptable to analyse conformational ensembles from various sources (homologous proteins, for instance).

To carry out the cavity detection with MDpocket, it is important to superimpose the PDB structures onto each other. To this end, solvent molecules and counter ions were stripped off the system prior to pdb export, and then structural alignments were carried out using ptraj from AmberTools (Case *et al.*, 2005).

### 2.2 fpocket parameters and output

MDpocket relies on the pocket detection program *fpocket*, which makes extensive use of Voronoi tessellation during cavity detection. This geometric approach allows retrieving without the  $\alpha$ -spheres (i.e. spheres that are in contact with exactly four atoms without any other atom situated within the sphere). The centre of the  $\alpha$ -sphere corresponds to a Voronoi vertex. A list of all Voronoi vertices (clustered into pockets) situated on the protein surface is provided in the output of *fpocket*.

The *fpocket* module is very flexible regarding the type of cavity to be detected. The flexibility is achieved through user accessible command line parameters that influence filtering and clustering of  $\alpha$ -spheres. The most important parameters are those that define the size of  $\alpha$ -spheres built up in a binding site (-m: minimum  $\alpha$ -sphere size; -M: maximum  $\alpha$ -sphere size). Moreover, filtering and clustering of  $\alpha$ -spheres can be modified using parameters -i (the minimum number of  $\alpha$ -spheres in the final pocket) and -n (the minimum number of  $\alpha$ -spheres close to each other for merging two binding sites into a single one).

Three different parameter sets for pocket detection have been assessed here in order to illustrate the scalability of the algorithm. Set 1 denotes the default *fpocket* parameter set (-m 3.0, -M 6.0, -i 30, -n 3), which is tailored for detection of small molecule (i.e. peptides, drug-like compounds) binding sites. Set 2 is intended to identify very small channels and pockets (-m 2.8, -M 6.0, -i 3, -n 2). Finally, Set 3 (-m 3.5, -M 5.5, -i 1, -n 2) is chosen to represent an  $\alpha$ -sphere with a physically meaningful minimum size, while retaining all the pockets (even tiny ones built by a single  $\alpha$ -sphere), it being thus better suited to identify very open cavities physically accessible to a water molecule and detection of continuous channels that can accommodate a water molecule.

### 2.3 MDpocket workflow: pocket detection

Pockets are detected based on the workflow depicted in Supplementary Figure S1:

- A 1 Å spaced grid is placed over the first snapshot of the set of superposed pdb files.
- fpocket* is run on every snapshot of the set of pdb files.
- For every snapshot, each  $\alpha$ -sphere is assigned to the grid point  $i$  closest to the  $\alpha$ -sphere centre. Note that several spheres can be assigned to a given point of the grid originating from the same snapshot or  $\alpha$ -from different snapshot. The number of counted  $\alpha$ -spheres assigned to each grid point is then normalized by the number of snapshots in order to generate the density map,  $\rho$  [Equation (1)].

$$\rho_i = \frac{1}{n} \sum_{i=1}^n d_{AS,i} \quad (1)$$

where  $n$  corresponds to the number of snapshots, AS stands for  $\alpha$ -sphere and  $d_{AS,i}$  is determined as

$$d_{AS,i} = \text{card}(AS_i \cap \{(x, y, z)_i \pm 0.5\}) \quad (2)$$

where the tolerance of 0.5 is considered for all grid points.

- (d) In addition, for every snapshot each grid point  $i$  is given an occupancy parameter  $\delta_i$ , which equals 1 or 0 depending on the previous assignment of  $\alpha$ -spheres to that point ( $\delta_i = 1$  if at least one  $\alpha$ -sphere has been assigned to the grid point  $i$ , otherwise  $\delta_i = 0$ ). Then, a frequency map,  $\phi$ , is generated by normalizing the sum of values  $\delta_i$  assigned to point  $i$  by the total number of snapshots [Equation (3)].

$$\Phi_i = \frac{1}{n} \sum_{i=1}^n \delta_i \quad (3)$$

The normalized pocket frequency map [Equation (3)] allows visualization of the opening frequency of a pocket during a MD trajectory. Thus, it indicates whether a given point in the grid is permanently accessible ( $\Phi_i = 1$ ), hindered ( $\Phi_i = 0$ ) or transiently accessible ( $0 < \Phi_i < 1$ ). In contrast, the pocket density map [Equation (1)] is intended to provide information about the environmental atom packing around the pocket. Both density and frequency maps can be visualized using VMD (Humphrey *et al.*, 1996), Chimera (Pettersen *et al.*, 2004) or PyMOL from the output files produced by MDpocket.

In contrast to the pocket detection on a single static structure, MDpocket provides information about the plasticity of pockets from the normalized frequency/density maps generated from a given ensemble of structures. This is a major difference to other approaches that assign discrete pocket IDs to track pockets during MD trajectories (Eyrich and Helms, 2007). In MDpocket, an accurate pocket ID identification and tracking is not necessary, rendering the detection protocol more generic and less error prone.

## 2.4 MDpocket workflow: pocket characterization

Frequency and density maps are valuable to explore pocket opening/closure. MDpocket also permits to characterize those pockets or binding sites by providing a variety of descriptors, which include the accessible surface area and volume of the pocket, the number of  $\alpha$ -spheres and the mean local hydrophobic density, which is an index of binding site druggability (Schmidtke and Barril, 2010).

To carry out the pocket characterization, the user can extract all grid points having a grid value equal or higher than a certain threshold from the previously calculated pocket frequency map (a default value of 0.5 is defined in MDpocket). Thus, visualization of the frequency map permits the user to select an area of interest (i.e. a transient channel or a binding site) using a graphical display tool, and the user-defined zone (saved as pdb file) can then be used as input for MDpocket in order to determine all pocket descriptors corresponding to the selected area for the whole ensemble.

## 2.5 MDpocket validation

The usefulness and accuracy of MDpocket have been calibrated considering three molecular systems studied previously in our group.

**HSP90:** a 78.5 ns trajectory run of the N-terminal domain of the heat shock protein 90 (HSP90) with explicit solvent (TIP3P water model) was first considered. The simulation was run in the NPT ensemble (1 atm, 298 K) using periodic boundary conditions and Ewald sums (grid spacing of 1 Å) for long-range electrostatic interactions. The parm99 force field and the Amber (Case *et al.*, 2005) package were also used. From this trajectory, 3925 equally spaced snapshots were extracted and analysed with MDpocket. Furthermore, an alternative ensemble of structures was built up by retrieving 88 X-ray crystallographic structures from the PDB (Supplementary Table S1), which were subsequently aligned using PyMOL.

**Mb:** the crystal structure of Mb (PDB entry 1VXD) (Yang and Phillips, 1996) was immersed in an octahedral box of TIP3P water molecules and the net charge of the system was neutralized with sodium ions. The final system contained ~21 000 atoms. The simulations were run using the PMEMD module of amber9 and the parm99 force field with special parameters for the haem residue (Bidon-Chanal *et al.*, 2006; Marti *et al.*, 2006).

The SHAKE algorithm was used to keep bonds involving hydrogen atoms at their equilibrium length, in conjunction with a 1 fs time step for the integration of the Newton's equations. Trajectories were collected in the NPT ensemble (1 atm, 298 K) using periodic boundary conditions and Ewald sums (grid spacing of 1 Å) for long-range electrostatic interactions. The systems were minimized using a multi-step protocol, involving first the adjustment of hydrogens, then the refinement of water molecules and finally the minimization of the whole system. The equilibration was performed by heating from 100 to 298 K in four 100 ps steps at 150, 200, 250 and 298 K. Finally, a 50 ns trajectory was obtained, collecting frames at 1 ps intervals. The MDpocket analysis was performed with 10 000 snapshots equally spaced in time.

**P38 Map kinase:** the PDB structure 1P38 was used as initial structure for a 50 ns MD trajectory. Leap was used to immerse the protein in an octahedral solvent box. The overall charge of the system was neutralized by the addition of counterions. The solvent box contained a mixture of water and 20% isopropanol molecules. In order to obtain more information about the equilibration protocol, refer to Seco *et al.* (2009). The production run was carried out at 1 atm and 300 K using periodic boundary conditions. In all, 5000 snapshots equally spaced in time have been used for the MDpocket analysis.

To assess whether MDpocket is able to give useful hints during the selection of receptor conformations for molecular docking, 32 X-ray crystallographic structures of P38 with DFG-in conformations and bound ligands were extracted from the PDB (Supplementary List S2), and aligned to all snapshots of the MDs using the C $\alpha$  atoms of residues 35–39, 45–50 and 100–104, which correspond to the stable part of the  $\beta$ -sheet lining the binding site. In order to extract the interaction energies for each DFG in ligand, the aligned ligand was extracted from the crystal structure and added to each snapshot of the MD trajectory to calculate the interaction energy. All energy calculations were performed using molecular operating environment (MOE) (Chemical Computing Group, 2009), and the default potential energy function with the merck molecular force field (MMFF) force field. No modifications or conformational changes were applied to the ligands near the residues in the binding site. Thus, this very crude interaction energy evaluation should mainly give insights into steric clashes that could occur in the ligand–protein complex, if the ligand is docked in a given conformation of the protein sampled during the MD trajectory.

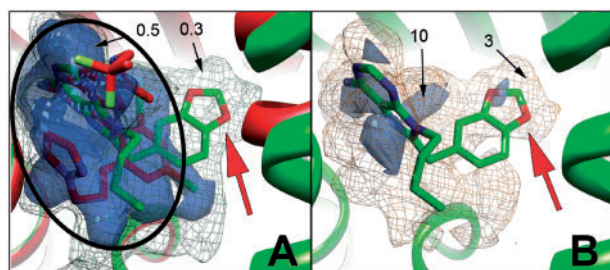
## 3 RESULTS

### 3.1 Spotting and interpreting variation on structural ensembles

Although protein dynamics and flexibility can be crucial for the recognition with other proteins or small ligands, it is still an often-neglected aspect in structural analysis. Insights into protein flexibility can be gained from the structural differences observed in X-ray structures or by using NMR-derived data. In both cases, expensive equipment is needed and sample preparation can be a real handicap. On the other hand, MDs represent a powerful theoretical tool to explore the dynamical behaviour of biomolecules and MDpocket is intended to take advantage of this fact to identify and characterize pockets and binding sites in proteins from the analysis of structural ensembles chosen to account for the intrinsic flexibility of proteins.

The capability of MDpocket is first explored by considering the heat shock protein 90 (HSP90), which was chosen because the available X-ray data reveals a transient opening of a hydrophobic subpocket connected to the known ATP binding site. This opening occurs when loop 2 (residues N105–A111) in the helix 4–loop 2–helix 5 motif reorganizes to an  $\alpha$ -helix (Wright *et al.*, 2004).





**Fig. 1.** HSP90 binding site derived from 88 X-ray crystallographic structures. (A) MDpocket pocket frequency map at 50% (blue iso-surface) and 30% (mesh). The red structure corresponds to a crystal structure where the sub-pocket is closed. The green structure has the sub-pocket open. The main pocket (black ellipse) is found in all snapshots. The subpocket (red arrow) is open in 35.2% of all X-ray structures, and the isosurface determined from MDpocket (mesh) appears at 35% of pocket appearance frequency. (B) MDpocket pocket density map (for clarity only the green structure is shown) at two levels of pocket density. The main pocket is found at low (3, yellow mesh) and high (10, blue surface) densities. The subpocket is also found at low (3) densities and a spot (blue surface) can also be seen at high densities despite the fact that the pocket does not open frequently.

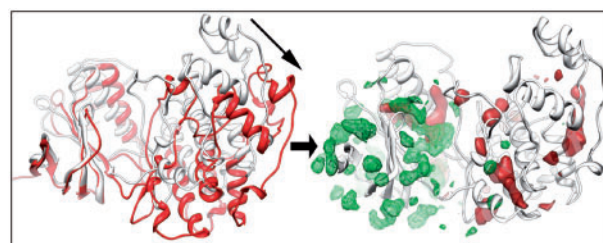
A total of 88 different PDB structures (Supplementary Table S1) were superimposed to the reference structure 1BYQ using PyMol. In this set, 31 structures have a straight helix conformation (corresponding to the open subpocket), whereas the rest of the conformations show predominance of loop 2 and thus the subpocket does not exist. Next, MDpocket was run on all superimposed structures.

Figure 1A shows the HSP90 binding site with a ligand in the main binding site (red protein and ligand) and another ligand filling the open subpocket (red arrow). The pocket frequency map, which gives the amount of time a pocket was found, is represented at two different isocontours (0.5, blue isosurface; 0.3, orange mesh). At the lower isovalue (0.3, corresponding to minimum 30% pocket opening on all conformations), the subpocket is open. The isovolume corresponding to the subpocket disappears above 35% and is not visible anymore for 50% of all conformations (blue isosurface). This result is in agreement with our prior knowledge that the pocket is open in ~35.2% of all conformations.

Figure 1B illustrates the pocket density map, which provides information on the density of  $\alpha$ -spheres in a pocket. The density map is represented at very high (10, blue isosurface) and low (3, orange mesh) isocontours. Interestingly, one can observe a highly dense isosurface in the subpocket (red arrow), although it just opens 35.2% in all conformations. Thus, the density map can give useful insights about the relative enclosure of different regions of a binding site regardless of its frequency of appearance.

### 3.2 Influence of the structural alignment on the results

The detection of pockets with MDpocket can be affected by the structural alignment of the frames included in the ensemble of structures. The influence of structural alignment on the results has been explored for 3925 snapshots taken from a 50 ns trajectory of P38 Map kinase, which was chosen due to the flexibility of the two lobes that define the ATP binding site (Fig. 2). Thus, Figure 2 shows how the alignment of the small lobe in snapshots taken at the



**Fig. 2.** Representative snapshots showing the potential effect of alignment on MDpocket results for P38 Map kinase. An important motion of one lobe versus the other is observed during the MD trajectory (all snapshots have been aligned using the small lobe as reference), white ribbon—beginning of the trajectory, red ribbon—end of the trajectory (left). Pocket frequency maps at 40% of pocket opening for (green) a trajectory aligned with residues lining the active site, and (red) a trajectory aligned with all residues (right).

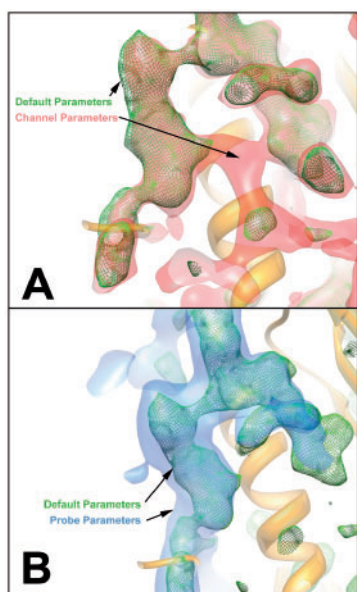
beginning and end of the trajectory leads to a large displacement in the bigger lobe, which reflects the relative motion between the two lobes along the trajectory.

Next, two scenarios have been considered. In the first case, the superimposition of all structures was done using the C $\alpha$  atoms of all protein residues (global alignment). Second, only the C $\alpha$  atoms of a stable part of the small lobe (the  $\beta$ -sheet lining the binding site; local alignment) were considered for superimposition. The MDpocket frequency maps at an isocontour of 0.4 reveals major differences between the two alignments regarding the pocket appearance. The green mesh in Figure 2 displays MDpocket results derived for the local alignment. Several conserved pockets can be observed on the small lobe and also the active site is well identified. However, no clear pocket is found on the bigger lobe. This is due to the large inter-lobe motions occurring during the MDs. The red isosurface reflect the same isocontour derived for the global alignment. In this case, pockets can be found on the bigger lobe, but the active site is just partially identified at this level of pocket frequency.

This example illustrates the sensitivity of the grid-based methodology implemented in MDpocket, especially when large motions between lobes or domains are involved, as one should expect that these changes will have a marked influence on the pockets. As a rule-of-thumb, it can be stated that if the aim of the MDpocket analysis is the study of one particular cavity or channel, a superimposition using the heavy atoms lining the pocket should be used. Otherwise, usage of all heavy atoms on the protein can be considered though mobile parts should be excluded for the alignment, as, pockets in contact with these mobile parts can yield underestimated pocket frequencies in MDpocket analysis.

### 3.3 Effect of *fpocket* parameters on the results

The proper choice of cavity detection parameters can be decisive for obtaining an accurate description of both tiny channels transiently formed in the protein matrix or permanent drug-like binding sites within the protein. In order to tackle these distinct situations, three different parameter sets (see Section 2) have been used to analyse the HSP90 in the region enclosing the binding site. Parameter set 1 (default) is intended to identify binding sites able to bind small molecular substrates or drug-like molecules. Results obtained for this parameter set are shown as green mesh in Figure 3A and B for the ensemble of 3925 HSP90 snapshots. Parameter set 2 is conceived



**Fig. 3.** MDpocket isosurfaces/meshes for (A) parameters set 1 (default parameters; isovalue 0.4) and set 2 (intended for channel parameters; isovalue 0.8), and (B) parameters set 1 (isovalue 0.4) and set 3 (water probe; isovalue 0.25).

for the identification of internal channels and small molecule binding sites (results shown as red isosurface in Fig. 3A). Finally, parameter set 3 is intended to support detection of cavities sterically able to host diatomic ligands or water molecules (results shown in Fig. 3B).

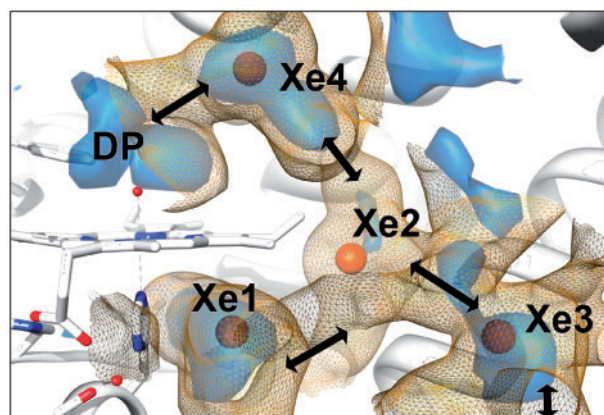
Inspection of Figure 3 reveals differences in the shape of the cavities/tunnels delineated by the isocontours generated from the three parameter sets. Thus, Figure 3A clearly shows the suitability of parameter set 2 to identify internal channels, which are nevertheless found as discontinuous regions when parameter set 1 is used. Interestingly, the putative channel observed on Figure 3A corresponds to a region that is known to be part of a larger loop, opening during the chaperoning cycle (Ali *et al.*, 2006). Likewise, Figure 3B shows that parameter set 3 discloses the solvent exposed part of the binding site, while very tiny internal channels or narrow parts of internal pockets are not identified here.

Hence, it can be concluded that putative gas or water migration pathways can be identified using parameter set 2. Indeed, a putative channel is found below the subpocket discussed earlier.

Even though parameter set 2 allows identification of tiny cavities and channels, they can be, however, physically meaningless if no molecule can fill them. Thus, calibration against pockets and channels derived using a specific probe mimicking the molecule of interest (i.e. a diatomic molecule or a water) can be valuable to gain insight into ligand migration. However, if the main interest is to track transient druggable cavities on the protein surface, default parameters (set 1) are more adequate. In any case, the user can easily adjust the parameters required for pocket detection, thus facilitating the exploration towards specific class of binding sites and channels.

### 3.4 Case 1: pocket detection in Mb

Mb is known to bind small diatomic ligands such as CO, O<sub>2</sub> or NO. Different internal cavities suited to hold small diatomic ligands



**Fig. 4.** Xenon atoms (orange spheres) from PDB structure 1J52 superimposed with MDpocket results on the Mb MD trajectory. Blue iso-surface: the pocket frequency map at 60%, which exhibits a close correspondence with all Xe binding sites. Orange iso-mesh: the pocket density map at 2 allows to discern putative migration channels from one Xe binding site to the other.

were first detected in soaking experiments of Mb crystals with Xe atoms (Tilton *et al.*, 1984). A large number of studies have been performed to characterize the ligand migration in Mb (Cohen *et al.*, 2006; Ostermann *et al.*, 2000; Schotte *et al.*, 2003; Scott and Gibson, 1997; Scott *et al.*, 2001; Tomita *et al.*, 2009). It is generally assumed that migration involves an intermittent passage between transient pockets, and the ligand entry to the distal cavity from the solvent phase. Here our aim is to show that MDpocket is a useful tool to detect those preferential Xe binding pockets, to identify the pathways that connect them in the interior of Mb and eventually detect the possible entrance for the ligand from the solvent phase from an ensemble of snapshots taken from an MD simulation.

The parameter set 2 was used to analyse 10 000 snapshots evenly taken over a 50 ns MD trajectory. As expected, the results show that frequently appearing pockets overlap with known Xe binding sites in the crystal structure (see blue isosurface in Fig. 4), which indicates that the cavities are present in the majority of the snapshots. Thus, they should generally be also detectable when only a single snapshot is used for the analysis. On the other hand, connections between cavities do not occur through long-lived channels, i.e. they are not present in the ensemble of crystal structures deposited in the PDB, which is in agreement with the notion that migration of diatomic ligands is triggered by transient opening of channels between pockets (Ostermann *et al.*, 2000; Schotte *et al.*, 2003). Using conformations coming from the MD trajectory, one can observe sparse opening of these channels with MDpocket. As these are characterized by a short lifetime, they are identified as less frequently appearing pockets. As seen in the HSP90 example, MDpocket is capable of detecting these transient channels and visualizing them using pocket density maps. Thus, the orange isomesh shown in Figure 4 delineates putative migration pathways between pocket sites. A more precise display can be achieved by examining the evolution in shape of isocontours taken at different values of the pocket density, as shown in Supplementary Figure S2. Here, the transient channel between the haem distal pocket (DP) and Xe4 is observed at higher densities. Transient channel opening is then found

between sites Xe4 and Xe2, and between Xe2 and Xe3. Finally, channel opening to Xe1 is found to be the less frequent event.

It is worth noting that the pattern of migration pathways provided by MDpocket analysis agrees with the experimental findings reported for CO migration using time-dependent X-ray crystallography (Schotte *et al.*, 2003; Tomita *et al.*, 2009). The experiments indicate that migration of CO from the DP to Xe4 is an initiating event in ligand migration appearing in the nanosecond timescale (Schotte *et al.*, 2003). Subsequent transitions involve the migration from Xe4 to sites 2 and 3. Furthermore, those studies also indicate that CO resides in pocket Xe1 for long periods. This fact can indicate that CO is either very stable in this pocket or that migration to other pockets is unfavoured by less frequent opening of transient channels. Our results support this latter possibility.

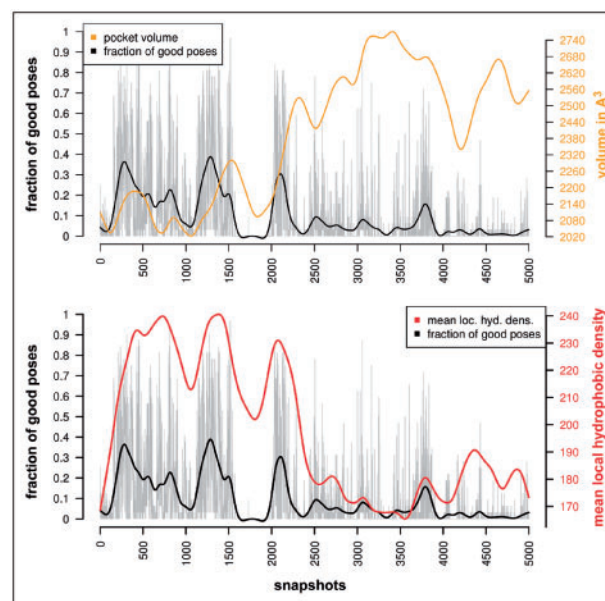
Further evidences of the transient character of channel opening are shown in Supplementary Figure S3. Here, pocket density maps are derived using either 10, 1000 or 10 000 conformations of Mb. What can be observed is that stable pockets (red arrows) are identified even using few snapshots. However, transient channels (green and blue arrows) can just be partially or even not found. Thus, usage of more conformational information allows for the visualization of these transient channels (green arrow).

The passage of diatomic ligands from the solvent into the distal haem cavity is another relevant phenomenon in globins. For Mb, it is generally accepted that HisE7 acts as gating residue for the entry pathway (Johnson *et al.*, 1989; Olson *et al.*, 1988; Perutz, 1970; Scott and Gibson, 1997). This gating mechanism is related to a rotation around the HisE7  $\text{Ca-C}\beta$  bond that places the residue outside the distal cavity. The opening event can be directly related to the appearance of a connecting pathway between the protein surface and the distal cavity detected with MDpocket. The frequency of appearance is thus a rough measure of the accessibility of the cavity, which in this case corresponds to  $\sim 30\%$ . It must be noted that this value is  $\sim 10\%$  higher than the alternate entry pathway through the Xe binding pockets. These results are in accordance with the experimental measurements that propose the HisE7 route as the main entry to the distal cavity (Scott *et al.*, 2001).

In conclusion, MDpocket allows easy and straightforward tracking and characterization of transient internal migration channels taking advantage of the wealth of conformational space sampled in molecular dynamics.

### 3.5 Case 2: MDpocket as a tool for docking

Docking of small molecules into binding sites is an outstanding tool in drug discovery (Jorgensen, 2004). While the flexibility of the ligand is generally addressed in modern docking software, very often only a rigid representation of the protein target is considered. Obviously, this can have a direct impact in limiting the enrichment in virtual screening studies. Thus, efforts are being undertaken to include protein motion in molecular docking (Henzler and Rarey, 2010). Here, we will show possible applications of MDpocket in ensemble docking strategies. It is known that docking results can be improved by using conformational ensembles of proteins (Barril and Morley, 2005; Novoa *et al.*, 2010) as a strategy to provide a better resolution on the motion of the binding site. One problem in using such ensembles is the selection of conformations to be included in a systematic docking approach (Rueda *et al.*, 2010). To test the suitability of MDpocket to guide the selection



**Fig. 5.** Fraction of binding poses without steric clashes (grey) of 32 binders on 5000 snapshots of the trajectory run for P38 Map kinase. Smoothed values are shown in black. Pocket volume and mean local hydrophobic density of the pocket are shown in orange and red, respectively. The mean local hydrophobic density is correlated with the fraction of good binding poses.

of those structures, the P38 Map kinase protein has been examined due to the known flexibility of its binding site. Thus, docking on different conformations should be considered on a target like P38 (Sherman *et al.*, 2006). In order to assess whether MDpocket can be used for efficient conformation selection for molecular docking, the ATP binding site was specifically tracked using MDpocket and characterized by means of two descriptors: the pocket volume and the mean local hydrophobic density. Whereas the volume reflects the fact that the size of the binding site is a major limitation to the binding of compounds, the second is a powerful predictor of the druggability of a binding site (Schmidtke and Barril, 2010).

During the MD trajectory run for P38 Map kinase, it can be observed that the P-loop lining the binding site is opening. This conformational rearrangement is relevant, as it allows increase of required space for docking certain P38 binders. However, a complete opening of the cavity yields a huge and very open pocket. Partially due to this opening of the loop, and furthermore the relative motion between the lobes, the pocket volume of the active site is increasing (Fig. 5). Though an open pocket can fit a larger variety of molecules without large steric clashes, a completely open binding site is not necessarily suitable to efficiently fit a given ligand.

To identify which conformations of the receptor can be theoretically considered for docking, 32 binders of P38 Map kinase with known crystallographic structure (Supplementary List S2) were superimposed to all snapshots of the MD trajectory using the  $\beta$ -sheet lining the binding site as reference. The ligand was then extracted from the superimposed crystal structure and inserted without altering its bioactive conformation into each snapshot of the trajectory.

Next, the interaction energy between the ligand and the protein was calculated. If the interaction energy is stabilizing (negative value), then the pose of the ligand (kept in the bioactive



conformation) is considered to be sterically acceptable in the current MD snapshot of the protein. It should be emphasized that this computational strategy was conceived with a 2-fold purpose. First, it provides a simple tool to identify whether the structural features of a given snapshot are suited to accommodate the ligand in its bioactive conformation. Second, it allows us to circumvent the flaws and limitations associated to ligand sampling and scoring inherent to classical docking approaches, which might lead to the prediction of poses dissimilar to the X-ray ones.

The fraction of acceptable poses out of the 32 known binders is tracked over time in Figure 5. The upper plot shows no significant correlation between the fraction of acceptable poses and the volume of the cavity. In fact, the results indicate that although the volume is notably enlarged after the first 2000 snapshots, as reflected in the increase of the root mean square deviation (RMSD) for all heavy atoms of residues lining the ATP binding site (Supplementary Figure S5), a bigger binding site is not necessarily the most suitable pocket to bind small molecules. This trend can be ascribed to the concomitant reduction of the putative interaction surface between ligand and receptor. Thus, the selection of interesting conformations suitable for molecular docking using only the volume of the binding site as indicator is not recommended. Therefore, the mean local hydrophobic density, a descriptor that reflects local densities of hydrophobic  $\alpha$  sphere clusters in a binding site and which has been shown to correlate with druggability (Schmidtke and Barril, 2010), was examined. As shown in the lower plot of Figure 5, there is a striking correlation between the mean local hydrophobic density and the fraction of good poses. A tentative explanation for this good correlation is that the mean local hydrophobic density catches situations in time, where hydrophobic patches are most accessible in the binding site and the binding site is compact. Accessibility of the hydrophobic surface is likely to have a beneficial effect on binding the rather hydrophobic drug-like molecules (Vieth *et al.*, 2004).

Overall, the preceding results suggest that tracking the mean local hydrophobic density during an MD trajectory on a pocket can give hints on the suitability of protein conformations for molecular docking. For instance, one can consider snapshots of the MD trajectory where the mean local hydrophobic density is maximized to use them as conformations for molecular docking. To the best of our knowledge, MDpocket is the first tool that might be able to identify conformations suited to bind small molecular binders, opening a variety of possible applications and rendering MD trajectories of proteins more accessible for molecular docking.

### 3.6 Comparison to existing methods

This section intends to propose a comparative analysis of MDpocket with other methods designed for the prediction of transient pockets and channels. With regard to pockets, the comparison is limited to EPOS<sup>BP</sup> (Eyrisch and Helms, 2007), whereas a larger variety of methods is considered for the detection of channels.

Only few tools have been reported to address the identification of transient drug binding sites. Using the PASS cavity detection algorithm (Brady and Stouten, 2000), EPOS<sup>BP</sup> defines pocket volumes via active site points and neighbouring probe positions identified by PASS. Using such an active site volume, all pocket-lining atoms (PLAs) are identified as being located  $>5$  Å from it. After execution of these steps on all snapshots of an MD trajectory, sets of PLAs can be compared with each other and clustered together

to find conserved cavities, which are assigned a pocket Id and whose pocket properties are tracked.

A major difference with MDpocket is that pockets identified with EPOS<sup>BP</sup> are mapped to a set of PLAs, while MDpocket maps pockets to a grid representation. While mapping cavities to PLAs might provide insensitivity to rotation and translation of the protein, it hinders the usage of the method on structural ensembles other than MD trajectories (despite the attempt to use residue names and atom names). Importantly, this is especially useful when analysing homologous structures for conserved cavities. On the other hand, EPOS<sup>BP</sup> allows tracking very few pocket descriptors: volume, depth and polarity. The usage of this latter descriptor might be debatable as it likely includes polar atoms in the binding pocket that do not contribute to the accessible surface area. MDpocket allows tracking a large set of pocket descriptors available via fpocket (for example, pocket size, polar and apolar surface areas, hydrophobicity and polarity measures, pocket density and average radius, local hydrophobic density, etc.). The availability of the source code of MDpocket also allows implementation of novel descriptors, but disclosure of EPOS<sup>BP</sup>'s and PASS's source code hinders this.

Using our example of 88 crystal structures of HSP90, we note two major advantages of MDpocket compared with EPOS<sup>BP</sup>: pocket tracking and performance. When the subpocket is open, the ATP binding site is identified by EPOS<sup>BP</sup> as a separate pocket with a distinct pocket Id (36.4% of cases), thus making it difficult to track large changes of cavity shape using a continuous representation. Computational performance of MDpocket is  $>25$ -fold faster making it better suited to analyse long MD trajectories; on this set, EPOS<sup>BP</sup> performed pocket detection and clustering (without analysis and property tracking) in 9 min versus 20 s for MDpocket on a single core of an Intel Q9550, 2.83 GHz.

A multitude of methods exist to tentatively detect transient channels and visualize them, such as Caver/Mole, MolAxis, implicit ligand sampling (ILS) and DyME.

In *Caver* (Beneš *et al.*, 2010), a grid is superimposed to the protein core and a starting position on this grid has to be defined. A value is associated to each grid point dependent on the radius of the biggest contact sphere that could be fit into the channel. Next a modified version of the Dijkstra shortest path detection algorithm is used to find the optimal path from the starting point inside the protein to the outside. *Caver* can be used on MD trajectories to identify conserved entry gorges for migration channels. However, the need to define the starting position on the grid might limit the capabilities of *Caver* as an exploratory tool to find transient channels.

More recently, the same authors published MOLE (Petrek *et al.*, 2007), which relies on the same algorithmic principles as *Caver*, but uses Voronoi tessellation of the interior of the protein to find ideal paths, has better execution speeds and allows analysis of bigger systems. However, the requirement for defining the starting point in the search of channels persists in MOLE.

*MolAxis* (Yaffe *et al.*, 2008) pursues to identify migration channels for small molecules inside the protein. Typical examples are given as channels observed in transmembrane proteins and cytochrome P450. Using the  $\alpha$  shape and the medial axis, *MolAxis* allows identifying the so-called corridors or probable migration pathways of small ligands. Similarly to *Caver*, *MolAxis* requires the starting position for searching channels to be defined. In comparison to *Caver*, *MolAxis* is significantly faster and thus facilitates the analysis of larger systems and large sets of MD trajectory snapshots.

Importantly, all the previously cited methods are able to perform channel detection on a single structure. For example, MolAxis allows tracking of channel radii during MD trajectories, but not the actual detection of transient channels as such, as the channel has to be predefined (or detected) and then tracked.

In *ILS* (Cohen *et al.*, 2008), all snapshots of the MD trajectory are superimposed to a reference structure, and then a suitable probe (i.e. a gaseous ligand) is moved through a regular grid inside the protein in order to determine the interaction energy with the protein, which is finally used to detect favourable migration paths. Like MDpocket, *ILS* performs a complete and exploratory cavity search. Results can be visualized as PMF maps using VMD. Unlike MDpocket, *ILS* is conceived for (i) detection of gas migration pathways and (ii) use on MD trajectories only (i.e. need of consistent topology between different protein conformations). Last, *ILS* itself is a purely exploratory tool, allowing the creation of PMF maps, but not allowing their analysis or the extraction of channel properties.

*DyME* was published during the writing of this article (Lin and Song, 2011). The initial steps of the *DyME* workflow and the need for a set of conformations are very similar to MDpocket. *DyME* detects all Voronoi vertices on and inside the protein and discards vertices with low clearances (radii of the  $\alpha$ -spheres in MDpocket). Remaining vertices are reduced to a maximum spanning tree. Next, vertices from the spanning tree are calculated for every conformation and conserved vertex clusters away from the bulk solvent (internal cavities) are identified. These clusters are then mapped back to the Voronoi vertex spanning tree of each conformation of the protein. Next, putative portal regions on the surface of the protein are identified, and last a so-called super graph is computed connecting stable cavities via maximum clearance channels observed at least once in an ensemble of protein conformations between each other and portal regions.

To show putative applications of *DyME*, the authors also used Mb as example. Thus, results obtained via MDpocket can be compared with *DyME*. The density grid produced by MDpocket (iso-value 1.8) is shown on Supplementary Figure S4 in mesh representation. Results obtained from *DyME* are superimposed as green pathways. As MDpocket results are more exploratory and contain more information, only channels around the Xe2, S1 and S2 pockets are shown for MDpocket. Interestingly, as with *DyME*, MDpocket identifies previously observed sites S1 and S2 (Bossa *et al.*, 2004). Furthermore, connections of these sites to the surface gates P7, P9 and P10 can also be observed as with *DyME*. Although not shown explicitly, gates like P1, P2, P4 and P6 are also identified by MDpocket. The overall coverage between *DyME* and MDpocket results is excellent.

Both *DyME* and MDpocket have several advantages over previously cited channel detection methods, notably the fact that both allow identification of the complete channel network and their flexibility of application. However, *DyME* infers the existence of small pockets via a clustering algorithm and the existence of connections between them via at least one occurrence of usually very low radius connection channels using Voronoi edges. MDpocket uses a more generic approach based on observations where  $\alpha$  spheres are found (and thus real voids) during time, how often and with what density (similar to cavity densities in *DyME*). Thus, MDpocket results correspond to physically more meaningful connection paths, while *DyME* paths can be true (if the clearance is high) but might also be inferred (if the clearance is low). This fact can be clearly

observed in Figure 7 of Lin and Song (2011), which shows the distribution of clearances for all channels for a 10 ns MD trajectory. While a significant amount of snapshots infer clearances of a radius of 1.5 Å, very few do have a clearance radius of 1.7 Å and above. Using MDpocket density and frequency maps and the corresponding detection parameters, results show actual voids of a minimum radius seen during a MD trajectory. Like *DyME*, MDpocket can also identify inferred channels by reducing the minimum size for  $\alpha$  spheres.

An apparent advantage of *DyME* is the measure of the channel radii. MDpocket, however, can track various properties of the channel (or pocket) and show their time evolution along the simulation. On the other hand, the computational efficiency of MDpocket is much better. Finally, MDpocket is stand-alone software distributed under the GNU GPL and is not based on commercial software like MATLAB.

## 4 DISCUSSION

MDpocket is a new tool intended to facilitate the qualitative and quantitative analysis of transient pockets and channels from conformational ensembles that account for the intrinsic dynamics of proteins. MDpocket is based on *fpocket*, a general-purpose pocket detection program, allowing adaptable, fast, free and reliable pocket detection. The accuracy of pocket prediction is mainly limited by two factors: (i) the grid-based nature of MDpocket, and (ii) the choice of parameters for filtering and clustering  $\alpha$ -spheres.

As MDpocket uses a grid-based methodology, the results can be affected by the necessary structural superimposition of the snapshots. Results derived for P38 Map kinase, where the two domains exhibit large relative motion with respect to each other, show a dependence on the set of atoms used for alignment. If the alignment is done using as reference one of the two subunits, pocket detection and averaging during time will be altered for all pockets on the other subunit. Consequently, pocket frequency and density on the second subunit are underestimated. In such a case, one should perform the structural alignment on the cavity of interest (the active site in our case), and bear in mind that pockets found on other places of the protein might not be representative of the whole trajectory.

The use of different parameter sets for pocket detection can have a major influence on results. Here, three parameter sets have been assessed, each addressing a given purpose. For exploratory MDpocket runs, probably water probe-sized  $\alpha$ -spheres (parameter set 3) are sufficient to give insights into the pocket behaviour during MD trajectories. However, if the main aim is to explore internal pathways in the protein matrix, parameter set 2 is better suited. This is illustrated by the analysis of the 50 ns MD simulation of Mb. MDpocket was capable of identifying crystallographically known Xe-binding pockets and infer putative connections between these binding sites in agreement with the experimental data. Furthermore, MDpocket allows for identification of the HisE7 migration pathway, which is considered to be the main entry of diatomic ligands to the distal haem cavity.

Exploration of binding sites adapted for compounds (i.e. drugs, substrates) larger than gaseous small molecules can be simply made by changing the search parameters (default parameter set). This versatility is illustrated by ‘simulating’ the docking of 32 ligands to P38 Map kinase. Virtual screening techniques often consider a rigid receptor conformation or at most a set of receptor conformations



(ensemble docking). In this latter case, it is often very difficult to choose representative receptor conformations for molecular docking (Rueda *et al.*, 2010). The results derived for P38 Map kinase indicate that the combined use of MDs and MDpocket with a descriptor related to druggability (mean local hydrophobic density) could be valuable to identify timeframes where a pocket of interest is in a good conformation to fit a drug-like molecule. These results strongly suggest that MDpocket could be used for an easy selection of protein conformations for ensemble docking strategies.

MDpocket is a versatile tool to render mechanistic studies of proteins involving protein motion easier. The software's adaptability and scalability allows applications in various domains like conformational selection for molecular docking or the study of structural plasticity of drug binding sites. Comparing MDpocket to existing methods, it is noteworthy that the program is the first tool addressing both channel and pocket detection and also characterization in one single framework. Contrary to most channel prediction methods, MDpocket follows an exploratory strategy that renders putative migration channels and binding sites. MDpocket is the only algorithm allowing easy selection of user-defined zones and further extraction of a large set of time-dependent descriptors of the selected zone. Finally, MDpocket is published within the fpocket suite of pocket detection programs, under the GNU GPL License and is freely available for download on <http://fpocket.sourceforge.net>.

## ACKNOWLEDGEMENTS

We thank Vincent Le Guilloux, Julien Maupetit and Pierre Tufféry for helpful discussions and support. We thank Allen Lin and Guang Song for providing results obtained in Mb with DyME.

**Funding:** Spanish Ministerio de Ciencia e Innovación (SAF2008-0559, SAF2009-08811); Generalitat de Catalunya (FI fellowship to P.S.; grant SCG-2009-294) for financial support.

**Conflict of Interest:** none declared.

## REFERENCES

- Ali,M.M.U. *et al.* (2006) Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, **440**, 1013–7.
- An,J. *et al.* (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics*, **4**, 752–761.
- Barril,X. and Morley,S.D. (2005) Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.*, **48**, 4432–4443.
- Beneš,P. *et al.* (2010) CAVER 2.1, software.
- Bidon-Chanal,A. *et al.* (2006) Ligand-induced dynamical regulation of NO conversion in Mycobacterium tuberculosis truncated hemoglobin-N. *Proteins*, **64**, 457–464.
- Borrelli,K.W. *et al.* (2005) PELE: protein energy landscape exploration. A novel Monte Carlo based technique. *J. Chem. Theory Comput.*, **1**, 1304–1311.
- Bossa,C. *et al.* (2004) Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys. J.*, **86**, 3855–3862.
- 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.
- Brooks,B.R. *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.
- Carrillo,O. and Orozco,M. (2008) GRID-MD-A tool for massive simulation of protein channels. *Proteins*, **70**, 892–899.
- Case,D.A. *et al.* (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
- Chemical Computing Group (2009) MOE (The Molecular Operating Environment) Version 2009.10, CCG.
- Cohen,J. *et al.* (2006) Imaging the migration pathways for O<sub>2</sub>, CO, NO, and Xe inside myoglobin. *Biophys. J.*, **91**, 1844–1857.
- Cohen,J. *et al.* (2008) Finding gas migration pathways in proteins using implicit ligand sampling. *Methods Enzymol.*, **437**, 439–457.
- Eyrich,S. and Helms,V. (2007) Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.*, **50**, 3457–3464.
- Glazer,D.S. *et al.* (2009) Improving structure-based function prediction using molecular dynamics. *Structure*, **17**, 919–929.
- Halgren,T. (2007) New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.*, **69**, 146–148.
- 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, 389.
- Henrich,S. *et al.* (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.*, **23**, 209–219.
- Henzler,A.M. and Rarey,M. (2010) In pursuit of fully flexible protein-ligand docking: modeling the bilateral mechanism of binding. *Mol. Inform.*, **29**, 164–173.
- Hess,B. *et al.* (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.
- Huang,B. (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics J. Integr. Biol.*, **13**, 325–330.
- Humphrey,W. *et al.* (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38, 27–28.
- Johnson,K.A. *et al.* (1989) Structure of myoglobin-ethyl isocyanide. Histidine as a swinging door for ligand entry. *J. Mol. Biol.*, **207**, 459–463.
- Jorgensen,W.L. (2004) The many roles of computation in drug discovery. *Science*, **303**, 1813–1818.
- Kim,D. *et al.* (2008) Pocket extraction on proteins via the Voronoi diagram of spheres. *J. Mol. Graph. Model.*, **26**, 1104–1112.
- Kleywegt,G.J. and Jones,T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **50** (Pt 2), 178–185.
- Kua,J. *et al.* (2002) Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach. *J. Am. Chem. Soc.*, **124**, 8260–8267.
- Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330, 307–308.
- Laurie,A.T.R. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
- Le Guilloux,V. *et al.* (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
- Leis,S. *et al.* (2010) In silico prediction of binding sites on proteins. *Curr. Med. Chem.*, **17**, 1550–1562.
- Liang,J. *et al.* (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.
- Lin,T.-L. and Song,G. (2011) Efficient mapping of ligand migration channel networks in dynamic proteins. *Proteins*, **79**, 2475–2490.
- MacKerel,A.D. Jr *et al.* (1998) CHARMM: the energy function and its parameterization with an overview of the program. In Schleyer,P.v.R. *et al.* (eds). *The Encyclopedia of Computational Chemistry*, Vol. 1. John Wiley & Sons, Chichester, pp. 271–277.
- Marti,M.A. *et al.* (2006) Dioxygen affinity in heme proteins investigated by computer simulation. *J. Inorg. Biochem.*, **100**, 761–770.
- Novoa,E.M. *et al.* (2010) Ensemble docking from homology models. *J. Chem. Theory Comput.*, **6**, 2547–2557.
- Olson,J.S. *et al.* (1988) The role of the distal histidine in myoglobin and haemoglobin. *Nature*, **336**, 265–266.
- Ostermann,A. *et al.* (2000) Ligand binding and conformational motions in myoglobin. *Nature*, **404**, 205–208.
- Pérot,S. *et al.* (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov. Today*, **15**, 656–667.
- Perutz,M.F. (1970) Stereochemistry of cooperative effects in haemoglobin. *Nature*, **228**, 726–739.
- 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.
- Petrek,M. *et al.* (2007) MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, **15**, 1357–1363.
- Pettersen,E.F. *et al.* (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
- Phillips,J.C. *et al.* (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
- Rueda,M. *et al.* (2010) Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inform. Model.*, **50**, 186–193.

- Schmidtke,P. and Barril,X. (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.*, **53**, 5858–5867.
- Schotte,F. *et al.* (2003) Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, **300**, 1944–1947.
- Scott,E.E. and Gibson,Q.H. (1997) Ligand migration in sperm whale myoglobin. *Biochemistry*, **36**, 11909–11917.
- Scott,E.E. *et al.* (2001) Mapping the pathways for O<sub>2</sub> entry into and exit from myoglobin. *J. Biol. Chem.*, **276**, 5177–5188.
- Seco,J. *et al.* (2009) Binding site detection and druggability index from first principles. *J. Med. Chem.*, **52**, 2363–2371.
- Sherman,W. *et al.* (2006) Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.*, **67**, 83–84.
- Sherman,W. *et al.* (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.*, **49**, 534–553.
- Spyrakis,F. *et al.* (2011) Histidine E7 dynamics modulates ligand exchange between distal pocket and solvent in AHb1 from *Arabidopsis thaliana*. *J. Phys. Chem. B*, **115**, 4138–4146.
- Tilton,R.F. *et al.* (1984) Cavities in proteins: structure of a metmyoglobin xenon complex solved to 1.9 Å. *Biochemistry*, **23**, 2849–2857.
- Tomita,A. *et al.* (2009) Visualizing breathing motion of internal cavities in concert with ligand migration in myoglobin. *Proc. Natl Acad. Sci. USA*, **106**, 2612–2616.
- Vieth,M. *et al.* (2004) Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.*, **47**, 224–232.
- Weisel,M. *et al.* (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, **1**, 7.
- Wright,L. *et al.* (2004) Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.*, **11**, 775–785.
- Yaffe,E. *et al.* (2008) MolAxis: efficient and accurate identification of channels in macromolecules. *Proteins*, **73**, 72–86.
- Yang,F. and Phillips,G.N. (1996) Crystal structures of CO-, deoxy- and met-myoglobins at various pH values. *J. Mol. Biol.*, **256**, 762–774.