Principal Component Analysis reveals correlation of cavities evolution and functional motions in proteins.

Nathan Desdouits, Michael Nilges, Arnaud Blondel

Journal of Molecular Graphics and Modelling 55 (2015) 13–24

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Principal Component Analysis reveals correlation of cavities evolution and functional motions in proteins

Nathan Desdouits, Michael Nilges, Arnaud Blondel *

Unité de Bioinformatique Structurale, Département de Biologie Structurale et Chimie, Institut Pasteur, CNRS - UMR3258, 25-28 rue du Docteur Roux, 75015 Paris, France

ARTICLE INFO

Article history:
Accepted 18 October 2014
Available online 25 October 2014

Keywords:
Protein cavities
Molecular dynamics
Cavity geometry evolution
Principal Component Analysis
Functional analysis
Drug design

ABSTRACT

Protein conformation has been recognized as the key feature determining biological function, as it determines the position of the essential groups specifically interacting with substrates. Hence, the shape of the cavities or grooves at the protein surface appears to drive those functions. However, only a few studies describe the geometrical evolution of protein cavities during molecular dynamics simulations (MD), usually with a crude representation. To unveil the dynamics of cavity geometry evolution, we developed an approach combining cavity detection and Principal Component Analysis (PCA). This approach was applied to four systems subjected to MD (lysozyme, sperm whale myoglobin, Dengue envelope protein and EF-CaM complex). PCA on cavities allows us to perform efficient analysis and classification of the geometry diversity explored by a cavity. Additionally, it reveals correlations between the evolutions of the cavities and structures, and can even suggest how to modify the protein conformation to induce a given cavity geometry. It also helps to perform fast and consensual clustering of conformations according to cavity geometry. Finally, using this approach, we show that both carbon monoxide (CO) location and transfer among the different xenon sites of myoglobin are correlated with few cavity evolution modes of high amplitude. This correlation illustrates the link between ligand diffusion and the dynamic network of internal cavities.

1. Introduction

1.1. Function, conformation and cavities: impact of dynamics

Proteins exert their function through interactions with other proteins, nucleic acids, substrates, etc. These interactions, in turn, depend on the protein conformation in a broad sense, which shapes the interaction interfaces, thus making functional interactions possible [1,2]. An important feature of protein shape is the existence of cavities inside them or grooves at their surface. A favorable geometry of these cavities and grooves allows surrounding amino-acids – here called the "pocket" – to mediate multiple contacts supporting interactions or catalysis. Protein cavities, due to this propensity to make numerous contacts, are a common subject of research, both for functional analysis and drug design. In drug design, virtual screening is performed with an active site or an allosteric pocket, where the binding of a molecule is expected to block (or activate) the protein function. To select the molecules that are most likely to bind, various "energy" or scoring terms can be used. Among these terms, shape complementarity clearly emerges as a key factor [3–5]. Small cavities can also reveal packing defects, which impact protein stability [6]. Furthermore, cavities are involved in the diffusion of water and small ligands in proteins, for example in myoglobin and cytochrome p450 [7–10].

Beyond static conformations, the dynamic evolution of structures also governs protein functions. Indeed, motions ranging from side chain rotations in active sites to major domain motions are involved in the control of protein function [11–13].

The relation between the conformation (atomic coordinates) and cavity geometry has been extensively studied [14–16]. Unfortunately, most studies are restricted to one or a few static structures. One exception is the case of globins, in which ligand diffusion is functionally important. In this case, dynamic analysis of cavities already revealed specific relations [17–20]. However, to the best of our knowledge, a quantitative and detailed analysis of the evolution of cavity geometry has rarely been performed until now. Most existing quantitative studies use exclusively simple descriptors of cavity geometry such as its surface area, its volume or the position of its geometric center [16]. Remarkably, dedicated

* Corresponding author. Tel.: +33 140613693; fax: +33 145688719.
E-mail addresses: nathan.desdouits@pasteur.fr (N. Desdouits), nilges@pasteur.fr (M. Nilges), ablondel@pasteur.fr (A. Blondel).

http://dx.doi.org/10.1016/j.jmgm.2014.10.011

Nomenclature

MD	molecular dynamics
PCA	Principal Component Analysis. Principal components are sometime called mode by analogy with vibrational analysis. Principal components are abbreviated PC
Cavity	an empty or water-filled volume at the surface or inside a protein that can contain a molecule, and is distinct from the bulk solvent. Here, a groove at the surface is also called cavity by extension
Pocket	amino-acids surrounding a cavity
atomic coordinates space	descriptor space for atomic structures, the coordinates of each atom.
cavities space	descriptor space for cavities. Here cavities are defined on a regular Boolean grid; points in cavity have a value of 1, points within protein atoms or bulk solvent have a value of 0
atomic trajectory	a collection of atomic coordinates for a protein (e.g. the usual output of a MD run)
cavity trajectory	a collection of cavity descriptors for a protein.
step space	index space pointing to steps of a trajectory; hence, one index points to the corresponding step of atomic coordinates and cavity descriptors

methods to study cavity dynamics have mostly been developed recently [21–26].

1.2. Definition of cavities

At first sight, it might seem easy to describe cavities in a protein structure. However, it requires sophisticated algorithms to make a relevant assignment due to the complex and somewhat subjective nature of cavities. Contrary to attributes such as secondary structure, which can be defined by atomic coordinates with little ambiguity, largely varying definitions can be proposed for cavities. The most common definitions are Lee and Richards' solvent accessible surface, Connolly surface, Voronoi tessellation, and alpha-spheres [27–29]. A vast number of very small cavities are present between the packed spherical atoms. These small void spaces have to be discriminated from relevant cavities that can enclose a ligand to avoid pointless complexity. Furthermore, limits have to be established to separate solvent accessible cavities (grooves) from the bulk solvent. These limits have to be drawn with practical but nonetheless subjective criteria since no real physical boundary exists. Finally, cavities can be encoded in different ways during computation, using grids, list of spheres, facets, etc. The encoding can influence the geometric description of cavities and how they can be manipulated.

1.3. Applications of cavity analysis

Cavities have been analyzed to explain biological processes. Noticeably, enzymatic cavities of numerous proteins have been thoroughly studied, in terms of volume, surrounding amino acid composition, and possible evolution during catalysis as in methane monooxygenase or lysozyme [30,31]. Furthermore, packing defects are important for protein function, as they give space for atomic motions, or even for water or ligand diffusion, as in cytochrome p450[9,10] and myoglobin [7]. These defects lead to a trade-off between protein stability and flexibility. Similarly, cavities between protein domains allow larger motions and structural transitions to appear [32]. Examples for such transitions can be found in the Dengue virus E protein [33] and during the EF-CaM association [34].

Drug design is another domain in which cavity analysis is important, because the cavity defines the shape in which the ligand has to fit during virtual screening. Virtual screening software packages often use cavity detection algorithms as a first step before actually placing the ligand [35,3]. A few software packages use geometric or energetic criteria to score cavities by druggability [36]. A major issue actively discussed in drug design is the modeling of pocket flexibility as it affects ligand binding [37,38]. In this context, tools to analyze the dynamics of cavity geometry should be essential to select relevant protein conformations in a virtual screening campaign. Furthermore, information gathered on the relation between the evolutions of cavities and those of structures can be used to take cavity flexibility into account or to model new pocket conformations.

Interestingly, results from virtual screening on modeled protein conformations are encouraging. For example, docking on MD structures can improve the results [39]. Other conformation building methods such as pressurization [40], fumigation [41] or SCARE [42] have been extensively used to build conformations *in silico* by various biased sampling methods. Nonetheless, the application of isotropic constraints or arbitrary bias on the protein structure in these methods does not make use of the unconstrained pocket flexibility, and thus can produce conformations that can be rather irrelevant. In this perspective, it would be interesting to select structures having a given cavity geometry (e.g. wider) from a preexisting set of conformations (e.g. from MD), but approaches performing this selection on comprehensive geometric criteria are still rare [39]. More specifically, providing analysis of the spontaneous evolution of cavity geometry in fine detail to guide selection, sampling or building procedures has, to the best of our knowledge, never been tried, and could bring a clear improvement to drug design projects involving virtual screening.

1.4. Work outline

In this article, we address the characterization of the evolution of cavity geometry in dynamical protein systems. For this, we performed a parallel Principal Component Analysis (PCA) on the protein structures and on the associated cavities. Hence, cavities were calculated on series of MD structures and encoded in a suitable format for the PCA analysis, which was adapted for the specificity of the cavity objects. This analysis was tested on different protein systems. To evaluate its robustness, we chose systems with different sizes and involving different types of functional motion (lysozyme, Dengue virus E protein, EF toxin of anthrax coupled with calmodulin, myoglobin; Fig. 1). We also compared different programs having slightly different definitions to detect cavities (gHECOM [43] and mkgrid, an in-house program).

This dual analysis characterizes how the cavity geometry evolutions correlate with that of the protein conformation. The first few principal components (PC) of structures and cavities displayed substantial temporal correlations, which faded in subsequent components. Those correlations support the significance of this parallel analysis, but also highlights specific information brought by direct analysis of cavity evolution. Interestingly, we found that it is also possible to build new protein conformation along cavity principal components. We found that these built conformations had cavity geometries remaining closer to the principal components than any of the cavities derived from the original trajectory. Hence, beyond an analysis tool, this methodology also proved to be a powerful building instrument.

We found that a limited number of PCA components can well describe the cavity evolution. This facilitates manipulations and allows us to apply the approach on different problems. For

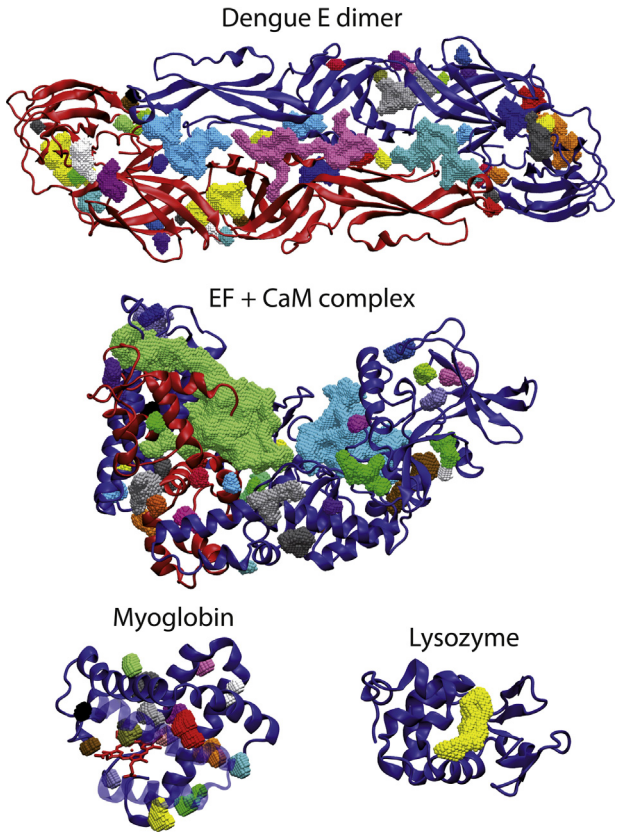

Fig. 1. Systems used in this study. The initial structure of the four systems used in this article and their mkgrid cavities.

example, we illustrate below how to cluster cavities according to their geometries in order to select a diverse set. We also applied PCA to myoglobin, an extensively characterized system. Our results are consistent results with those previously described on functional cavities (distal pocket and xenon cavities) [44]. Interestingly, the frequency of cavity occurrence mapped the carbon monoxide binding free energy landscape [45]. We found that the hopping of carbon monoxide between cavities correlated with the first few cavity components. This underlines the important role of cavity flexibility in the diffusion processes of small ligands in proteins.

These results support the relevance of the analysis and demonstrate the convenience of the "step-space/cavity-space/structure-space" parallel formalisms to analyze and grasp functional cavity evolution.

2. Theoretical methods

2.1. Cavity definition

We used programs defining cavities on Boolean grids, which is essential for subsequent analyses. A cavity is defined as the volume accessible to a small solvent probe (sphere of 1.4 Å radius) delimited by a "solvent accessible surface" [27], but inaccessible to the "bulk solvent" delimited with a larger probe. This allowed us to detect "grooves", peripheral cavities or clefts, which can be important for the biological activity of proteins and which were also considered as "cavities" here. Adjacent cavity grid points were clustered into connected independent cavities, and small cavities (<12 Å^3) were discarded. Cavity trajectories were obtained by calculation of cavities on each conformation of atomic trajectories. Details on calculations are provided in Section 3.

2.2. Principal Component Analysis in step space

By definition, a Principal Component Analysis (PCA) is performed by diagonalization of the covariance matrix V of a list of descriptors D. In the case of atomic coordinates, the descriptors D^{atoms} are the atomic coordinates of each atom for each step, which have a size of $3n^{atoms} \times s$, with s the number of steps in the trajectory. For cavity sets, the descriptors D^{cav} are the states of each grid points for each step of the trajectory. The state is 1 if the grid point is located in a cavity for this step, else 0. To reduce the size of the descriptors, only the grid points that are in state 1 in at least one step in the trajectory are considered. This list of grid points of size n^{cav} is called the domain of definition of the cavity sets. Thus, D^{cav} has a size of $n^{cav} \times s$. Note that, to make a relevant analysis of the internal motions of a protein, it is necessary to align the structures. This alignment is particularly crucial for the cavity analysis since their descriptors are defined on a static grid. The covariance matrix used in PCA is calculated on the centered descriptors M, defined for all steps i as $M_i = D_i - C$, with $C = \langle D_i \rangle$. Noticeably, this centering switches the cavity sets representation from Booleans to real numbers.

The usual definition of PCA uses the covariance matrix in the descriptor space: $V_{desc} = 1/s \cdot M \cdot M^T$, of size $n \times n$. The diagonalization of the covariance matrix yields a list of eigenvectors v_i, which are the directions of evolution of the principal components *PCi*. The eigenvalues, λ_i, both give the variances explained by v_i and the magnitudes of *PCi*. Since the principal components are already in the descriptor space, they can be directly visualized.

PCA can also be performed with the covariance matrix in the step space: $V_{step} = 1/s \cdot M^T \cdot M$, of size $s \times s$ thanks to favorable mathematic identities shown in SI. Interestingly, although V_{desc} and V_{step} have different dimensions, they share the same non-zero eigenvalues, and their eigenvectors are easily related (see SI). Thus, eigenvectors in coordinate space v_i can be recalculated from their corresponding eigenvector in step space. It is advantageous to perform the calculation in the step space when $s < n^{cav}$, which is true most of the time.

Remarkably, performing the calculations in step space enables the comparison of principal components for cavity sets and atomic coordinates, as the matrices have the same sizes and refer to the same entities, i.e. the steps of the trajectory.

2.3. Projection and reconstruction of cavities

To project a cavities or atomic coordinates vector R on a given component v_j, the average coordinate C is subtracted from it for consistency with the PCA calculation. This yields the formula: $p_j = (R - C) \cdot v_j$.

Like for the atomic coordinates, cavities can be reconstructed along a principal component, v_j. For that, a vector of real numbers, $G^r_{i,j}(\lambda) = C^{cav}_i + \lambda \cdot v_{i,j}$, is calculated on each point i of the definition domain on the grid. Then a vector of Booleans, G^b is built on the same grid as $G^b_{i,j}(\lambda) = 0$ for $G^r_{i,j}(\lambda) < c_{off}$ and $G^b_{i,j}(\lambda) = 1$ otherwise. This Boolean truncation is noted: $G^b_{i,j}(\lambda) = (G^r_{i,j}(\lambda) \geq c_{off})$. Tests monitoring the I_{self} curves (see below) obtained with various c_{off} values and protein systems suggested to use a c_{off} value of 0.5, which was kept for all subsequent calculations.

2.4. Atomic coordinates built along step space components

Reconstruction of atomic trajectories can be made along any vector in the steps space. Hence, atomic structures can be built along cavity principal components in step space, v^{cav}_i as: $R^{atoms}(\lambda) = C^{atoms} + \lambda \cdot M^{atoms} \cdot v^{cav}_i$. Values of λ are usually limited so that the projection along the component does not exceed the maximum

projection observed in the trajectory. However, to probe the limits of the method, additional structures were built beyond that limit to test how distorted and unrealistic they would be. Then, cavities can be calculated on the reconstructed structures with the same program, gHECOM or mkgrid, and compared with the targeted theoretical cavities, $G_i^r(\lambda) = C^{cav} + \lambda \cdot M^{cav} \cdot v_i^{cav}$ or their Boolean truncation, $G_{i,j}^b(\lambda) = (G_{i,j}^r \geq c_{off})$.

2.5. Comparison of principal components from different spaces

When PCAs are performed in the step dimension, cavity and atomic components or eigenvectors can readily be compared because they have the same dimensions, and are expressed as linear combination of the same trajectory steps. Comparison can be quantified with $RMSIP_n$ (Root Mean Square Inner Product [46,17]):

$$RMSIP_n = \left(\sum_{k=1}^{n} \frac{1}{n} \sum_{l=1}^{n} \left(v_l^{cav} \cdot v_k^{coord} \right)^2 \right)^{1/2} \quad (1)$$

We use a value of $n = 10$ as in Ref. [46,17] as it covers the relevant components in our applications.

2.6. Synthetic cavities

Two synthetic cavity trajectories were generated as a control to test the effect of discretization and truncation on PCA. One was generated by stretching a centered cubic cavity from 10 Å to 20 Å by 0.05 Å steps on both sides along the *x*-axis and discretization on a 0.5 Å step size grid. The same procedure was applied on a spherical cavity (radius of 5 Å) to generate the second trajectory.

2.7. Quantification of the Boolean truncation effect

PCA can readily be applied to atomic positions since they are expressed as real quantities. However, for Boolean objects such as cavities in our description, it appeared convenient to use intermediate real representations for the PCA calculations followed by a Boolean truncation. This however, introduces some non-linearity. Additionally, s is often smaller than n^{cav} and then, the s eigenvectors of the cavity trajectory, $v_{step,i\in[1,s]}^{cav}$, form an incomplete basis set that only spans a subspace of the entire cavity n^{cav}-dimension space. As a result, Boolean truncation of a vector given by a linear combination of v_i can "breach" out of its original s-dimension sub-space. Here, we specify a target cavity as a displacement, Δ^{cav}, from the average cavity of the original trajectory, C^{cav}: $G^r = C^{cav} + \Delta^{cav}$, followed by a truncation to Boolean space, resulting in $G^b = (G^r \geq 0.5)$. Four related indices ranging from 0 to 1, named I_{self}, I_{comp}, I_{sub} and I_{other}, are used to quantify the deviation between the target displacement Δ^{cav} and the effective one, $G^b - C^{cav}$.

I_{self} is the fraction of the effective displacement along the direction of Δ^{cav}. To simplify notations here, we choose v_1 to be along Δ^{cav}; $v_1 = \Delta^{cav} / \| \Delta^{cav} \|$:

$$I_{self} = \frac{(G^b - C^{cav}) \cdot v_1}{\| G^b - C^{cav} \|}$$

where $\|X\|$ denotes the norm of vector X in the full n^{cav} dimension space. The larger the I_{self} values, the closer the cavities to the component used in the reconstruction. A value of 1 indicates colinearity with v_1 and a value of 0 indicates orthogonality.

I_{sub} is the fraction of the effective displacement in the subspace defined by the original trajectory:

$$I_{sub} = \left(\sum_{i=1}^{s} \left(\frac{G^b - C^{cav}}{\| G^b - C^{cav} \|} \cdot v_i \right)^2 \right)^{1/2}$$

I_{comp} is the fraction of the effective displacement in the subspace defined by the other components within the original trajectory subspace:

$$I_{comp} = (I_{sub}^2 - I_{self}^2)^{1/2}$$

Finally, I_{other} is the fraction of the effective displacement orthogonal to Δ^{cav}:

$$I_{other} = (1 - I_{self}^2)^{1/2}$$

2.8. Measure of similarity of two cavities

We defined a similarity index between two cavities i and j by applying the Jaccard index to their Boolean descriptors:

$$S_{ij} = \frac{G_i^b \cap G_j^b}{G_i^b \cup G_j^b}$$

The more classical normalized dot product measure that can apply to real-valued cavities was also used when appropriate:

$$S_{ij}^r = \frac{1}{||G_i^b|| \, ||G_j^b||} \sum_{l=1}^{n^{cav}} G_{i,l}^b G_{j,l}^b$$

3. Simulation methods

3.1. Protein systems and dynamics

Three 120 ns trajectories were calculated with different seeds and recorded every 10 ps for hen egg white lysozyme and myoglobin. The NAMD program [47] was used with the CHARMM22 force field.

For the lysozyme (PDB entry: 2LYZ), the protein was solvated in a rectangular solvent box of size 70.4Å ×52.3 Å ×49.0 Å containing 5192 water molecules in addition to the 101 crystal water. The positive charges of the protein were neutralized by adding 8 chloride ions. Atoms with high energy (>10 kcal/mol) were first minimized by 100 steps of steepest descent in the CHARMM program. Then the solvent was heated and equilibrated at 300 K (Langevin coupling constant of 100 ps^{-1}) in a 10 ps molecular dynamics with the protein fixed. The system was then equilibrated by performing 1 ns molecular dynamics at 310 K with NAMD. The Particle Mesh Ewald algorithm was used to calculate electrostatics. The Langevin coupling constant was set to 0.1 ps^{-1}. The cutoff distance was set to 12 Å for nonbonded interactions, and the time step was 1 fs. The production simulations were run with the same parameters.

For the myoglobin trajectories, the CHARMM test case "mbco4958" was used [48], which is composed of a pre-solvated system (4958 water molecules) with a carbon monoxide molecule (CO) located in the heme distal binding site. The whole system (protein, heme, CO and water) was in a 55.5 Å cubic box. The proximal histidine 93 was bound to the heme iron, while the CO was kept free. The 1 ns equilibration of the system and the three 120 ns simulations were performed with the same protocol as for the lysozyme simulations described above.

Additionally, we calculated a 10 ns trajectory of dengue E protein (PDB entry: 1OKE). The system was solvated with 39,852 water molecules in a rectangular box of size 180Å ×90 Å ×85 Å. The system was simulated with the same parameters as for the lysozyme, except for the temperature (300 K) and the Langevin coupling constant (1 ps^{-1}).

Finally, the last 10 ns from a previously published 15 ns MD of EF-CaM complex with 2 bound calcium ions were also used [34].

Table 1
Number of atoms and lengths of the trajectories used in this article.

Protein system	# protein atoms	Traj. lengths
Dengue E protein (dimer)	12,258	10 ns
Anthrax EF-CaM complex	9942	15 ns
Sperm whale myoglobin (MbCO)	2534	120 ns ×3
Hen egg white lysozyme	1960	120 ns ×3

For analysis, water molecules and ions were removed from all trajectories, and each structure was aligned onto the first structure of the corresponding trajectory by least-square fit on all heavy atoms (except CO). See Table 1 for a summary of the trajectories sizes and lengths.

3.2. Cavity detection and protein volume

Cavities were calculated on Boolean grids with either gHECOM [43] or mkgrid, a program developed in the laboratory, which limits cavity extension based on curvature and produce smaller cavities. The grid spacing was set to 0.5 Å for both programs. Hence, we approximate the cavity volume as the sum of the volumes (0.125 $Å^3$ each) of the grid cells marked as cavities. The solvent probe sphere had a radius set to 1.4 Å. The "bulk solvent" probe sphere radius was set to 10 Å for mkgrid, and to 3 Å for gHECOM, to limit the size of the cavity sets to tractable levels. Both programs were run on each conformation of the atomic trajectories without water or ions (and without CO in the case of myoglobin). Analyses were usually performed on cavity sets of 1000 or 10,000 steps, with equally spaced time steps whenever possible, although we have tested the analysis with up to 36,000 steps on myoglobin.

The protein *envelope* was defined as the set of grid points that are not considered as bulk in any step of the trajectory (large probe radius of 10 Å). Similarly, we approximated the protein volume for each step as the set of all grid points that are not in the volume accessible to the solvent (small probe radius of 1.4 Å).

3.3. Cavities selected for single cavity analysis

Two cavities of the Dengue E dimer were selected for local analysis: the β-OG cavity between domains I and II, and an interdormain cavity located between domains I and III of one monomer and the tip of domain III of the other. The ATP cavity of EF was also selected. To facilitate the selection we used mkgrid, as the detected cavities are less extended and fuse less often than with gHECOM. The pocket surrounding a cavity is defined as the list of residues that are within 1.5 Å of any grid points of the cavity in at least one of the conformations. PCAs were performed on the new subsets (selected cavity and pocket) and their PCs compared, as for the whole system.

To measure the importance of the alignment, the trajectories were aligned on the pockets defined above and the cavities recalculated and re-extracted, yielding 3 new cavity sets and their corresponding realigned pocket trajectories. PCAs on these aligned sets were calculated and compared as above.

3.4. Selection of conformations with diverse cavity geometries

The Dengue E dimer β-OG cavity defined above was used for the conformation selection application. The 10,000 available structures were used for this analysis. To reduce the dimensionality of the cavities, the projection of each step of the cavity set onto their 100 first principal components was taken, reducing the cavity descriptor vector size from 18,131 down to 100. This allowed us to efficiently use a k-means algorithm to divide the cavity set into 4 clusters. For each cluster, the cavity with the highest overlap with the centroid was chosen as its representative cavity. To make a comparison, conformations were also selected with the same procedure, except that the PCA was performed on the structure of the pocket instead of the cavity descriptors. The representative cavity was then derived from the structure with the lowest RMSD to the centroid. Overlap between the representative cavities was used as a similarity measure between clusters. The average overlap between the cavities of a cluster and the representative cavity was used as an internal similarity measure.

3.5. Localization of CO within binding sites

Myoglobin MD calculations were started with one CO molecule in its canonical position next to the heme iron and the distal pocket (DP). The three 120 ns trajectories were inspected visually to discard frames where the CO had escaped from myoglobin. The remaining frames were concatenated in a single trajectory and their corresponding cavities calculated.

We identified the myoglobin site in which CO binds with the DBSCAN algorithm [49] along the trajectory. Trajectory steps were labeled by the site identity, i. An epsilon distance of 1.0 Å to define edges and a density threshold of 10 (MinPts, minimal number of highly connected points to nucleate clusters) were selected empirically to minimize the number of CO positions that are considered as noise by DBSCAN. These positions were discarded in subsequent analysis.

3.6. Myoglobin internal cavity defined by CO positions

To focus the analysis on the myoglobin cavities that can host CO (distal pocket, xenon cavities 1–4), the trace of the CO in the trajectory was collected. The 24,800 CO positions were traced with the 8 vertices of each cube hosting the CO oxygen atom center. Then, the analysis was restricted to the subset of cavities that had at least one grid point in this CO trace.

To relate the cavities with the position of the CO, a residence cavity, A_i, is calculated by averaging the sets of internal cavities for the steps where CO is in binding site i. The evolution of the internal cavities when CO moves from site i to site j is then calculated as $(A_j - A_i)$ and is called the $i \rightarrow j$ transfer cavity. These residence and transfer cavities, expressed in step space, can be related to the principal components by projection.

To evaluate the statistical significance of these projections, a null hypothesis was built. We calculated the probability of each possible $i \rightarrow j$ transfer from our existing data and generated 20 random binding site series according to the same transfer probabilities. These series were then projected on principal components to provide the null hypothesis projection.

4. Results

4.1. Dynamic properties of the cavities

The volume of cavity sets largely depended on software and parameters: gHECOM produced larger volumes despite the use of a much smaller bulk probe (Fig. S1). Nonetheless, the instantaneous total volume of the cavity set varied widely during the MD runs with both programs (Fig. S1.a). The domain of definition of cavities was found to expand rapidly up to ten times the total cavity volume for single structures (Fig. S1.b). Depending on the system and the cavity detection program, it represented between 7.6% and 65.8% of the total protein envelope and could span a volume equivalent to that of the protein (Table 2).

Our description of the mean cavity occupancy of myoglobin (Fig. S2) is very close to that of previous studies [18,19].

The time autocorrelation function of cavities showed small values in the nanosecond time scale, and for some systems, already at

Table 2
Volume of the global protein envelope and the domain of definition of cavities. Volume is given in Å^3 for mkgrid and gHECOM cavities. The ratio of the domain of definition volumes to the protein envelope is given in parentheses, in percentage. The protein volumes are calculated as explained in Section 3 and is averaged over all steps. The RMSD plateau is the average value over the last 2/3rd of the trajectory.

System	Traj. length	RMSD plateau (Å)	Protein envelope (Å^3)	Domain of definition (Å^3)				Average protein volume (Å^3)
				mkgrid	(%)	gHECOM	(%)	
Dengue	10 ns	1.78	239,459	36,314	(15.2)	128,016	(53.5)	127,688
EF-CaM	10 ns	1.80	201,669	37,555	(18.7)	102,954	(51.1)	101,773
Lysozyme 1	100 ns	1.91	47,833	9,992	(20.9)	20,486	(42.9)	20,197
Lysozyme 2	100 ns	1.42	42,716	7,621	(17.8)	17,405	(40.7)	20,070
Lysozyme 3	100 ns	1.46	45,079	8,022	(17.8)	18,324	(40.6)	20,021
Lysozyme 4	10 ns	1.17	37,260	4,379	(11.8)	13,847	(37.2)	20,072
Myoglobin 1	100 ns	1.31	49,308	5,447	(11.0)	27,066	(54.9)	26,113
Myoglobin 2	100 ns	1.20	48,636	5,177	(10.6)	31,265	(64.3)	26,045
Myoglobin 3	100 ns	1.33	49,047	5,245	(10.7)	32,254	(65.8)	26,092
Myoglobin 4	10 ns	1.12	45,234	3,436	(7.6)	28,342	(62.7)	26,050

the picosecond time scale (Fig. S3, e.g. autocorrelation values below 0.5). Noticeably, the exponential decay observed revealed good and regular equilibration and diffusion properties. Hence, cavities displayed a steady diffusive behavior that can be exploited to sample cavity geometries.

4.2. First characterizations of the cavity evolution modes

We performed PCA on the atomic and cavity trajectories for four protein systems, but also on the 3 isolated cavities from 2 of the systems (see Section 3). An example of a cavity set principal component (PC) is given in Fig. 2. According to the PCA formulation, cavity principal components map the zones in which the cavity presence evolves during the trajectory and specify the relative direction and amplitude of those variations. The latter are quantified by eigenvalues. The sign of the eigenvectors is arbitrary in PCA calculation. Nonetheless, zones having the same sign evolve correlatively and zones having opposite signs evolve anti-correlatively. Hence, components can be viewed as a description of the cavities evolution between their negative and positive areas in the course of the trajectory (Fig. 2).

The first cavity components appeared smoother than those having a larger rank, see Fig. S4.a. This was quantified by spatial autocorrelation functions, which showed a global decrease along the eigenvector rank, but also significant variations and a substantial level of noise, see Fig. S4.b and c. We tested whether this noise came from the Boolean truncation involved in the calculation, or from the basis set, which, derived from the actual cavities along the trajectory, could intrinsically be rugged. For that, we analyzed synthetic cavities stretched linearly along one axis in a motion resembling to a single mode of evolution (Fig.S5.a). Noticeably, the eigenvalue spectra (Fig.S5.b) showed a more elaborate picture illustrating the importance of the basis set and the difficulty to render motion of object presenting sharp edges (Fig.S5.c).

Hence, the PCA on the synthetic cavities (Fig. S5) revealed and quantified some potential limits of the method due to the Boolean nature of the cavities encoding and limitations in the basis set definition restricted to s dimension, largely smaller than the full dimension of the cavity space, n^{cav}, in most cases. This should be

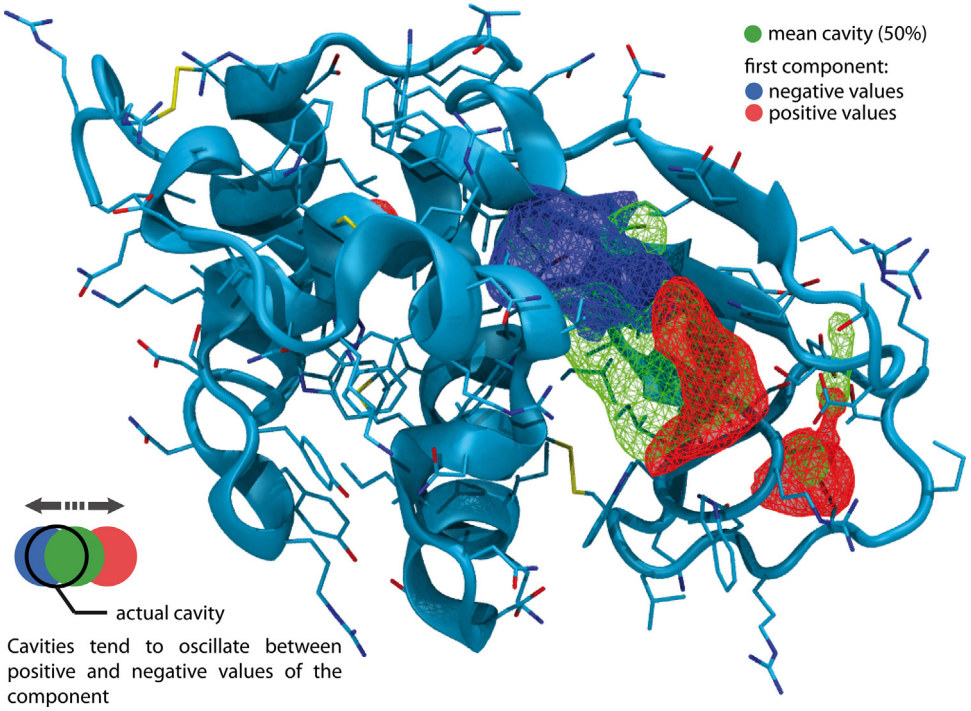

Fig. 2. First component of a 100 ns lysozyme trajectory. In green mesh, the 50% isosurface of the mean cavity occupancy. The first cavity component is shown by blue (resp. red) meshes representing isosurfaces cut at negative (resp. positive) intermediate values.

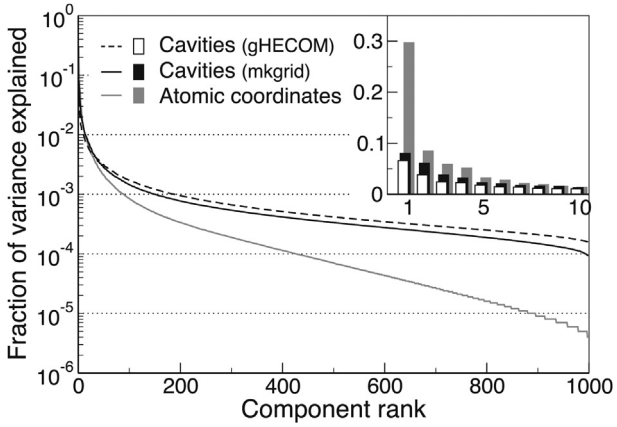

Fig. 3. Explained variance per component. Contributions of components to the global variance in a 100 ns trajectory of lysozyme. mkgrid: solid line (inset: black boxes); gHECOM: dashed line (inset: white box) and atomic coordinates: gray line (steps on lower right are due to numerical round-off, inset: gray boxes). *Y*-scale is logarithmic for the main graph and linear for the inset.

kept in mind when interpreting cavity PCA and comparing principal components of cavities and atomic coordinates.

4.3. Cavity and atomic component comparison

Comparison of eigenvalues revealed that the first components accounted for a larger fraction of the variability for the atomic coordinates than for the cavities (gHECOM or mkgrid) (Fig. 3), once again emphasizing the 'noisy' nature of cavities.

The dot product between cavities and atomic coordinate components expressed in the step space quantifies their correlation. Dot product values organized in matrices along the respective eigenvalue ranks showed a high correlation between high rank components (in absolute value since the sign of eigenvectors is arbitrary; Fig. 4). This correlation faded for lower rank components (global matrix in the lower-right corner).

Quantification by the RMSIP showed large overlap between cavity and atomic coordinate components (Figs. 4 and S6/7). RMSIP increased with the size of the cavities and were, hence, larger for gHECOM-detected cavities. The correlation for the three isolated cavities with their respective pockets (Fig. S7) are slightly lower than the correlations of their respective complete systems but remained highly significative, especially for the high rank components. Noticeably, the correlations were at least as high as that for lysozyme or myoglobin.

Hence, correlations appeared to be dependent on the systems, in some instances on the cavity detection program, but not significantly on the length of the trajectory. Dot product matrices and RMSIP of the three single cavities showed very little dependence of the correlations with alignment.

The significant level of correlation between the atomic and cavity components indicates that the limits suggested by the synthetic cavities study mentioned in the previous subsection and Fig. S5 are not too restrictive in practical cases. This correlation also suggests that beyond a convenient method to characterize the evolution and diversity of cavity, this approach could be a useful tool to relate atomic positions and cavity geometry.

4.4. Cavity reconstructions

Cavities can be rebuilt along the direction of a principal component, v_j, starting from the average cavity, *C*. The calculation is first made with real values on a grid and then converted to Booleans by truncation with a threshold value (see Methods). Fig. S8 shows the effect of Boolean truncation on the linearity of cavity reconstruction for synthetic cavities as a basic control, and then on lysozyme cavities in MD as an example. Noticeably, at low extension along the direction of the component (projection close to zero), the reconstruction is nearly orthogonal to that component, while it gradually aligns to it when the absolute value of the projection increases. Interestingly, for extensions that are similar or slightly larger than that of the original data set, the reconstructions align quite well with the component, thus closely resembling the target geometry.

4.5. Reconstruction of atomic coordinates using cavities principal components

Beyond building cavities aiming at a target geometry, we tested whether it was now possible to identify/build atomic structures with cavities closely approaching a target geometry.

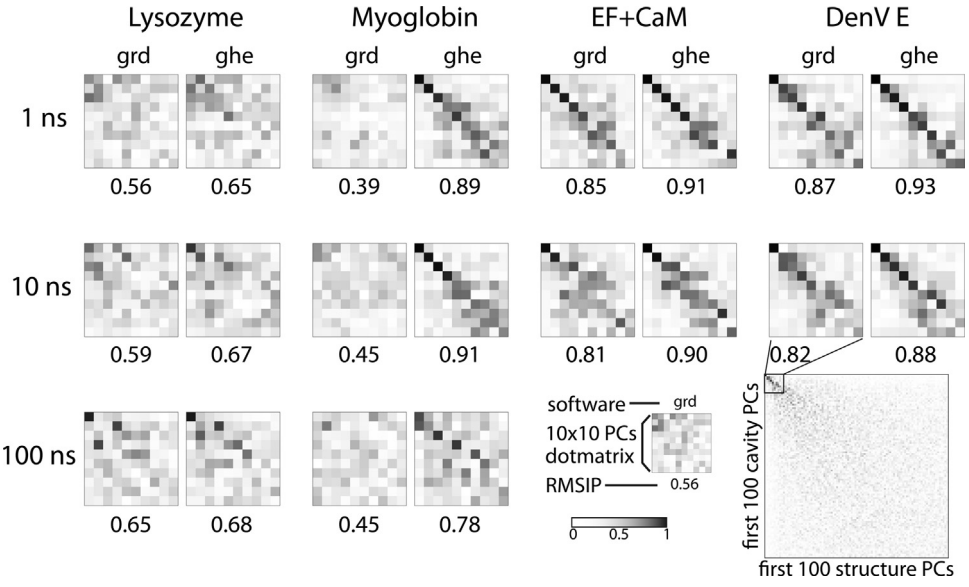

Fig. 4. Similarities of cavities and atomic coordinates components. Dot product matrix of the first ten principal components of cavities and atomic coordinates for the four studied protein systems. Trajectories of 1 ns, 10 ns and 100 ns (when available), were analyzed with both mkgrid (grd) and gHECOM (ghe). Absolute values are given in gray scale from white (0, no similarity) to black (1, identity). RMSIP values are given under each matrix.

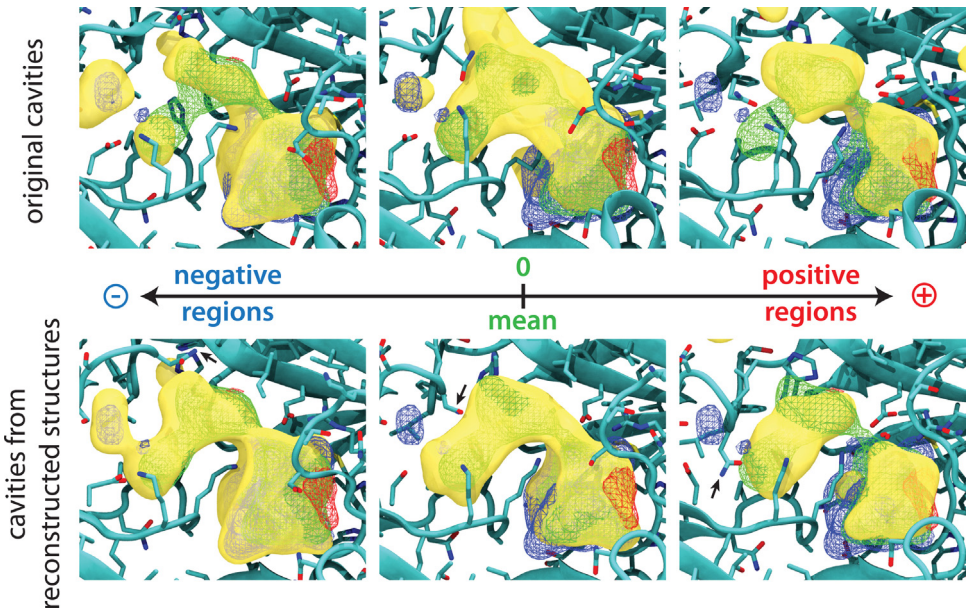

Fig. 5. Original cavities vs. Cavities of reconstructed structures. Cavity derived from a 10 ns MD trajectory of Dengue E protein analyzed by mkgrid. Top: Original cavities from the trajectory that best matches evolution along the first cavity component (highest I_{self} and almost longest extension along the component). Bottom: Cavities of structures reconstructed along the same cavity component at the maximum extensions observed in the trajectory. Negative, null, and positive projection valued cavities are displayed respectively on the left, middle, and right of the horizontal axis. Cavities, average cavity, positive and negative regions of the cavity component are displayed with yellow, green, red and blue meshes respectively.

A first possibility is to search for such a conformation within a preexisting set such as a MD trajectory. Projection on cavity PCA components can readily be used for that. An illustration is given on top of Fig. 5 and in the next section of this article.

The high similarity between atomic and cavity components suggests another possibility: use of cavity components to build such structures. To test this approach we built atomic trajectories along cavity components with the formulation of the PCA in the step space.

Fig. 5 shows a comparison of the original cavity sets that are closest to the first cavity component, and the cavity sets of the reconstructed structures along that same component (from a 10 ns trajectory of Dengue E). Noticeably, the cavity sets from reconstructed structures appeared closer to the PC than the best matching cavity sets from the original trajectory.

This observation was confirmed by quantification of the similarities between the model cavities derived from components and the cavities of the original or reconstructed structures (Fig. S9.a). Noticeably, the geometry of cavities from the reconstructed structures (dashed line) appeared to be closer to the principal component (solid line) than any of the original cavities (dots) in the projection range of the original trajectory. Extrapolation beyond the original trajectory range gives more divergent results for the cavities of the reconstructed structures (black dashed lines vs. plus signs in Fig. S9.a). This is not surprising since cavity geometry is not a linear function of the atomic coordinates in general. The same calculation was repeated for all the systems and summarized by the maximum value of I_{self} for sake of conciseness in Fig. S9.b. This relation appeared dependent on the system, as original cavity sets of smaller systems appeared closer to the principal component than the cavities of the reconstructed structures. Nonetheless, on systems such as myoglobin, cavities from cavity-component reconstructed structures, reconstructed cavities and original cavities can come very close to the targeted cavity component. Moreover, Fig. S9.b shows almost no time scale dependence for these results, suggesting that moderately long dynamics could be sufficient to perform such an analysis.

4.6. Example of application: selecting conformations with diverse cavity geometry

As an illustration, we used PCA on cavities to select a set of conformations of the dengue E β-OG pocket that is representative of different cavity geometries. The pocket delineation yielded 56 residues, for an average of 9.2 surrounding residues per conformation. PCA helped to reduce the dimensionality of the cavity space from several thousands (the number of grid points) down to 100, capturing 74% of the total cavity variance. Then, k-means clustering was used to select diverse cavity geometries. For comparison, the same clustering procedure was used with the atomic structure of the pocket. The first 100 components of the pocket captured 91% of the total structural variance. Fig. 6 shows the resulting representative cavities and the cluster averages, and table S1 their respective volumes. In this example, cavities of the conformations selected by making use of structures have smaller overlaps and are thus slightly more diverse than the ones selected using cavities. Interestingly, cavities of the conformations selected with cavity PCA had a higher intra-cluster average overlap showing a more self-consistent and representative clustering than with structural PCA. Lastly, cavities selected by PCA on cavities appeared to have quite larger volumes and volume ranges (125 ± 57 Å^3) than cavities selected by PCA on structures (105 ± 49 Å^3).

4.7. Example of application: CO migration and cavity evolution

The relation between CO migration and the evolution of the internal cavities of myoglobin was analyzed on MD trajectories (see Methods). CO was within myoglobin for 248 ns (24,800 trajectory steps) out of the 360 ns recorded trajectories. In the first MD run, CO visits the distal pocket (DP), the xenon binding site 4 (Xe4), then goes back to DP and stays there until the end of the 120 ns. In the second MD run, CO visits DP, and then oscillates between Xe4, Xe2 and Xe1, before entering Xe3 almost at the end of the 120 ns. Finally, in the third MD run, CO visits DP, then a small pocket over DP (considered as being part of DP by the DBSCAN algorithm), and exits through a gate

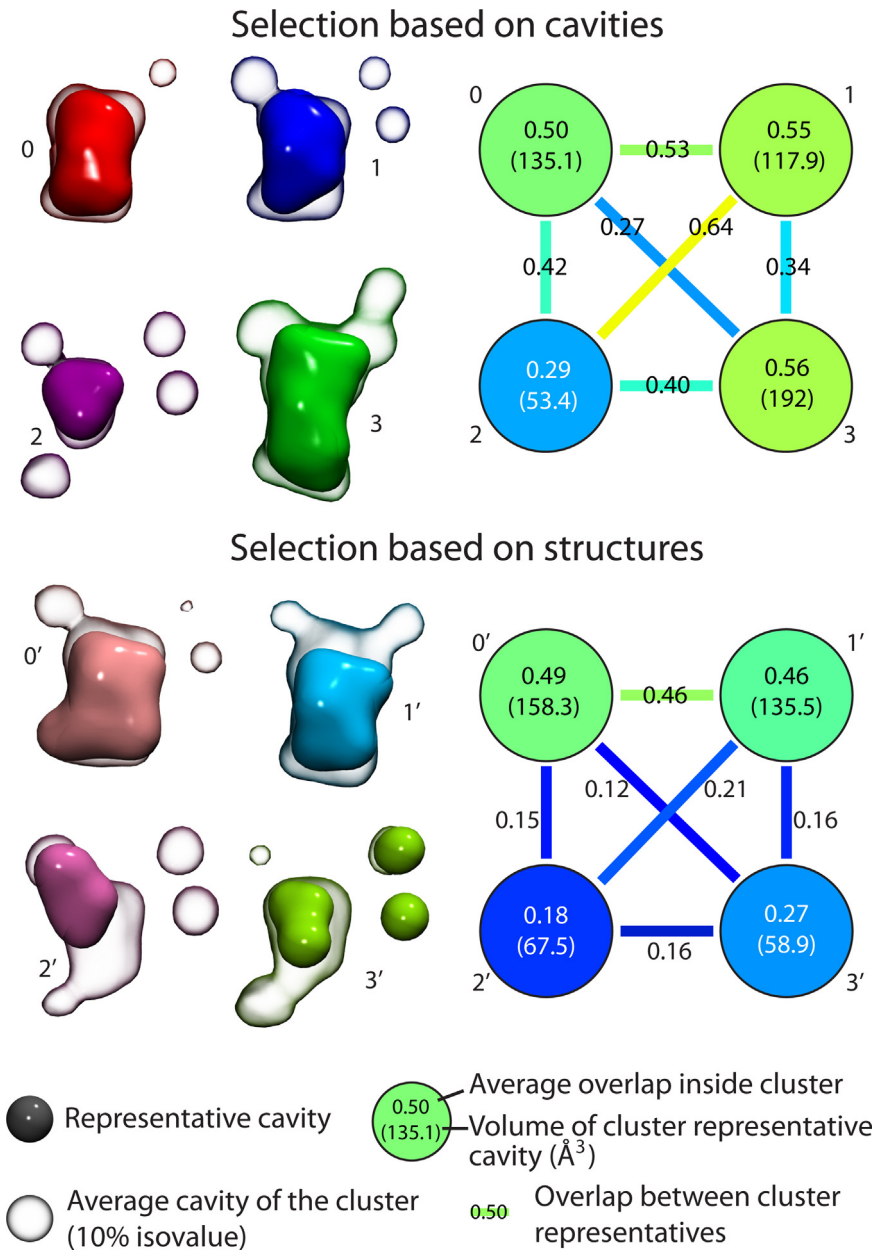

Fig. 6. Diversity assessment of cavities selected by cavity shape or structures. Representative cavities (opaque) and average cavities (10% isosurface) of 4 cavity clusters (left: cavity based clustering, right: structures based) of the dengue E β-OG pocket (gHECOM). Diagrams give the measure of the average overlaps of clusters cavities with their centroids in circles and the overlaps between the different centroids (lines) for the two types of clustering. Color code goes from blue (0: no overlap) to red (1: perfect overlap, identity) and corresponds to the value given over the symbol or next to it.

formed by the residues F46, H48, L49, M55, D60 and L61 after 8 ns. Steps were grouped according to the site in which CO was bound. We present the results according to an implicit (DP-Xe4-Xe2-(Xe1)-Xe3) order, which is compatible with the transitions observed in the dynamics, but should not affect the observations. Table 3 gives the number of conformations for each binding site and for a so-called "Noise" cluster as given by the DBSCAN algorithm.

Table 3
Number of steps with CO bound for each site, as defined by DBSCAN.

Site	DP	Xe4	Xe2	Xe1	Xe3	"Noise"
# conf.	12,224	1789	1559	8954	231	43

Fig. 7 shows the CO oxygen positions accumulated along the trajectories. Cavities were calculated with mkgrid on these trajectories and analyzed as a unique set.

Internal cavities found when CO is bound to each specific site were grouped and averaged. These average cavities were then projected along the cavity principal components of the global set of 24,800 conformations (Fig. S10, absolute projection value given). Comparison of these projections with the RMS value of the projections of all the conformations on each PC (gray zone in Fig. S10) showed that the first PCs (1 to 10–15) were at least as represented as expected. Consequently, the average cavity sets can be well represented with those first PCs. The first PC is highly represented for all the cavity sets. Other specificities are more difficult to identify except for a strong contribution of the 3rd PC in the internal cavity set of the somewhat central Xe2 binding site. Nonetheless,

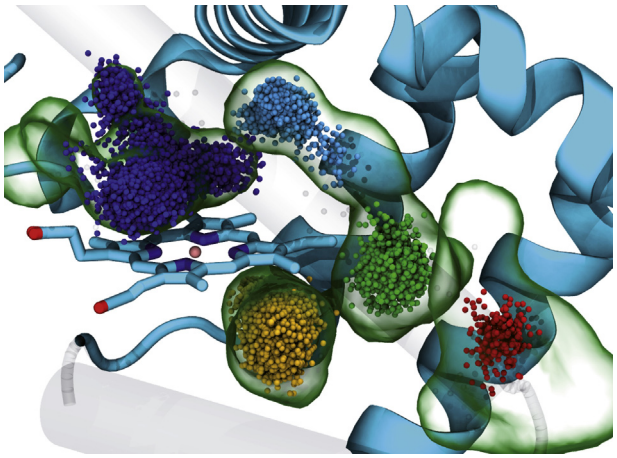

Fig. 7. Definitions of carbon monoxide binding sites in myoglobin. Positions of the oxygen atom of carbon monoxide in myoglobin during the 248 ns concatenated trajectory (24,800 positions). Color code corresponds to the DBSCAN binding site definition: distal pocket (DP) in blue, Xe4 in cyan, Xe2 in green, Xe1 in yellow, Xe3 in red. The 3% isosurface of the average cavity is shown in green ghost contour.

the first 10–15 PCs appeared to be good indicators of the CO position.

The average cavity set differences were then calculated for pairs of adjacent binding sites. They are shown in Fig. 8a and display the average evolution of the cavities geometry when the CO position is switched from one site to one of its neighbors. The observed changes were significant, thus, substantiating the notion of "cavity transfer modes". Fig. 8b shows the absolute value of the projections of those average cavity set differences on the global components. Noticeably, the contribution of the first PC, which was present in all the individual cavity sets, seems to play a more specific role with a strong contribution for the (Xe1–Xe2) transition and a smaller one for the other transitions. This finding was not *a priori* obvious, because it both requires that the contributions have the same amplitude and the same sign in the compared average cavity sets. Noticeably, those differences of average cavity sets had comparable or larger amplitudes than the root mean square projections on the first few (~10) principal components (gray zone). Hence, they appeared to synthetically depict key geometrical cavity variations. This correspondence fades rapidly for subsequent components (Fig. 8.b, inset), possibly even faster than for the individual average cavity sets (Fig. S10). Hence, the first few components appeared to account quite well for the impact of CO diffusion on cavities.

This finding is quantitatively substantiated in Table 4 which shows that the deviation of the average cavity differences is of the same order than that of the global cavities trajectory, and much larger than that of random binding site series.

Table 4
Amplitude of the transfer cavities. The norm of the transfer cavity is given (see Section 3 and Fig. S10 for definition and interpretation). The rightmost column gives the ratio of that norm to the total RMSD of the cavities trajectory. Transfer cavity norm for random series as defined in Fig. 8 is given in brackets for reference.

Site 1	Site 2	Transfer cavity norm	Ratio to total RMS deviation (%)
DP	Xe4	7.1 (3.1)	35.0
Xe4	Xe2	12.8 (3.9)	63.5
Xe2	Xe1	13.9 (4.6)	69.1
Xe2	Xe3	15.5 (8.9)	76.7

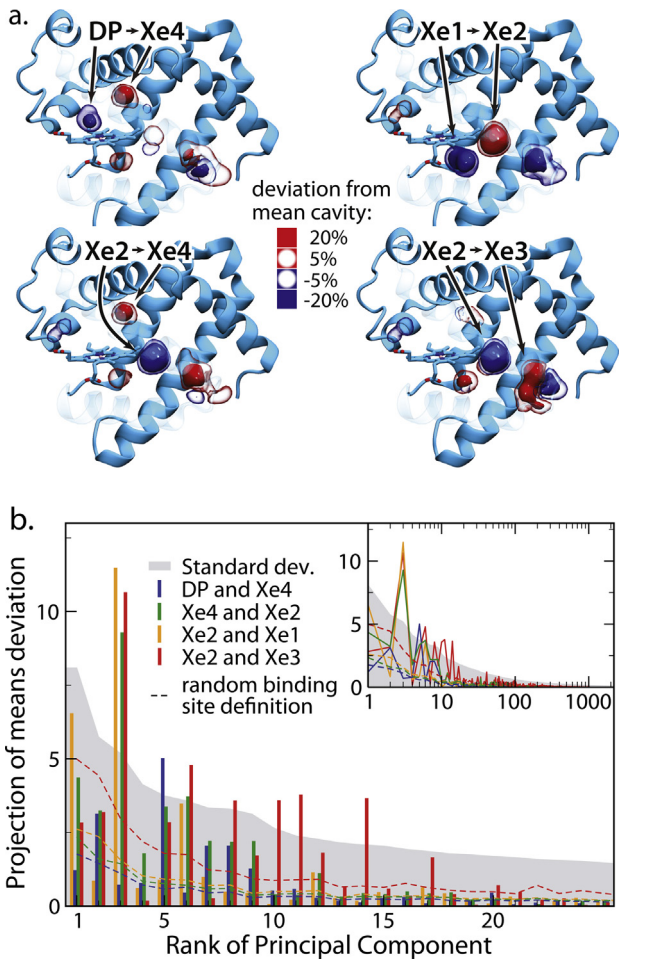

Fig. 8. Location and amplitude of the cavity transfer modes. (a) Average cavity set difference between conformations where CO is bound to either site of a pair of adjacent CO binding sites. From top to bottom, left to right: Xe4–DP, Xe4–Xe2, Xe2–Xe1, and Xe3–Xe2. Positive isosurfaces of the deviation from the global average cavity are shown in red and negative isosurfaces in blue (plain surface: 20%, contour: 5%). (b) Absolute value of projections of those average cavity set differences on cavity principal components (see Section 3 and Fig. S10). Difference between DP and Xe4 sites (DP-Xe4) is shown in blue, Xe4-Xe2 in green, Xe2-Xe1 in yellow and Xe2-Xe3 in red. Root mean square deviations of projections are represented by a gray shade. Dashed lines give the null hypothesis expected absolute value of the projection from 20 random binding site series as defined in the methods (same color code).

5. Discussion

The geometrical evolution of cavities was analyzed for four different proteins systems subjected to molecular dynamics. The first finding was that cavities evolve widely during MD, and wander almost everywhere, covering in some examples a cumulative volume comparable to that of the solvent excluded volume of the protein (Fig. S1 and Table 2). Using linear descriptors we could decompose this apparently chaotic evolution into interpretable modes with Principal Component Analysis. Favorable mathematical relations allowed us to compare principal components of atomic coordinates and cavities, unveiling the high correlation between the two (Fig. 4). This suggested that this approach allows a parallel monitoring and handling of cavity evolution and atomic displacement. Indeed, cavity components could be used to identify conformations with given cavity geometry from its molecular trajectory (ex: Fig. 5). With this approach, we could even design new conformations driven by cavities components, in some cases resembling the targeted geometry even

more than any of the conformation from the trajectory (Figs. 5 and S9a).

We have presented two examples of application as proof of concept for the use of cavities dynamics for drug design and functional analysis of protein. In the first application, PCA was used to reduce cavity descriptor dimension to efficiently cluster and select cavities within a large set of conformations (Fig. 6). The second application uses PCA as a discovery tool to unveil the very dynamic nature of internal cavities of myoglobin and their relations with the diffusion of CO in the protein (Fig. 8).

In light of these results, PCA on cavities appeared to have a strong potential in improving the relevance of virtual screenings in drug design, as well as in versatile analyses of protein function.

5.1. Limitations of the usage of PCA on cavities

PCA on cavities could present several limitations due to the nature of the cavities descriptors. First, since cavities are defined at absolute positions in space, the analyses and especially PCA can be very sensitive to the alignment of the protein system. In fact, misaligned conformations or very mobile domains could produce a lot of uninformative variance, which could impact the shape and contribution of high rank cavity components. Interestingly, in the presented work this alignment aspect did not appear to cause detectable difficulties. Despite being central to the PCA approach described here, the grid format can raise data size problems. On large systems with many conformations (>10,000), cavities can take up to several gigabytes of memory, requiring powerful hardware according to current standards for their manipulation. One direction of development would be to differentiate "solvent" cavities and "vacuum" cavities. They could have a substantially different status in analyzing functional binding or in a drug design application. However, this would require a profound enhancement of the computational methods developed here despite a similar level of theoretical ingenuity. Finally, PCA is classically formalized in real valued vector space. Its application, here, to Boolean vectors required truncation, which resulted in some mathematical "leaks" towards high rank components. However, the high correlation between cavities and structures components, as well as the relevance of cavity geometries resulting from conformation reconstruction, showed that this theoretical limitation had surprisingly little impact on the use of cavities components in the presented practical cases.

5.2. PCA unveils hidden cavity evolution and its relation with protein function

Cavity evolution along time proved to be very dynamic, with small and large cavities frequently appearing or disappearing at almost any place in the protein. Despite this apparent noisy behavior, cavities appeared to follow more global variation schemes, which could only be identified with a method such as PCA. This opens new avenues to inspect protein dynamics, through the study of cavities evolution. Furthermore, use of the step space relates cavities and structures and allows unveiling the link between cavity geometry and functional state. Hence, despite the fact that cavities could appear as a deformed, blurred and thus poorly informative derivation of the atomic position, the discovery of a strong connection between internal cavity modes and CO diffusion in myoglobin highlights the relevance of this analysis. Noticeably, study of O_2 diffusion in myoglobin should also be highly informative, but the smaller available biophysical data and the quadrupolar nature of that ligand requiring involved simulation methods reduce its tractability.

Thus, this approach appears to be a useful tool that can complement traditional analysis of protein dynamics. We believe, for instance, that the functional analysis of cavities with PCA opens new opportunities to better understand protein systems in which cavities play a major role. For example in the cytochrome p450 family, such an analysis could help refine our understanding of the pathways and mechanism for the exit of water molecules and ligands [9,10,50]. Similarly, in the permease family, it could give insights on the recognition and transport mechanisms [51–53]. Additionally, this kind of analysis can be used to search for allosteric cavities in proteins exhibiting large motions [34].

5.3. Strong relation between structure and cavities should open new opportunities in drug design

Comparison of the principal components of structures and cavities highlighted the strong correlation of their evolution. Using this fact and the interchangeability between structure and cavities components, we could show that it is possible to build new protein conformations with cavities having a given geometry. Nonetheless, structures resulting from this linear reconstruction starting from an average structure should be checked to avoid distortions.

The method presented here is expected to be quite valuable in drug design, following the examples of conformation building methods, such as pressurization [40], fumigation [41] or SCARE [42] that have shown promise. Among the advantages of our approach, the use of an existing conformation sample strengthens the relevance of the conformer selection, and singles out the flexible zones of the receptors. It can be used to build conformations with specific features encompassed in the dynamical components of the cavity. Hence, based on the component trends, a cavity extending toward interesting residues, diversely opened cavities, merging of existing small cavities, can be built in a controlled manner. As a result, this procedure seems very promising to improve relevance and diversity in drug design projects.

6. Conclusions

Principal Component Analysis (PCA) of cavities, despite some technical limitations, proved to be powerful and robust and opens new opportunities to visualize and explore the dynamics of protein cavities. This technique is a powerful and precise way to decompose and classify dynamical evolution of cavity geometry, which is especially useful given the very volatile and noisy behavior of cavities. Subtle functional mechanisms involving cavities can easily be unveiled and analyzed as shown here on the role of internal cavities evolution in the diffusion of carbon monoxide in myoglobin. We also expect that this tool can help in improving the relevance of virtual screening for drug design applications. Selection and/or reconstruction of conformations using knowledge of cavity geometry dynamics are examples of applications to establish strategies in drug design. It is also a simple yet sound way to fully exploit the geometric diversity to incorporate flexibility into virtual screening, for example, to reduce risks in drug design projects by increasing the diversity of hit scaffolds. Applications in drug design will provide more information on their relevance.

Acknowledgement

This work has been supported by the AXA Research Fund.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2014.10.011.

References

[1] S.J. Wodak, J. Janin, Structural basis of macromolecular recognition, in: J. Janin, S.J. Wodak (Eds.), Protein Modules and Protein–Protein Interaction, Volume 61 of Advances in Protein Chemistry, Academic Press, Amsterdam, 2002, pp. 9–73.
[2] K.D. Corbett, T. Alber, The many faces of Ras: recognition of small GTP-binding proteins, Trends Biochem. Sci. 26 (12) (2001) 710–716.
[3] R.L. DesJarlais, R.P. Sheridan, G.L. Seibel, S.J. Dixon, I.D. Kuntz, R. Venkataraghavan, Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure, J. Med. Chem. 31 (4) (1988) 722–729.
[4] R. Chen, Z. Weng, A novel shape complementarity scoring function for protein-protein docking, Proteins 51 (3) (2003) 397–408.
[5] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites, J. Mol. Gr. Modell. 21 (4) (2003) 289–307.
[6] A. Elisabeth Eriksson, W.A. Baase, X.-J. Zhang, D.W. Heinz, M. Blaber, E.P. Baldwin, B.W. Matthews, Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect, Science 255 (5041) (1992) 178–183.
[7] R. Elber, M. Karplus, Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin, J. Am. Chem. Soc. 112 (25) (1990) 9161–9175.
[8] R. Elber, Ligand diffusion in globins: simulations versus experiment, Curr. Opin. Struct. Biol. 20 (2) (2010) 162–167.
[9] R.C. Wade, P.J. Winn, I. Schlichting, Sudarko, A survey of active site access channels in cytochromes {P450}, J. Inorg. Biochem. 98 (7) (2004) 1175–1182.
[10] Y. Miao, J. Baudry, Active-site hydration and water diffusion in cytochrome p450cam: a highly dynamic process, Biophys. J. 101 (6) (2011) 1493–1503.
[11] A.G. Watts, I. Damager, M.L. Amaya, A. Buschiazzo, P. Alzari, A.C. Frasch, S.G. Withers, *Trypanosoma cruzi* trans-sialidase operates through a covalent sialyl-enzyme intermediate: tyrosine is the catalytic nucleophile, J. Am. Chem. Soc. 125 (25) (2003) 7532–7533.
[12] S. Liu, J.S. Chang, J.T. Herberg, M.-M. Horng, P.K. Tomich, A.H. Lin, K.R. Marotti, Allosteric inhibition of *Staphylococcus aureus* D-alanine:D-alanine ligase revealed by crystallographic studies, Proc. Natl. Acad. Sci. U. S. A. 103 (41) (2006) 15178–15183.
[13] R.D. Vale, R.A. Milligan, The way things move: looking under the hood of molecular motor proteins, Science 288 (5463) (2000) 88–95.
[14] J. Simon, Hubbard, Patrick Argos, Cavities and packing at protein interfaces, Protein Sci. 3 (12) (1994) 2194–2206.
[15] Shrihari Sonavane, Pinak Chakrabarti, Cavities and atomic packing in protein structures and interfaces, PLoS Comput. Biol. 4 (9) (2008) e1000188, 09.
[16] Susanne Eyrisch, Volkhard Helms, What induces pocket openings on protein surface patches involved in protein–protein interactions? J. Comput. Aided Mol. Des. 23 (2) (2009) 73–86.
[17] Cecilia Bossa, Andrea Amadei, Isabella Daidone, Massimiliano Anselmi, Beatrice Vallone, Maurizio Brunori, Alfredo Di Nola, Molecular dynamics simulation of sperm whale myoglobin: effects of mutations and trapped CO on the structure and dynamics of cavities, Biophys. J. 89 (1) (2005) 465–474.
[18] J.Z. Ruscio, D. Kumar, M. Shukla, M.G. Prisant, T.M. Murali, A.V. Onufriev, Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin, Proc. Natl. Acad. Sci. U. S. A. 105 (27) (2008) 9204–9209.
[19] M.A. Scorciapino, A. Robertazzi, M. Casu, P. Ruggerone, M. Ceccarelli, Breathing motions of a respiratory protein revealed by molecular dynamics simulations, J. Am. Chem. Soc. 131 (33) (2009) 11825–11832.
[20] M. Gabba, S. Abbruzzetti, F. Spyrakis, F. Forti, S. Bruno, A. Mozzarelli, F. Javier Luque, C. Viappiani, P. Cozzini, M. Nardini, F. Germani, M. Bolognesi, L. Moens, S. Dewilde, CO rebinding kinetics and molecular dynamics simulations highlight dynamic regulation of internal cavities in human cytoglobin, PLOS ONE 8 (1) (2013) e49770, 01.
[21] Susanne Eyrisch, Volkhard Helms, Transient pockets on protein surfaces involved in protein–protein interaction, J. Med. Chem. 50 (15) (2007) 3457–3464.
[22] P. Schmidtke, A. Bidon-Chanal, F. Javier Luque, X. Barril, MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories, Bioinformatics 27 (23) (2011) 3276–3285.
[23] M. Krone, M. Falk, S. Rehm, J. Pleiss, T. Ertl, Interactive exploration of protein cavities, Comput. Gr. Forum 30 (3) (2011) 673–682.
[24] Alexander Metz, Christopher Pfleger, Hannes Kopitz, Stefania Pfeiffer-Marek, Karl-Heinz Baringhaus, Holger Gohlke, Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein–protein interface, J. Chem. Inf. Model. 52 (1) (2012) 120–133.
[25] P. Ashford, D. Moss, A. Alex, S. Yeap, A. Povia, I. Nobeli, M. Williams, Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets, BMC Bioinformatics 13 (1) (2012) 39.
[26] D.B. Kokh, S. Richter, S. Henrich, P. Czodrowski, F. Rippmann, R.C. Wade, TRAPP: a tool for analysis of transient binding pockets in proteins, J. Chem. Inf. Model. 53 (5) (2013) 1235–1252.
[27] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (3) (1971) 379–400.
[28] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (4612) (1983) 709–713.
[29] J. Liang, C. Woodward, C. Woodward, H. Edelsbrunner, H. Edelsbrunner, Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design, Protein Sci. 7 (9) (1998) 1884–1897.
[30] S.J. Lee, M.S. McCormick, S.J. Lippard, U.-S. Cho, Control of substrate access to the active site in methane monooxygenase, Nature 494 (7437) (2013) 380–384, 02.
[31] B.W. Matthews, Structural and genetic analysis of the folding and function of t4 lysozyme, FASEB J. 10 (1) (1996) 35–41.
[32] S.J. Hubbard, P. Argos, A functional role for protein cavities in domain:domain motions, J. Mol. Biol. 261 (2) (1996) 289–300.
[33] Y. Modis, S. Ogata, D. Clements, S.C. Harrison, Structure of the dengue virus envelope protein after membrane fusion, Nature 427 (6972) (2004) 313–319.
[34] E. Laine, C. Goncalves, J.C. Karst, A. Lesnard, S. Rault, W.-J. Tang, T.E. Malliavin, D. Ladant, A. Blondel, Use of allostery to identify inhibitors of calmodulin-induced activation of *Bacillus anthracis* edema factor, Proc. Natl. Acad. Sci. U. S. A. 107 (25) (2010) 11277–11282.
[35] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, A geometric approach to macromolecule–ligand interactions, J. Mol. Biol. 161 (2) (1982) 269–288.
[36] S. Pérot, S. Olivier, M.A. Miteva, A.-C. Camproux, B.O. Villoutreix, Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery, Drug Discov. Today 15 (15–16) (2010) 656–667.
[37] C. B-Rao, J. Subramanian, S.D. Sharma, Managing protein flexibility in docking and its applications, Drug Discov. Today 14 (7–8) (2009) 394–400.
[38] J.D. Durrant, J.A. McCammon, Computer-aided drug-discovery techniques that account for receptor flexibility, Curr. Opin. Pharmacol. 10 (6) (2010) 770–774.
[39] C.D. Wassman, R. Baronio, Ö. Demir, B.D. Wallentine, C.-K. Chen, L.V. Hall, F. Salehi, D.-W. Lin, B.P. Chung, G. Wesley Hatfield, A. Richard Chamberlin, H. Luecke, R.H. Lathrop, P. Kaiser, R.E. Amaro, Computational identification of a transiently open l1/s3 pocket for reactivation of mutant p53, Nat. Commun. 4 (January) (2013) 1407.
[40] I.M. Withers, M.P. Mazanetz, H. Wang, P.M. Fischer, C.A. Laughton, Active site pressurization: a new tool for structure-guided drug design and other studies of protein flexibility, J. Chem. Inf. Model. 48 (7) (2008) 1448–1454.
[41] R. Abagyan, I. Kufareva, The flexible pocketome engine for structural chemogenomics, in: E. Jacoby (Ed.), Chemogenomics, Volume 575 of Methods in Molecular Biology, Humana Press, Clifton, NJ, 2009, pp. 249–279.
[42] G. Bottegoni, I. Kufareva, M. Totrov, R. Abagyan, A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE), J. Comput. Aided Mol. Des. 22 (5) (2008) 311–325.
[43] T. Kawabata, Detection of multiscale pockets on protein surfaces using mathematical morphology, Proteins 78 (5) (2010) 1195–1211.
[44] R.F. Tilton, I.D. Kuntz, G.A. Petsko, Cavities in proteins: structure of a met-myoglobin xenon complex solved to 1.9. ang, Biochemistry 23 (13) (1984) 2849–2857.
[45] L. Maragliano, G. Cottone, G. Ciccotti, E. Vanden-Eijnden, Mapping the network of pathways of CO diffusion in myoglobin, J. Am. Chem. Soc. 132 (3) (2010) 1010–1017.
[46] A. Amadei, M.A. Ceruso, A. Di Nola, On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations, Proteins 36 (4) (1999) 419–424.
[47] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD, J. Comput. Chem. 26 (16) (2005) 1781–1802.
[48] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, CHARMM: the biomolecular simulation program, J. Comput. Chem. 30 (10) (2009) 1545–1614.
[49] M. Ester, H.-P. Kriegel, Jörg Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise KDD, vol. 96, 1996, pp. 226–231.
[50] L. Benkaidali, F. André, B. Maouche, P. Siregar, M. Benyettou, F. Maurel, M. Petitjean, Computing cavities, channels, pores and pockets in proteins from non-spherical ligands models, Bioinformatics 30 (6) (2014) 792–800.
[51] J. Abramson, I. Smirnova, V. Kasho, G. Verner, H. Ronald Kaback, S. Iwata, Structure and mechanism of the lactose permease of *Escherichia coli*, Science 301 (5633) (2003) 610–615.
[52] I. Smirnova, V. Kasho, J. Sugihara, H. Ronald Kaback, Opening the periplasmic cavity in lactose permease is the limiting step for sugar binding, Proc. Natl. Acad. Sci. U. S. A. 108 (37) (2011) 15147–15151.
[53] V. Chaptal, S. Kwon, M.R. Sawaya, L. Guan, H. Ronald Kaback, J. Abramson, Crystal structure of lactose permease in complex with an affinity inactivator yields unique insight into sugar recognition, Proc. Natl. Acad. Sci. U. S. A. 108 (23) (2011) 9361–9366.