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