BBA - Proteins and Proteomics 1865 (2017) 1416–1422

Contents lists available at ScienceDirect

BBA - Proteins and Proteomics

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

Conformations of the HIV-1 protease: A crystal structure data set analysis

Luigi Leonardo Palese

University of Bari "Aldo Moro", Department of Basic Medical Sciences, Neurosciences and Sense Organs (SMBNOS), Bari 70124, Italy

ARTICLE INFO

Keywords:
HIV-1
Protease
Structure
Protease inhibitor
Resistance
PCA

ABSTRACT

The HIV protease is an important drug target for HIV/AIDS therapy, and its structure and function have been extensively investigated. This enzyme performs an essential role in viral maturation by processing specific cleavage sites in the Gag and Gag-Pol precursor polyproteins so as to release their mature forms. This 99 amino acid aspartic protease works as a homodimer, with the active site localized in a central cavity capped by two flexible flap regions. The dimer presents closed or open conformations, which are involved in the substrate binding and release. Here the results of the analysis of a HIV-1 protease data set containing 552 dimer structures are reported. Different dimensionality reduction methods have been used in order to get information from this multidimensional database. Most of the structures in the data set belong to two conformational clusters. An interesting observation that comes from the analysis of these data is that some protease sequences are localized preferentially in specific areas of the conformational landscape of this protein.

1. Introduction

Proteins are complex macromolecules whose functions are intimately rooted in their dynamics [1,2]. Substrate recognition and release, enzymatic activities, allosteric regulation and protein-protein interactions require conformational transitions. The in-depth analysis of the functionally important motions of a protein requires the knowledge of the multiple conformations that it can assume, and molecular dynamics is among the most popular tools to perform such task [3]. This is nowadays an extremely powerful an mature technique, which can lead to the prediction of experimentally observable quantities [4–6]. Since this technique can generate very large data sets of conformations of the system under analysis starting from a single structure, it is the method of choice when only few experimental conformations are available for a particular protein. But the improvements in NMR spectroscopy and in X-ray crystallography permit to consider, at least for some molecules, also the experimentally determined conformations as a source of data for the analysis of the functionally relevant motions [7,8]. The conformational ensembles obtained by NMR spectroscopy often contain conformers that can reveal important aspects of the protein dynamics. Beside the multiple conformers in the NMR-determined structures, today a large set of proteins has a considerable number of different conformations determined by X-ray diffraction. The number of protein structures deposited in the Protein Data Bank (PDB) [8] continues to grow at an impressive rate [9]. To date, more than 129,000 different structures are reported in this database, and among these structures 108,599 are protein structures obtained by X-ray crystallography (April 2017). The redundant subsets of structures deposited in the PDB are an extremely interesting tool to gain insight into the protein dynamics. Obviously, these static X-ray structures may not directly reflect the overall protein dynamics, but they surely represent, at least, potential conformations in the dynamical landscape of proteins.

Howsoever they came into being, these data sets of multiple conformations describe changes that occur in (very) high dimensional spaces. The large number of atoms that constitute a typical macromolecule, each described by three degrees of freedom in the sampled conformations, and the complexity of the motions, necessarily lead to consider some dimensionality reduction techniques in order to discern the key features of macromolecular dynamics. One widely used technique for dimensionality reduction is principal component analysis (PCA), which is a statistical method based on covariance (or also correlation) analysis [5,10-16]. PCA is a linear transformation that projects the original space of correlated variables into a new space of uncorrelated ones, which are called principal components (PCs). Because most of the system's variance is usually contained in a small subset of the PCs, this technique has been widely used in the dimensionality reduction task. PCA is nowadays of routine use in the analysis of the conformations sampled by molecular dynamics experiments, particularly in its essential dynamics version [17]. Moreover, PCA has been used in the analysis and classification of structures in NMR ensembles [18] as well as in ensemble of structures obtained by X-ray diffraction [19,20], suggesting that there are some relations between the collective motion predicted by the first PCs obtained from experimental ensembles and the normal modes calculated by elastic network

E-mail address: luigileonardo.palese@uniba.it.

http://dx.doi.org/10.1016/j.bbapap.2017.08.009
Received 29 April 2017; Received in revised form 22 July 2017; Accepted 10 August 2017
Available online 26 August 2017
1570-9639/

models.

Here the results of the PCA analysis performed on a large crystallographic data set of the HIV-1 protease (retropeptidase) will be reported. This enzyme plays a critical role in the life cycle of HIV [21,22], and a large number of structures of this are available in the PDB. HIV-1 protease [23] is a C2 symmetric homodimer (each subunit containing 99 residues) with a single active site located at the dimer interface. The active site contains a catalytic aspartate (D25) in the sequence signature DTG. Three regions can be recognized in each monomer: the terminal region (residues 1–4 and 95–99), which is important in the dimerization of the active enzyme; the core region (residues 10–32 and 63–85 of each monomer), which participates to the dimerization and to the catalytic site; and the flap region consisting of two solvent exposed loops (residues 33–43 of each chain) and two flexible, glycine rich β-hairpins flaps (residues 44–62 of each chain). This last region is important in the substrate (and inhibitor) binding. The flexible flaps cap the catalytic triad, D25, T26, G27, and close upon the substrate (or inhibitor) at the active site in the ligated state.

The HIV-1 protease has been the target of extensive drug discovery efforts, which have led to several inhibitors approved for clinical use, which are effective on the wild-type enzyme [24]. But the high replication rate of the virus, combined with the error-prone mechanism of the viral reverse transcriptase, rapidly generate a pool of mutant viruses, often resistant to the protease competitive inhibitors. It was demonstrated that nearly one-half of HIV-1 protease positions are under selective drug pressure [25]. This arms race between virus and researchers has produced an exceptional wealth of protease structures in the PDB. It's interesting to note that these structures are relative to several related sequences, so this ensemble of structures offers an exceptional opportunity not only to obtain information about the possible motions of the HIV-1 protease, but also to determine if a particular sequence has a defined bias towards a particular conformation. In the past years, crystallographic data sets (containing up to 164 structures of this enzyme) have been analyzed [19,20]. In this work we will describe the results of the analysis of a HIV-1 protease data set containing 552 dimer structures.

2. Materials and methods

2.1. Data set

The X-ray structures of the HIV-1 protease were obtained from the PDB [8,26,27]. Structures sharing 90% identity with the Consensus B sequence (Stanford HIV database) [28–31] were initially considered (581 structures). We restricted the database to the method used to obtain the structure (X-ray diffraction), to the protein dimers only (so excluding monomeric entries), and to the enzyme type (3.4.23.16, HIV-1 retropepsin). For successive analysis, multiple conformations of the α-carbon atoms were removed from the pdb files; pdb files containing an incorrect number of α-carbon atoms were excluded from the subsequent analysis. The maximum accepted refinement resolution was 3.1 Å. After these selection steps 552 HIV-1 protease structures, as dimer, were included in the analysis and we will refer to them as the data set.

The above described data set was employed as a starting point for two more sets. The first was composed by all the monomers present in the principal data set. The second one was a dimer data set containing only the structures that met the following criteria [32]: R_{free} value reported; at least $R_{observed}$ or R_{work} or R_{all} reported for the structure; $7 < (R_{free} - R) < 2$; resolution at least 2.6 Å. It should be noted that these criteria, and particularly the higher cutoff of $(R_{free} - R)$ parameter, may discard also high resolution structures. However, since this last data set was designed as a check against the overfitting, the use of these criteria was inevitable. Using this selection criteria, 395 dimer structures were included in this data set, to which we will refer hereafter as the high quality (HQ) data set.

All the above described data sets are reported in [33].

2.2. Data analysis

The structures contained in a data set were aligned to a common reference by a rotation and translation matrix obtained by Tcl (www.tcl.tk) scripting in VMD [34]. The atomic coordinates of the superposed structures were stored in a pdb file (again by a Tcl scrip in VMD). For the analysis, the Cartesian coordinates of α-carbon atoms of the superposed structures in the data set were extracted and arranged as a matrix by a Tcl script in VMD followed by a vi (www.vim.org) editing step, in order to obtain the data in a format readable by the numerical analysis software (see below). The data are arranged such that each matrix row represents a sample, and each column represents a degree of freedom.

Before proceeding with the analysis of data, a preprocessing step that can be described as

$$x_{std}^{i} = \frac{x^{i} - \mu_x}{\sigma_x}$$

was applied [35], where μ_x is the sample mean of a particular degree of freedom column and σ_x the corresponding standard deviation. This normalization was performed using the appropriate scikit-learn [36] built-in function. For PCA, the truncated SVD algorithm implemented in the scikit-learn software package was used [36,37]. For the monomer data set, which contains a sufficiently large number of entries, PCA was also calculated by a classical method that requires the knowledge of the true correlation matrix, which has been described in detail elsewhere [5,14,15,35,38,39]. Briefly, after the centroid subtraction, the covariance matrix of the data set was obtained as

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$$

where $\langle ... \rangle$ represents the average over all the samples in the data set. The correlation matrix is calculated from this matrix as

$$P_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

and this square symmetric matrix was diagonalized as

$$R^T P R = \Lambda$$

using standard numerical routines [40,41], where R is an orthonormal transformation matrix (whose column vectors are the eigenvectors of P), the superscript ^{T} means transposition and Λ is a diagonal matrix whose elements are the eigenvalues. After sorting the columns of the eigenvector matrix R and eigenvalue diagonal matrix Λ in order of decreasing eigenvalues, the empirical matrix was projected onto the eigenvectors to give the principal components.

The random projection algorithm was carried out exactly as the PCA, except for the fact that the square symmetric correlation matrix was replaced by a random symmetric one [42]. This random symmetric matrix M is defined as

$$M = \frac{G + G^T}{2}$$

where G is a normal distributed random square matrix, so that M belongs to the Gaussian Orthogonal Ensemble [14-16,43]. Thus, it is like a PCA with relaxed constraints relatively to the matrix to be used in calculating the new orthonormal reference system, where only the matrix symmetry is preserved. The inverse participation ratio was calculated as

$$I_k = \sum_{\alpha=1}^{N} (v_{\alpha}^{(k)})^4$$

which describes how many components of an eigenvector k (of dimension N) significantly contribute to its length [15,44]. As a measure of the magnitude of the contribution of a single amino acid in a particular eigenvector, a fluctuation index defined as