



ELSEVIER

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Analysis of the conformations of the HIV-1 protease from a large crystallographic data set



Luigi Leonardo Palese

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

ARTICLE INFO

Article history:

Received 26 August 2017

Accepted 28 September 2017

Available online 6 October 2017

ABSTRACT

The HIV-1 protease performs essential roles in viral maturation by processing specific cleavage sites in the Gag and Gag-Pol precursor polyproteins to release their mature forms. Here the analysis of a large HIV-1 protease data set (containing 552 dimer structures) are reported. These data are related to article entitled "Conformations of the HIV-1 protease: a crystal structure data set analysis" (Palese, 2017) [1].

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Specifications Table

Subject area	<i>Chemistry, Biology.</i>
More specific subject area	<i>Biochemistry, HIV-1 protease structure.</i>
Type of data	<i>Table (csv files), text file, figure, animated figures.</i>
How data was acquired	<i>Input data for analysis were obtained as pdb files from public database.</i>
Data format	<i>Raw: pdb files (as text files). Analyzed: table (csv files), text file, graph, animated GIF.</i>
Experimental factors	<i>Raw pdb files were checked for quality.</i>

DOI of original article: <http://dx.doi.org/10.1016/j.bbapap.2017.08.009>

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

<http://dx.doi.org/10.1016/j.dib.2017.09.076>

2352-3409/© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Experimental features	<i>The pdb files included in the database were analyzed by different computational protocols.</i>
Data source location	<i>Not applicable.</i>
Data accessibility	<i>Analyzed data are within this article.</i>

Value of the data

- The described data set includes a very large number of the public available structures of the HIV-1 protease.
 - The database can be useful in the drug design and analysis studies.
 - The evidence that preferential conformations are adopted by different sequences could represent an interesting benchmark for the computational prediction and fine tuning of protein structures.
-

1. Data

1.1. Data sets

The large HIV-1 protease data set used in the analysis is reported in csv format (file name HIV-1_dataset.csv). Data in this file are arranged in columns (headers in the first row): the first column reports the PDB id of each entry; the second column refers to the internal sequence id; the last two columns report the calculated first and second principal component projections, respectively (calculated by the truncated SVD method [1]). The high quality structures are listed in the file HIV-1_HQ_dataset.csv. In the file are reported the PDB id, the available quality data (R observed, R all, R work, R free, refinement resolution, and the R difference); last column reports the sequence cluster id.

The full set of fluctuations (see [1]) is reported in the file fluctuations.csv. Each row in this file represents an eigenvector (297 eigenvector describe the monomer), and each amino acid is reported as a column (99 amino acid compose the monomer).

The first and second principal modes calculated for the monomer data set are reported as animated GIF image (see [1] for details). Some relevant modes are reported as nmd file [1–3].

Supplementary material related to this article can be found online at: [doi:10.1016/j.dib.2017.09.076](https://doi.org/10.1016/j.dib.2017.09.076).

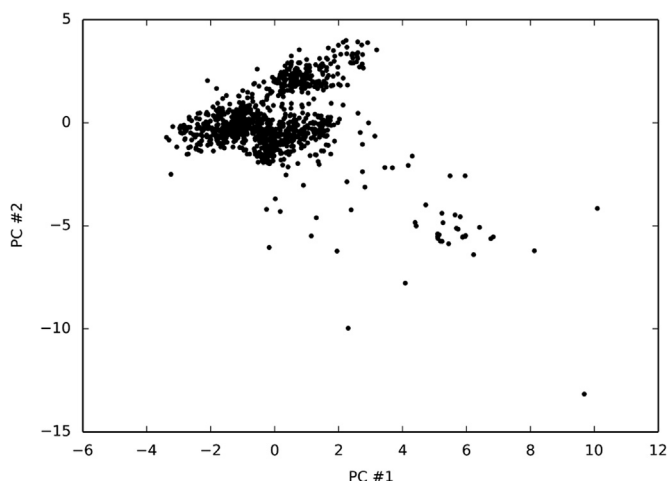


Fig. 1. The PCA of the monomer structures calculated by the covariance matrix method.

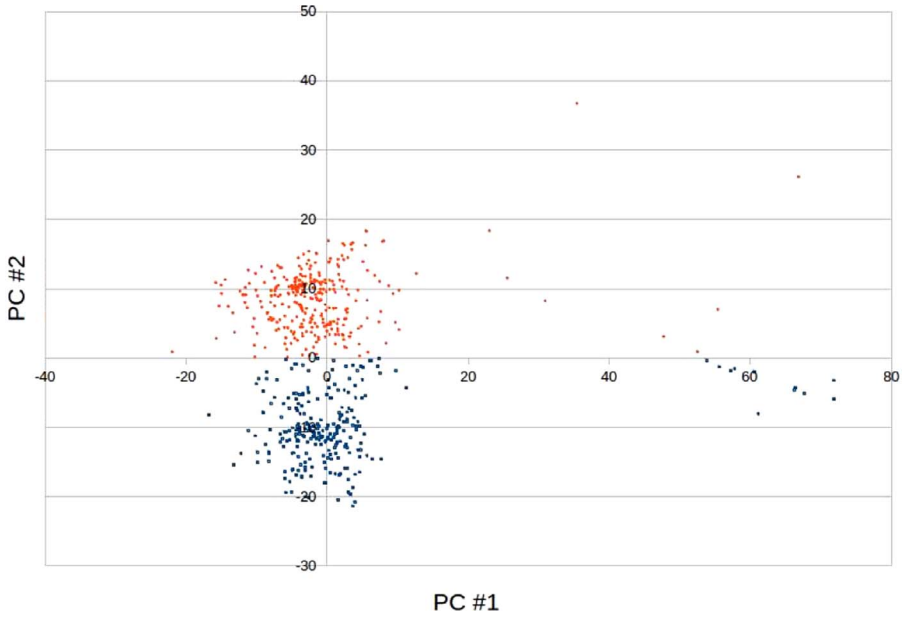


Fig. 2. PCA projection of the dimer data set. The entries are colored in blue if their second PC was negative, in red if positive.

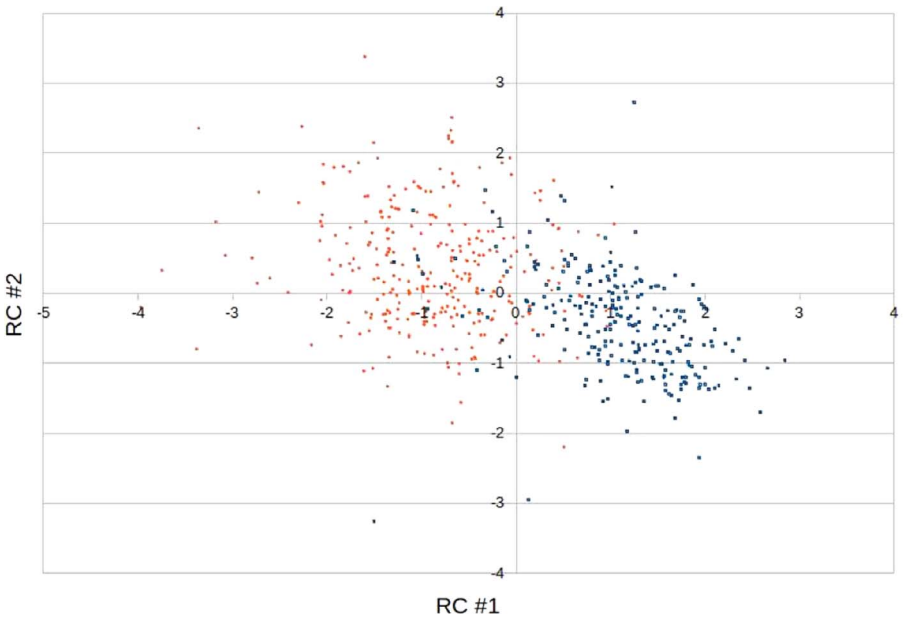


Fig. 3. Random projection of the dimer data set. Color code for each entry is the same as in Fig. 2.

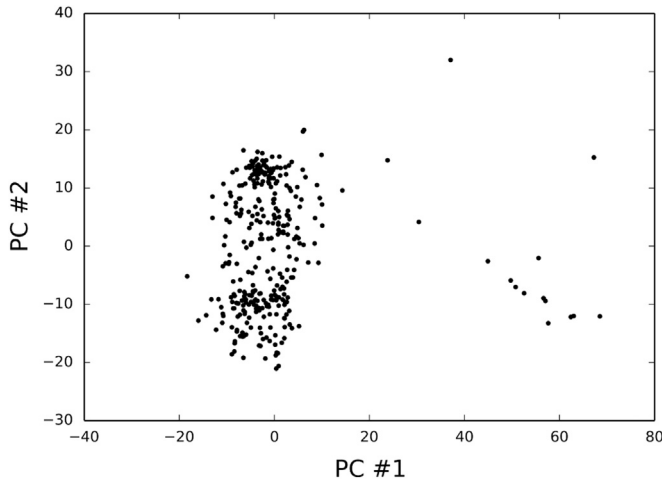


Fig. 4. PCA of the HQ dimer data set (truncated SVD method).

PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Consensus B
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 101
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 229
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 239
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 502
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 663
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 745
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 1633
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 1848
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFVKVRQYDQILIEICGHKAIGTVLVGPTFPVNIIGRNLLTQIGCTLNF	Cluster 6241

Fig. 5. Some sequence groups of the HIV-1 protease data set (see [1]).

Some results of the analysis reported in [1] on the above described data set are reported as Figs. 1–4. The reader could refer to [1] for full details.

2. Relevant sequence clusters in the data set

Some of the sequence clusters of the HIV-1 protease data set discussed in [1] are reported in Fig. 5; differences respect to the Consensus B sequence (Stanford HIV database) [2–5] are in red.

3. Experimental design, materials and methods

The structures sharing the 90% identity with the Consensus B sequence (Stanford HIV database) [4–7] were initially considered. The X-ray structures of the HIV-1 protease were obtained from the PDB [8–10]. A total number of 581 structures in the PDB met this criterion. The structures obtained by X-ray, of dimeric form, classified with an E.C. number 3.4.23.16 (HIV-1 retropepsin), and with a refinement resolution better of at least 3.1 Å were further selected. The number of alpha-carbon atoms in the downloaded pdb files was checked by the bash *grep* function after deleting the multiple conformations by the bash *sed* command. Few structures requested a further manual editing step. Finally 552 HIV-1 protease structures, as dimer, were included in the data set.

The structures contained in a data set were aligned to a common reference by Tcl (www.tcl.tk) scripting in VMD [3]. The new atomic coordinates were stored in a pdb file. For the analysis, the Cartesian coordinates of alpha-carbon atoms of the superposed structures of the data set were extracted and arranged in a matrix form by a Tcl script in VMD. Bracket in the obtained text file were removed in vi (www.vim.org). The result was that the coarse grained data conformations were

arranged in a matrix such that each row represented a sample, and each column a degree of freedom. This data matrix was analyzed by methods described in [11–25], as reported in [1].

Acknowledgements

The author thanks Prof. S. Scacco for generous support.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.dib.2017.09.076>.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.dib.2017.09.076>.

References

- [1] L.L. Palese, Conformations of the HIV-1 protease: a crystal structure data set analysis, *Biochim., Biophys. Acta*, 1865, (2017) 1416–1422.
- [2] A. Balkan, L.M. Meireles, I. Bahar, ProDy: protein dynamics inferred from theory and experiments, *Bioinformatics* 27 (2011) 1575–1577.
- [3] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, *J. Mol. Graphics* 14 (1996) 33–38.
- [4] S.-Y. Rhee, M.J. Gonzales, R. Kantor, B.J. Betts, J. Ravela, R.W. Shafer, Human immunodeficiency virus reverse transcriptase and protease sequence database, *Nucleic Acids Res.* 31 (2003) 298–303.
- [5] R.W. Shafer, Rationale and uses of a public HIV drug-resistance database, *J. Infect. Dis.* 194 (2006) S51–S58.
- [6] S.-Y. Rhee, R. Kantor, D.A. Katzenstein, R. Camacho, L. Morris, S. Sirivichayakul, L. Jorgensen, L.F. Brigido, J.M. Schapiro, R.W. Shafer, International Non Subtype B HIV-1 Working Group, HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes, *AIDS* 20 (2006) 643–651.
- [7] R.W. Shafer, D.R. Jung, B.J. Betts, Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries, *Nat. Med.* 6 (2000) 1290–1292.
- [8] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, *Nucl. Acids Res.* 28 (2000) 235–242.
- [9] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide protein data bank, *Nat. Struct. Biol.* 10 (2003) (980–980).
- [10] P.W. Rose, A. Prlić, A. Altunkaya, C. Bi, A.R. Bradley, C.H. Christie, L. Di Costanzo, J.M. Duarte, S. Dutta, Z. Feng, et al., The RCSB protein data bank: integrative view of protein, gene and 3D structural information, *Nucleic Acids Res.* 45 (2017) D271–D281.
- [11] S. Raschka, *Python Machine Learning*, Packt Publishing, Birmingham, UK, 2015.
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
- [13] N. Halko, P.-G. Martinsson, J.A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Rev.* 53 (2011) 217–288.
- [14] R. Bro, A.K. Smilde, Principal component analysis, *Anal. Methods* 6 (2014) 2812–2831.
- [15] F. Bossis, L.L. Palese, Amyloid beta (1–42) in aqueous environments: effects of ionic strength and E22Q (Dutch) mutation, *Biochim. Biophys. Acta* 1834 (2013) 2486–2493.
- [16] L.L. Palese, Random matrix theory in molecular dynamics analysis, *Biophys. Chem.* 196 (2015) 1–9.
- [17] L.L. Palese, Correlation analysis of Trp-cage dynamics in folded and unfolded states, *J. Phys. Chem. B* 119 (2015) 15568–15573.
- [18] J. Shlens, A Tutorial on Principal Component analysis, arXiv preprint arXiv:1404.1100, 2014.
- [19] S. Van Der Walt, S.C. Colbert, G. Varoquaux, The NumPy array: a structure for efficient numerical computation, *Comput. Sci. Eng.* 13 (2011) 22–30.
- [20] T.E. Oliphant, *Python for scientific computing*, *Comput. Sci. Eng.* 9 (2007) 10–20.
- [21] L.L. Palese, A Random Version of Principal Component Analysis in Data Clustering, arXiv preprint arXiv:1610.08664, 2016.
- [22] F. Pérez, B.E. Granger, IPython: a system for interactive scientific computing, *Comput. Sci. Eng.* 9 (2007) 21–29.
- [23] J.D. Hunter, Matplotlib: a 2D graphics environment, *Comput. Sci. Eng.* 9 (2007) 90–95.
- [24] L.L. Palese, Protein dynamics: complex by itself, *Complexity* 18 (2013) 48–56.
- [25] F. Bossis, L.L. Palese, Molecular dynamics in cytochrome c oxidase Mössbauer spectra deconvolution, *Biochem. Biophys. Res. Commun.* 404 (2011) 438–442.