

Original Article

Application of PolyPRep tools on HIV protease polyproteins using molecular docking

Aplicação da ferramenta PolyPRep em poliproteínas de protease de HIV usando docking molecular

M. F. R. Dias^{a,b,*} , F. L. L. Oliveira^a , V. S. Pontes^a  and M. L. Silva^{a,c} 

^aInstituto Nacional de Metrologia, Qualidade e Tecnologia – INMETRO, Laboratório de Macromoléculas, Diretoria de Metrologia Aplicada às Ciências da Vida – DIMAV, Duque de Caxias, RJ, Brasil

^bSecretaria Estadual do Espírito Santo – SEDU/ES, Vitória, ES, Brasil

^cUniversidade Federal do Rio de Janeiro – UFRJ, Instituto de Biodiversidade e Sustentabilidade – NUPEM, Macaé, RJ, Brasil

Abstract

In recent years, the development of high-throughput technologies for obtaining sequence data leveraged the possibility of analysis of protein data *in silico*. However, when it comes to viral polyprotein interaction studies, there is a gap in the representation of those proteins, given their size and length. The prepare for studies using state-of-the-art techniques such as Machine Learning, a good representation of such proteins is a must. We present an alternative to this problem, implementing a fragmentation and modeling protocol to prepare those polyproteins in the form of peptide fragments. Such procedure is made by several scripts, implemented together on the workflow we call PolyPRep, a tool written in Python script and available in GitHub. This software is freely available only for noncommercial users.

Keywords: polyprotein, fragmentation, modelling, 3D structures.

Resumo

Nos últimos anos, o desenvolvimento de tecnologias de alto rendimento para obtenção de dados sequenciais potencializou a possibilidade de análise de dados proteicos *in silico*. No entanto, quando se trata de estudos de interação de poliproteínas virais, existe uma lacuna na representação dessas proteínas, devido ao seu tamanho e comprimento. Para estudos utilizando técnicas de ponta como o Aprendizado de Máquina, uma boa representação dessas proteínas é imprescindível. Apresentamos uma alternativa para este problema, implementando um protocolo de fragmentação e modelagem para preparar essas poliproteínas na forma de fragmentos de peptídeos. Tal procedimento é feito por diversos scripts, implementados em conjunto no workflow que chamamos de PolyPRep, uma ferramenta escrita em script Python e disponível no GitHub. Este software está disponível gratuitamente apenas para usuários não comerciais.

Palavras-chave: poliproteínas, fragmentação, modelagem, estrutura 3D.

1. Introduction

Recent advancements on structural bioinformatics allow scientists to perform interaction studies on a wide range of pathogen protein structures (Chen et al., 2018). This enhances the process of gathering information, for example Rational Drug Design protocols (Rishton, 2003). Coupled with the current high performance of computational resources, High-throughput *in silico* methods for the study of interactions between several proteins and a single receptor paves the road for the development of pharmacological leads more efficiently (Cronk, 2013). Nevertheless, the problem resides when this study is performed on polyproteins. Such proteins are huge protein chains composed of functional subunits, and generally separated from the main body on developmental stages of virion maturation (Su et al., 2018). HIV-1's life cycle poses a

good example of such mechanism, in which the polyproteins Env, Gag and Gag-pol are carried outside the host cell by the immature virion. Structural analysis protocols, such as molecular docking or molecular dynamics simulation, require at least an atomic coordinate file for both ligand (substrate) and receptor ("scissor") (Könnnyü et al., 2013; Perez et al., 2010). Many of the polyproteins of infectious organisms have no resolved structure. *In silico* studies increases the potential to discover new therapeutic drugs while reducing extensive lab work (Muhammad et al., 2021). The problem resides in the fact that polyproteins possess huge structures (1400+ residues long), which are often changeable. This represents a problem on resolving experimentally their structures (Su et al., 2018). To overcome such problems and model those interactions,

*e-mail: marfedias@gmail.com

Received: November 14, 2020 – Accepted: October 16, 2021



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

we present PolyPRep, a simple tool/library, written in Python that accomplishes the fragmentation, labelling (cleavage interface) and linear 3D structure modelling for polyproteins. This modelling enables performing *in silico* protocols on polyproteins. This fragmentation protocol has been applied successfully over HIV-1 polyproteins, Gag and Gag-Pol, therefore allows the structural analyses and can be applied to other protein sequences. Nonetheless, these studies should be used with care and reported compounds should be tested in lab experiments to prove of *in silico* studies.

2. Software Description

The tool consists of the workflow with modules that are tied together by a framework for the execution of the fragmentation protocol.

2.1. Input files

The user accesses the database sequences for the polypeptide of interest (like NCBI's Genbank) and fetch the sequence file. The user must specify an interface file (.fas extension) containing the sequences of cleavage sites on Pearson Fasta format. A run_example.py file (provided alongside the library) can be used to execute the protocols. The user can either use this file or code their own file in order to execute the methods. The advantage of using this protocol is customizability since the user can adapt everything to their own scripts and programs. All the usability info is provided on the README file, accompanying the repository (GitHub, Inc., 2020).

2.2. Fragmentation and modeling procedure

This step consists of an interface search that creates a "Site Cluster" (no actual Mathematical Clustering method is applied), which consists of sequences containing the cleavage interface found plus a head and tail (cutoff) of neighboring amino acids on the sequence. The cutoff is defined by the chosen size of fragments or by the user. The output is a series of fasta files for each set (fragment size, and label, i.e. POSITIVES_4aa). The next step is the construction of 3D structures of the fragments. The current version of the software uses MODELLER (Sali and Blundell, 1993) as the structure building tool. Once the aim of this step of the study is to produce "docking able" structures (which will be stochastically modified), we used Comparative Modelling as the model building paradigm. At this stage the *model_builder* class uses a dummy (full gap) alignment, demanding the software to create random loop structures, which will be optimized by MODELLER's Structure Optimization toolset and will be sterically concise.

2.3. Batch preparation

The last step is preparing the structures for the desired protocol (Docking, Molecular Dynamics, etc.), which requires a careful protonation and electrostatics preparation (partial charges), depending on the software or protocol. PolyPRep is able to convert fragment structure file formats, using the OpenBabel suite (O'Boyle et al., 2011).

2.4. Output

The user can choose whether to use only fragmentation, Modeller, or directly OpenBabel. PolyPRep organizes workflow outcomes depending on the module. The Fragmentation module produces fragment libraries in the form of Fasta Files for each fragment size and label. Modeling module produces a series of structures. A log file is generated containing statistics about the sequence space (number of sequences, redundant fragments, interfaces found, etc.). The modeling procedure outcomes are 3D models formatted as PDB files. The OpenBabel program prepares the file for molecular docking analyses.

3. Results

To test its functionality, we applied the developed tool (Figure 1) to the large scale docking of HIV-1 Gag-pol polyprotein fragments against HIV-1 aspartyl protease (HIV-PR). Both polyproteins are cleaved by HIV-PR during viral maturation to prepare the enzymatic repertoire for the virion to infect a new host cell. There is a total of 12 cleavage interfaces annotated (Könnyű et al., 2013). The purpose was to enhance the sampling rate between positive and negative classes for more thorough analyses of the interactions between the substrate and HIV-PR. We outlined the protocol to produce six libraries of fragments, with sequence lengths ranging from 3 to 8 residues. From the polyprotein sequence (ID) we could build a total of 11,492 negative and 297 positive sequences (22, 33, 44, 55, 66 and 77, fragments respectively of fragment sizes from 3 to 8). It took the modeling procedure an average time of execution of 5 minutes per library (each fragment length and both pos. and neg.). It took the preparation protocol a maximum 2 minutes to execute on each group. Several MODELLER parameters can be tuned during this step, such as structure optimization method and model candidate numbers (using modeller's DOPEScore as filtering criteria). Even though those structures will be severely modified during both molecular dynamics and molecular docking protocols, we offer the user the option to optimize the fragments as they best fit their need. PolyPRep smartly organizes the files produced, alongside log files from those protocols, in an easy-navigable manner, fitting to the file system and environment of choice (Linux and Windows).

We docked each fragment against HIV-PR structure PDBID 1F7A (Prabu-Jeyabalan et al., 2000), obtained from the Protein Data Bank, using utilizing o programa Autodock VINA (4.2) (Morris et al., 2009). This structure was solved on a good resolution (2.1 Å), having a 10-amino acid peptide on its active site. The molecular docking energy values were significant, all with negative values (Table 1) and corroborate the poses obtained (Figure 2), showing that the models are effective. This shows that our protocol can create structure models that are suited for some of the *in silico* procedures widely used for structural analyses on protein interactions. As docking protocols are widely applied on drug design studies, we hope our software will be of help on such studies. This tool was registered at INPI – "PolyPRep – Fragmentation and Modelling 3D structures for Polyproteins" – BR512020001609-0 (Silva and Dias, 2020).

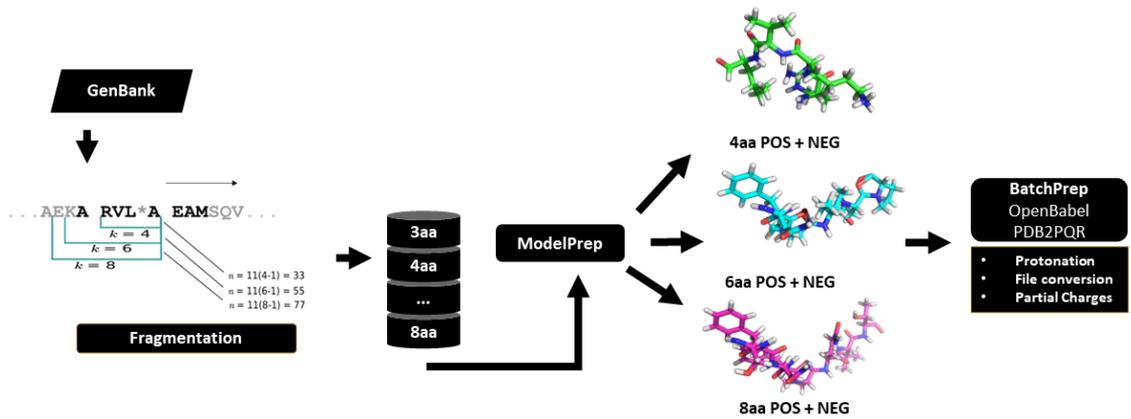


Figure 1. Polyprotein sequence fragmentation and labeling workflow. User input consists of a protein sequence in Fasta format, gathered from commonly used biological sequence databases (i.e. UniProt), and a configuration file containing cleavage interfaces. PolyPreP performs a search for cleavage interfaces and constructs positive clusters (sequences that contain cleavage sites), labeling such sequences. Negative fragments are built from a sequence space excluding positive clusters. After sequence preparation, each fragment dataset has its 3D structure modelled and prepared according to the chosen protocol. * means the site of cleavage by HIV-PR.

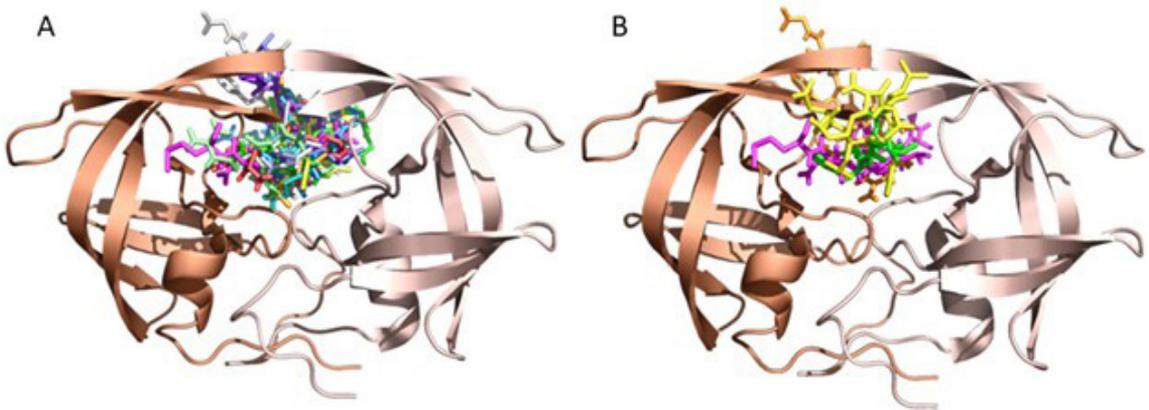


Figure 2. Representation of docking between peptides, created by PolyPreP, derived from the Gag and Gag-Pol polyproteins with HIV-protease 1F7A. A) 11 peptide docking consisting of peptides made up of 8 amino acids. B) docking with the models formed by 4, 6 and 8 amino acid residues.

Table 1. Energy values, using molecular docking, obtained through the interaction between HIV-PR 1F7A and peptides formed by 4, 6 and 8 amino acid residues (kcal/mol).

Peptides	Energy	Energy	Energy
	kcal/mol	kcal/mol	kcal/mol
	4aa	6aa	8aa
KARVL*AEAM (pep1)	-6.87	-10.8	-12.5
SQNY*PIVQ (pep2)	-6.63	-8.31	-4.41
ATIM*MQRG (pep3)	-4.84	-3.18	-5.86
RQAN*FLGK (pep4)	-9	-4.81	-6.53
RQAN*FLRE (pep5)	-9	-4.04	-3.67
PGNF*LQSR (pep6)	-8.4	-4.92	-4.8
DLAF*LQGK (pep7)	-9.78	-8.51	-2.38
SFNF*PQVT (pep8)	-9.8	-6.04	-4.62
TLNF*PISP (pep9)	-7.39	-5.82	-7.57
AETF*YVDG (pep10)	-5.84	-7.06	-8.71
RKVL*FLDG (pep11)	-8.31	-1.96	-8.87

Source: the authors. aa = means amino acid; * means the site of cleavage by HIV-PR.

Acknowledgements

This work was supported by the CNPq (*Centro Nacional de Desenvolvimento Científico e Tecnológico*), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and Inmetro (Instituto Nacional de Metrologia, Qualidade e Tecnologia).

References

- CHEN H, GUO W, SHEN J, WANG L, and SONG J., 2018. Structural principles analysis of host-pathogen protein-protein interactions: a structural bioinformatics survey. *IEEE Access*, vol. 6, pp. 11760-11771.
- CRONK, D. High-throughput screening. In R.G. HILL and H.P. RANG, eds. *Drug discovery and development*. USA: Elsevier, 2013, pp. 95-117. <http://dx.doi.org/10.1016/B978-0-7020-4299-7.00008-1>.
- KÖNNYÜ, B., SADIQ, S.K., TURÁNYI, T., HÍRMONDÓ, R., MÜLLER, B., KRÄUSSLICH, H.G., COVENEY, P.V. and MÜLLER, V., 2013. Gag-Pol processing during HIV-1 virion maturation: a systems biology approach. *PLoS Computational Biology*, vol. 9, no. 6, pp. e1003103. <http://dx.doi.org/10.1371/journal.pcbi.1003103>. PMID:23754941.
- MORRIS, G.M., HUEY, R., LINDSTROM, W., SANNER, M.F., BELEW, R.K., GOODSSELL, D.S. and OLSON, A.J., 2009. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785-2791. <http://dx.doi.org/10.1002/jcc.21256>. PMID:19399780.
- MUHAMMAD, S., MAQBOOL, M.F., AL-SEHEMI, A.G., IQBAL, A., KHAN, M., ULLAH, S., and KHAN, M.T., 2021. A threefold approach including quantum chemical, molecular docking and molecular dynamic studies to explore the natural compounds from *Centaurea jacea* as the potential inhibitors for COVID-19. *Brazilian Journal of Biology*, vol. 83, e247604. <https://doi.org/10.1590/1519-6984.247604>.
- O'BOYLE, N.M., BANCK, M., JAMES, C.A., MORLEY, C., VANDERMEERSCH, T. and HUTCHISON, G.R., 2011. Open Babel: an open chemical toolbox. *Journal of Cheminformatics*, vol. 3, no. 1, pp. 33. <http://dx.doi.org/10.1186/1758-2946-3-33>. PMID:21982300.
- PEREZ, M.A.S., FERNANDES, P.A. and RAMOS, M.J., 2010. Substrate recognition in HIV-1 protease: a computational study. *The Journal of Physical Chemistry. B*, vol. 114, no. 7, pp. 2525-2532. <http://dx.doi.org/10.1021/jp910958u>. PMID:20121080.
- PRABU-JEYABALAN, M., NALIVAIIKA, E. and SCHIFFER, C.A., 2000. How does a symmetric dimer recognize an asymmetric substrate? a substrate complex of HIV-1 protease. *Journal of Molecular Biology*, vol. 301, no. 5, pp. 1207-1220. <http://dx.doi.org/10.1006/jmbi.2000.4018>. PMID:10966816.
- RISHTON, G.M., 2003. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today*, vol. 8, no. 2, pp. 86-96. <http://dx.doi.org/10.1016/S1359644602025722>. PMID:12565011.
- SALI, A. and BLUNDELL, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, vol. 234, no. 3, pp. 779-815. <http://dx.doi.org/10.1006/jmbi.1993.1626>. PMID:8254673.
- SILVA, ML, DIAS MFR. *Polyprep*. Brasil. INPI: BR512020001609-0. 2020.
- SU, C.T., KWONG, C., VERMA, C.S. and GAN, S.K., 2018. Modeling the fulllength HIV-1 Gag polyprotein reveals the role of its p6 subunit in viral maturation and the effect of non-cleavage site mutations in protease drug resistance. *Journal of Biomolecular Structure & Dynamics*, vol. 36, no. 16, pp. 4366-4377. <http://dx.doi.org/10.1080/07391102.2017.1417160>. PMID:29237328.