International Journal of Biomedical Data Mining

ISSN: 2090-4924

OPEN ACCESS Freely available online

Short communication

PolyPRep: A Simple Tool for Fragmentation and Modelling 3D Structures of the Unresolved Polyproteins.

Fernando Limoeiro[1,2], Maria Fernanda Ribeiro Dias[2,3*], Vinicius Santos de Pontes[1] and Manuela Leal da Silva[1,2]

[1]*Universidade Federal do Rio de Janeiro,*

[2]*Instituto Nacional de Metrologia, Qualidade e Tecnologia,*

[3]*Secretaria Estadual do Espírito Santo*

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. To 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.

INTRODUCTION

Recent advancements on structural bioinformatics allow scientists to perform interaction studies on a wide range of pathogen protein structures (1). This enhances the process of gathering information, for example Rational Drug Design protocols (2). 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 (3). 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 (4). 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") (5,6). Many of the polyproteins of infectious organisms have no resolved structure. The problem resides in the fact that such polyproteins possess huge structures (1400+ residues long), which are often changeable. This represents a problem on resolving experimentally their structures (4). To overcome such problems and model those interactions, 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, and, therefore, allows the structural analyses.

SOFTWARE DESCRIPTION

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

Input files

The user can access of sequences database for the polyprotein of interest (like NCBI's Genbank) and fetch the sequence file. The user must specify a interface file (.fas extension) containing the

Correspondence to: Maria Fernanda Ribeiro Dias, Instituto Nacional de Metrologia, Qualidade e Tecnologia, Brazil, E-mail: marfedias@gmail.com

Received: November 03, 2020; **Accepted:** August 30, 2021; **Published:** September 10, 2021

Citation: Dias MFR (2021) PolyPRep: A simple tool for fragmentation and modelling 3D structures of the unresolved Polyproteins. Int J Biomed Data Min. 10:p168

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 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.

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 (7) as the structure building tool. Because the aim 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.

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 (8).

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 produced 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.

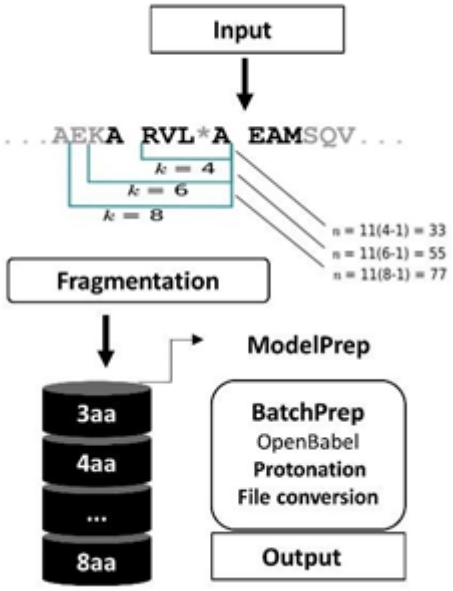

Figure 1: Polyprotein sequence fragmentation and labeling workflow.

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.

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 (5). The purpose was to enhance the sampling rate between positive and negative classes for more thorough analyses of the interactions between 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 11492 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 (9), obtained from the Protein Data Bank. This structure was solved on a good resolution (2.1 A), having a 10-amino acid peptide on its active site.

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.

REFERENCES

1. Chen, H., Guo, W., Shen, J., Wang, L., & Song, J., Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey. IEEE Access, (2018), 6, 11760–11771.
2. Rishton, G. M., Nonleadlikeness and leadlikeness in biochemical screening. Drug Discovery Today, (2003), 8(2), 86–96.
3. Cronk, D. High-throughput screening. In Drug Discovery and Development (2013), 95–117. Elsevier.
4. Su, C. T.-T., Kwoh, C.-K., Verma, C. S., & Gan, S. K.-E. 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 and Dynamics, (2017), 36(16), 4366–4377.
5. Könnyű B, Sadiq SK, Turányi T, Hírmondó R, Müller B, Kräusslich H-G, Gag-Pol Processing during HIV-1 Virion Maturation: A Systems Biology Approach. PLoS Comput Biol (2013), 9(6): e1003103.
6. Pettit, S. C., Henderson, G. J., Schiffer, C. A., & Swanstrom, R. Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. Journal of virology, (2002), 76(20), 10226–10233
7. A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. (1993), 234, 779-815.
8. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. Open Babel: An open chemical toolbox. Journal of Cheminformatics, (2011), 3(1), 33.
9. Prabu-Jeyabalan, M., Nalivaika, E., & Schiffer, C. A. How does a symmetric dimer recognize an asymmetric substrate? a substrate complex of HIV-1 protease. Journal of Molecular Biology, (2000), 301(5), 1207–1220.