CapiPy: python based GUI-application to assist in protein immobilization

David Roura Padrosa[1,*], Valentina Marchini[1] and Francesca Paradisi[1]

[1]Department of Chemistry and Biochemistry, Freiestrasse 3. 3012, Bern, Switzerland.

*To whom correspondence should be addressed.

Protein immobilization while widespread to unlock enzyme potential in biocatalysis, remains tied to the trial an error approach. Nonetheless, several databases and computational methods have been applied for protein characterization and their study. CapiPy is a user-friendly application for protein model creation and subsequent analysis with special focus on the ease of use and interpretation of the results to help in the decision of which immobilization approach would be ideal for a protein of interest. The package has been tested with two separate random sets of 150 protein sequences from Uniprot with more than a 70% overall success rate.
The package is free to use under the GNU General Public License v3.0. All necessary files can be downloaded from https://github.com/drou0302/CapiPy. All external requirements are also freely available, with some restrictions for non-academic users.

1 Introduction

The use of enzymes to replace typical catalysts has attracted enormous attention as an alternative to traditional chemical synthesis for the many advantages it presents (Sheldon and Woodley 2018; Tamborini et al. 2018). Nonetheless, stabilization of this macromolecules is often needed to avoid the loss of catalytic activity. To overcome these limitations, immobilization of proteins has emerged as one of the most promising techniques to stabilize them and allow their reuse (Bommarius and Paye 2013; Romero-Fernández and Paradisi 2020). Although the great expectations and its widespread use, this technique is still very unpredictable and thus, the optimization must be done for each protein individually via a trial and error approach. This greatly affects not only the real its applicability but also increases the experimental effort needed. CapiPy, the package presented here, allows scientists with know-how of protein immobilization, to unlock the potential of bioinformatics as a support tool for their experimental design.

2 CapiPy description

CapiPy is a python based tool with graphical interface, based on Biopython (Cock et al. 2009) and using PySimpleGUI, that can be easily installed and run in an Anaconda environment. It is freely available, well documented and designed to be user-friendly, even for inexpert users incorporating executable files (batch or sh, depending on the system). Although it is thought to be utilized sequentially certain functionalities are also provided as a stand-alone to maximize its flexibility. The codebase and installation have been tested in all newest versions of the main operating systems (Windows 10, MACOS Catalina and Ubuntu 20.02 LTS) and the instructions are clearly indicating the necessary packages and software installation to ensure smooth CapiPy operation.

Blast and Modelling

The script uses BLAST (Camacho et al. 2009) to search against the Protein Data Bank database only. The blast result is stored, and the best hit is used (after user confirmation) as a template to create a model using MODELLER (Webb and Sali 2014). In addition, if the template presents more than one chain in the original structure, the created model is cloned and superimposed to match the quaternary structure. The goodness of the final multimeric model is assessed with the RMSD compared to the template protein.

Active site identification

To identify the active site, both UniProt (Bateman 2019) and the M-CSA (Ribeiro et al. 2018) databases are used. Either the retrieved sequence form UniProt or the best hit from the subset of M-CSA with similar EC number is aligned using ClustalW2. To avoid false results, if the identity between both sequences is lower than a 15%, no active site is calculated while if it falls between 15-40%, a warning is embedded in the final result. The results are stored in a text file with all the retrieved information.

Surface and residue clustering

This script is designed to, simply from a structural representation of the protein of interest, calculate the minimum bounding box and volume of the protein (Guardado-Calvo 2018) and the exposure of each residue of the structure. Exposition is calculated using the half-sphere exposure measure (Heffernan et al. 2016; Song et al. 2008) to classify all residues in three categories: buried, semi-exposed or exposed. The exposed subset is then analyzed to find clusters (defined as a group of 3 or more residues with their Cα at less than 10Å) classified in one of 5 categories: positively charged, negatively charged, hydrophobic, lysine clusters or histidine clusters. This clusters, are saved in a PyMOL script file for easy visualization. Additionally, the possible interaction of any of this clusters to either the interchain interface, the active site residues or any user-specified residue can be assessed. If the specified residues are at less than 10Å distance, a warning is printed into the file.

Immobilization literature retrieval

The last script included in the package combines the information obtained from UniProt, by blasting the query sequence against the Swiss-Prot database, to retrieve the keywords to be searched in PubMed. Therefore, the details of the 20 most relevant scientific publications are conveniently stored in an Excel compatible file.

CapiPy performance

CapiPy has been tested with 2 sets of 150 randomly retrieved sequences from UniProt database with known EC number. No bias depending on the size, quaternary assembly or EC code was detected (supplementary dataset), with more than 70% overall success rate. The most common bottlenecks identified were the lack of high homology models either for the modelling, superposition or active site identification.

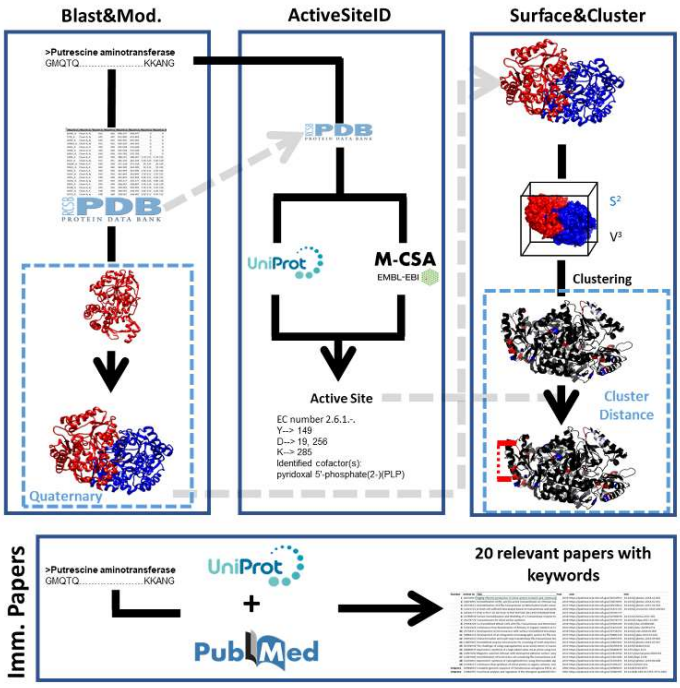

Figure 1: Scheme of CapiPy's workflow. The information and general pipeline of each module is detailed in the scheme above. Transparent grey lines indicate the information can be shared between different scripts.

3 Conclusions

CapiPy, for its ease of installation, handling and result interpretation, constitutes a useful tool for scientists in the biocatalysis field with special focus on protein immobilization. Further work is now under development to expand CapiPy's functionality and compatibility with broadly used software, such as PyMOL or UCSF Chimera as well as better linkage between the experimental data and the generated by the package.

4 Acknowledgements

The authors wish to thank all members of the Paradisi Research group for their insights on the program design and implementation.

5 Funding

This work has been supported by the SNSF (200021_192274).

Conflict of Interest: none declared.

6 References

Bateman, Alex. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47(D1): D506–15.

Bommarius, Andreas S., and Mariétou F. Paye. 2013. "Stabilizing Biocatalysts." *Chemical Society Reviews* 42(15): 6534.

Camacho, Christiam et al. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10(1): 421.

Cock, Peter J.A. et al. 2009. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics." *Bioinformatics* 25(11): 1422–23.

Guardado-Calvo, Pablo. 2018. *Python Script to Calculate and Draw a Minimal Bounding Box for a given Protein. Version 2*.

Heffernan, Rhys et al. 2016. "Highly Accurate Sequence-Based Prediction of Half-Sphere Exposures of Amino Acid Residues in Proteins." *Bioinformatics* 32(6): 843–49.

Ribeiro, António J M et al. 2018. "Mechanism and Catalytic Site Atlas (M-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites." *Nucleic Acids Research* 46(D1): D618–23.

Romero-Fernández, María, and Francesca Paradisi. 2020. "Protein Immobilization Technology for Flow Biocatalysis." *Current Opinion in Chemical Biology* 55: 1–8.

Sheldon, Roger A., and John M. Woodley. 2018. "Role of Biocatalysis in Sustainable Chemistry." *Chemical Reviews* 118(2): 801–38.

Song, Jiangning, Hao Tan, Kazuhiro Takemoto, and Tatsuya Akutsu. 2008. "HSEpred: Predict Half-Sphere Exposure from Protein Sequences." *Bioinformatics* 24(13): 1489–97.

Tamborini, Lucia, Pedro Fernandes, Francesca Paradisi, and Francesco Molinari. 2018. "Flow Bioreactors as Complementary Tools for Biocatalytic Process Intensification." *Trends in Biotechnology* 36(1): 73–88.

Webb, Benjamin, and Andrej Sali. 2014. "Comparative Protein Structure Modeling Using MODELLER." *Current Protocols in Bioinformatics* 47(1): 5.6.1-5.6.32.