Bioinformatics, 2023, **39(10)**, btad613
https://doi.org/10.1093/bioinformatics/btad613
Advance Access Publication Date: 5 October 2023
Applications Note

Structural bioinformatics

SBILib: a handle for protein modeling and engineering

Patrick Gohl[1], Jaume Bonet [1], Oriol Fornes [2], Joan Planas-Iglesias [3,4], Narcís Fernandez-Fuentes [5], Baldo Oliva [1,*]

[1]Department of Medicine and Life Sciences, SBI-GRIB, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain
[2]Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC V5Z 4H4, Canada
[3]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic
[4]International Clinical Research Center, St Anne's University Hospital Brno, 656 916 Brno, Czech Republic
[5]Institute of Biological, Environmental and Rural Science, Aberystwyth University, Aberystwyth SY23 3DA, United Kingdom

*Corresponding author. Department of Medicine and Life Sciences, (SBI-GRIB), Universitat Pompeu Fabra, C/Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain. E-mail: baldo.oliva@upf.edu (B.O.)

Associate Editor: Lenore Cowen

Abstract

Summary: The SBILib Python library provides an integrated platform for the analysis of macromolecular structures and interactions. It combines simple 3D file parsing and workup methods with more advanced analytical tools. SBILib includes modules for macromolecular interactions, loops, super-secondary structures, and biological sequences, as well as wrappers for external tools with which to integrate their results and facilitate the comparative analysis of protein structures and their complexes. The library can handle macromolecular complexes formed by proteins and/or nucleic acid molecules (i.e. DNA and RNA). It is uniquely capable of parsing and calculating protein super-secondary structure and loop geometry. We have compiled a list of example scenarios which SBILib may be applied to and provided access to these within the library.

Availability and implementation: SBILib is made available on Github at https://github.com/structuralbioinformatics/SBILib.

1 Introduction

Macromolecular structure and interaction profiles are essential for the understanding of protein function in health and disease (Yue *et al.* 2005). Binding affinity may be measured at the interface of protein–protein and protein–DNA interaction (Ma *et al.* 2002), and therefore these sites are diagnostic of interaction stability. On the therapeutic front, protein loop similarity offers the potential for grafting, yielding acceptor proteins with beneficial properties (Jones *et al.* 1986, Tang *et al.* 2019). When the loop presented by one protein is replaced by a biologically active loop of another protein ("loop grafting") it is possible to transfer that loop's function, if biological activity of the loop is maintained by ensuring similar loop geometry or flexibility (Smith *et al.* 1995, Schenkmayerova *et al.* 2021). These methods, and many more, rely on standalone bioinformatic packages and databases such as ArchDB (Bonet *et al.* 2014a,b), or on methods that require adapting available bioinformatics packages to purpose written code for analysis [e.g. Biopython (Cock *et al.* 2009), pdb-tools (Rodrigues *et al.* 2018), etc.]. In addition to handling 3D structures, these packages must also be able to handle amino acid sequences and secondary structures (i.e. sequence and local regular conformation), in particular for their alignment and comparison. We have developed a Python package to address these functionalities, as well as introduce super-secondary structure (structures composed of one or more adjacent regular secondary structures) handling functionality. The Structural BioInformatics library (SBILib) is designed to facilitate analysis of macromolecular structures and interactions, as well as to integrate the results from external tools such as BLAST (Camacho *et al.* 2009), DSSP (Kabsch and Sander 1983, Touw *et al.* 2015), or CD-HIT (Fu *et al.* 2012) for protein sequence analyses. Its design strength lies in providing several common analysis tools under one umbrella, resulting in a streamlined approach to structural bioinformatics projects, including the prediction of loop conformations (Fernandez-Fuentes *et al.* 2006), redesign of super-secondary protein structures (Bonet *et al.* 2014a,b), quality assessment of protein folds (Aguirre-Plans *et al.* 2021), or analyses of protein–protein and protein–DNA interactions. For example, SBILib is a core component of ArchDB, Frag'r'Us (Bonet *et al.* 2014a,b), MODPIN (Meseguer *et al.* 2020), InteractoMIX (Mirela-Bota *et al.* 2021), and ModCRE (Fornes *et al.* 2022). We would like to place particular emphasis on the extended functionality of SBILib as it compares to other Python packages such as atomium (Ireland and Martin 2020), ProDy (Zhang *et al.* 2021), Biotite (Kunzmann *et al.* 2023) with regards to protein modeling and engineering. Here, we present this Python library which we have made available on our GitHub repository along with a user manual and tutorial.

Received: 28 March 2023; **Revised:** 22 September 2023; **Editorial Decision:** 2 October 2023; **Accepted:** 4 October 2023

2 SBILib architecture

2.1 Installation

The SBILib library can be installed through pip (pip install SBILib) or downloaded from GitHub. Installation instructions are provided in the README file. SBILib dependencies include: BLAST, CD-HIT, DSSP, NumPy (Harris *et al.* 2020), and SciPy (Virtanen *et al.* 2020). After installation, SBILib can be used both interactively through the Python interpreter or the command line in custom scripts.

2.2 Input parsing

The library is designed to process protein and nucleic acid (i.e. DNA and RNA) structural information both in Protein Data Bank (PDB) and *mmCIF* formats. This information can be retrieved directly from the PDB (Berman *et al.* 2000) via accession codes or the PDBlink module provided within SBILib, or from a local file. The PDBlink module also enables the automatic retrieval and management of PDB files locally. Additionally UniProt IDs may be used to retrieve models from the AlphaFold Protein Structure Database (Varadi *et al.* 2022) using the AlphaFoldlink module.

2.3 Integration of external software

The library can handle the results from external software (BLAST, CD-HIT, DSSP) and integrates them as internal objects or methods. External software must be installed locally, and users can manually specify their system address in the configuration file (SBILib/external/configSBI.txt).

3 SBILib capabilities

3.1 Parsing

SBILib provides PDB, DSSP, and BLAST parsing. The SBILib's BLAST module takes amino acid sequences as input to search for potential homologs. The results are stored as a Python object. Automatic parsing and storage of BLAST results are handled by additional modules. In addition to filtering based on BLAST statistics (e.g. E-value, coverage, % identity, etc.), users may automate the selection of hits above the Rost's curve of significant sequence identity (Rost 1999), such as required for structure homology modeling (Krieger *et al.* 2003).

3.2 Formatted alignments

Protein sequences and BLAST objects can also be used to generate alignments in different formats. For instance, alignments in PIR format for each hit from searching a query sequence in a database of template structures using BLAST can be used as input for MODELLER (Eswar *et al.* 2008) for homology modeling.

3.3 Secondary and super-secondary structures

Secondary structure calculation through the integration of DSSP allows SBILib to be used in postmodeling processing of de novo protein structure prediction and the prediction of loops. Protein loops are flexible, disordered regions of a protein connecting two regular secondary structures. Remodeling of the protein backbone for protein design applications often takes place in these regions due to their flexibility. Within SBILib, loop geometry is automatically calculated and can be used to search for similar super-secondary structures using *smotifs* (Fernandez-Fuentes *et al.* 2010). The ability to parse these super-secondary structures is unique to the library. This functionality may be leveraged for downstream applications such as protein loop grafting (Fig. 1).

3.4 Protein–protein and protein–DNA interfaces

Residue–residue contacts (between amino acids, nucleotides, or mixed) are handled as lists of objects. Each object has references to positions in the chain object of a PDB object. Distances are calculated using SciPy.

4 Examples

The GitHub repository includes several examples describing the use of the library. These include reading protein files in various formats, analyzing basic protein information, handling the results from a BLAST search, viewing the geometry and conformation of smotifs (i.e. features that were used in the classification of loops and their prediction from sequence),

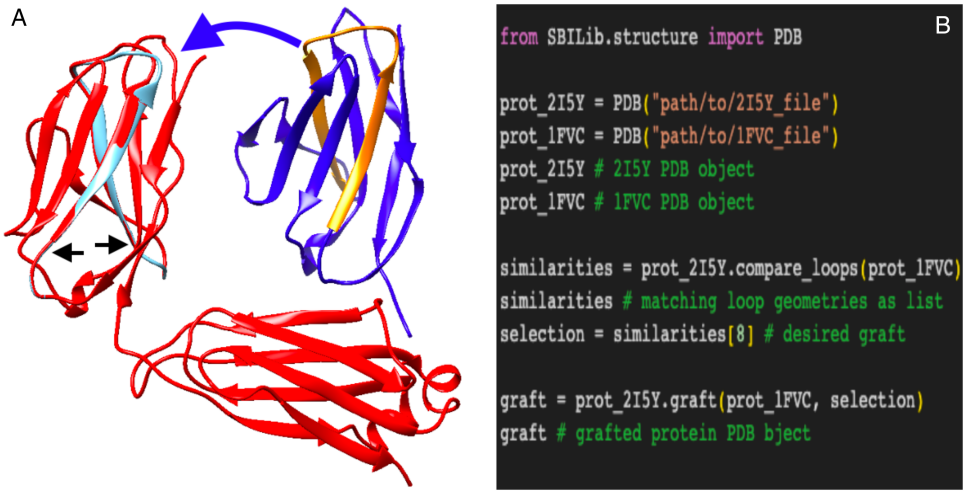

Figure 1. A technical demonstration of SBILib's grafting function. (A) Anti-HIV-1 antibody 17B (pdb:2I5Y) chain L grafted (red) and ungrafted (light blue) superimposed, and ANTI-P185-HER2 ANTIBODY 4D5 (pdb:1FVC) chain A (dark blue). Due to structural identity only the 2I5Y loop and a few flanking residues are shown in the superimposition (light blue). The light chain loop 1FVC:A:19–38 (orange) was grafted in the place of the light chain loop 2I5Y:L:19–38 (beginning and end marked with black arrows) to produce a grafted protein (red). (B) Workflow followed to produce the grafted protein.

as well as investigating protein interactions with other biomolecules (Github: Scenarios.ipynb & README.md).

5 Conclusion

SBILib is an open-source Python library for the analysis of protein folds, as well as macromolecular structures and interactions. It can be both implemented programmatically in large bioinformatic projects or used as a stand-alone tool for routine protein structure analyses. The loop handling features of the library can be used in the field of structure-based computational protein design for loop grafting or remodeling purposes. The SBILib package is integral to various current projects and as such is expected to see updates and additions as its functionality within those projects evolve.

Conflict of interest

None declared.

Funding

This work was supported by grants [PID2020-113203RB-I00] and "Unidad de Excelencia María de Maeztu" [ref: CEX2018-000792-M], funded by the MCIN and the AEI [DOI: 10.13039/501100011033]. J.P.-I. acknowledges support from the Czech Ministry of Education [EXCELES LX22NPO5102].

Data availability

Supplementary information is available at https://github.com/structuralbioinformatics/SBILib. doi: 10.5281/zenodo.8402071.

References

Aguirre-Plans J, Meseguer A, Molina-Fernandez R *et al.* SPServer: split-statistical potentials for the analysis of protein structures and protein–protein interactions. *BMC Bioinformatics* 2021;**22**:4–13.

Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.

Bonet J, Planas-Iglesias J, Garcia-Garcia J *et al.* ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res* 2014a;**42**: D315–9.

Bonet J, Segura J, Planas-Iglesias J *et al.* Frag'r'Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design. *Bioinformatics* 2014b;**30**:1935–6.

Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421–9.

Cock PA, Antao T, Chang JT *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422–3.

Eswar N, Eramian D, Webb B *et al.* Protein structure modeling with MODELLER. Methods Mol Biol. 2008:426:145-59. doi: 10.1007/978-1-60327-058-8_8.

Fernandez-Fuentes N, Dybas JM, Fiser A. Structural characteristics of novel protein folds. *PLoS Comput Biol* 2010;**6**:e1000750.

Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. *Nucleic Acids Res* 2006;**34**: W173–6.

Fornes O, Meseguer A, Aguirre-Plans J *et al.* ModCRE: a structure homology-modeling approach to predict TF binding in cis-regulatory elements. bioRxiv, 2022, https://doi.org/10.1101/2022.04.17.488557.

Fu L, Niu B, Zhu Z *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.

Harris CR, Millman KJ, Van Der Walt SJ *et al.* Array programming with NumPy. *Nature* 2020;**585**:357–62.

Ireland SM, Martin AC. Atomium—a python structure parser. *Bioinformatics* 2020;**36**:2750–4.

Jones PT, Dear PH, Foote J *et al.* Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* 1986;**321**:522–5.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolym Original Res Biomol* 1983;**22**:2577–637.

Krieger E, Nabuurs SB, Vriend G. Homology modeling. *Methods Biochem Anal* 2003;**44**:509–23.

Kunzmann P, Müller TD, Greil M *et al.* Biotite: new tools for a versatile Python bioinformatics library. *BMC Bioinformatics* 2023;**24**:236.

Ma XH, Wang CX, Li CH *et al.* A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Eng* 2002;**15**:677–81.

Meseguer A, Dominguez L, Bota PM *et al.* Using collections of structural models to predict changes of binding affinity caused by mutations in protein–protein interactions. *Protein Sci* 2020;**29**:2112–30.

Mirela-Bota P, Aguirre-Plans J, Meseguer A *et al.* Galaxy InteractoMIX: an integrated computational platform for the study of protein–protein interaction data. *J Mol Biol* 2021;**433**:166656.

Rodrigues JP, Teixeira JM, Trellet M *et al.* pdb-tools: a swiss army knife for molecular structures. *F1000Res* 2018;**7**:1961.

Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.

Schenkmayerova A, Pinto GP, Toul M *et al.* Engineering the protein dynamics of an ancestral luciferase. *Nat Commun* 2021;**12**: 3616.

Smith JW, Tachias K, Madison EL. Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin αIIbβ3 (*). *J Biol Chem* 1995;**270**:30486–90.

Tang H, Shi K, Shi C *et al.* Enhancing subtilisin thermostability through a modified normalized B-factor analysis and loop-grafting strategy. *J Biol Chem* 2019;**294**:18398–407.

Touw WG, Baakman C, Black J *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* 2015;**43**:D364–8.

Varadi M, Anyango S, Deshpande M *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022; **50**:D439–44.

Virtanen P, Gommers R, Oliphant TE *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 2020;**17**:352–272.

Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005;**353**:459–73.

Zhang S, Krieger JM, Zhang Y *et al.* ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics* 2021;**37**:3657–9.