Bioinformatics, 34(22), 2018, 3942–3944
doi: 10.1093/bioinformatics/bty474
Advance Access Publication Date: 21 June 2018
Applications Note

OXFORD

Structural bioinformatics

BioStructMap: a Python tool for integration of protein structure and sequence-based features

Andrew J. Guy[1,2,*], Vashti Irani[1,3], Jack S. Richards[1,3,4,5] and Paul A. Ramsland[1,2,6,7]

[1]Life Sciences, Burnet Institute, Melbourne, VIC 3004, Australia, [2]Department of Immunology, Monash University, Melbourne, VIC 3004, Australia, [3]Department of Medicine, University of Melbourne, Melbourne, VIC 3050, Australia, [4]Department of Infectious Diseases, Monash University, Melbourne, VIC 3004, Australia, [5]Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, VIC 3050, Australia, [6]Department of Surgery, Austin Health, University of Melbourne, Heidelberg, VIC 3084, Australia and [7]School of Science, RMIT University, Bundoora, VIC 3083, Australia

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on December 4, 2017; revised on May 8, 2018; editorial decision on June 8, 2018; accepted on June 18, 2018

Abstract

Summary: A sliding window analysis over a protein or genomic sequence is commonly performed, and we present a Python tool, BioStructMap, that extends this concept to three-dimensional (3D) space, allowing the application of a 3D sliding window analysis over a protein structure. BioStructMap is easily extensible, allowing the user to apply custom functions to spatially aggregated data. BioStructMap also allows mapping of underlying genomic sequences to protein structures, allowing the user to perform genetic-based analysis over spatially linked codons—this has applications when selection pressures arise at the level of protein structure.
Availability and implementation: The Python BioStructMap package is available at https://github.com/andrewguy/biostructmap and released under the MIT License. An online server implementing standard functionality is available at https://biostructmap.burnet.edu.au.
Contact: andrew.guy@burnet.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Consideration of three-dimensional (3D) protein structure is important in many areas of research, including antibody-antigen interactions, protein–protein interactions and drug interactions with proteins. For example, antibody recognition of a dominant epitope can lead to selection pressures on residues associated with that epitope; these residues may be distant in the linear protein sequence despite being spatially connected. In immunology, these non-linear sequence-structure relationships are referred to as discontinuous or conformational epitopes. There are a number of pre-existing online tools that allow for visualization and mapping of pre-defined features onto protein structures (Ashkenazy *et al.*, 2010; Baker and Porollo, 2016; Porollo and Meller, 2007; Segura *et al.*, 2017), however none of these tools allow for application of a 3D sliding window over a protein structure using user-defined functions. There are many settings in which sliding window analysis is applied to genomic or protein sequences, and we demonstrate that this sliding window approach can be extended to 3D protein structures.

2 Materials and methods

We present here a Python package named BioStructMap that allows mapping of sequence-associated data onto a protein structure. This tool also allows for the application of a 3D 'sliding window' over a protein structure. The user can apply a variety of functions to spatially aggregated data, mapping the result back to the central residue within each window. The user must provide sequence-aligned data, a reference sequence and PDB format coordinates over which to

process data. For each residue in the structure, all residues within a user-defined radius are selected. Data corresponding to these residues (i.e. specific characteristics of interest for these residues) is then passed to a function that returns a numerical value, which is then mapped back to the central residue (Supplementary Fig. S1). A number of pre-defined functions are included in the BioStructMap package. Users can also supply their own function for data processing. Data is output as a Python dictionary of residues and associated values, written to a PDB file in the B-factor column, or as a text file. Results can be viewed using PyMOL or similar programs.

BioStructMap uses the Biopython Bio.PDB module for handling PDB files, and can accept both PDB and mmCIF files as input. Sequence alignments are performed using either the NCBI BLAST+ package or the Biopython Bio.pairwise2 module. Alignment of DNA sequences to protein sequences is performed using Exonerate (Slater and Birney, 2005) which allows handling of intron-containing sequences and reverse-sense translation. Calculation of Tajima's D is performed using the Python DendroPy package (Sukumaran and Holder, 2010).

The source code for BioStructMap is available on GitHub (https://github.com/andrewguy/biostructmap) or via the Python Package Index. A simple web-server interface is also available at https://biostructmap.burnet.edu.au, using the JavaScript NGL viewer for visualization of protein structures (Rose and Hildebrand, 2015). Results can be viewed in the browser or downloaded as PDB files. Further details on BioStructMap use are available in Supplementary Material.

3 Usage example

In areas of endemic malaria, immune selection pressure on the malaria parasite can lead to balancing selection, in which low-frequency alleles are maintained at a higher proportion than would otherwise be expected under a neutral model of selection. Tajima's D (Tajima, 1989) is one statistic that has been used to identify regions under balancing selection within the malaria genome, and has previously been applied as a sliding window over genes of interest (Arnott *et al.*, 2013, 2014). We have also previously applied the BioStructMap tool to key vaccine candidates from *P. falciparum* and *P. vivax*, incorporating protein structural information into calculations of selection pressures and diversity (Guy *et al.*, 2018a, b). We illustrate here one of the potential uses for the BioStructMap tool, applying a 3D sliding window calculation of Tajima's D over the protein structure of *Plasmodium falciparum* EBA-175 Region II (RII), a leading malaria vaccine candidate (Fig. 1). This approach groups data that are spatially connected but are distant in the linear sequence. Nucleotide sequences for EBA-175 RII were extracted from GenBank, originally deposited from a study examining signatures of selection in *P. falciparum* strains from Kenya and Thailand (Verra *et al.*, 2006). Since known structures contain a number of unresolved residues, ModPipe (Eswar, 2003) was used to generate a comparative structural model for EBA-175 RII. A radius of 15 Å was selected for each window as this is the typical maximum-dimension for an antibody-antigen interface (Ramaraj *et al.*, 2012). When analyzed, a surface exposed loop with a high spatially derived Tajima's D value is identified in both Kenyan and Thai isolates. Importantly, this region is involved in the dimerization of EBA-175 RII around its glycophorin A binding partner on the surface of the human red blood cell (Tolia *et al.*, 2005), and antibodies that target the dimerization interface of EBA-175 RII have previously been shown to be highly effective at inhibiting parasite entry into red blood cells (Chen *et al.*, 2013). A region within the F1 domain is also identified as having high Tajima's D values within Thai samples, but to a much lesser extent in Kenyan samples. Further experimental work would be required to validate this region as a target of functional antibody responses.

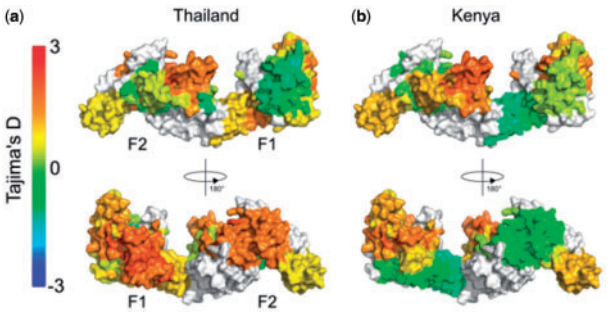

Fig. 1. Tajima's D calculation applied as a 3D sliding window over the protein structure of *P. falciparum* EBA-175 RII. The F1 and F2 domains are indicated on the monomeric structure. Nucleotide sequences were obtained from *P. falciparum* isolates from (**a**) Thailand ($n = 48$) and (**b**) Kenya ($n = 39$) (Verra *et al.*, 2006). The BioStructMap Python package was used to apply Tajima's D calculations using a 3D sliding window with a radius of 15 Å. The structural model is available via ModBase, accession number: ed998157a605f5e58e-d66e198e0ae1ab. Structures were visualized with PyMOL

4 Concluding remarks

The BioStructMap package and associated web interface allow for visualization of sequence-aligned data over a 3D protein structure, as well as allowing the incorporation of protein structural information into sequence-based metrics using a 3D sliding window approach. This tool is applicable to a variety of problems, including identification of regions under various forms of genetic, immunological or drug selection pressure and spatial mapping of residue characteristics that may affect immunogenicity, solubility, binding interaction, etc. The tool is easily extensible, allowing users to define their own functions to apply to spatially aggregated data.

Acknowledgements

The authors thank Dyson Simmons and Andrew Walter for support with development and deployment of the BioStructMap web-based server.

Funding

This work was supported by the National Health and Medical Research Council (NHMRC) of Australia [APP1037722 & APP1125788 to J.S.R.], and an Australian Post-graduate Award to A.J.G. Burnet Institute received funding from the NHMRC Independent Research Institutes Infrastructure Support Scheme, and the Victorian State Government Operational Infrastructure Support Scheme.

Conflict of Interest: none declared.

References

Arnott,A. *et al.* (2014) Distinct patterns of diversity, population structure and evolution in the AMA1 genes of sympatric *Plasmodium falciparum* and *Plasmodium vivax* populations of Papua New Guinea from an area of similarly high transmission. *Malar. J.*, **13**, 233.

Arnott,A. *et al*. (2013) Global population structure of the genes encoding the malaria vaccine candidate, *Plasmodium vivax* apical membrane antigen 1 (Pv AMA1). *PLoS Negl. Trop. Dis.*, **7**, e2506.

Ashkenazy,H. *et al*. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.

Baker,F.N. and Porollo,A. (2016) CoeViz: a web-based tool for coevolution analysis of protein residues. *BMC Bioinformatics*, **17**, 119.

Chen,E. *et al*. (2013) Structural and functional basis for inhibition of erythrocyte invasion by antibodies that target *Plasmodium falciparum* EBA-175. *PLoS Pathog.*, **9**, e1003390.

Eswar,N. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.*, **31**, 3375–3380.

Guy,A.J. *et al*. (2018a) Proteome-wide mapping of immune features onto *Plasmodium* protein three-dimensional structures. *Sci. Rep.*, **8**, 4355.

Guy,A.J. *et al*. (2018b) Structural patterns of selection and diversity for *Plasmodium vivax* antigens DBP and AMA1. *Malar. J.*, **17**, 183.

Porollo,A. and Meller,J. (2007) Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D. *BMC Bioinformatics*, **8**, 316.

Ramaraj,T. *et al*. (2012) Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta*, **1824**, 520–532.

Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.

Segura,J. *et al*. (2017) 3DBIONOTES v2.0: a web server for the automatic annotation of macromolecular structures. *Bioinformatics*, **33**, 3655–3657.

Slater,G.S.C. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Sukumaran,J. and Holder,M.T. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tolia,N.H. *et al*. (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. *Cell*, **122**, 183–193.

Verra,F. *et al*. (2006) Contrasting signatures of selection on *the Plasmodium falciparum* erythrocyte binding antigen gene family. *Mol. Biochem. Parasitol.*, **149**, 182–190.