Bioinformatics, 34(13), 2018, 2302–2304
doi: 10.1093/bioinformatics/bty086
Advance Access Publication Date: 14 February 2018
Applications Note

Structural bioinformatics

StructureMapper: a high-throughput algorithm for analyzing protein sequence locations in structural data

Anssi Nurminen[1] and Vesa P. Hytönen[1,2,*]

[1]Faculty of Medicine and Life Sciences and BioMediTech, University of Tampere, 33520 Tampere, Finland and [2]Fimlab Laboratories, 33520 Tampere, Finland

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 27, 2017; revised on January 31, 2018; editorial decision on February 12, 2018; accepted on February 13, 2018

Abstract

Motivation: StructureMapper is a high-throughput algorithm for automated mapping of protein primary amino sequence locations to existing three-dimensional protein structures. The algorithm is intended for facilitating easy and efficient utilization of structural information in protein characterization and proteomics. StructureMapper provides an analysis of the identified structural locations that includes surface accessibility, flexibility, protein–protein interfacing, intrinsic disorder prediction, secondary structure assignment, biological assembly information and sequence identity percentages, among other metrics.

Results: We have showcased the use of the algorithm by estimating the coverage of structural information of the human proteome, identifying critical interface residues in DNA polymerase γ, profiling structurally protease cleavage sites and post-translational modification sites, and by identifying putative, novel phosphoswitches.

Availability and implementation: The StructureMapper algorithm is available as an online service and standalone implementation at http://structuremapper.uta.fi.

Contact: vesa.hytonen@uta.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A large portion of bioinformatics tools that are being used in protein research function by processing only primary amino acid sequences. While this approach can be adequate in particular cases, many of these tools could benefit from additional information about proteins under study. Proteins and enzymes gain their functionality through their three-dimensional structure and this information often remains underutilized due to a lack of appropriate tools. The Protein Data Bank (PDB; Berman *et al.*, 2000) contains ~131 000 protein structures and it is currently growing at a rate of ~11 000 new structures deposited every year (http://www.rcsb.org). By random sampling, we have estimated the current coverage of structural data of the human proteome (including partial structures, and homologous structures from other species) to lie between 20% and 25%. These structural data can be used as an important source of information to support the predictions of sequence-based bioinformatics tools and to help the evaluation of experimental data arising from high-throughput experiments. However, efficient utilization of structural data requires large amount of manual work, and very often, the use of several different algorithms.

To solve this problem, we have developed an open-source algorithm, StructureMapper and demonstrated its efficacy in extracting relevant structural information from large datasets that cannot be inferred from the primary amino acid sequence alone. Tools such as PROMALS3D (Pei *et al.*, 2008) and Expresso (Armougom *et al.*, 2006) are available for aligning sequences to structures. However, none of the available tools provide the capability to locate and visualize specific amino acids within the alignments, provide information of the structural properties of the amino acids of interest, or can

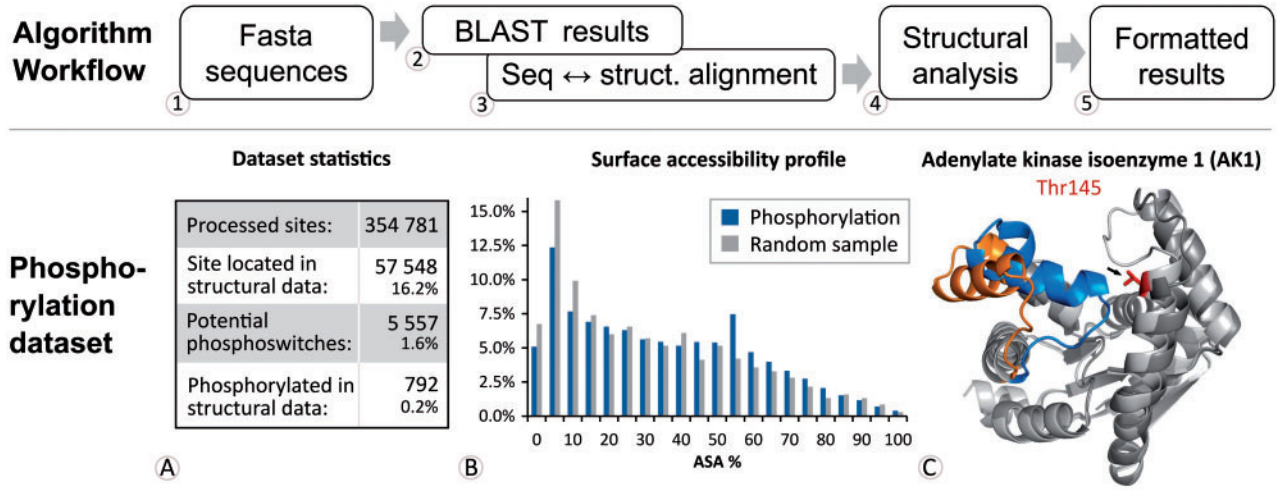

Fig. 1. The StructureMapper algorithm and its application to the PhosphoSitePlus phosphorylation dataset. Top: StructureMapper processes the inputted sequences in five distinct steps. Bottom: (**A**) Putative phosphoswitches were identified by StructureMapper in the PhosphoSitePlus dataset and the PDB database. (**B**) A surface accessibility profile of the 354 781 experimentally verified phosphorylation sites (blue/dark). In a comparison between a randomized sample of 5000 threonine, serine and histidine residues from the human proteome (grey/light), the profiles look very similar, having a linear downward trend in ASA, with the exception of a ~2.5% increase of highly solvent exposed sites at ASA between 50% and 55%. (**C**) Example case of a putative phosphoswitch identified by the StructureMapper algorithm. PDB structures 1Z83 (blue/dark) and 3ADK (orange/light) show two alternative conformations of AK1 (~12 Å apart) in the immediate spatial proximity of phosphorylated residue Thr145 (arrow/red). The ASA difference of Thr145 is 41.1% between the structures. Grey coloring is used for parts of the structures that superimpose nearly identically (Color version of this figure is available at *Bioinformatics* online.)

process sequences in a high-throughput manner that is able to utilize parallel processing for maximum efficiency.

The StructureMapper is designed to do all this in a turn-key manner that does not require expert bioinformatics skills or custom scripting. Therefore, no directly comparable algorithms are currently available, and StructureMapper may open numerous new ways into exploitation of 3D structural data in biological research. A more detailed comparison of StructureMapper to existing tools is provided in Supplementary Table S1.

2 Results and discussion

To demonstrate the capabilities of StructureMapper, we have analyzed two unique datasets. The first dataset consists of post-translational modifications (PTMs) (Supplementary Figs. S1 and S2), and in particular, reversible phosphorylation sites (PhosphoSitePlus; Hornbeck *et al.*, 2015) that were analyzed to identify putative, novel phosphoswitches (Fig. 1 and Supplementary Material). The second dataset (Supplementary Material and Figs. S3 and S4) consists of experimentally verified protease cleavage sites and a comparison of their structural properties against a random sampling of similar sites in the annotated human proteome (Uniprot; Apweiler *et al.*, 2004). Furthermore, StructureMapper was used to identify potentially deleterious mutations in the structure of DNA polymerase gamma (Supplementary Material and Fig. S6).

The processing of the input sequences and marked points of interest (POIs) in the sequences advances in five distinct steps (Fig. 1). In step one, the inputted sequences and the marked points of interest (POIs) in the amino acids sequences are prepared for BLAST searches (Altschul *et al.*, 1990) to identify structures with sufficient similarity (adjustable criteria) in step two.

In step three, when multiple system cores are available, StructureMapper begins the process of simultaneously locating the POIs within the homologous structures identified by pairwise sequence alignments that are created for the queried sequence and the sequence that is extracted from the structure file. StructureMapper utilizes the Biopython library (Cock *et al.*, 2009) for creating the alignments. If the assigned alignment score of each sequence-structure pair falls below a specified threshold (low reliability), the structure is not used in further processing.

After the positions of the amino acids of interest have been located in the homologous structures, they are examined for their features (step four). For the calculation of accessible surface area (ASA), a custom algorithm is used (available as a standalone Python algorithm). The ASA-algorithm employs a 'rolling-ball' method (Shrake and Rupley, 1973) and the algorithm is able to effectively parallelize the processing with each additional processor. The ASA values are reported as percentages of the surface area of the residue that is available for contact with solvent atoms. In the analysis, established methods for assigning secondary structure information (DSSP; Kabsch and Sander, 1983) and intrinsically disordered region predictions (DISOPRED3; Jones and Cozzetto, 2015) are used. StructureMapper can construct any biological assemblies specified for the POI, and is able to determine if the POI is located at a protein-protein interface (homomer/heteromer). In step five, StructureMapper compiles the results into a tab-delimited result file that can be easily processed further or viewed in a spreadsheet application.

For a randomized sample of 5000 unspecified amino acids of the annotated human proteome (Apweiler *et al.*, 2004) StructureMapper could locate the amino acids of interest in a homologous structure in 1146 cases (data not shown), making the overall estimated coverage of structural data of human proteins 22.9% in the PDB.

The StructureMapper algorithm is well suited for use in the quality assessment and refinement of experimental data. Possible use cases can include evaluation of prediction algorithms and generating profiles for data obtained using experimental methods such as phosphoproteomics. StructureMapper has been designed so that it works on any modern operating system and it can be integrated to work as a part of an algorithm pipeline. StructureMapper has been released with a MIT software license and the source code made available on the StructureMapper online server.

Funding

Funding from Finnish Funding Agency for Innovation (Tekes; project THERDIAB diary #1843/31/2014) and the Academy of Finland (grant no. 290506) are gratefully acknowledged.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Apweiler,R. *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **32**, 115D–1119.

Armougom,F. *et al*. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res*., **34**, W604–W608.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Cock,P.J.A. *et al*. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Hornbeck,P.V. *et al*. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res*., **43**, D512–D520.

Jones,D.T. and Cozzetto,D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.

Kabsch,W. and Sander,C. (1983) Dictionary o f protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Pei,J. *et al*. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res*., **36**, 2295–2300.

Shrake,A. and Rupley,J. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol*., **15**, 351–371.