Bioinformatics, 32(3), 2016, 474–476
doi: 10.1093/bioinformatics/btv574
Advance Access Publication Date: 6 October 2015
Applications Note

Databases and ontologies

The new protein topology graph library web server

Tim Schäfer[1], Andreas Scheck[1], Daniel Bruneß[1], Patrick May[2] and Ina Koch[1,*]

[1]Molecular Bioinformatics, Institute of Computer Science, Johann Wolfgang Goethe-University Frankfurt am Main, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main, Germany and [2]Genome Analysis, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on July 7, 2015; revised on September 10, 2015; accepted on September 27, 2015

Abstract

Summary: We present a new, extended version of the Protein Topology Graph Library web server. The Protein Topology Graph Library describes the protein topology on the super-secondary structure level. It allows to compute and visualize protein ligand graphs and search for protein structural motifs. The new server features additional information on ligand binding to secondary structure elements, increased usability and an application programming interface (API) to retrieve data, allowing for an automated analysis of protein topology.

Availability and implementation: The Protein Topology Graph Library server is freely available on the web at http://ptgl.uni-frankfurt.de. The website is implemented in PHP, JavaScript, PostgreSQL and Apache. It is supported by all major browsers. The VPLG software that was used to compute the protein ligand graphs and all other data in the database is available under the GNU public license 2.0 from http://vplg.sourceforge.net.

Contact: tim.schaefer@bioinformatik.uni-frankfurt.de; ina.koch@bioinformatik.uni-frankfurt.de

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

1 Introduction

The huge amount of three-dimensional (3D) protein structures available in databases like the PDB (Berman *et al.*, 2000) call for tools for automated analysis and intuitive visualization of protein structures. Databases such as SCOP (Andreeva *et al.*, 2008) and CATH (Sillitoe *et al.*, 2012) classify global protein topology based on semi-automated methods and expert knowledge. Some automated methods such as FSSP (Holm and Sander, 1997) rely on global 3D alignments and pairwise comparison. 3D Complex (Levy *et al.*, 2006) focuses on protein complexes. ProSMoS (Shi *et al.*, 2007), TOPS (Michalopoulos *et al.*, 2004), TOPS+ (Veeramalai and Gilbert, 2008), Pro-Origami (Stivala *et al.*, 2011) and the Protein Topology Graph Library (PTGL, May *et al.*, 2010) all work on the super-secondary structure level and apply different methods to find substructures. Ligand information is of great interest to determine protein function and in drug discovery, but most databases ignore information on ligand binding to secondary structure elements (SSEs).

The PTGL (Koch *et al.*, 2013; May *et al.*, 2004) is a database that uses a graph model to describe proteins on the chain level. The graphs are based on 3D atom data from the PDB and the SSE assignments of the DSSP algorithm (Kabsch and Sander, 1983). A *protein ligand graph* (PLG) is a labeled, undirected graph that models one protein chain. Each vertex represents an SSE of that chain or a ligand assigned to the chain in the PDB file, and the edges model contacts between SSEs including their relative spatial orientation. A hard sphere model is used to define atom contacts. A rule set determines whether two SSEs are in a contact, based on the number of atom contacts between them. For details on how the contacts and orientations are computed, see Schäfer *et al.* (2012). Connected components of PLGs are called *folding graphs* (FGs). PTGL defines

four different linear notation strings for FGs. Users can search the database for protein structural motifs, which is based on fast text matching in the linear notations. The PTGL provides intuitive two-dimensional visualizations of the FGs for all linear notations.

We rewrote the PTGL from scratch, implemented ligand support and other new features and updated all data to the April 2015 version of the PDB, which contains information on more than 230 000 protein chains. The data will be updated on a regular basis by our automated update procedure.

2 Features

The PTGL web server allows to search for PLGs, FGs and linear notation strings of FGs based on PDB identifiers, keywords, ligands or structural motifs. Several common motifs like the beta propeller or jelly roll are pre-implemented. Querying by custom linear notation strings enables the search for arbitrary arrangements of SSEs. Table 1 lists the counts for several motifs and their co-occurrence in protein chains with different ligands. In addition to the alpha, beta and alpha-beta graphs of the previous PTGL release, the current version supports ligands and introduces three additional graph types: alpha ligand graphs contain alpha helices and ligands, beta ligand graphs contain beta strands and ligands and alpha-beta ligand graphs contain alpha helices, beta strands and ligands.

Table 1. Motifs in the database and their co-occurrence with ligands

Motif	Motif type	Motif count		Ligand types	
		Abs.	Percentage	Abs.	Percentage
Globin fold	All-alpha	9799	4.17	954	5.68
Four helix bundle	All-alpha	15 531	6.61	1247	7.42
Beta propeller	All-beta	1512	0.64	190	1.13
Immunoglobin fold	All-beta	8797	3.74	769	4.58
Up and down barrel	All-beta	1192	0.50	193	1.14
Jelly roll	All-beta	14 266	6.07	944	5.62

The database contains a total of 234 735 protein chains and 16 790 ligand types.

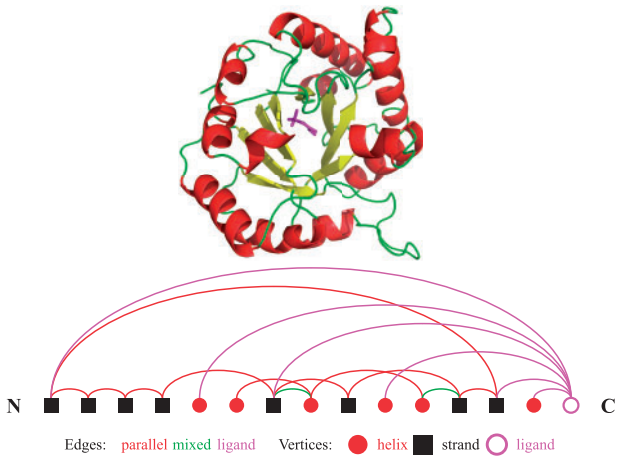

Fig. 1. Visualization of a structure and its FG. The reduced notation of the largest FG of 7TIM chain A is depicted below a 3D view of the structure. Each vertex in the graph represents an SSE or a ligand of the protein chain. The edges model spatial contacts. Note the parallel beta barrel of size 8 and the ligand in the center of the barrel, which exhibits contacts to several SSEs (Color version of this figure is available at *Bioinformatics* online.)

These graphs enable users to search for substructures that bind ligands. Note that the PTGL is a chain-level database, and currently only intra-chain contacts are computed. The PLGs can be exported in standard graph file formats like Graph Modeling Language (GML, Himsolt, 1997), e.g. to be analyzed by software like Gephi (Bastian *et al.*, 2009), GraphViz (Ellson *et al.*, 2004) or Cytoscape (Cline *et al.*, 2007). The PTGL server visualizes PLGs and FGs as two-dimensional schematic diagrams, and the resulting images can be exported in vector and bitmap formats. Figure 1 depicts an example visualization of the structure of triosephosphate isomerase with phosphoglycolohydroxamic acid bound (Davenport *et al.*, 1991).

The PTGL provides links to the PDB and CATH. We also offer an interactive 3D visualization of the graphs by JSmol (Hanson *et al.*, 2013). Clients can use the web frontend to browse the PTGL data using any recent browser. For large-scale analyses of protein topologies, the REST API can be queried. It serves PG and FG data in JavaScript Object Notation and Extensible Markup Language formats and supports retrieval of graph visualizations. Online help and the API documentation are available on the web site, and example calls are listed in the Supplementary Information of this article.

3 Implementation

All protein graph data available on the PTGL 3 server were computed from PDB and DSSP data by an extended version of the VPLG software (Schäfer *et al.*, 2012). This was done offline during the server installation or update procedure that ran in parallel for many PDB files on a computer cluster. The update procedure involved downloading PDB data, generating DSSP data, running VPLG and copying the result files and database from the computer cluster to the PTGL server. It was controlled by Bash and Python scripts. Supplementary Figure S1 in the Supplementary Material depicts an overview of the update procedure and the server architecture. Data on the PLGs, FGs and linear notations were stored in a PostgreSQL database, including images and GML files. An entity-relationship diagram of the database schema and a visualization of the PTGL server architecture is provided in Supplementary Figure S2.

4 Summary

The PTGL uses graphs to model protein structures and ligands on the super-secondary level. The graphs were computed from 3D atom data from the RCSB PDB and the secondary structure assignments of the DSSP algorithm. The PTGL works on the chain level. In the future, we will extend it to also support multi-chain proteins and protein complexes.

The new version of the PTGL was rewritten from scratch and was designed to be easy to maintain and extend. It includes the latest PDB data, ligand information and a RESTful application programming interface.

The PTGL supports fast searching for protein chains by substructures, ligands and other properties. The different protein graph types allow for the assessment of the overall structure of a chain including its ligands. The four different visualizations of FGs can be used to analyze and compare protein substructures in greater detail. The new PTGL is a powerful and updated tool which enables investigations of protein topology on a large scale. It is useful in applications like protein classification, drug design and studies of evolutionary relationships between proteins.

Acknowledgements

We would like to thank Norbert Dichter and Jens Einloft for technical help and discussions, and Jörg Ackermann for reading the manuscript.

Funding

This work was partly supported by a grant from Friedrich Naumann Foundation for Freedom.

Conflict of Interest: none declared.

References

Andreeva,A. *et al.* (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**(suppl 1), D419–D425.

Bastian,M. *et al.* (2009) Gephi: an open source software for exploring and manipulating networks. In: *International AAAI Conference on Weblogs and Social Media*. Vol. 8, pp. 361–362.

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

Cline,M.S. *et al.* (2007) Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.*, **2**, 2366–2382.

Davenport,R.C. *et al.* (1991) Structure of the triosephosphate isomerase-phosphoglycolohydroxamate complex: an analog of the intermediate on the reaction pathway. *Biochemistry*, **30**, 5821–5826.

Ellson,J. *et al.* (2004) Graphviz and dynagraph—static and dynamic graph drawing tools. In: Junger,M. and Mutzel,P. (eds.) *Graph Drawing Software*. Springer-Verlag, pp. 127–148.

Hanson,R.M. *et al.* (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to Proteopedia. *Israel J. Chem.*, **53**, 207–216.

Himsolt,M. (1997) GML—graph modelling language. *Technical report*. University of Passau.

Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

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

Koch,I. *et al.* (2013) Hierarchical representation of super-secondary structures using a graph-theoretical approach. In: Kister,A.E. (ed.) *Protein Supersecondary Structure*. Springer, New York, pp. 7–33.

Levy,E.D. *et al.* (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, e155.

May,P. *et al.* (2004) PTGL—a web-based database application for protein topologies. *Bioinformatics*, **20**, 3277–3279.

May,P. *et al.* (2010) PTGL: a database for secondary structure-based protein topologies. *Nucleic Acid Res.*, **38**, D326–D330.

Michalopoulos,I. *et al.* (2004) TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res.*, **32**(suppl 1), D251–D254.

Schäfer,T. *et al.* (2012) Computation and visualization of protein topology graphs including ligand information. In: Böcker,S. *et al.* (eds),*German Conference on Bioinformatics 2012*, volume 26 of *OpenAccess Series in Informatics (OASIcs)*. Dagstuhl, Leibniz-Zentrum für Informatik, Germany, pp. 108–118.

Shi,S. *et al.* (2007) Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. *Bioinformatics*, **23**, 1331–1338.

Sillitoe,I. *et al.* (2012) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, **41**, D490–D498.

Stivala,A. *et al.* (2011) Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics*, **27**, 3315–3316.

Veeramalai,M. and Gilbert,D. (2008) A novel method for comparing topological models of protein structures enhanced with ligand information. *Bioinformatics*, **24**, 2698–2705.