D326–D330 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
doi:10.1093/nar/gkp980

PTGL: a database for secondary structure-based protein topologies

Patrick May[1,*], Annika Kreuchwig[2], Thomas Steinke[3] and Ina Koch[4,5,*]

[1]Max Planck Institute for Molecular Plant Physiology, Bioinformatics, Am Muehlenberg 1, 14476 Potsdam-Golm, [2]Leibniz-Institut fuer Molekulare Pharmakologie, Structural Bioinformatics, Robert-Roessle-Strasse 10, 13125 Berlin, [3]Zuse Institute Berlin, Computer Science Research, Takustrasse 7, 14195 Berlin, [4]Beuth University for Technology Berlin, FB VI, Bioinformatics, Seestrasse 64, 13347 Berlin and [5]Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Ihnestrasse 73, 14195 Berlin, Germany

Received August 15, 2009; Revised October 14, 2009; Accepted October 15, 2009

ABSTRACT

With growing amount of experimental data, the number of known protein structures also increases continuously. Classification of protein structures helps to understand relationships between protein structure and function. The main classification methods based on secondary structures are SCOP, CATH and TOPS, which all classify under different aspects, and therefore can lead to different results. We developed a mathematically unique representation of protein structure topologies at a higher abstraction level providing new aspects of classification and enabling for a fast search through the data. Protein Topology Graph Library (PTGL; http://ptgl.zib.de) aims at providing a database on protein secondary structure topologies, including search facilities, the visualization as intuitive topology diagrams as well as in the 3D structure, and additional information. Secondary structure-based protein topologies are represented uniquely as undirected labeled graphs in four different ways allowing for exploration under different aspects. The linear notations, and the 2D and 3D diagrams of each notation facilitate a deeper understanding of protein topologies. Several search functions for topologies and sub-topologies, BLAST search possibility, and links to SCOP, CATH and PDBsum support individual and large-scale investigation of protein structures. Currently, PTGL comprises topologies of 54 859 protein structures. Main structural patterns for common structural motifs like TIM-barrel or Jelly Roll are pre-implemented, and can easily be searched.

INTRODUCTION

The number of 3D protein structures determined by methods of nuclear magnetic resonance and X-ray crystallography is increasing rapidly. Currently, over 54 000 structures are deposited in the Protein Data Base (PDB) (1). Protein structure classification helps to understand the relationship between protein structure and function. Results can be used for protein structure prediction and for the exploration of evolutionary studies. Protein topologies are also important for the study of protein folding processes, in particular of folding pathways (2,3).

Many proteins share similar geometric features in the conformation of the protein backbone (4). With a few exceptions, the 3D structure of proteins can be characterized by patterns of secondary structure elements (SSEs) (5). Structural protein topology can be defined as the relationship between the sequential ordering of SSEs and their spatial organization. Protein topology is one of the principal properties by which protein structures can be classified, categorized and compared. The main protein structure classifications based on SSEs are provided by SCOP (6), CATH (7) and TOPS (8,9).

Super-secondary structure motifs, such as Greek-key or α–β–α motifs, describe the interaction and position of SSEs. The arrangements of SSE motifs open the possibility for a topological description of protein structures. The first theoretical work on protein topologies refer to β structure topologies (10,11) or α topologies (12). Later, Koch and co-workers (13,14) defined a protein graph incorporating helices as well as strands.

The simplest representations of protein topologies are schematic diagrams of protein folds illustrating the SSEs and their spatial neighborhoods. Richardson (4) derived the first protein topology diagrams as a cartoon representation of a biological point of view. Also several β-motifs

*To whom correspondence should be addressed. Fax: +49 (0)331 5678136; Email: may@mpimp-golm.mpg.de
Correspondence may also be addressed to Ina Koch. Tel: +49 (0)30/8413-1168; Fax: +49 (0)30/8413-1152; Email: koch_i@molgen.mpg.de

of SSEs were described, such as the Greek-key or Jelly Roll motif (15). This cartoon representation is widely used, so, for example, in the protein classification databases, CATH and SCOP. Another, but more complicated, representation is the hydrogen bond diagram that provides a graphical representation based on the hydrogen bonds between spatially neighbored SSEs (11). The protein structure databases CATH, SCOP and PDBsum (16,17) do not provide a direct comparison of protein topologies. CATH and SCOP use knowledge on sequence similarity and protein function within their classification workflow, whereas Protein Topology Graph Library (PTGL) only relies on the underlying graph representation and does not explicitly consider the order of SSEs. The PDB sum database provides for every CATH domain a single topological diagram generated using the PROMOTIF tool (11), but again no global topology comparison is possible.

The database TOPS (8,9) generates topology diagrams automatically and provides the possibility to search for secondary structure motifs. TOPS uses a graph-based description of protein topologies, which is similar to ours. They also provide a linear notation, which is restricted to only one description type. Due to the underlying graph definition, which explicitly preserves the sequential order of SSEs, similar topologies with different sequential order of SSEs cannot be found. PTGL does not explicitly consider the order of SSEs in the underlying graph definitions.

We developed a web-based database application, PTGL (18), for representation and retrieval of protein topologies and additional protein data combined with online search tools for data interrogation by sequence similarity and keyword queries. We provide unique linear notations of four descriptions for protein structures on different abstraction levels based on graph theory. The main idea is to mine known protein structures as protein graphs. This enables us to represent complex information such as protein structures by relatively simple, but unique schematic descriptions.

GRAPH NOTATIONS

Protein topologies, i.e. the relationship between the ordering and connectedness of SSEs, can generally be expressed in terms of graph theory. A 'protein graph' is defined as a labeled undirected graph for a single PDB chain, where the vertices correspond to SSEs (helices, h and strands, e), and the edges represent spatial adjacencies of SSEs. The SSEs are defined according to the assignment of the DSSP algorithm (19). The edges are defined through contacts between SSEs using van der Waals contacts (20). Vertices of the protein graph are enumerated from the N- to the C-terminus. According to this direction, two spatially neighbored SSEs can occur in a parallel (p), antiparallel (a), or mixed (m) neighborhood. A protein graph consists of one or more connected components, which we call 'folding graphs', because they often represent single folding units or domains. According to the considered SSE type we distinguish between α–β, α and β graphs. Folding graphs, containing vertices with a degree greater than two, are called 'bifurcated'.

To explore topological aspects of SSEs in proteins, we provide four types of diagrams for every type of notation, KEY, ADJ, RED and SEQ. KEY diagrams are similar to those defined by Richardson (14). SSEs are ordered according to their spatial arrangement. Helices and strands are drawn as red cylinders and black arrows, respectively. The arcs describe sequential neighborhoods between SSEs (Figure 1). In the ADJ, RED and SEQ diagrams, SSEs are arranged according to their sequential order from the N- to C-Terminus. Helices are drawn as red circles and strands as black squares. The arcs between SSEs indicate spatial neighborhoods. Edges in the ADJ and RED diagrams are colored according to their labels, red for 'parallel', green for 'mixed' and blue for 'antiparallel' neighborhoods. In the ADJ diagrams, also SSEs of other folding graphs are considered, whereas in RED diagrams only SSEs of the folding graph under consideration are involved. Edges in the SEQ diagrams are drawn in black and indicate sequential neighborhoods. In dependence of the diagram type, four linear notations can be derived, which can be used for searching and classification. An example for a protein graph and the corresponding folding graphs in KEY notation is given in Figure 1.

CONTENT AND USAGE

PTGL is stored locally in an object-relational PostgreSQL database running on a Linux server. Programs for export, import and visualization have been implemented in Perl and C. PTGL input data are acquired from the PDB. The DSSP program is running locally on the file server. Additional data as header, title, molecule and information about ligands, cofactors and hetero atoms are derived from the associated PDB file.

The online search tool has five query browsers: (i) the simple keyword query searches for keywords, which can be connected by Boolean operators; (ii) the more customizable query searches in selected fields and tables, e.g. for sub-topologies in all β graphs or all graphs representing a certain CATH classification; (iii) a third batch query form accepts lists of proteins; and (iv) a sequence search based on local BLAST (21) searches against the sequences stored in the PTGL main table. The BLAST outputs are stored for 3 days, and the user has access via a request ID (5). The recent search type is the motif browser that allows the user to search for pre-defined common structural motifs, such as TIM-barrel and Jelly Roll motifs (Figure 2).

A selection of subsets of homologous protein structures has been incorporated. The clustering of the polypeptide chain sequences is done according to pre-calculated PDB cluster tables based on the CD-HIT algorithm (22).

The 'Query result browser' shows all protein entries found. Then, the user can select the proteins of interest together with the graph type (α–β, α or β) and notation type. Chains fulfilling the search criteria are marked

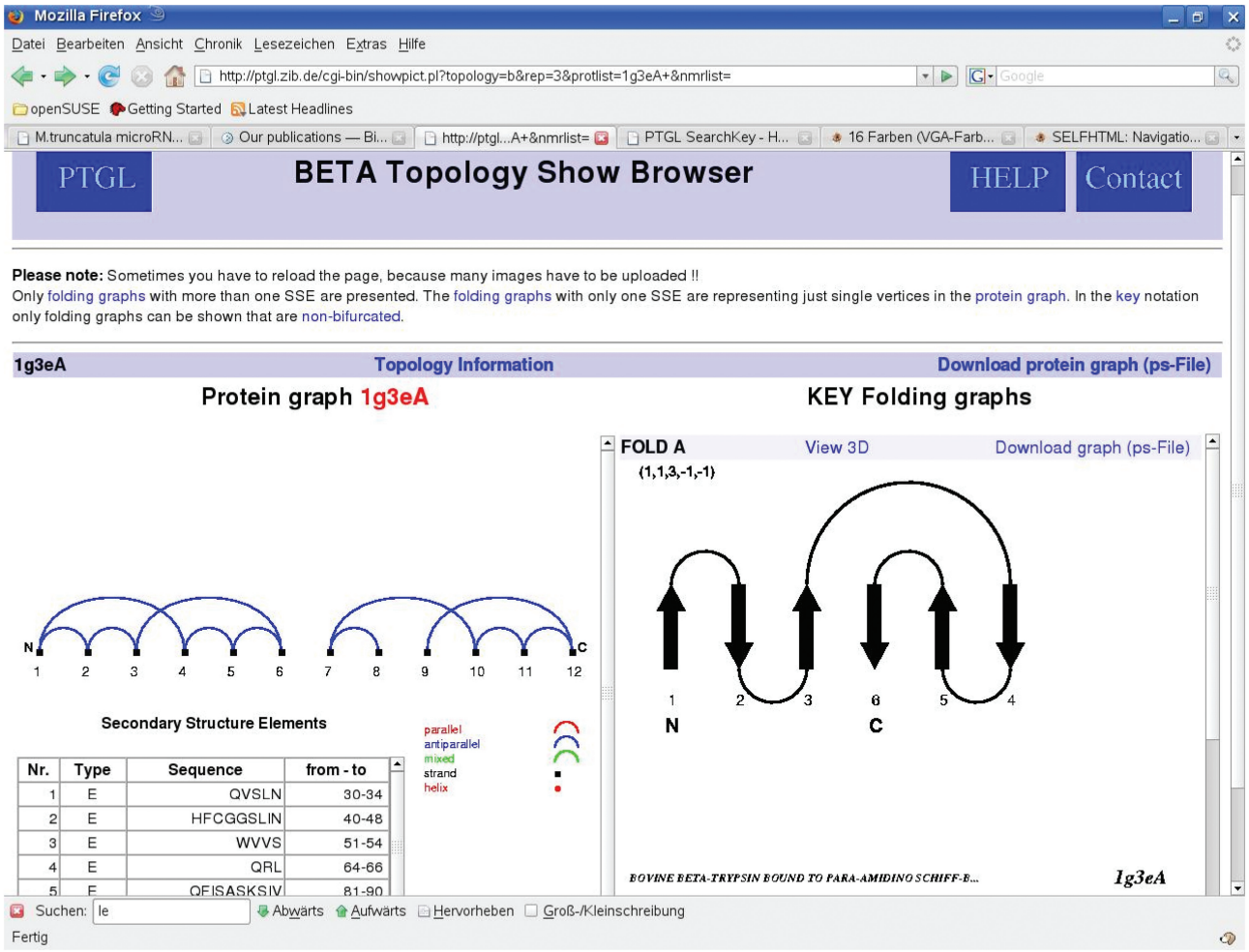

Figure 1. The 'Topology Browser' for protein 1G3E chain A. On the left side, the β protein graph is shown together with additional information for the SSEs. On the right side, the β folding graph A in KEY notation is shown. The linear notation is '(1,1,3,−1,−1)' representing a barrel structure.

in grey. Links are provided to CATH, SCOP and PDBsum. In the 'Topology browser', for every selected PDB chain the topology diagrams of the protein graph and all folding graphs, consisting of more than one SSE, are represented according to the selected notation together with a table compiling all SSEs (Figure 1).

All diagrams of protein and folding graphs are generated automatically on the fly as PostScript files that are converted to the Portable Network Graphics format for representation on the web. The graph descriptions and SSE information are available as ASCII files.

The topologies are also represented as 3D images which can be animated, using Jmol (http://jmol.sourceforge.net). Additionally, users can upload their custom PDB files to generate the different linear notations and diagrams. All available graphs are compiled and searchable via the 'Content' page, where general statistics is provided for every graph and notation type, including the numbers of current entries with their links to the 'Query Browser'.

The 'Help' and 'User Guide' pages in PTGL (http://ptgl.zib.de/ptglhelp.html) give an exhaustive introduction with examples how to use and how to link to PTGL. They can be used as tutorial.

PTGL currently holds topology information of 54 859 proteins and 2 094 546 SSEs. Only proteins with at least one defined SSE according to DSSP a protein structure resolution <3.5 Å, and a sequence length of at least 20 amino acids have been considered. Table 1 shows the total number of folding graphs for every graph type. The number of unique α–β graphs is less than the sum of unique α and unique β graphs, because in α–β graphs, per definition, additional contacts between helices and strands are considered.

CONCLUSIONS

PTGL is an online database tool for retrieval and search of specific protein topologies using four linear notations, which are based on a unique graph-theoretical description of protein topology. Furthermore, PTGL provides searching facilities using sequence similarity in proteins. Protein topologies are represented in four different types of schematic diagrams and as 3D images. The database is useful for any kind of theoretical protein structure analysis, protein structure prediction and protein function prediction. Our database, PTGL, represents a

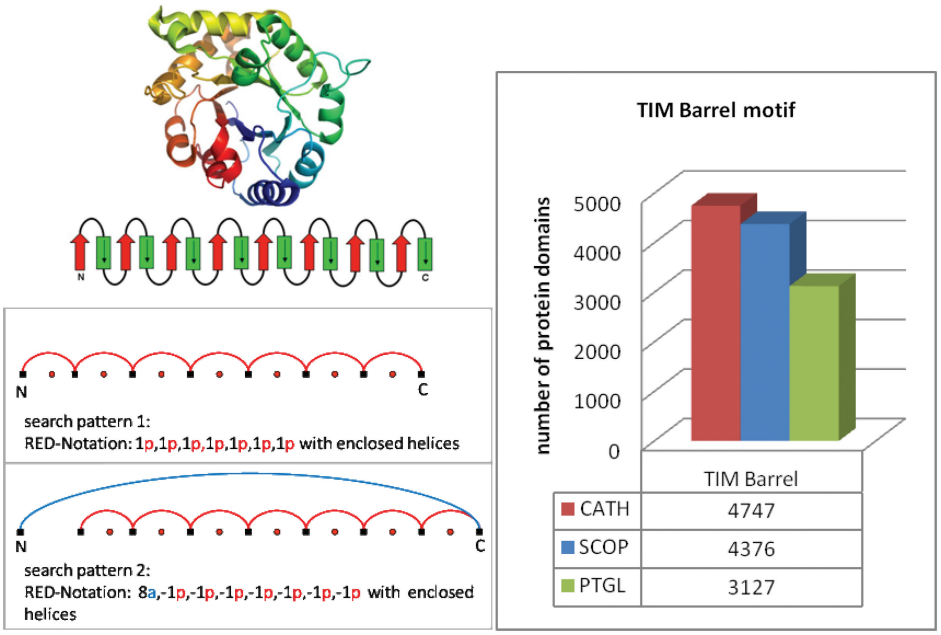

Figure 2. The TIM barrel motif as 3D image (upper left), the corresponding KEY notation (middle left) and two search patterns (bottom left), both described in RED as β graphs. For visualization purposes of the TIM barrel motif, the helices are shown in the β graph image. On the right site, the different occurrences of the motif in CATH, SCOP and PTGL are depicted. The differences between the PTGL and CATH or SCOP mainly rely on applying a stronger, but more precisely definition for the motif, which limits the risk for false positives.

Table 1. PTGL content

Graph type	Total	Different	Non-bifurcated	Barrels
α–β	606 816	77 127	2432	666
α	636 820	29 984	680	191
β	199 122	13 418	2246	696

The first column indicates the type of the folding graph. 'Total' gives the total amount of folding graphs of the particular graph type; 'different' the amount of different graphs among the total number; 'non-bifurcated' the number of different non-bifurcated graphs including barrels; and 'barrels' the number of different barrel structures.

useful extension to the existing protein structure topology databases.

PTGL was originally introduced in 2004 (18); the graph representations and corresponding topology diagrams were proposed in Koch and co-workers (13,14). Now, we added several new search functionalities as the pre-defined consensus notations for the most common structural motifs, the possibility to upload and to analyze custom protein structures, and additional functional annotation, such as Enzyme Commission numbers and links to the PDBsum database. Furthermore, we updated the database to the recent PDB, SCOP and CATH releases, and moved to a new web location (http://ptgl.zib.de).

FUNDING

German Federal Ministry of Education and Research by the FORSYS BMBF grant (GoFORSYS Grant No. 0313924 to P.M.). Funding for open access charge: Max Planck Society.

REFERENCES

1. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
2. Przytycka,T., Srinivasan,R. and Rose,G.D. (2002) Recursive domains in proteins. *Protein Sci.*, **11**, 409–417.
3. Zaki,M.J., Nadimpally,V., Bardhan,D. and Bystroff,C. (2004) Predicting protein folding pathways. *Bioinformatics*, **20**, I386–I393.
4. Richardson,J.S. (1977) Beta-sheet topology and the relatedness of proteins. *Nature*, **268**, 495–500.
5. Chothia,C. and Finkelstein,A.V. (1990) The classification and origins of protein folding patterns. *Annu. Rev. Biochem.*, **59**, 1007–1039.
6. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
7. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
8. Michalopoulos,I., Torrance,G.M., Gilbert,D.R. and Westhead,D.R. (2004) TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res.*, **32**, D251–D254.
9. Veeramalai,M.D. and Gilbert,D.R. (2008) A novel method for comparing topological models of protein structures enhanced with ligand information. *Bioinformatics*, **24**, 2698–2705.
10. Artymiuk,P.J., Grindley,H.M., Poirette,A.R., Rice,D.W., Ujah,E.C. and Willett,P. (1994) Identification of beta sheet motifs, of psi loops, and of patterns of amino acid residues in three dimensional protein structures using a subgraph isomorphism algorithm. *J. Chem. Inf. Comput. Sci.*, **229**, 707–721.

11. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF—a program to identify structural motifs in proteins. *Protein Science*, **5**, 212–220.
12. Grigoriev,I.V., Mironov,A.A. and Rakhmaninova,A.B. (1994) Inter-helical contacts determining the architecture of alpha-helical globular proteins. *J. Biomol. Struct. Dyn.*, **12**, 559–572.
13. Koch,I., Kaden,F. and Selbig,J. (1992) Analysis of protein sheet topologies by graph-theoretical methods. *Prot. Struct. Funct. Genet.*, **12**, 314–324.
14. Koch,I., Lengauer,T. and Wanke,E. (1996) An algorithm for finding maximal common subtopologies in a set of proteins. *J. Comp. Biol.*, **3**, 289–306.
15. Brandon,C. and Tooze,J. (1999) *Introduction to Protein Structure*. Garland, London/New York.
16. Laskowski,R.A., Chistyakov,V.V. and Thornton,J.M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.
17. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.
18. May,P., Barthel,S. and Koch,I. (2004) PTGL-Protein Topology Graph Library. *Bioinformatics*, **20**, 3277–3279.
19. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
20. Kaden,F., Koch,I. and Selbig,J. (1990) Knowledge-based prediction of protein structures. *J. Theor. Biol.*, **147**, 85–100.
21. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
22. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.