BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 23 2011, pages 3315–3316
doi:10.1093/bioinformatics/btr575

Structural bioinformatics

Advance Access publication October 12, 2011

Automatic generation of protein structure cartoons with Pro-origami

Alex Stivala[1,*], Michael Wybrow[2], Anthony Wirth[1], James C. Whisstock[3] and Peter J. Stuckey[1,*]

[1]Department of Computer Science and Software Engineering, The University of Melbourne Parkville Campus, Victoria 3010, [2]Clayton School of Information Technology and [3]Department of Biochemistry and Molecular Biology, ARC Centre of Excellence in Structural and Functional Microbial Genomics, Monash University Clayton Campus, Clayton, Victoria 3800, Australia

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Protein topology diagrams are 2D representations of protein structure that are particularly useful in understanding and analysing complex protein folds. Generating such diagrams presents a major problem in graph drawing, with automatic approaches often resulting in errors or uninterpretable results. Here we apply a breakthrough in diagram layout to protein topology cartoons, providing clear, accurate, interactive and editable diagrams, which are also an interface to a structural search method.

Availability: Pro-origami is available via a web server at http://munk.csse.unimelb.edu.au/pro-origami

Contact: a.stivala@pgrad.unimelb.edu.au; pjs@csse.unimelb.edu.au

Received and revised on August 30, 2011; accepted on October 8, 2011

Comparisons between structurally related proteins are aided by simplified topology maps of protein structure [e.g. TOPS cartoons (Westhead *et al.*, 1999)]. Such diagrams are particularly useful when comparing very distantly related folds [e.g. Rosado *et al.* (2007)]. The automatic generation of such cartoons is a challenging problem, requiring as it does the simplification of a complex 3D object into a 2D representation, with the (often competing) requirements of correctness, maximizing the amount of information conveyed, ease of comprehension and aesthetic appeal. To the best of our knowledge, there are three existing systems for the automatic generation of such diagrams from atomic co-ordinates in PDB files. HERA (Hutchinson and Thornton, 1990) generates detailed hydrogen-bonding diagrams. TOPS generates topological cartoons with a 'top-down' view in a style similar to the hand-drawn diagrams of Sternberg and Thornton (1977). A recent addition to the PDBsum website (Laskowski, 2009) generates topology diagrams for protein domains in an 'exploded' style derived from HERA. These styles of diagram are not always sufficient for the purposes of structural biologists and in particular do not accurately convey the relative position of β-sheets and α-helices. One solution is to draw suitable diagrams manually, and the TopDraw program (Bond, 2003) is a simple diagram editor designed to aid the manual drawing of protein structure cartoons in a perpendicular view, looking 'side-on' to secondary structure elements (SSEs).

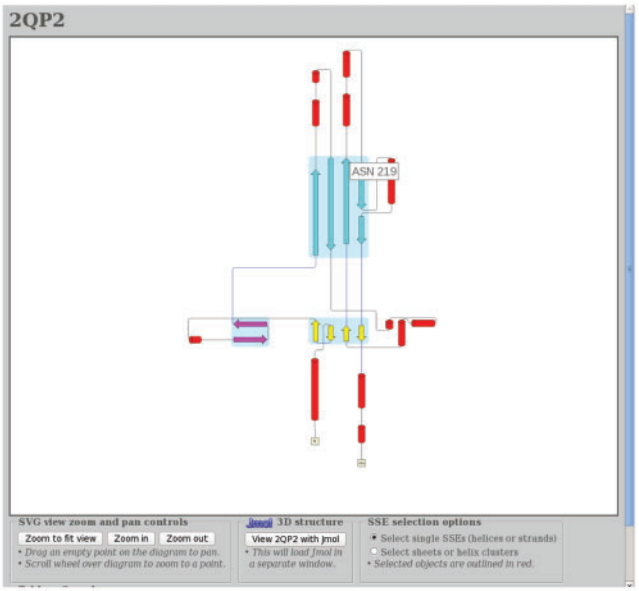

Fig. 1. Pro-origami cartoon of the MACPF domain of Plu-MACPF (Rosado *et al.*, 2007), PDB identifier 2QP2. The orientation of the C-sheet (yellow strands) relative to the β-sheet (turquoise strands) has been chosen so that the strands immediately following each other in sequence are oriented in the same direction, despite this not necessarily reflecting their real orientation in 3D space. Helices are aligned with strands, except for clusters of helices, or where doing so would compromise readability of the cartoon: one helix between two strands of the β-sheet is drawn beside the sheet, and the large N-terminal helix is not placed between the B and C sheets.

Pro-origami is an approach for automatically generating protein structure cartoons in a systematic fashion, similar to those that might be manually produced via human cognition and depicted using TopDraw. β-strands and α-helices are drawn as arrows and cylinders, respectively, proportional in length to the number of residues they contain. β-sheets are shown as groups of β-strands positioned relative to each other according to their actual relative positions (Fig. 1). Most importantly, Pro-origami provides a solution to the most complex part of the problem; achieving a layout that

*To whom correspondence should be addressed.

remains comprehensible while maintaining as much of the 3D structural information as possible.

The first stage of the Pro-origami process is, optionally, to decompose the protein into domains, using either a CATH (Pearl *et al.*, 2005) CDF file or the DDOMAIN program (Zhou *et al.*, 2007). Secondary structure is determined using either the secondary structure assignments in the PDB file, or the DSSP (Kabsch and Sander, 1983) or STRIDE (Frishman and Argos, 1995) programs, from which hydrogen bond information is also obtained. Constraints derived include groupings of strands within sheets and the relative positions of strands in a sheet, separation between adjacent strands in a sheet, and the alignment of strands and helices on horizontal or vertical axes. β-strands are grouped into sheets using algorithms similar to those employed by TOPS, making use of hydrogen bond and 3D geometrical information. β-barrels are displayed 'flattened' into sheets by breaking the bridge relationship between two of the strands. These strands are marked in the cartoon to indicate this situation. Sheets are positioned relative to each other according to the contact map and tableau (Kamat and Lesk, 2007) computed for the protein, which defines the relative orientation of secondary structure elements. Helices are positioned using a heuristic which attempts to make the diagram as easy as possible to read, usually by placing them aligned on the axis of the closest strand.

The initial positions of the cartoon elements and the set of constraints and connectors are used as the input to the Dunnart constraint-based diagram editor (Dwyer *et al.*, 2009). Dunnart then lays out the diagram according to the constraints, with specialized coding to ensure that elements do not overlap. Connectors between the elements are routed so as to not collide with elements and to try to reduce unnecessary crossings (Wybrow *et al.*, 2010).

The Pro-origami web server allows diagrams to be generated from any PDB file, and a copy of the PDB (Berman *et al.*, 2000) is stored on the server. Alternatively, a PDB file may be uploaded to the server. The resulting diagram may be either downloaded or viewed as an SVG or bitmap file. For browsers that support JavaScript and SVG, the diagram is interactive, showing the residue at the part of the diagram over which the mouse pointer is positioned.

In addition to providing protein structure cartoons, Pro-origami acts as a user interface to a protein substructure searching algorithm (Stivala *et al.*, 2009). The user can select a set of SSEs on the cartoon to act as a motif query, and have the matching SSEs highlighted in the Pro-origami cartoons of the matched structures.

One of the aims of Pro-origami is to make cartoons easy to interpret visually. One of the criteria for easy interpretation is that the sequence of helices and strands should be easy to follow by eye along the connectors, and Pro-origami attempts to minimize the number of cases in which connectors run too close to each other, a situation we describe as a connector 'overlap', which Pro-origami detects and attempts to correct automatically. We ran Pro-origami on a database of non-redundant protein domains, resulting in a total of 16 602 cartoons. Pro-origami produces a cartoon with at least one 'overlap' for only 2.29% of these domains. On our server (Dell PowerEdge R200), the average time taken to generate a cartoon in this set is 2.3 s.

Pro-origami automates the process of generating protein topology diagrams, showing the relative size of strands and helices, the alignment of strands in a sheet and the general positioning of sheets and helices with respect to one another. Such information is not completely conveyed by the topological style of TOPS cartoons or HERA (PDBsum) diagrams. Our approach also reduces unnecessary connector crossings and generates diagrams that are interactively editable with an editor (Dunnart). The latter approach permits preservation of meaningful constraints such as the positions of strands in a sheet, and the connectors between SSEs, while the user edits the diagram. Pro-origami also allows interactive graphical construction of queries to a protein substructural search system.

ACKNOWLEDGEMENT

Discussions with Dr Tim Dwyer and Prof. Kim Marriott have assisted our work. Dr Arun Konagurthu provided us with source code to generate protein tableaux.

Funding: We thank the Australian Research Council and NICTA for support. ARC Federation Fellow and Honorary National Health and Medical Research Council Principal Research Fellow (to J.C.W.). Australian Postgraduate Award (to A.S.).

Conflict of Interest: none declared.

REFERENCES

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

Bond,C.S. (2003) TopDraw: a sketchpad for protein structure topology cartoons. *Bioinformatics*, **19**, 311–312.

Dwyer,T. *et al.* (2009) Dunnart: A constraint-based network diagram authoring tool. In Tollis,I.G. and Patrignani,M. (eds) *Graph Drawing 2008*, Vol. 5417 of *Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, pp. 420–431.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Hutchinson,E.G. and Thornton,J.M. (1990) HERA — a program to draw schematic diagrams of protein secondary structures. *Proteins*, **8**, 203–212.

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

Kamat,A.P. and Lesk,A.M. (2007) Contact patterns between helices and strands of sheet define protein folding patterns. *Proteins*, **66**, 869–876.

Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D335–D359.

Pearl,F. *et al.* (2005) The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.

Rosado,C.J. *et al.* (2007) A common fold mediates vertebrate defense and bacterial attack. *Science*, **317**, 1548–1551.

Sternberg,M.J.E. and Thornton,J.M. (1977) On the conformation of proteins: The handedness of the connection between parallel β-strands. *J. Mol. Biol.*, **110**, 269–283.

Stivala,A. *et al.* (2009) Tableau-based protein substructure search using quadratic programming. *BMC Bioinformatics*, **10**, 153.

Westhead,D.R. *et al.* (1999) Protein structural topology: Automated analysis and diagrammatic representation. *Protein Sci.*, **8**, 897–904.

Wybrow,M. *et al.* (2010) Orthogonal connector routing. In Eppstein,D. and Gansner,E.R. (eds) *Graph Drawing 2009*, Vol. 5849 of *Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, pp. 219–231.

Zhou,H. *et al.* (2007) DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. *Protein Sci.*, **16**, 947–955.