Knot localization in proteins

Eric J. Rawdon*[1], Kenneth C. Millett†, Joanna I. Sułkowska‡§ and Andrzej Stasiak‖

*Department of Mathematics, University of St. Thomas, 2115 Summit Avenue, St. Paul, MN 55105, U.S.A., †Department of Mathematics, University of California Santa Barbara, 552 University Road, Santa Barbara, CA 93106, U.S.A., ‡Center for Theoretical Biological Physics, University of California San Diego, 9500 Gilman Drive, San Diego, CA 92037, U.S.A., §Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland and ‖Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne-Dorigny, Switzerland

Abstract

The backbones of proteins form linear chains. In the case of some proteins, these chains can be characterized as forming linear open knots. The knot type of the full chain reveals only limited information about the entanglement of the chain since, for example, subchains of an unknotted protein can form knots and subchains of a knotted protein can form different types of knots than the entire protein. To understand fully the entanglement within the backbone of a given protein, a complete analysis of the knotting within all of the subchains of that protein is necessary. In the present article, we review efforts to characterize the full knotting complexity within individual proteins and present a matrix that conveys information about various aspects of protein knotting. For a given protein, this matrix identifies the precise localization of knotted regions and shows the knot types formed by all subchains. The pattern in the matrix can be considered as a knotting fingerprint of that protein. We observe that knotting fingerprints of distantly related knotted proteins are strongly conserved during evolution and discuss how some characteristic motifs in the knotting fingerprints are related to the structure of the knotted regions and their possible biological role.

Introduction

From the first discovery of knots in the backbones of proteins in their native state, researchers have been interested in finding the exact location of these knots with the hope of understanding how knots form in proteins and determining the biological function of the knots. In the present review, we discuss recent efforts to classify the full knotting profile in protein backbones that, for the purpose of their topological characterization, are represented by linear polygons with vertices corresponding to the sequential alpha carbons. We do not discuss the topological analysis of closed circuits formed in part by covalent bonds along protein backbones, but also involving inter- and intra-chain covalent bonds of cysteine bridges or covalently bound metal atoms [1–3].

In the first systematic study aimed at detecting knotting in polypeptide chains, Mansfield [4,5] analysed the approximately 400 protein structures known at the time and found only two proteins that were knotted. However, these two knots were very shallow (Mansfield called these "loose knots"), i.e. removing just a few of the terminal amino acids from the analysed protein would result in unknotted chains. The fact that only two of the ~400 protein structures analysed could be considered as knots (and these were very shallow) led biologists to believe that it was not possible that protein folding pathways could result reproducibly in the formation of deep knots. However, this changed with the discovery of deeply knotted proteins [6] in which up to 70 amino acids could be removed from the nearest end of the protein and the remaining portion of the protein structure still could be considered as knotted. The existence of deeply knotted proteins demonstrated that protein folding can result in the formation of robust knots and opened the possibility that knotted regions can have a special biological role.

Yet there were more mysterious knotting behaviours to be discovered. Taylor [6] describes the location and depth of the knot core of the full protein when the entire chain is knotted. However, this paper did not report on any knotting behaviour in subchains outside of that core. King et al. [7] were the first to analyse the full knotting spectrum of subchains within some proteins. They discovered slipknotted proteins (i.e. proteins whose entire chain is unknotted, but which contains some subchains which are knotted) as well as proteins with multiple knotted cores of the same or various knot types. It became clear that the knotting of the entire protein, or even just the knotting in the core of the entire chain, was not sufficient to describe the entirety of the knotting of the protein.

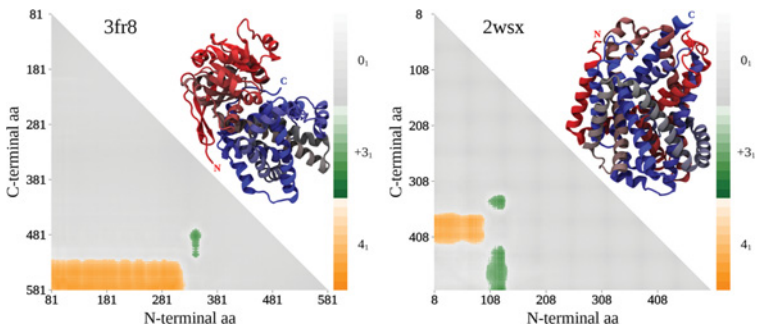
A matrix presentation of the entirety of knotting within proteins was introduced in [7]. This presentation encodes the knotting of the protein and all of its subchains in a matrix image, and provides a visual means to inspect the knots formed by all subchains. In particular, one can see regions forming different types of knots occurring in different patterns. The knotted cores and their relative arrangements provide insights into the native fold of the protein, and could suggest why the knotting exists in these proteins. In the remainder of the present article, we describe the matrix presentation for knotting in proteins, discuss challenges in describing the knotted cores and show how common patterns of knotting within protein families suggest a deep connection between knotting and function.

Key words: knotting fingerprint, protein backbone, protein knot, slipknotted protein.

[1]To whom correspondence should be addressed (email ejrawdon@stthomas.edu).

Figure 1 | Matrix presentations for the proteins with PDB codes 3FR8 (left) and 2WSX (right)

Each square cell in the matrix shows the knot type of one subchain of the protein. The N-terminal amino acid position of that subchain is indicated on the *x*-axis, and its C-terminal amino acid position is indicated on the *y*-axis. Thus the lower-left-hand corner shows the knot type of the entire chain and cells near the diagonal correspond to very short subchains of the protein. The intensity of the colour within each cell corresponds to the percentage of closures forming the given dominant knot type for the subchain. The colour bar on the right shows the knot types obtained for the protein as well as a gradient for the colouring intensity by steps of 10%. For chiral knot types, the + and − signs indicate the right- and left-handed forms respectively.



Characterizing the knotting in proteins

Before we discuss the knotting patterns, we must be clear about how we classify knotting in an open chain. Indeed, defining the knot type of an open chain is an interesting problem in itself and different algorithms are discussed in another article in this issue of *Biochemical Society Transactions* [8]. In that article, we present the uniform closure method [9–11], whereby the knotting of an open chain is classified as a distribution of knot types obtained by connecting the free ends of the open chain to points uniformly chosen on a large sphere enveloping the chain. The dominant knot type is then labelled as the knot type of the chain. For the remainder of the present article, we use that strategy to classify the types of knots in open chains. Once we have agreed on how to define the knotting of an open chain, we can compute the knot type of the entire protein chain and of all of its subchains. King et al. [7] defined a matrix image presentation for encoding the knotting of all subchains of a given protein. In [11], we built on this presentation using the uniform closure method [8–10].

Figure 1 explains how to identify the knotting within the subchains of a protein from its matrix presentation. Knotted proteins, such as ketol-acid reductoisomerase from rice (PDB code 3FR8) have the lower-left-hand corner of the matrix presentation (corresponding to the knot type of the entire protein chain) coloured, whereas slipknotted proteins, such as the carnitine transporter from *Escherichia coli* (PDB code 2WSX), contain coloured regions elsewhere in the matrix, but have a grey lower-left-hand corner (signifying an unknotted arc). In addition to knots and slipknots, we sometimes observe isolated regions in the matrix presentation containing the same knot type, as in the two trefoil regions for 2WSX.

Knotting fingerprints

For a given protein, we use the term knotting fingerprint to denote the entirety of the knotting information present in the matrix presentation, including types of knots present and the regions' sizes and shapes. For knotted or slipknotted proteins, one sees at least one roughly rectangular region of knotting. Each such region corresponds to a nested set of subchains with a particular knot type. The shortest subchain within the region (the subchain corresponding to the cell closest to the diagonal) of the particular knot type defines the knotted or slipknotted core. The roughly rectangular knotting regions are not always fully filled in, since the boundaries are transition areas where one can have two or more knot types with similar probabilities.

Because of the nearly constant distance between sequential α carbons and the steric exclusion of polypeptide chains, proteins behave essentially as smooth thick tubes [12]. For this reason, the knotting fingerprints of proteins are rather tame, i.e. changing a subchain length by one amino acid can result in a change of knot type that could be produced by at most one intersegmental passage [13]. Such a behaviour would not be expected for random chains, for example, where shortening a subchain by one segment can result in a change of knot type that would require more than one intersegmental passage. By observing some knotting fingerprints, we can see why the search for the knotted core sometimes gives a different result when it is carried out using a top-down approach in contrast with a bottom-up approach [14]. In the top-down approach, one removes terminal vertices until the knot type detected for the entire protein is no longer present, whereas in the bottom-up approach, one searches for the shortest sequence of amino acids forming the same knot as the entire protein. The knotting fingerprint of the DehI protein (PDB code 3BJX) (Figure 2A) shows that the top-down approach would give a 6_1 core size that is approximately 40 amino acids larger than the 6_1 core size detected using the bottom-up approach.

Also note that the precise pattern in the knotting fingerprint depends on the algorithm used to determine the knot type of an open chain. When the termini of an open

Figure 2 | **Matrix presentations for the proteins with PDB codes 3BJX (left) and 2AXC (right)**

Left: the determination of the knotted core for the DehI protein (PDB code 3BJX) varies by approximately 40 amino acids depending on whether one uses a top-down or bottom-up approach since there are two distinct regions forming the 6_1 knot. Right: the N-terminal translocation domain for colicin E7 in *E. coli* (PDB code 2AXC) shows four different trefoil regions.

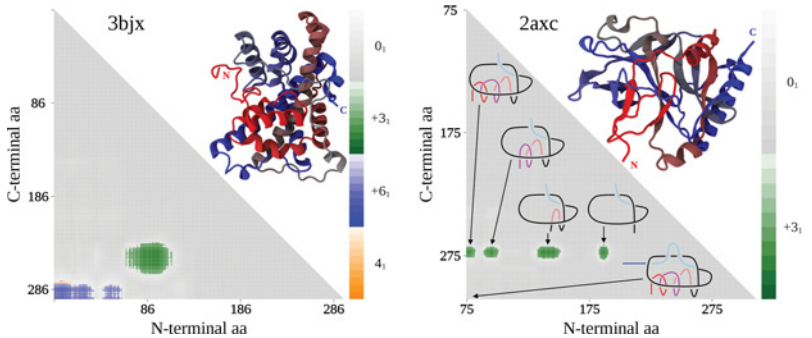
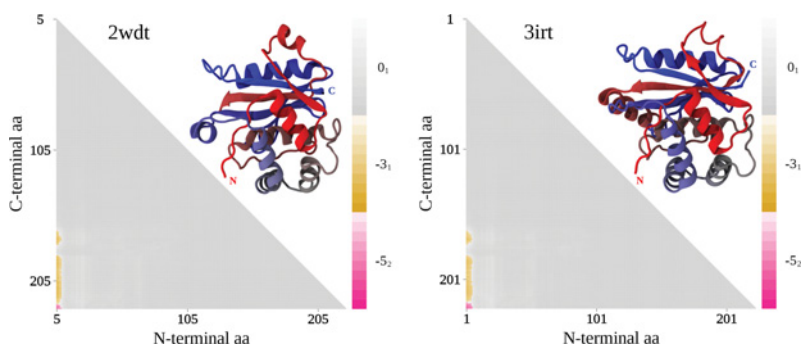


Figure 3 | **The ubiquitin C-terminal hydrolases from *Plasmodium* (PDB code 2WDT) (left) and humans (PDB code 3IRT) (right) have almost identical knotting fingerprints despite only 32% sequence similarity**



chain are on the 'outside' of a chain (which is typically the case for the entire protein chain), the different algorithms generally agree in assigning a knot type to the chain. However, when the termini are 'near the centre' of a chain (which happens extensively when analysing subchains of proteins), the algorithms can disagree in assigning a knot type to the chain. The single closure algorithms (such as chain simplification) [8] often require some 'choices' to be made in order to assign an appropriate closure, and thus knot type, for the chain. The stochastic algorithms, such as the uniform closure procedure used here, require no 'choices' in these ambiguous situations since they uniformly sample from potential knotted states and thus remain unbiased. For example, in [11], we found that the LeuT(Aa) protein (PDB code 2A65), contains subchains forming 3_1 and 4_1 knots. King et al. [7] found 3_1, 4_1 and 5_2 knots for 2A65. The 'choices' one, inevitably, is forced to make when using single closure algorithms is a serious deficiency in the approach, and thus we believe that the stochastic algorithms provide a more solid characterization of the knotting within subchains. However, note that, despite the differences in knot detection algorithms, generally there is only a small fraction of subchains for which the different algorithms disagree.

We then can define a notation for the knotting regions present in the protein. We begin with a K or S, representing that the protein is either knotted or slipknotted respectively. This is followed by a list of the knot types corresponding to the regions (with multiplicity if there is more than one region with a given knot type) in decreasing order of knotted core length within the regions. For example, in Figure 1, the protein 3FR8 is of type $K4_1\ 3_1$ and protein 2WSX is $S3_1\ 4_1\ 3_1$. This naming is not sensitive to the size, shape or placement of the regions in the knotting fingerprint. In particular, there are many proteins that are described simply as $K3_1$ or $S3_1$, but whose matrix presentations look much different.

One might assume that the knotting fingerprints are unique to each protein. However, we found that many knotting fingerprint motifs reappear throughout our calculations. Furthermore, we found that proteins with the same function in different organisms showed similar knotting fingerprints despite large differences in the amino acid sequences. For example, the matrix presentation for the ubiquitin C-terminal hydrolases 3IRT (human), 1CMX (yeast) and 2WDT (*Plasmodium*) are nearly identical (Figure 3) despite very low sequence identities (ranging from 25% to 32%). This knotting fingerprint motif has persisted through hundreds of

millions of years of evolutionary separation, suggesting that the knotting is indeed critical to the function of the protein.

We present several similar cases in [11]. Although the exact function of the knotting is not yet established, the case of cloacins and S-pyocins provides possible clues to this mystery. Cloacins and S-pyocins are toxins which are released by some bacteria. They enter other bacterial cells via membrane translocation. The knotting fingerprint in Figure 2(B) for the N-terminal translocation domain for colicin E7 in *E. coli* (PDB code 2AXC) shows four isolated regions of 3_1 knots (see the schematic drawing in Figure 2B to see how these regions are created). The protein forms a large loop strapping together several β-strands. In following the chain, one alternates between being on different sides of the strapping loop. If a subchain terminates on one side of the loop, it forms a trefoil knot, but if it terminates on the other side of the loop, an unknot is formed. Since the large loop embraces a significant portion of the protein, one is tempted to conjecture that this embracing stabilizes the relevant parts of the proteins. This is consistent with results of many researchers (see, e.g., [15–19]).

Conclusions

Today, basic information about knotted proteins is easily accessible through many webpages [20–23] that allow researchers to determine the knot type of a protein as well as to locate knotted positions along the backbone. Furthermore, the entirety of the knotting and slipknotting within the proteins can now be visualized using the matrix presentation. The future analysis of the knotting fingerprint motifs within the matrix presentations will yield new clues into the critical link between the geometrical configuration and the function of proteins.

Acknowledgements

We thank the Isaac Newton Institute for Mathematical Sciences for sponsoring the Topological Aspects of DNA Function and Protein Folding workshop and for hosting our stays at the Institute.

Funding

E.J.R. was supported by the National Science Foundation (NSF) [grant number 1115722]. J.I.S. was supported by the Foundation of Polish Science [grant number PHY-0822283] and by the Center for Theoretical Biological Physics sponsored by the National Science Foundation [grant number MCB-1214457]. A.S. was supported by the Swiss National Science Foundation [grant number 31003A-138367].

References

1 Trabi, M. and Craik, D.J. (2002) Circular proteins: no end in sight. Trends Biochem. Sci. **27**, 132–138
2 Boutz, D.R., Cascio, D., Whitelegge, J., Perry, L.J. and Yeates, T.O. (2007) Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. J. Mol. Biol. **368**, 1332–1344
3 Cao, Z.B., Roszak, A.W., Gourlay, L.J., Lindsay, J.G. and Isaacs, N.W. (2005) Bovine mitochondrial peroxiredoxin III forms a two-ring catenane. Structure **13**, 1661–1664
4 Mansfield, M.L. (1994) Are there knots in proteins? Nat. Struct. Biol. **1**, 213–214
5 Mansfield, M.L. (1997) Fit to be tied. Nat. Struct. Biol. **4**, 166–167
6 Taylor, W.R. (2000) A deeply knotted protein structure and how it might fold. Nature **406**, 916–919
7 King, N.P., Yeates, E.O. and Yeates, T.O. (2007) Identification of rare slipknots in proteins and their implications for stability and folding. J. Mol. Biol. **373**, 153–166
8 Millett, K.C., Rawdon, E.J., Stasiak, A. and Sułkowska, J.I. (2012) Identifying knots in proteins. Biochem. Soc. Trans. **41**, 533–537
9 Millett, K.C., Dobay, A. and Stasiak, A. (2005) Linear random knots and their scaling behavior. Macromolecules **38**, 601–606
10 Millett, K.C. and Sheldon, B.M. (2005) Tying down open knots: a statistical method for identifying open knots with applications to proteins. Ser. Knots Everything **36**, 203–217
11 Sułkowska, J.I., Rawdon, E.J., Millett, K.C., Onuchic, J.N. and Stasiak, A. (2012) Conservation of complex knotting and slipknotting patterns in proteins. Proc. Natl. Acad. Sci. U.S.A. **109**, E1715–E1723
12 Banavar, J.R., Hoang, T.X., Maddocks, J.H., Maritan, A., Poletto, C., Stasiak, A. and Trovato, A. (2007) Structural motifs of biomolecules. Proc. Natl. Acad. Sci. U.S.A. **104**, 17283–17286
13 Darcy, I.K., Scharein, R.G. and Stasiak, A. (2008) 3D visualization software to analyze topological outcomes of topoisomerase reactions. Nucleic Acids Res. **36**, 3515–3521
14 Tubiana, L., Orlandini, E. and Micheletti, C. (2011) Probing the entanglement and locating knots in ring polymers: a comparative study of different arc closure schemes. Prog. Theor. Phys. Suppl. **191**, 192–204
15 Sayre, T.C., Lee, T.M., King, N.P. and Yeates, T.O. (2011) Protein stabilization in a highly knotted protein polymer. Protein Eng., Des. Sel. **24**, 627–630
16 Sułkowska, J.I., Sułkowski, P., Szymczak, P. and Cieplak, M. (2008) Stabilizing effect of knots on proteins. Proc. Natl. Acad. Sci. U.S.A. **105**, 19714–19719
17 Virnau, P., Mirny, L.A. and Kardar, M. (2006) Intricate knots in proteins: function and evolution. PLoS Comput. Biol. **2**, 1074–1079
18 Sułkowska, J.I., Sułkowski, P. and Onuchic, J.N. (2009) Jamming proteins with slipknots and their free energy landscape. Phys. Rev. Lett. **103**, 268103
19 Bornschlogl, T., Anstrom, D.M., Mey, E., Dzubiella, J., Rief, M. and Forest, K.T. (2009) Tightening the knot in phytochrome by single-molecule atomic force microscopy. Biophys. J. **96**, 1508–1514
20 Comoglio, F. and Rinaldi, M. (2011) A topological framework for the computation of the HOMFLY polynomial and its application to proteins. PLoS ONE **6**, e18693
21 Kolesov, G., Virnau, P., Kardar, M. and Mirny, L.A. (2007) Protein knot server: detection of knots in protein structures. Nucleic Acids Res. **35**, W425–W428
22 Lai, Y.L., Chen, C.C. and Hwang, J.K. (2012) pKNOT v.2: the protein KNOT web server. Nucleic Acids Res. **40**, W228–W231
23 Lua, R.C. (2012) PyKnot: a PyMOLtool for the discovery and analysis of knots in proteins. Bioinformatics **28**, 2069–2071

Received 5 November 2012
doi:10.1042/BST20120329