Identifying knots in proteins

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

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

Abstract

Polypeptide chains form open knots in many proteins. How these knotted proteins fold and finding the evolutionary advantage provided by these knots are among some of the key questions currently being studied in the protein folding field. The detection and identification of protein knots are substantial challenges. Different methods and many variations of them have been employed, but they can give different results for the same protein. In the present article, we review the various knot identification algorithms and compare their relative strengths when applied to the study of knots in proteins. We show that the statistical approach based on the uniform closure method is advantageous in comparison with other methods used to characterize protein knots.

Introduction

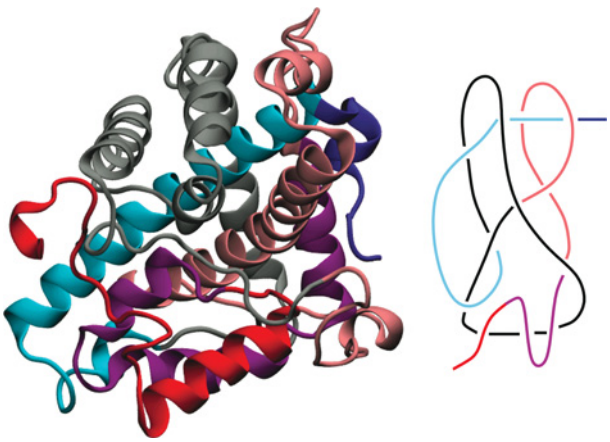
With the realization in the 1960s that many molecules consist of long polymeric chains, proposals were put forward stating that they would be knotted with probability increasing to 1 as their length increased to infinity [1,2]. At the same time, linking and knotting were introduced into the statistical mechanical models of polymers [3,4]. A bit later on, the theory of reptation added to our understanding of polymeric properties and the consequences of entanglement [5]. Initially, the consideration of knots formed by proteins was limited to those cases where disulfide bridges or covalently closed metal atoms formed natural closed circuits determined by these covalent bonds and the protein backbone [6,7]. More recent interest has been concentrated on the open knots formed entirely by the protein backbones. The critical issue is how to mathematically identify and characterize knotting in open chains using a topological formalism that can be applied to these open protein knots. As a consequence, a number of strategies have been considered. The first approach used was the determination of a primitive path associated to a polymer chain [8], whereby one keeps the ends of the open polymer chain fixed and implements a procedure that shortens the chain so as to concentrate the knotting and other manifestations of entanglement without violating excluded volume constraints. However, as the initial focus was on interactions between distinct polymer chains, researchers did not fully appreciate that, whereas chain shortening preserves the knot type of closed chains, this is not necessarily the case for open chains [9,10]. In 1994, Mansfield [11] reported the first systematic studies of the approximately 400 known protein structures deposited in the PDB. Mansfield used a double stochastic closure method combined with direct observation to evaluate evidence of knotting. Noting that the termini of these structures are preferentially located near the surface of the protein structures, Mansfield later employed a preferential closure method to provide evidence of knotting in the proteins MAT [(*S*)-adenosylmethionine synthetase] and CAB (carbonic anhydrase B) [12]. These protein knots are shallow and a small displacement of an end of the polypeptide chain could unknot them. In 2000, using an enhanced primitive path method, Taylor [13] identified the presence of a 'deeply embedded' knot in a protein structure, thereby giving the first identification of a robustly knotted protein structure. In the present article, we explore the methods and variations that have been applied to the study of knotting in proteins. With the addition of new structures, researchers have discovered many more interesting examples of knots and slipknots (knotted segments contained within larger unknotted segments [14]). We use the protein DehI (PDB code 3BJX), in which the Stevedore's knot has been identified [15], as it provides an excellent example of a challenging open knot. It and its simplified knot structure are shown in Figure 1.

Open chain knot identification strategies

Knot theorists define the knotting of an open chain in terms of a pair consisting of the chain and a three-dimensional ball containing the knotted chain in its interior and meeting its boundary sphere in exactly the termini of the chain [16]. Unfortunately, there are infinitely many such a balls giving different 'knots', so selecting the 'right one' is, in fact, the critical challenge. Thus a useful way to pose this question is 'how can one reasonably sample the space of formed knots so as to find the most dominant one?' One can look to topological methods used to study polygonal ring models of knots [17]. Reidemeister [17] describes their topological

Key words: knot, Millett–Dobay–Stasiak (MDS), primitive path, open chain, slipknot.
Abbreviation used: MDS, Millett–Dobay–Stasiak.
[1]To whom correspondence should be addressed (email Andrzej.Stasiak@unil.ch).

Figure 1 | **Protein DehI (PDB code 3BJX) with its associated colour-coded knot structure showing the presence of the Stevedore's knot**



equivalence by using the existence of triangles meeting the polygon in exactly one or two edges. These provide ways to increase or decrease the number of edges that do not change the knot type of the closed polygonal rings. This observation leads to an elementary curve-shortening strategy used to determine the primitive path by successive shortening of the polygonal chain. One searches for adjacent edges in the polygon such that the associated (solid) triangle is not pierced by a distal portion of the protein backbone. The pair of edges is then replaced by the third edge of the triangle, thereby shortening the chain. For an open polygon, the initial and terminal vertices are fixed in this method. For closed chains, the knot type is preserved by these operations so they can be used to reduce the complexity of the configuration [18]. Once no further simplifications are possible, one is left with a much simpler configuration whose structure can be determined visually or by using one of the methods described below. For example, when the end vertices are on the boundary of the convex hull, one can close the chain using an arc lying outside this convex hull so as to not introduce (or lose) knotting.

A variation of this method was used by Taylor [13] to identify a deep figure-eight knot, 4_1, in acetohydroxy acid isomeroreductase (PDB code 1YVE) in his study of the approximately 3440 structures deposited in the PDB by 2000. In Taylor's smoothing method [13], one begins at the N-terminus and, sequentially, considers three successive vertices, defines a new vertex as their average position, and forms the two triangles determined by the first two vertices and the new vertex and by the last two vertices and the new vertex. If neither of these solid triangles has a distal intersection with the structure, replace the middle vertex with the new vertex and continue with the next sequence until the end is reached, and then continue, repeating from the N-terminus until the result stabilizes. This method may, as was shown by a simple example [8,9], give different results if one begins at the C-terminus.

Although strategies such as elementary curve shortening and Taylor's smoothing algorithm do 'simplify' chains to something that might be easier to analyse, they do not simplify every unknotted conformation to a segment connecting the N- and C-termini as one might hope. In addition, as was shown in [8,9], the order in which the simplification of the chain occurs can affect the knot type determined by the algorithm. One might wonder to what extent the order of simplification affects knot type determination in proteins specifically. We explored this problem by carrying out successive elementary curve-shortening moves at random edges of proteins until no more curve-shortening moves were possible (we call this the random elementary curve-shortening method). For the deeply knotted 1YVE protein, the figure-eight knot was identified each of the 50 times we applied the algorithm (Table 1). Applying this method to 3BJX, we found 6_1 half of the time and 0_1 half of the time. This shows that the knot type identification in proteins also can be affected by the order of the simplification moves, and suggests that simplification is not as robust as one might imagine. If the termini lie on the boundary of the convex hull of the conformation, both methods do not change the knotting and therefore lead to the same identification of the associated knot type. If, however, one of the termini lies within the interior of the convex hull, one is faced with the task of unambiguously determining the closure. For example, one might wish to identify the 'correct' single closure that expresses the knotting of the conformation.

There are several single closure strategies that we review next. The first, and the simplest, is the direct closure method in which one simply connects the termini by a straight segment. As protein termini often lie close to the surface of the protein structure, but not necessarily close to each other, the closure segment frequently passes through the 'centre' of the protein. In such a case, the closing segment can produce a knot from an essentially unknotted protein structure (see the case of protein 2A65 in Table 1). For 3BJX, however, one still finds the 6_1 knot (Table 1).

Other members of this class of single closure methods employ a specific algorithm to determine a closing edge or sequence of edges beginning at the termini and ending on a large sphere containing the structure (from which a standard closure will give a well-defined result). For example, one can select a random edge direction and take parallel edges starting from the termini to an enclosing sphere. The closure can be accomplished by using any arc on the sphere to connect the new vertices; for example, one could take a great circle closure [19]. Compared with the direct closure method, these unbiased parallel edges are less likely to result in the added edges passing through the centre of the protein when the two termini are situated on opposite sides of the protein. Still, in the random edge direction method, the added edges can pass through portions of the centre of the protein and affect the knotting structure. Furthermore, different choices of edge directions are likely to result in different knot types.

To reduce this degree of uncertainty, other methods have been proposed. One of these is the radial method, given by extending a ray, based at the centre of mass of the structure,

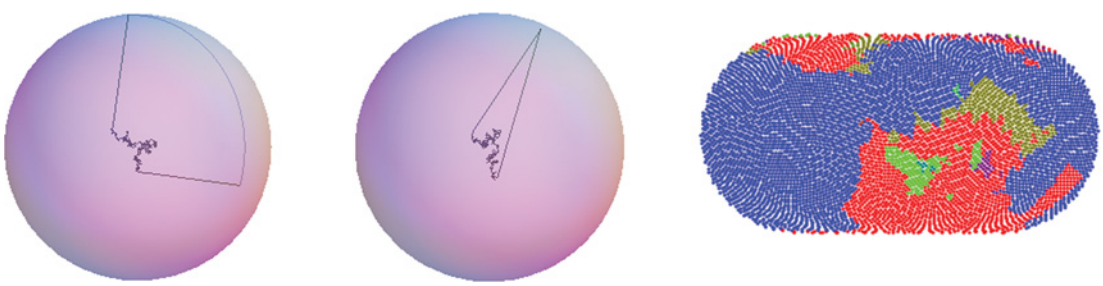
Table 1 | **Results of different knot detection methods applied to selected proteins**

The primitive path method reports the results of a single set of 50 applications of the elementary curve-shortening method; applications of the random MDS method using ten different sets of 6400 randomly selected points on a sphere; applications of the uniform closure method using ten random rotations of a set of 6400 uniformly distributed points on the sphere; the direct method joins the termini; and the radial method adds initial and terminal edges to a sphere using the rays from the centre of mass passing through the termini.

PDB code	Primitive path	Random MDS	Uniform closure	Direct	Radial
1XD3	3_1 84%	5_2 67.7 ± 0.3%	5_2 67.61 ± 0.00%	0_1	5_2
1DMX	0_1 80%	3_1 66.8 ± 0.3%	3_1 67.22 ± 0.00%	0_1	3_1
1FUG	3_1 50%	3_1 85.9 ± 0.2%	3_1 85.75 ± 0.00%	3_1	3_1
1YVE	4_1 100%	4_1 71.9 ± 0.4%	4_1 71.56 ± 0.01%	4_1	4_1
1KOP	0_1 100%	0_1 44.4 ± 0.5%	0_1 44.26 ± 0.01%	4_1	0_1
1ZTU	4_1 60%	4_1 81.7 ± 0.2%	4_1 81.74 ± 0.01%	4_1	4_1
2VEA	0_1 64%	4_1 92.5 ± 0.2%	4_1 92.45 ± 0.00%	4_1	4_1
3BJX	6_1 50%	6_1 64.1 ± 0.1%	6_1 63.99 ± 0.01%	6_1	6_1
2A65	0_1 58%	0_1 85.5 ± 0.4%	0_1 85.32 ± 0.01%	6_1	0_1

Figure 2 | **Two closure methods and an Eckert IV knot distribution representation**

The double and single stochastic closure methods are illustrated on the left and in the middle. On the right, an Eckert IV projection of the knot types resulting from 64 000 uniformly distributed single point closures of the protein structure (PDB code 3BJX): 64% blue is 6_1, 27% red is the unknot, 6% dark green is 4_1, and 2.5% light green is 3_1 (compare with Table 1).



from each of the termini to an enclosing sphere [20–23]. When applied to 3BJX, the radial method identifies the Stevedore's knot (Table 1). In yet another method, protein repulsion closure, one focuses on the specific nature of a protein [14] and follows a backbone-smoothing procedure (similar to curve shortening) and generates a sequence of small segments from the termini as though they are repelled by the protein to reach the exterior of the protein structure, at which time one closes the chain in the complement of the convex hull of the protein. Similarly [24], in the minimally interfering closure method, one extends terminal segments to the closest points on the boundary of the convex hull when the termini are closer to this boundary than they are to each other (otherwise one would use the direct closure). One can imagine other variations that exploit specific knowledge of the protein structure and which would appear attractive. A common concern is that each depends on a specific structure and may have unanticipated consequences when applied to a large family of dissimilar structures. This is, indeed, a key facet of the challenge of creating a method that will apply equally well to each of the subchains as well as the entire chain independent of their specific structure.

The double stochastic closure method [10,20] represents a rather different strategy. It consists of adding an edge from each of the termini to independent random points on a large sphere containing the chain and then connecting these points, on the sphere, by a segment of a great circle (Figure 2). The knot type of the associated ring is then determined for each closure. Applying this double stochastic closure method to 3BJX, in a 1000-closure sample, we found 22 different knot types. The Stevedore's knot, 6_1, appeared at 47.5%, the unknot at 35%, 4_1 at 6.2%, 3_1 at 4.1%, and the remainder at much smaller proportions. Thus this method provides weaker evidence of the demonstrated knotting [15] when compared with the other methods discussed in the present article (Table 1).

In 2005, concerned with the dependence of the Taylor method on the specific shortening sequence, we proposed a single stochastic method: the MDS (Millett–Dobay–Stasiak) method [8,9]. In the MDS method, one again employs a large ball containing the protein (or other open chain structure). The boundary sphere approximates the 'sphere at infinity' (Figure 2). At a random set of points on the sphere, ones whose connection edges with the termini defines

a non-singular polygon, we used the HOMFLY (Hoste–Ocneanu–Millett–Fryed–Lickorish–Yetter) knot polynomial [25] to identify the associated knot type. This knot type is defined except on a compact one-dimensional set of measure zero, is locally constant and takes on only finitely many values. As a consequence, one can define rigorously the proportion of the closures that gives a specific knot type, thereby defining the knotting spectrum of the configuration as shown in a histogram of the distribution of knot types [8,9]. If a knot type occurred more than 50% of the time, we proposed that the protein has that knot type in the MDS method [8,9]. Currently, however, we select the knot type occurring most often as the knot type of the configuration. In the case of 3BJX, our simulations show that 6_1 occurs approximately $64.1 \pm 0.1\%$ of the time, thereby providing confirming evidence of the known knotting [15].

To estimate the proportion of closures that give a specific knot type using the MDS method, we employed a Monte Carlo method that randomly selects closure points on the sphere, in some cases as many as 10 000, in order to give accurate estimates [9]. A more effective strategy, often used in numerical analysis, is to employ a carefully constructed finite set of points that are close to being uniformly distributed on the sphere, giving the uniform closure method. How many such points are necessary to give an accurate estimate? In Figure 2, we show the Eckert IV projection of the spherical distribution of knot types (there are 12 distinct ones) for 64 000 uniformly distributed points [26]. For comparison, the uniformly and randomly generated datasets of 49, 100, 169, 400 and 6400 points give proportions of 6_1s: $64.7 \pm 2.2\%$, $63.8 \pm 0.9\%$, $64.7 \pm 0.7\%$, $64.1 \pm 0.5\%$ and $64.1 \pm 0.1\%$ and $59.6 \pm 4.2\%$, $64 \pm 1.9\%$, $62.4 \pm 3.1\%$, $65.1 \pm 1.4\%$ and $64.0 \pm 0.0\%$ respectively. Although there is some expected difference between the uniform and random results (as there would be between successive random estimates), the data demonstrate the stability of the MDS method in providing a consistent identification.

Discussion

With more than 75 000 structures currently deposited in the PDB, a number that is rapidly increasing, and the interest in the possible function of protein knots (both global knots and substructures such as slipknots), the ability to unambiguously assess the presence of knotting is of increasing importance. The occurrence of protein knots raises many evolutional and functional questions for which compelling data are necessary. As experimental techniques are still unable to determine the knotting mechanism [27,28], computer simulations can still shed new light on the folding landscape [29–33] and the dynamic of optimization of chain structures [30] to guide efficient knotting. To do so one requires accurate and efficient methods to detect and identify knotted and slipknotted structures.

The present review of the various strategies commonly employed in this analysis of protein structures, their strengths and uncertain aspects, as well as a comparison of their relative effectiveness has led us to prefer the version of the stochastic method in which one uses a set of uniformly distributed points on the sphere to estimate the knotting distribution, as is shown in Figure 2. Although our analysis may lead one to conclude that, for all practical purposes, it may be possible to give an adequate estimation with fewer points, it seems prudent to use as many as 100 uniform closure points to give a measure of numerical confidence.

Another facet, in addition to the precision of the method, is that of computational effectiveness. In undertaking a detailed analysis of the local knotting structure of a given protein, one must analyse all subchains of the structure [27,28]. This is a formidable computational task when applied to all of the protein structures available. The uniform closure method is computationally attractive. In another article in this issue of *Biochemical Society Transactions* [28], we review the implementation of the uniform closure method to study the knotting found within the complete subchain array, the presentation of the resulting knotting data and what one can learn from the analysis of this 'knotting fingerprint' associated with the protein structure.

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 Delbrück, M. (1962) Knotting problems in biology. Proc. Symp. Appl. Math. **14**, 55–58
2 Frisch, H.L. and Wasserman, E. (1961) Chemical topology. J. Am. Chem. Soc. **18**, 3789–3795
3 Edwards, S.F. (1967) Statistical mechanics with topological constraints: I. Proc. Phys. Soc. **91**, 513–519
4 Edwards, S.F. (1968) Statistical mechanics with topological constraints: II. J. Phys. A: Gen. Phys. **1**, 15–28
5 de Gennes, P.-G. (1971) Concept de reptation pour une chaine polymerique. J. Chem. Phys. **55**, 572
6 Crippen, G. (1974) Topology of globular proteins. J. Theor. Biol. **45**, 327
7 Liang, C. and Mislow, K. (1994) Knots in proteins. J. Am. Chem. Soc. **116**, 11189–11190
8 Edwards, S.F. (1977) The theory of rubber elasticity. Br. Polym. J. **9**, 140–143
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 Mansfield, M.L. (1994) Are there knots in proteins? Nat. Struct. Biol. **1**, 213–214
12 Mansfield, M.L. (1997) Fit to be tied. Nat. Struct. Biol. **4**, 166–167
13 Taylor, W. (2000) A deeply knotted protein structure and how it might fold. Nature **406**, 916–919
14 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
15 Bolinger, D., Sułkowska, J.I., Hsu, H.-P., Mirny, L.A., Kardar, M., Onuchic, J.N. and Virnau, P. (2010) A Stevedore's protein knot. PLoS Comput. Biol. **6**, e1000731

16 Kirby, R.C. and Lickorish, W.B.R. (1979) Prime knots and concordance. Math. Proc. Camb. Phil. Soc. **86**, 437–441
17 Reidemeister, K. (1927) Elementare Begründung der Knotentheorie. Abh. Math. Semin. Univ. Hamburg **5**, 24–32
18 Koniaris, K. and Muthukumar, M. (1991) Self-entanglement of ring polymers. J. Chem. Phys. **95**, 2873–2881
19 Janse van Rensburg, E.J., Sumners, D.W., Wasserman, E. and Whittington, S.G. (1992) Entanglement complexity of self-avoiding walks. J. Phys. A: Math. Gen. **25**, 6557–6566
20 Lua, R.C. and Grosberg, A.Y. (2006) Statistics of knots, geometry of conformations, and evolution of proteins. PLoS Comp. Biol. **2**, e45
21 Virnau, P., Mirny, L.A. and Kardar, M. (2006) Intricate knots in proteins: function and evolution. PLoS Comp. Biol. **9**, e122
22 Virnau, P., Mallam, A. and Jackson, S. (2011) Structures and folding pathways of topologically knotted proteins. J. Phys.: Condens. Matter **23**, 033101
23 Marcone, B., Orlandini, E., Stella, A.L. and Zonta, F (2005) What is the length of a knot in a polymer? J. Phys. A: Math. Gen. **38**, L15–L21
24 Tubiana, Orlandini, L.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
25 Millett, K.C. and Lickorish, W.B.R. (1987) A polynomial invariant for oriented links. Topology **26**, 107–141
26 Sloan, I.H. and Womersley, R.S. (2004) Extremal systems of points and numerical integration on the sphere. Adv. Comp. Math. **21**, 102–125
27 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
28 Rawdon, E.J., Millett, K.C., Sułkowska, J.I. and Stasiak, A. (2012) Knot localization in proteins. Biochem. Soc. Trans. **41**, 538–541
29 Mallam, A.L. and Jackson, S.E. (2012) Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. Nat. Chem. Biol. **8**, 147–153
30 King, N.P., Jacobitz, A.W., Sawaya, M.R., Goldschmidt, L. and Yeates, T.O. (2010) Structure and folding of a designed knotted protein. Proc. Natl. Acad. Sci. U.S.A. **107**, 20732–20737
31 Sułkowska, J.I., Sułkowski, P. and Onuchic, J. (2009) Dodging the crisis of folding proteins with knots. Proc. Natl. Acad. Sci. U.S.A. **106**, 3119–3124
32 Sułkowska, J.I., Noel, J.K. and Onuchic, J.N. (2012) Energy landscape of knotted protein folding. Proc. Natl. Acad. Sci. U.S.A. **109**, 17783–17788
33 Li, W., Terakawa, T., Wang, W. and Takada, S. (2012) Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. Proc. Natl. Acad. Sci. U.S.A. **109**, 17789–17794

Received 15 November 2012
doi:10.1042/BST20120339