BIOINFORMATICS

Vol. 22 no. 14 2006, pages e252–e259
doi:10.1093/bioinformatics/btl236

Rapid knot detection and application to protein structure prediction

Firas Khatib*, Matthew T. Weirauch and Carol A. Rohl**

Department of Biomolecular Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064

ABSTRACT

Motivation: Knots in polypeptide chains have been found in very few proteins, and consequently should be generally avoided in protein structure prediction methods. Most effective structure prediction methods do not model the protein folding process itself, but rather seek only to correctly obtain the final native state. Consequently, the mechanisms that prevent knots from occurring in native proteins are not relevant to the modeling process, and as a result, knots can occur with significantly higher frequency in protein models. Here we describe Knotfind, a simple algorithm for knot detection that is fast enough for structure prediction, where tens or hundreds of thousands of conformations may be sampled during the course of a prediction. We have used this algorithm to characterize knots in large populations of model structures generated for targets in CASP 5 and CASP 6 using the Rosetta homology-based modeling method.

Results: Analysis of CASP5 models suggested several possible avenues for introduction of knots into these models, and these insights were applied to structure prediction in CASP 6, resulting in a significant decrease in the proportion of knotted models generated. Additionally, using the knot detection algorithm on structures in the Protein Data Bank, a previously unreported deep trefoil knot was found in acetylornithine transcarbamylase.

Availability: The Knotfind algorithm is available in the Rosetta structure prediction program at http://www.rosettacommons.org

Contact: bort@soe.ucsc.edu

1 INTRODUCTION

In a formal topological sense, knots in protein chains cannot be defined because the protein backbone, disregarding disulfide bridges and other sources of backbone crosslinks, does not form a closed loop. Jane Richardson (1977) was the first to define a knotted protein chain as one which cannot be fully extended to a straight line if one were to grab the N- and C-terminus in each hand and pull. Few protein structures have been observed to contain knots in their backbones (Nureki *et al*., 2002), and in most cases where knots have been observed, they tend to be simple overhand knots near one terminus (Mansfield, 1994). These knots could in theory form by threading a short section of the polypeptide chain through a loop formed by another backbone section. Such knots disappear if a few residues are trimmed from the terminal ends (Taylor, 2000). Deep knots, in contrast, occur far from the protein chain termini and have been rarely observed.

Because knots in protein structures are rare, protein structure prediction methods should generally avoid introducing knots into the polypeptide backbone. Most structure prediction methods do not, however, check for knots. Additionally, few protein structure prediction methods model the kinetic protein folding process, so the entropic mechanisms that have been cited as explanations for the relative absence of knots in protein structures (Taylor, 2000) are not likely to prevent the introduction of knots in the modeling process. In fact, algorithms used for structure prediction do introduce knots in the polypeptide backbone, as demonstrated by predictions made for the Comparative Assessment of Methods for Structure Prediction (CASP) experiments (Moult *et al*., 1995). In the CASP 4 protein structure prediction experiment, one submitted model was assessed as being reasonably accurate in terms of atomic coordinates, but was also described by the CASP assessors as an ‘‘impossible structure’’ because it contained a trefoil knot (Tramontano *et al*., 2001). In the most recent CASP 6 experiment, the assessors reported that knotted models were still being submitted and that such knotted models submitted for comparative modeling targets were rejected out of hand without additional assessment (Tress *et al*., 2005a).

Knots in polypeptides can be difficult to detect by visual inspection alone, as evidenced by the fact that the assessors accepted some knotted CASP 6 models, presumably because it was not apparent that these models contained knots. Algorithms for automated knot detection have been reported (Taylor, 2000) but are too slow for general use in structure prediction, where tens or hundreds of thousands of conformations may need to be examined in the course of a single structure prediction. Here we present Knotfind, a rapid algorithm for knot detection, and report its application in the context of the Rosetta homology-based structure prediction method (Bradley *et al*., 2003; Rohl *et al*., 2004a). Additionally, the algorithm was applied to experimentally-determined protein structures in the Protein Data Bank (PDB; Berman *et al*., 2000) identifying a previously unreported deep trefoil knot.

2 METHODS

2.1 Knot-detection algorithm

The Knotfind algorithm considers only Cα atoms in a single protein chain and progressively ‘eliminates’ atoms from the Cα trace to simplify the chain. Triples of consecutive Cα atoms, *i*-1, *i*, *i*+1, are considered, ordered by

*To whom correspondence should be addressed.

**Current address: Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck & Co., Inc., 401 Terry Avenue N., Seattle, WA 98195

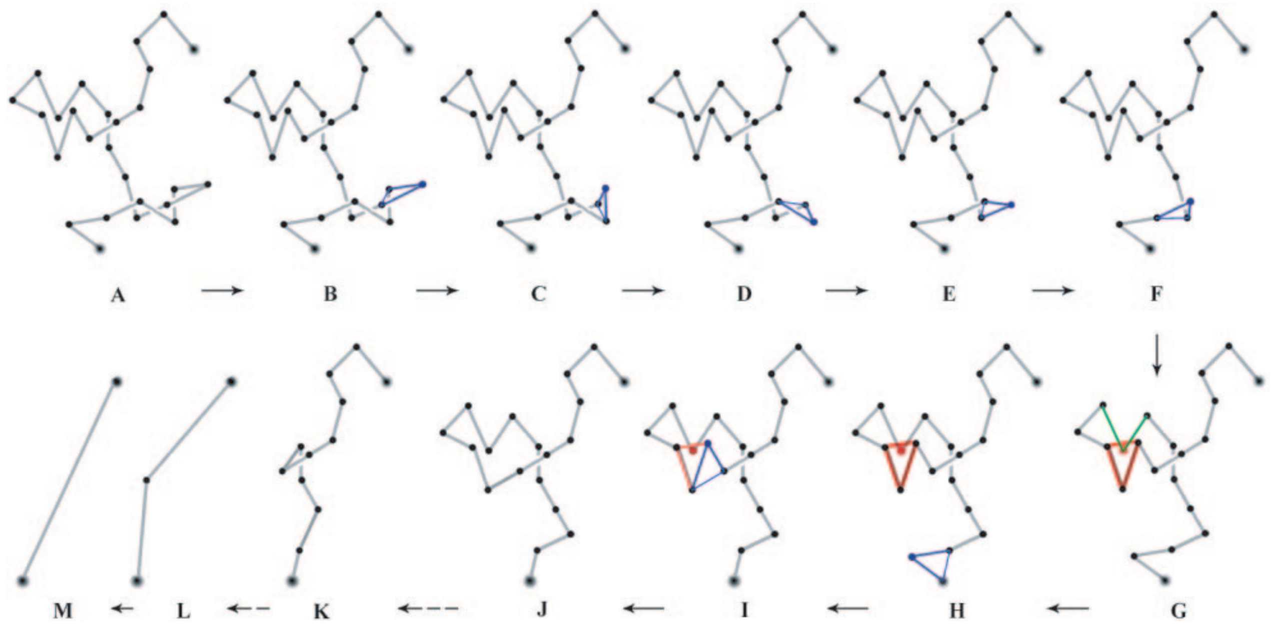

Fig. 1. Schematic illustration of the Knotfind algorithm. Starting from the initial Cα trace (trace A), atoms are progressively eliminated from the chain, effectively simplifying it to a straight line. For steps shown in traces B to F, the central Cα atom in the triple describing the most acute triangle (triangle shown in blue) is removed. In trace G, the most acute triangle cannot be simplified because two line segments (green) pass through this triangle; removing the central atom would effectively result in passing the red chain segment through the green segments. Since the red triple cannot be simplified, the triple forming the second most acute triangle (blue) is targeted for simplification in trace H. In trace I, the red triple still cannot be simplified, but the triple forming the next most acute triangle (blue) can be, yielding trace J. Trace K is obtained by seven additional atom removals and trace L by nine additional simplifications.

increasing Cartesian distance between atoms *i*-1 and *i*+1. For an individual triple, if no line segments connecting consecutive Cα atoms *j*, *j*+1 (for all *j* < *i*-1 and *j* > *i*+1) cross through the triangle defined by *i*-1, *i*, *i*+1, then Cα *i* is removed from the chain. If any line segment connecting two consecutive Cα atoms intersects the triangle, however, then no simplification of this triple is made and the algorithm proceeds to the triple with the next shortest *i*-1, *i*+1 distance. After any Cα is eliminated from the chain, the algorithm returns to the triple with the shortest *i*-1, *i*+1 distance. This procedure is repeated until the last triple in the distance list has been selected and simplified, if possible. When the algorithm terminates, if the only atoms left in the chain are the N- and C-terminal Cα atoms such that chain has been simplified to a straight line, the protein contains no knots (Figure 1). If, instead, the chain cannot be fully simplified to a single extended segment, the chain contains one or more knots and the remaining Cα atoms in the chain define the knotted region. In cases where a knot is detected, the algorithm is repeated using an alternate scheme to order the triples for simplification in which the area of the triangle defined by each *i*-1, *i*, *i*+1 triple is used in place of the *i*-1, *i*+1 interatomic distance to reduce false positives.

To determine if a line segment intersects a triangle, the algorithm first ensures that the plane containing the triangle and the line containing the line segment are not parallel, and then determines if both endpoints of the line segment lie on the same side of the plane. For segments that intersect the plane of the triangle, the algorithm determines if the intersection point lies within the triangle, relying on the fact that the sum of the internal angles of a point inside a triangle is 2π. Thus, any point lying outside the triangle will have smaller angle sums (http://astronomy.swin.edu.au/∼pbourke/geometry/linefacet). An effective line width of 0.0003 is used in order to handle round off errors on arccosines in computing angle sums.

The Knotfind algorithm has been implemented in the Rosetta structure prediction program available at http://www.rosettacommons.org and in the Undertaker program (Karplus *et al*., 2005).

2.2 Protein structures

The PISCES server was used to identify 9,553 protein chains in the RCSB PDB as of February 12, 2006 with less than 90% sequence identity, with x-ray structures of resolution better than 3.0Å and no R-factor filtering (R ≤ 1.0) (Wang *et al*., 2003). This list was supplemented with four protein chains that have previously been reported to be knotted (1dmxA, 1fugA, 1yveI, and 2btv), but which did not meet the resolution or sequence identity cutoffs, giving a total list of 9,557 chains that were examined using the Knotfind algorithm. Coordinate files were obtained from the RCSB PDB and ATOM records were compared to the sequence as defined in the SEQRES header to define regions of missing density. Missing density leading to a significant chain discontinuity (i.e. multiple residues not at a chain terminus) can make identification of a knot ambiguous because Cα atoms surrounding the missing density are artificially connected in a Cα trace. Consequently, structures with missing density that were reported by the algorithm to be knotted were visually inspected for confirmation to eliminate those that did not actually contain a knot. Among the 9,557 chains checked, seven knotted structures detected by Knotfind could be attributed to significant missing density: 1gkuB, 1jr1A, 1mqsA, 1o6lA, 1u2zA, 1yc0A, and 2bm0A.

2.3 Rosetta decoy sets

For predicted structures, models generated during the course of structure predictions made for CASP 5 and CASP 6 were utilized. Many structure prediction methods, including Rosetta, generate large numbers of possible model structures, referred to as ‘decoys’, from which a final best model is then selected. Decoy structures for CASP 5 and CASP 6 targets were generated using the Rosetta homology-based modeling method (Bradley *et al*., 2003; Rohl *et al*., 2004a) during the process of the CASP experiments. The CASP 5 decoy sets were generated by the Baker group (Group 2) and exclude decoy sets for any targets for which the de novo Rosetta prediction method was used (Bradley *et al*., 2003). A total of 45,366 decoys were examined here. Decoy sets for individual targets included between 199 and 4,019 decoys. Decoy populations for CASP 6 were those generated during the course of predictions made by the Rohl group (Group 079), also using the Rosetta homology-based method. A total of 119,543 decoys were examined. Decoy sets for individual targets included between 883 and 11,934 decoys. In addition to manually generated Rosetta decoys, models generated by the automated Robetta server, which utilizes the Rosetta method, were also examined. Robetta predictions (Group 101) for CASP 6 targets were obtained from the Robetta server (Chivian *et al*., 2003; Kim *et al*., 2004) and are also available from the CASP 6 website (http://predictioncenter.org/casp6).

All Rosetta decoy sets used here, including models from the Robetta server, use the same basic Rosetta homology-based structure prediction method which has been described elsewhere (Bradley *et al*., 2003; Chivian *et al*., 2005). In brief, predictions begin from an alignment to a parent protein of known structure. Coordinates for aligned regions are taken directly from the parent structure and serve as a fixed template. Coordinates for structurally variable regions (SVRs), corresponding to both gaps in the alignment as well as regions of uncertain alignment, are constructed by assembling short fragments of known structure. These fragments are selected from the database of known protein structures based on similarity of sequence and predicted and known secondary structures. For short SVRs, geometric fit to the template is also considered. The selected fragments are combined using a Monte Carlo simulated annealing search by means of a knowledge-based potential function derived from the observed distributions of residues in known protein structure along with a gap penalty to ensure chain continuity in the final model. A more detailed description of the Rosetta approach and the potential function (Rohl *et al*., 2004b), and the SVR modeling method (Rohl *et al*., 2004a) are described in detail elsewhere. Differences between the CASP 5 and CASP 6 SVR modeling methods are described below.

For CASP 5 decoys, a library of possible conformations was selected via a database search for SVRs shorter than 17 residues. For each decoy, a random conformation for each short SVR was selected and then long SVRs were modeled by fragment assembly in the context of the template. For CASP 6 decoys, a library of possible conformations was generated for every SVR, regardless of length using a combination of database search and fragment assembly. For short SVR regions with 7 or fewer residues, conformations were selected directly from the database and used without further modification. For SVRs in the length range of 8-12 residues, conformations were assembled from 3-9 residue fragments in the context of the entire fixed template. For SVRs greater than 12 residues in length, a reduced template of four residues, two on each side of the SVR, was extracted, and the long SVR was modeled in the context of this reduced template by fragment assembly. For each SVR, regardless of length, 100-200 conformations were initially selected or generated and each of these library conformations was then checked using the Knotfind algorithm to eliminate those that resulted in knots when grafted onto the fixed template. Additionally, conformations with significant steric clashes with the template or large chain discontinuities were discarded. Complete models were then constructed by combining conformations from these libraries, using a Monte Carlo simulated annealing search to optimize the Rosetta centroid-based energy function.

2.4 Undertaker decoy sets

Undertaker decoys for CASP 6 targets were graciously provided by Kevin Karplus. A total of 2,373 decoys were examined. Decoy sets for individual targets included between 6 and 115 decoys. The Undertaker program combines fragment assembly and other methods with coordinate information extracted from alignments to a parent structure in order to generate models for proteins (Karplus *et al*., 2005). Rosetta and Undertaker are substantially different in terms of the optimization strategies and cost functions used, but share substantial similarity in their approach to conformation modification, which includes fragment assembly. Undertaker differs from the Rosetta-based strategy employed for construction of the Rosetta decoy sets used here in that regions modeled on the basis of homology to a parent of known structure are not treated as a fixed template in Undertaker, but instead are subject to conformational modifications.

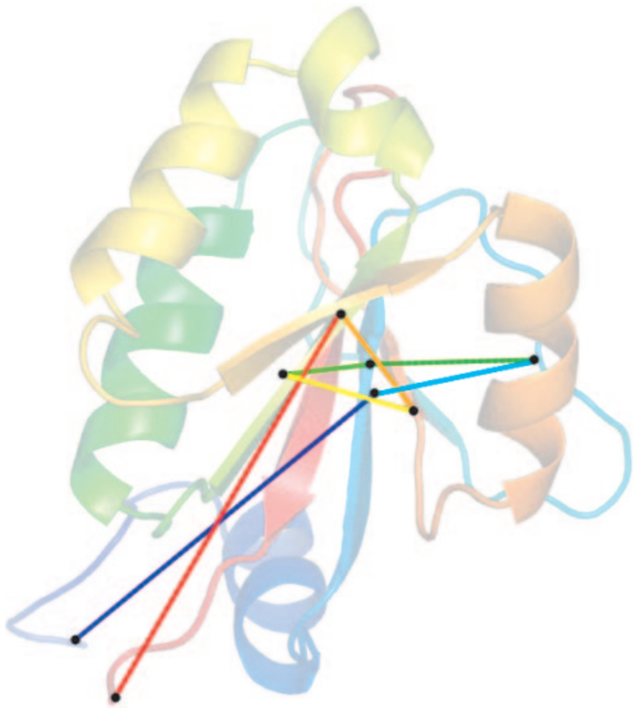

Fig. 2. The trapped state obtained for 1ogdA when simplifying triples in decreasing order of acuteness. Protein chain 1ogdA is shown as a ribbon and the final state resulting from the Knotfind algorithm, applied with triples ordered according to *i*-1, *i*+1 distance, is shown overlaid with the eight unsimplified Cα's indicated by black spheres. At this point, no triple can be simplified, yet this protein chain does not contain a knot. When triples are considered in order of their area by Knotfind, the chain completely simplifies.

3 RESULTS

3.1 Knotfind algorithm

The Knotfind algorithm attempts to simulate the process of pulling the protein chain from both ends in order to determine if the chain contains a knot. As described by Richardson's operational definition, an unknotted chain can be completely pulled into a fully extended conformation. In the presence of a knot, however, the chain cannot be fully extended without one segment of the chain being passed through another segment of the backbone. In the Knotfind algorithm, the chain pulling is modeled by progressively removing atoms from the chain. For each atom removal, all other segments of the backbone are checked to ensure that removal of an atom does not effectively cause one segment of the backbone to pass through another.

Simplifying the chain in a series of discrete steps allows the Knotfind algorithm to be fast, but leaves open the possibility that the chain trace can become trapped in a partially simplified state that does not contain a knot but cannot be further simplified according to

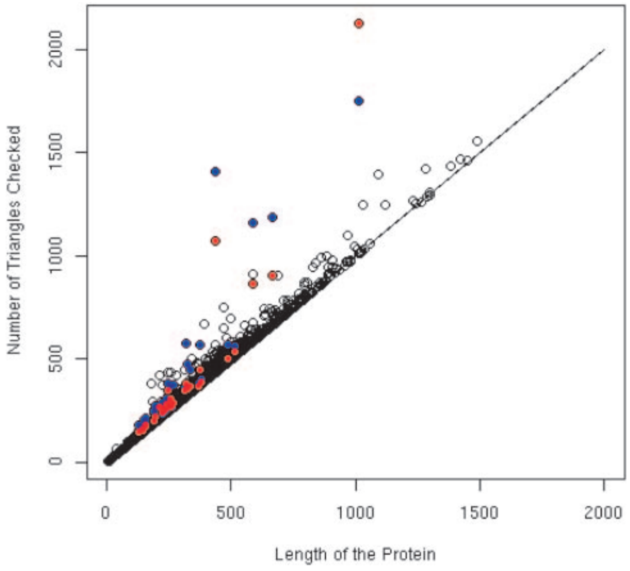

Fig. 3. The number of triples checked by Knotfind as a function of protein chain length for the 9553 protein chains taken from the PISCES server. Colored dots indicate protein chains reported as knotted by one of the triple ordering schemes used by Knotfind (blue: *i*-1, *i*+1 distance; red: triangle area. See Methods).

the Knotfind algorithm. To minimize the possibility of such false positives, triples are considered for simplification in order of the *i*-1, *i*+1 distance, allowing the simplification to start with the most local backbone features before simplifying more global features. Using this strategy, only one false positive is observed among the 9,557 protein chains examined here. The trapped chain configuration for this chain, 1ogdA, is shown in Figure 2. As described in the Methods above, chains that cannot be fully simplified in the first pass of the algorithm are subjected to a second check during which triples are ordered according the area of the triangle that each defines, considering smallest area first. When used in isolation, this area-based ranking method resulted in four false positives (1e2kA, 1y6vA, 2a65A, 2c5aA). When the two methods are applied sequentially, no false positives are observed in the set of PDB chains examined, or in any of the knotted decoy structures that were visually inspected.

One of the main advantages of the Knotfind algorithm is its speed. Despite using two different triplet-ordering schemes in cases where the first scheme does not result in a completely simplified chain, the algorithm as implemented in Rosetta requires on average less than 0.01 seconds for a single chain. When using Rosetta to evaluate the 9,553 chains from the PISCES server, incorporating the Knotfind algorithm added fewer than 90 seconds to the overall run time on an Intel(R) Xeon(TM) CPU 2.80GH compared to evaluating the chains using Rosetta without Knotfind. Most of the time in Knotfind is spent determining if triples can be simplified by establishing if any line segments intersect the triangles defined by each triple. The number of triples checked depends linearly on the length of the chain in the absence of a knot, while knotted chains require that more triangles be tested for simplification than would be expected on the basis of chain length (Figure 3).

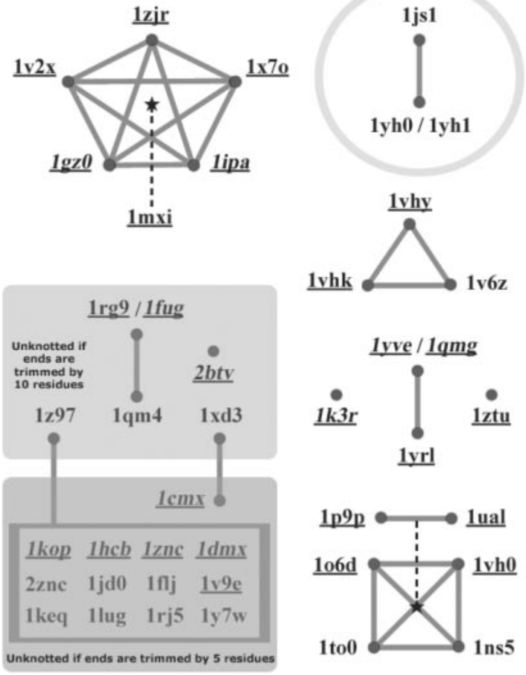

Fig. 4. Relationships between knotted proteins detected by Knotfind. Protein chains are referenced by their PDB codes. Protein pairs sharing sequence similarity (BLASTp evalue < 1E-05) are indicated by solid lines. Structural similarity (MAMMOTH evalue < 1E-07) is indicated by dotted black lines. The pair of sequence similar proteins circled in the top right corner represents a knotted protein fold that has not been previously reported. The 12 chains in the box on the lower left are all sequentially similar to each other. Knotted chains that become unknotted when both ends are trimmed by five residues are grouped in the shaded lower left corner. The shaded box above it contains chains that become unknotted when the ends are trimmed by ten residues. Underlined PDB codes have been previously reported as knotted in the articles describing the experimental structure determination (Badger *et al*., 2005; Lim *et al*., 2003; Elkins *et al*., 2003; Komoto *et al*., 2004; Ahn *et al*., 2003; Nureki *et al*., 2004; Saito *et al*., 2004; Mosbacher *et al*., 2005; Tyagi *et al*., 2005; Pleshe *et al*., 2005; Wagner *et al*., 2005). Articles describing experimental structure determination have not yet been published for 1lug, 1ns5, 1to0 or 1v6z. PDB codes in italics indicate proteins reported by Taylor as being knotted (1cmxA, 1dmxA, 1fugA, 1hcb, 1kopA, 1yveI, 1zncA, 2btvB in Taylor, 2000) (1ipaA, 1k3r, 1qmgA in Taylor *et al*., 2003a) and 1gz0 (Taylor *et al*., 2003b). Note that (Taylor *et al*., 2003a) additionally reports six ''accession numbers for knotted proteins'' that were not found to be knotted here either by the Knotfind algorithm or by visual inspection. One case, 1g0z, is likely a typographical error for 1gz0, which is later reported as knotted in (Taylor *et al*., 2003b). The other five proteins, 1mt6, 1mvh, 1h3i, 1ml9, and 1mlv were later reported to not contain true knots according to the algorithm of Taylor in (Taylor *et al*., 2003b).

3.2 Knots in protein structures

The Knotfind algorithm was initially applied to protein structures in the RCSB PDB. Twenty-one deeply knotted proteins were found in the collection of 9,553 protein chains taken from the PDB (∼0.2%). In addition, eighteen proteins were identified to contain shallow knots which disappear after trimming five to ten residues from the termini. Of the twenty-one deeply knotted proteins detected,

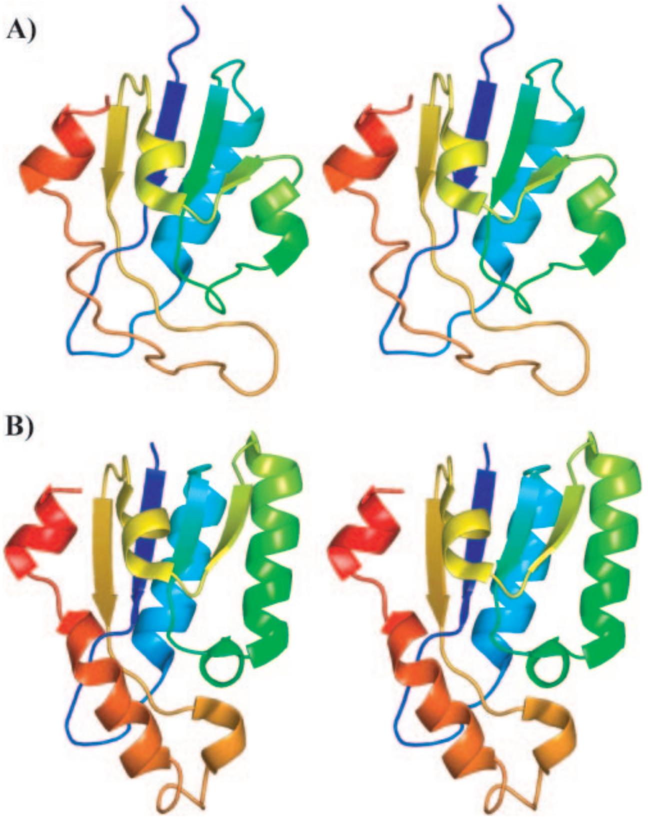

Fig. 5. Stereo view of a previously unreported deep trefoil knot in acetylornithine transcarbamylase. (A) Residues 165-266 of 1js1 chain X (324 residues total) contain a deep trefoil knot where the loop between the yellow strand and red helix threads through the loop comprised of the blue strand and cyan helix. (B) Residues 171-285 of 1yh1 chain A (336 residues total) also contain a deep trefoil knot where the loop between the yellow strand and orange helix threads through the blue loop.

most have been previously reported, or a knot in a protein with sequence or structural similarity has been previously reported. Novel deep trefoil knots were detected, however, in two acetylornithine transcarbamylases, 1js1, and 1yh0/1yh1, which are similar in sequence and structure to each other. These proteins are not sequentially or structurally similar to any other previously reported knotted proteins (Figure 4). The trefoil knot in 1js1 and 1yh1 is shown in stereo in Figure 5.

Interestingly, the knot in aceytlornithine transcarbamylase is found in the acetylornithine binding domain, where two loops, a proline rich loop and the 240s loop, appear to be threaded through one another (residues 173-183 and 236-259 respectively in 1js1 (Shi *et al*., 2002); residues 177-188 and 252-278 respectively in 1yh0/1yh1 (Shi *et al*., 2005)). These two loops are presumably responsible for specificity for acetylornithine relative to the unacetylated substrates preferred by the structurally similar, but unknotted, enzymes ornithine transcarbamylase (36% sequence similarity) and aspartate transcarbamylase (40% sequence similarity). The 240s loop in acetylornithine transcarbamylase lacks the essential binding motifs found in the ornithine and aspartate transcarbamylases. Shi *et al*., (2002) hypothesize that the conformational rigidity of the proline-rich loop, which contains four prolines not found in ornithine transcarbamylase, may be responsible for excluding ornithine from the active site by preventing movement of the 240s loop towards the active site.

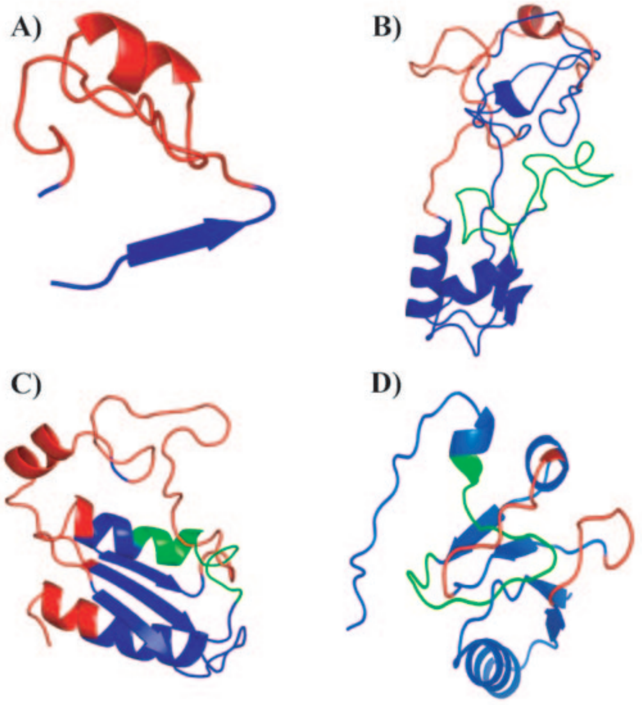

Fig. 6. Examples of knots observed in Rosetta decoys. (A) A Type 1 SVR knot from a T195 CASP 5 decoy (only residues 181-217 are shown), where a knot is entirely localized within a single SVR (residues 188-215, red). The local template structure is shown in blue. (B) A Type 2 SVR knot from T261 (only residues 58-207 are shown), where an SVR (residues 164-189, shown in green) threads through a template region (blue). This model was submitted as Robetta's top ranked model for the target. (C) A Type 3 SVR knot from a T195 CASP 5 decoy (only residues 181-299 are shown), where two SVRs thread through one another. The SVR comprising residues 188-215 in shown in green while the SVR spanning residues 242-253 is shown in red. Template regions are shown in blue. (D) T202 model 1 submitted by Robetta for CASP 6 (only residues 1-101 are shown). An SVR (residues 69-85, shown in green) threads through both the template (blue) and through another SVR (residues 49-56 in red), making this both a Type 2 and Type 3 SVR knot.

3.3 Knots in Rosetta decoys

Approximately 5% of the CASP 5 decoys were found to have knots (2,163/45,366). During the course of CASP 5, a high frequency of occurrence of knots had been observed for certain targets, requiring a significant effort in manual inspection to discard those models containing knots (Rohl *et al*., 2004a). This non-uniform distribution of knotted decoys was confirmed, as some targets showed a high percentage of knotted conformations, while others had virtually none (Figure 7A).

To gain a better understanding of the origin of knots in CASP 5 decoys, we also manually inspected the 291 knotted decoys for target T195 which showed the highest frequency of knot formation. SVRs judged to be responsible for knot formation fell into one of three different categories (Figure 6): 1) a single SVR contained a knot that was entirely localized to this SVR (4 examples) 2) a SVR

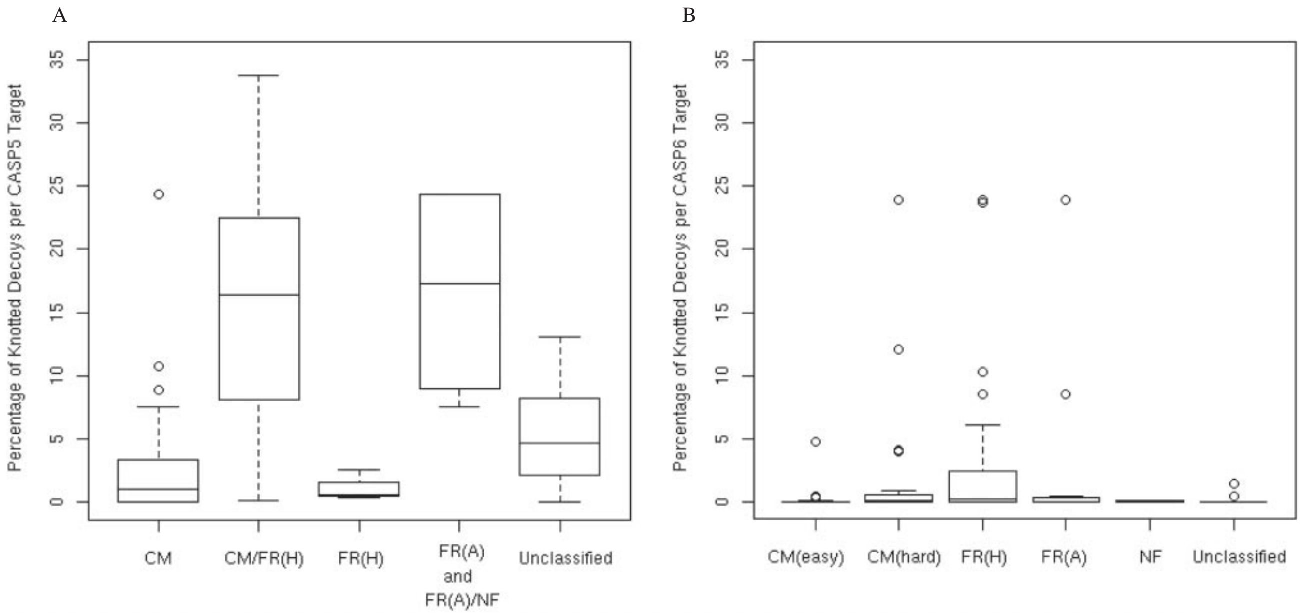

Fig. 7. Frequency of knotted decoys using Rosetta for targets in CASP. The frequencies of knots in decoys sets for A) CASP 5 and B) CASP 6 targets are shown as boxplots. Targets have been binned by difficulty using the assignments defined by CASP assessors (Kinch *et al*., 2003; Tress *et al*., 2005b). In cases where multiple domains of one CASP target have different classifications, the decoy set for the target is included in each classification. Categories, in order of generally increasing difficulty, are comparative modeling (CM); fold recognition, homologous (FR(H)); fold recognition, analogous (FR(A)); and new fold (NF).

threaded through a template region (157 examples), and 3) two SVRs wrapped around one another (138 pairs). In this analysis, SVRs of three residues or less were treated as part of the fixed template due to the fact that their conformation is highly constrained by the geometric constraints imposed by the template.

Additionally, we analyzed CASP 6 models submitted by the automated Robetta server in which the methods used in CASP 5 Rosetta CM predictions were implemented (Bradley *et al*., 2003). Robetta submitted three models in CASP 6 that contained knots which it ranked as its best predictions for T202, T203, and T261 (Figure 6). Seven additional Robetta models which were ranked below the top model also contained knots (T199 Model 3; T202 Model 5; T211 Model 2; T208 Model 2; T235, Domain 1, Model 2 and 4; T261 Model 2). Visual analysis of these structures was consistent with the hypothesis that knot formation was related to SVR modeling, and all knots in Robetta models could be classified as Type 2 and Type 3 as defined above.

Based on our analysis of CASP 5 decoys, we modified our SVR modeling procedure in an attempt to reduce the frequency of knot formation in CASP 6 decoy sets. Libraries of conformations were generated for each SVR and these libraries were screened to eliminate any conformations that resulted in knots when grafted onto the template structure in the absence of all other SVRs. Applying this procedure to T195, we found that pre-filtering the conformational libraries reduced the frequency of knots from 25% (971/3,966) to approximately 20% (745/3,737) by detecting all Type 1 and Type 2 SVRs. Additionally we applied this modified protocol in CASP 6 predictions and found decrease in the overall frequency of knotted decoys (∼1%, 1,343/119,543) relative to that observed for CASP 5 decoys (5%). Knot frequencies in individual decoy sets are shown in Figure 7.

3.4 Knots in Undertaker CASP 6 decoys

In order to assess the extent to which knot formation is specific to the modeling strategy used by Rosetta, we also examined decoy sets generated by the SAM-TO4 group (Group 166) method for CASP 6 targets using the Undertaker program (Karplus *et al*., 2005). In CASP 6, most decoys created by the Undertaker program were knot-free, but decoys sets for a few targets had a high frequency of knots. The highest occurrence of knots in Undertaker decoy sets was found for T228 (12% of decoys knotted), T237 (9% knotted) and T218 (8%). On these same targets, the Rosetta decoy sets had knot frequencies of 10%, 6%, and 0%, respectively. In general, however, there was little or no correlation between the knot frequency in Undertaker decoys and Rosetta decoys across the CASP 6 targets (unpublished data). The Undertaker method does not explicitly model regions of the backbone as either part of a template or a SVR. However, it is similar to Rosetta in that it introduces chain breaks at points corresponding to gaps in the alignment. Visual inspection of Undertaker decoys indicated that the majority of the knots in Undertaker decoys could be explained by threading that occurred while resolving gaps in the backbone or when merging two domains that were modeled separately (K.Karplus, personal communication).

4 DISCUSSION

4.1 Efficacy of the knotfind algorithm

An algorithm for knot detection has been previously described by Taylor (2000) and applied to detect knots in protein structures in the PDB. Taylor's algorithm progressively smoothes the protein backbone: at each iteration, each atom in the backbone is moved incrementally toward the midpoint of the line segment formed by the N- and C-terminally adjacent atoms, subject to a clash check that ensures the protein backbone does not pass through itself. Knotfind shares the basic approach of trying to straighten out the protein chain, but does so in a stepwise fashion that avoids the need to compute new atom positions and enables the algorithm to converge rapidly.

An additional benefit of not modifying atomic coordinates during the course of the algorithm is that when a knot is detected in a protein chain, the knot can be localized in the structure without the need to interpret a smoothed or distorted chain trace. In cases where chains cannot be completely simplified to an extended segment, the coordinates of the remaining Cα atoms can be used to facilitate the visual identification and analysis of the knot.

One caveat with the Knotfind algorithm is that its performance with respect to false positives and false negatives has not been rigorously proven. The triplet-ordering schemes used here are selected to attempt to minimize the possibility of false positives by first simplifying local backbone features. Notably, triplet-ordering schemes that do not target local backbone features preferentially over global features tend to result in higher occurrence of false positives. For example, considering triples from N- to C-terminal order results in seven false positives (1e2kA, 1e2wA, 1k7hA, 1ohfA, 1p6xA, 1y6vA, 2a65A). Combining two triplet ordering schemes eliminates all false positives in the set of PDB chains examined here, suggesting that false positives, while possible, are likely to be rare.

4.2 Application of knotfind to Rosetta homology-based structure prediction

The detailed analysis of knotted Target 195 decoys suggests that three sources of knots can be generated by Rosetta's comparative modeling approach: SVRs that knot with themselves, SVRs that thread themselves with the template, and pairs of SVRs that thread through each other. The first type of knot is likely introduced by the high gap penalty used to ensure chain continuity. In our experience, the introduction of such knots is rare, perhaps not surprisingly as significant steric clashes generally accompany such knots. Reductions in the gap penalty accompanied by more efficient methods of loop closure, such as the cyclic coordinate descent method (Canutescu and Dunbrack, 2003) can be used to reduce the likelihood of introducing such knots during the modeling process.

Knots of Type 2 and Type 3 are not localized to a single region of the backbone, but instead are attributed to one section of the chain threading through another. In the Rosetta-based method, such knots can be introduced into models because SVR conformations are selected from databases or are initially modeled only in the context of local stem geometry. When such conformations are combined with a fixed template structure, or with models for other SVR regions, threadings can occur which are difficult or impossible to resolve. To reduce the occurrence of such knots, we filtered libraries of SVR conformations during the generation of CASP 6 targets in order to eliminate conformations that were threaded through the fixed template structure and observed a significant reduction in knotted percentage. Interestingly, in some cases such filtering could also be used to guide alignment choice. For example, if all or nearly all conformations for a particular SVR, selected on the basis of fitting the geometric restraints imposed by the template, result in a knot, the original alignment to the parent structure is likely incorrect as it implies structurally unfeasible gaps.

While the frequency of knots was significantly reduced by filtering with the Knotfind algorithm in CASP 5 compared to CASP 6, decoy sets for some CASP 6 targets still show significant occurrence of knots. Since this filtering step only considered single SVRs in the context of the fixed template, knots that are introduced by pairs of SVRs threading through one another (Type 3), are not detected and are expected to still occur in CASP 6 decoy sets. For CASP 6 predictions, these knotted decoys were eliminated from the final decoy population in the model selection process using the Knotfind algorithm. Such knots however, could be eliminated earlier in the modeling process by pairwise examination of SVR conformations in the libraries, or by checking complete models early in the optimization process.

4.3 General application to structure prediction

The Knotfind algorithm can be applied to the benefit of many structure prediction approaches. The most obvious application of the Knotfind algorithm is the screening of final models to ensure that a knotted decoy is not selected. Such screening is particularly important in an automated method such as Robetta where an expert does generally not examine final predictions manually. The speed of the Knotfind algorithm makes it appropriate not just for post-filtering decoy populations to eliminate knotted structures, but also for application during the protein structure prediction process, either as a filter as described here or as part of a scoring scheme used during optimization. For example, Knotfind is now implemented in Undertaker as a cost function that is only used when the potential for knots is high as determined by an expert predictor.

The causes of high knot frequency in some modeling problems is likely to be specific to the particular method used and the structural details of the protein being modeled. On the basis of comparison of knot formation in Rosetta and Undertaker decoys, it seems likely that the introduction of chain breaks during the modeling process is a contributing factor to increased probability of knot formation. Additionally, the location of such chain breaks and the size of the gap introduced at each discontinuity are likely to be important factors as well. This conclusion suggests that a knot detection algorithm is likely not only to be applicable to homology based methods that must model gaps implied by alignments, but in any protein modeling method that introduces chain breaks during the modeling process, including for example *de novo* prediction methods that have recently been demonstrated to be capable of prediction accuracies better than 1Å for small proteins (Bradley *et al*., 2005).

ACKNOWLEDGEMENTS

We thank Josue Samayoa, David Bernick, and Craig Lowe for the decoy structures predicted for CASP 6, Dylan Chivian and David Baker for CASP 5 decoy sets, and Kevin Karplus for Undertaker decoy sets and discussion.

REFERENCES

Ahn,H.J. *et al*. (2003) Crystal structure of tRNA(m(1)G37)methyltransferase: insights into tRNA recognition. *EMBO J*., **11**, 2593–603.

Badger,J. *et al*. (2005) Structural analysis of a set of proteins resulting from a bacterial genomics project. *Proteins*, **60**, 787–796.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.

Bradley,P. *et al*. (2003) Rosetta predictions in CASP 5: successes, failures, and prospects for complete automation. *Proteins*, **53** (Suppl 6), 457–468.

Bradley,P. *et al*. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **5742**, 1868–71.

Canutescu,A.A. and Dunbrack,R.L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci*., **5**, 963–72.

Chivian,D. *et al*. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53**, 524–533.

Chivian,D. *et al*. (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins*, **61** (Suppl 7), 183–92.

Elkins,P.A. *et al*. (2003) Insights into Catalysis by a Knotted TrmD tRNA Methyltransferase. *J. Mol. Biol*., **333**, 931–949.

Karplus,K. *et al*. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins*, **61** (Suppl 7), 135–142.

Kim,D.E. *et al*. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*., **55**, 656–677.

Kinch,L.N. *et al*. (2003) CASP5 target classification. *Proteins*, **53** (Suppl 6), 340–51.

Komoto,J. *et al*. (2004) Crystal structure of the S-adenosylmethionine synthetase ternary complex: a novel catalytic mechanism of s-adenosylmethionine synthesis from ATP and MET. *Biochemistry*, **43**, 1821–1831.

Lim,K. *et al*. (2003) Structure of the YibK methyltransferase from Haemophilus influenzae (HI0766): a Cofactor Bound at a Site Formed by a Knot. *Proteins*, **51**, 56–67.

Mansfield,M.L. (1994) Are there knots in proteins. *Nat Struct Biol*., **1**, 213–214.

Mosbacher,T.G. *et al*. (2005) Structure and function of the antibiotic resistance-mediating methyltransferase AviRb from Streptomyces viridochromogenes. *J Mol Biol*., **3**, 535–45.

Moult,J. *et al*. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–v.

Nureki,O. *et al*. (2002) An enzyme with a deep trefoil knot for the active-site architecture. *Acta Crystallogr D Biol Crystallogr*., **58**, 1129–1137.

Nureki,O. *et al*. (2004) Deep Knot Structure for Construction of Active Site and Cofactor Binding Site of tRNA Modification Enzyme. *Structure*, **4**, 593–602.

Pleshe,E. *et al*. (2005) Structure of a class II TrmH tRNA-modifying enzyme from Aquifex aeolicus. *Acta Crystallograph Sect F Struct Biol Cryst Commun*., **61**, 722–728.

Richardson,J.S. (1997) Beta-Sheet topology and the relatedness of proteins. *Nature*, **268**, 495–500.

Rohl,C.A. *et al*. (2004a) Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins*, **55**, 656–677.

Rohl,C.A. *et al*. (2004b) Protein structure prediction using Rosetta. *Methods Enzymol*., **383**, 66–93.

Saito,R. *et al*. (2004) Structure of bovine carbonic anhydrase II at 1.95 A resolution. *Acta Crystallogr D Biol Crystallogr*., **60**, 792–5.

Shi,D. *et al*. (2002) Crystal structure of a transcarbamylase-like protein from the anaerobic bacterium Bacteroides fragilis at 2.0 A resolution. *JMB*, **320**, 899–908.

Shi,D. *et al*. (2005) Crystal Structure of *N*-Acetylornithine Transcarbamylase from *Xanthomonas campestris*. *J. Biol. Chem*., **280**, 14366–14369.

Taylor,W.R. *et al*. (2000) A deeply knotted protein and how it might fold. *Nature*, **406**, 916–919.

Taylor,W.R. *et al*. (2003a) Protein knots: A tangled problem. *Nature*, **421**, 25.

Taylor,W.R. *et al*. (2003b) A knot or not a knot? SETting the record 'straight' on proteins. *Comput Biol Chem*., **27**, 11–15.

Tramontano,A. *et al*. (2001) Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, **45** (Suppl 5), 22-38.

Tress,M. *et al*. (2005a) Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins*, **61** (Suppl 7), 27–45.

Tress,M. *et al*. (2005b) Domain definition and target classification for CASP6. *Proteins*, **61** (Suppl 7), 8–18.

Wang,G. *et al*. (2003) Jr. PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Tyagi,R. *et al*. (2005) The crystal structure of a bacterial class II ketol-acid reductoisomerase: domain conservation and evolution. *Protein Sci*., **14**, 3089–3100.

Wagner,J.R. *et al*. (2005) A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. *Nature*, **438**, 325–331.