Protein Structure Geometry: Knot or Not? An Algorithm for Detecting Knots in Proteins

Bardia Ghayoumi, Mira Mastores, Abby Saenz

Winter 2018

1 Introduction

Continued improvements in the technology to elucidate protein structure and folding have led to the discovery that some proteins contain knots as part of their native $3D$ conformation. An open chain consisting of anywhere from twenty to several thousands of amino acids, proteins achieve their unique, native 3D folded structure, denoted as the protein's tertiary structure, via folding caused by weak intermolecular forces. These structures are deeply complex, thus it is not difficult to imagine the polypeptide chains becoming entangled on the path to native conformation. Due to the complexity of protein structure, identifying knotted regions of a chain demanded the development of computational methods. Considered to be one of the hardest problems of the modern age [1], protein folding provides key insights into the behavior of any living system. Proteins are the minuscule biological machines that allow for a living system to persist. Inextricably linked to function, if a protein's structure can be determined, its chemical properties are available to be analyzed. Unfortunately, there is a dearth of tertiary structures compared to the number of protein sequences available online, due to the great difficulty and expense of X-Ray Crystallography, a pivotal technique in structural biology. Still considered relatively rare, knotted proteins account for about 1% of all known protein structures in the Protein Data Bank (PDB), yet more may be discovered as new structures arise and the algorithms for detecting them improve in accuracy. [3]

Knotted peptide chains occur less frequently than random chance would predict, indicating that nature has developed a mechanism to specifically avoid knots. [5] Indeed, knotted proteins take longer to fold and thus appear to be evolutionarily disadvantageous. Despite their rarity, evolution has chosen to conserve specific knotting patterns across species, indicating that knotted proteins must offer some functional advantage to offset the less efficient folding. A recent paper by Tumanski [6] examined three deeply knotted proteins with widely differing structures and biological activity, in an attempt to identify how knots contribute to their function. They noted that the amino acids around the knot participated in an increased number of chemical interactions. Those bordering the knot displayed higher kinetic stability, as defined by the relative rate of unfolding, as well as decreased solute accessibility. Interestingly, deep knots, located far away from either end of a protein, have little effect on the relative rate of denaturation. [1] Other researchers have noted that in many cases, knots are found near the active sites of enzymes, leading to the assumption that knots could lend important stability to the binding of substrates in catalysis. [5] One of the most well studied examples of this is the RNA methyltransferase from thermophilic bacteria, whose active site requires a knot for optimal function, and whose extreme environment would likely select for more stable proteins. [7]

In order to computationally discover knots in proteins, it is important to distinguish how knotted proteins differ from what is meant by a true, topological knot. For our purposes, a knotted protein is defined as the amino acid backbone passing through a self formed loop; an apt analogy would be to consider the protein as a piece of string. Distinct from these are cysteine knots, which occur after protein folding, when one disulfide bond passes through a loop created by two others.[10] Also not considered topological knots, slipknots occur when a protein chain forms a knot during folding, but bends back on itself to effectively untie the knot when the entire structure is considered. Slipknots may not show up when subjected to computational analysis if the algorithms written to detect them do not adjust for them by examining the protein structure in parts, rather than as a whole. [8]

A mathematician's knot differs from this conventional idea of a knot. Mathematically, a knot is the embedding of a circle in three-dimensional Euclidean Space, R3, that is preserved under continuous deformations called isotopies. These true, topological knots are defined by having a property of being closed - that is, their "ends" are connected and cannot be untangled to produce a simple loop without being torn.[1] Physical properties such as the length or the thickness of the knot are not considered under this definition, although they may contain biological significance. Given this definition, a knot only makes sense for a closed loop, and the unjoined termini on the ends of a protein chain simply do not fit the criteria. However, several methods have been proposed to create an artificial, closed loop, and our algorithm relies heavily on this caveat.

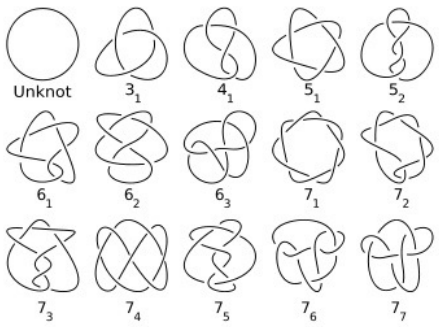


Figure 1: Topological representation of knots.

One method takes advantage of the fact that because they are charged and hydrophilic, the termini of most protein chains are found at the protein surface. This method begins by joining the ends of the peptide chain to form a closed loop, allowing the protein to fit the formulas defined by mathematical knot theory. One such algorithm used by Grosberg [5] and Millet [2] involves closing the loop by encasing the protein in a sphere, then connecting the surface hugging termini to a point on the sphere. The knot can thereafter be subjected to the well established mathematical operations characterizing knots to identify knot type. [2]

An issue with physically closing the loop occurs for proteins whose termini are not shallow, or are tangled up in the knot themselves. An algorithm which closes the loop may therefore unravel the knot or create a new one. For this reason, Taylor devised the second method, in which the termini of the protein are held fixed in space, and the protein chain is repetitively straightened. The parts of the chain which are blocked from becoming smoothed must contain a knot. [4] Holding the ends fixed allows us to treat the protein chain as if it were a loop. In this paper we will describe how we have attempted to recreate Taylor's chain-smoothing algorithm and reproduce his results.

In the field of protein modelling, lattice models depict proteins in their simplest form, as each amino acid represented by a bead connected through peptide bonds, represented as sticks connecting

the beads on a 3D cube lattice. An off lattice model generalizes the lattice so that each amino acid is related through a series of tangent spheres. [12] To visualize the protein structure before and after running it through our algorithm, we utilized a $C\alpha$ off lattice model, with each bead representing the $C\alpha$ carbon of that amino acid. The positions of each amino acid are determined experimentally by X-ray Crystallography or NMR spectroscopy, and a database of protein structures (The Protein Data Bank, or PDB) exists to store all currently known structural information about a given protein.

We selected five proteins of varying residue length whose structures were already known to either contain or not contain a knot, and used their amino acid sequence positions available on the PDB as test cases for our algorithm. Of the proteins without knots, 5pti (Bovine pancreatic trypsin inhibitor) has 58 residues and 1timA (triose phosphate isomerase) has 247 residues. Of the knotted proteins, 1yveI (acetohydroxy acid isomeroreductase) has 513 residues, 2efv (Hypothetical Protein(MJ0366) from Methanocaldococcus jannaschii) has 82 residues, and 3bjx (haloacid dehalogenase) has 303 residues.

2 Methods

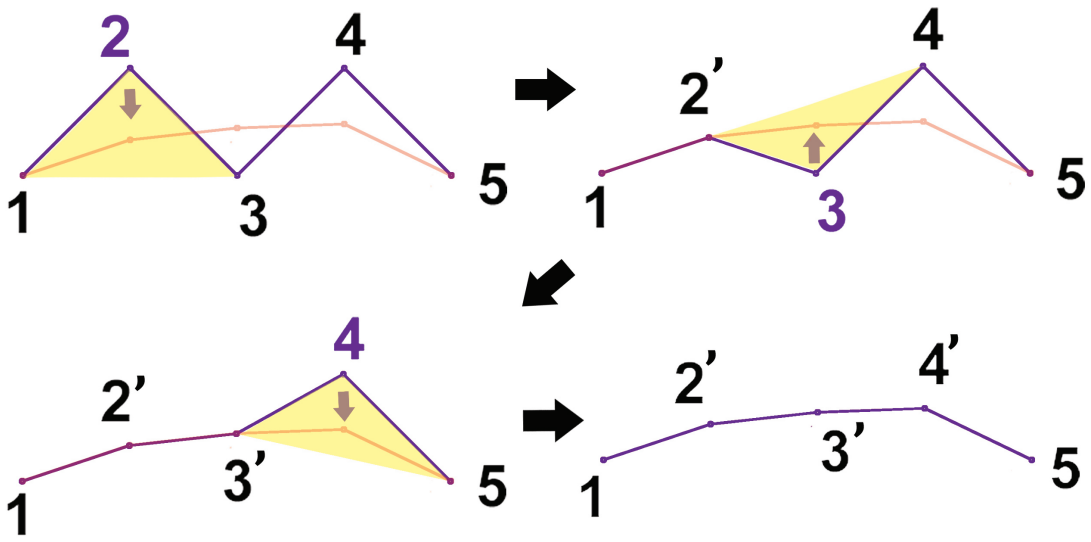


Figure 2: Visualization of one iteration of the flattening algorithm. Colored points symbolize point B of the yellow triangle ABC. Each iteration, B is moved to the centroid of ABC, B', if it is not blocked.

- A triangle ($\triangle ABC$) was made out of every triplet of three amino acids in a protein
- For every triangle, if $\triangle ABC$ is not flat, we attempt to move point B to B', which is the average of all three vertices
    - Flatness is characterized by the distance from point B to the line segment AC. Due to floating point arithmetic, a user defined threshold must be specified in order to define the strictness of a flat triangle. If $\angle ACB$ or $\angle CAB > 90°$, this distance is the length of line segment AB or CB, depending on which side point B is located. Otherwise, the distance is given by the magnitude of $(\vec{AB} \times \vec{AC})$ divided by $|\vec{AC}|$, which is the area of a parallelogram solving for height.
- Two sub-triangles are made: $\triangle AB'B$ and $\triangle CB'B$
- For every amino acid pair (labeled as DE) in the protein sequence, check to see if any form a line segment intersecting either sub-triangle.
    - In order to determine if a line segment intersects a triangle, the point projected from the parametric line formed by DE onto the plane defined by $\triangle ABC$ is found by solving the plane equation for the parametric parameter t. If $0 < t < 1$, then the line segment itself intersect the plane ABC at point P. To determine if P is inside $\triangle ABC$, point P

is defined relative to point A as: $P = A + \alpha(\vec{AB}) + \beta(\vec{AC})$. If α and β are both positive and sum to less than 1, then P is said to be within $\triangle ABC$. [11]

- If any DE are blocking $\triangle ABC$, increment $N_{could\_have\_moved}$
- If no DE are blocking $\triangle ABC$, update B to B' and increment N_{moved}
- Repeat while the termination condition is not met
    - The algorithm is complete once $N_{moved} = 0$. This indicates that an iteration was done where no new moves were made.
    - If $N_{could\_have\_moved} = 0$, then that means that all triangles are sufficiently flat with no blocks, indicating the protein has no knot present
    - If $N_{could\_have\_moved} \neq 0$, then there is a blockage in the protein, suggesting a knot might be present.

3 Results

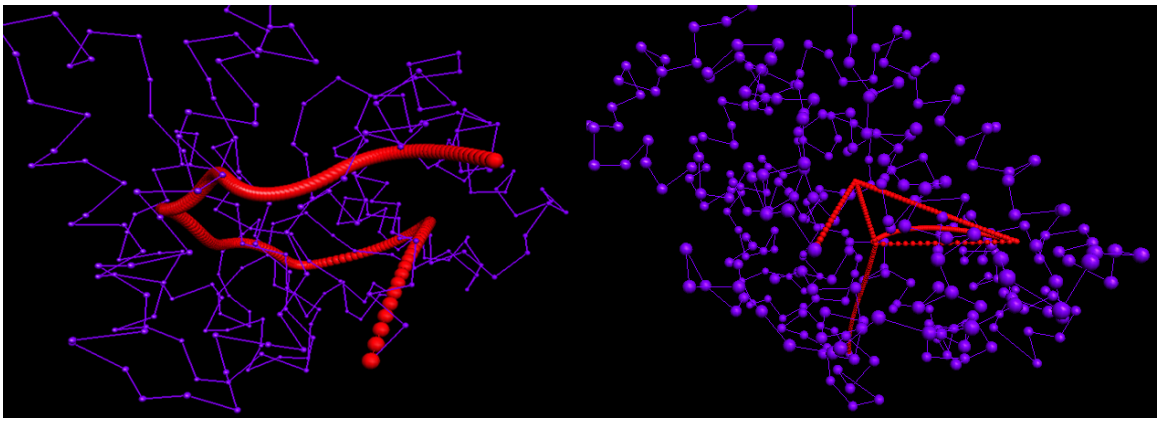


Figure 3: Chain Smoothing Algorithm output for: [Left] 1timA (unknotted), [Right] 3bjx (knot present).

Our tests on the proteins 1timA, 1yvel, 2efv, 3bjx, 5pti suggest that our algorithm is sufficient for detecting knots in proteins, with some limitations. Utilizing a 3D visualization software, we compared the original structure of our protein with the smoothened chain as outputted by our program. This 3D graphic provided us with confirmation that our algorithm was both straightening the chain and detecting knots correctly. (Figure 3)

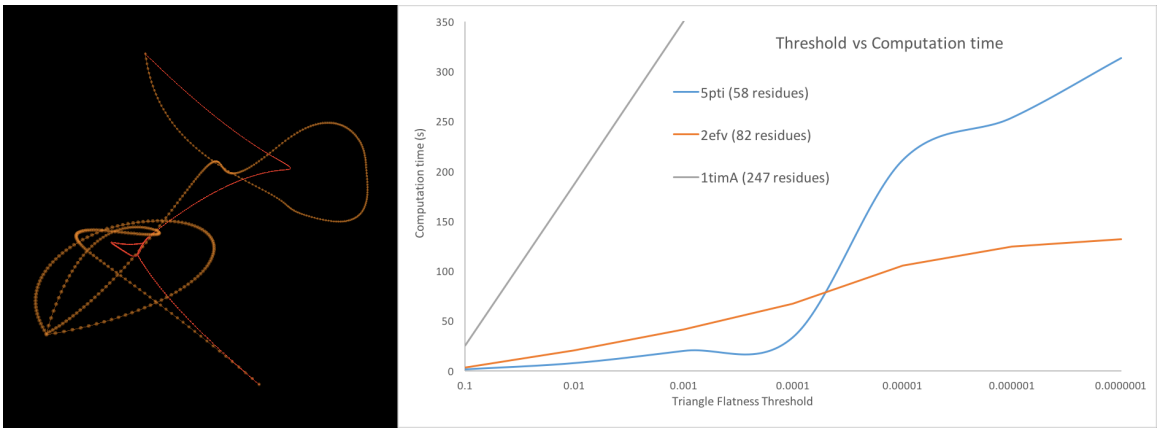
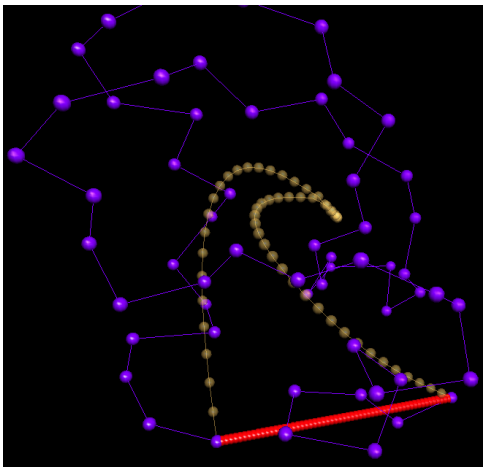
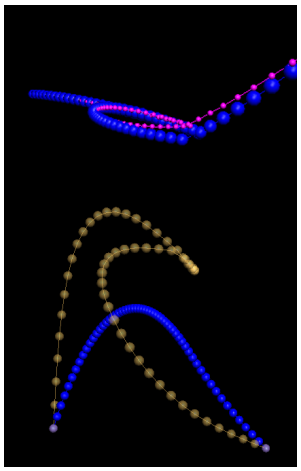


Figure 4: (Left) A threshold of 0.01Å (red) compared to 0.001Å (orange) for 1 yvel (Right) Threshold vs. Computation Time curves for 5pti, 2efv, and 1timA. The curves for the rate of change of computation time differs nonlinearly for each protein, implying that the time complexity of our algorithm is due to unforseen factors

A threshold of 0.01Å was sufficient for the proteins 1timA, 1yvel, 2efv, and 3bjx. In these cases, applying a stricter, more precise threshold of 0.001Å improved the quality of the algorithm output - as shown in the figure below, more of the chain becomes straightened and the area of the chain involved in the block (knot) is decreased. For all proteins, increasing the threshold level does increase computation time, however the rate of increase in computation time also depends upon the length of the protein chain and complexity of the structure. Because it relies on so many different factors, the



(a) 5pti native structure (purple) compared to a threshold of 0.01Å (yellow) and 0.00001Å (red). Note elimination of the tangle.



(b) [Top] 2efv computed N to C (pink) and reversed (blue) [Bottom] 5pti computed N to C (orange) and reversed (blue) Note the elimination of the "tangle"

time complexity of our program is difficult to predict which may be considered a practical drawback.

For the protein 5pti, an issue appeared in which our algorithm failed to fully straighten the chain at one location, resulting in a kink or bend in the visualization of the chain smoothing (see Figure 5(a)). Taylor noted this issue in his algorithm as well, citing it as a tangle in the protein chain [4]. From the perspective of the algorithm, several attempted moves became gridlocked and blocked each other. This issue was resolved by applying a threshold level of 0.00001Å, though this dramatically increased execution time. (see Figure 3).

Throughout the current literature on knotted proteins, one of the biggest criticisms of the Taylor chain smoothing method is that the output changes depending on which terminal, N or C, is fed through the algorithm first [2][5]. To see whether this problem persisted in our program, we ran the sequences of two proteins, one with a knot (5pti) and one without a knot (2efv), through our program in reverse with a threshold of 0.01Å. Indeed, the output for both algorithms was different. In the case of 5pti, the tangled issue completely disappeared in the reversed sequence. For 2efv, the output of the reversed sequence was only slightly off from the original, yet any difference between termini is significant. A discrepancy in results based off the direction that the algorithm operates makes our program less trustworthy, as it could cause a knot to be missed or falsely identified.

4 Further Discussion

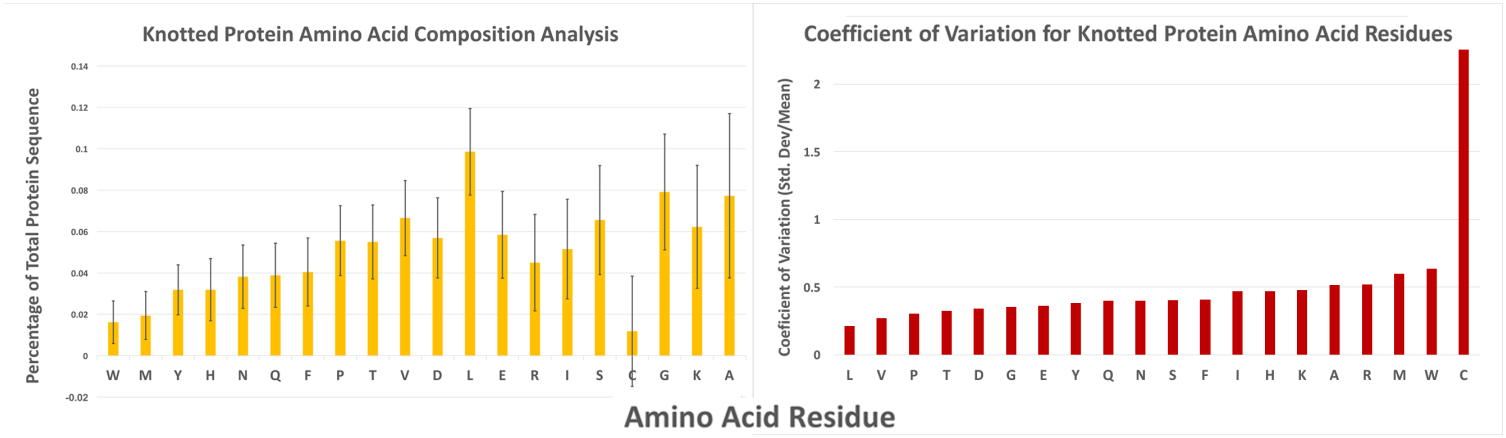


Figure 6: Statistical Analysis of the PDB for Knotted Proteins

One limitation of our program is that it ignores the biochemical interactions between the amino acids that form the protein structure by treating all residues as equal, and simplifying the complex structure to a set of 3D coordinates. In an attempt to deepen our understanding of the biological interactions that govern knotted proteins, we conducted a statistical analysis of the residue types involved in knotted proteins, to try and find if certain residues are conserved between knotted proteins.

From the PDB, we gathered the FASTA sequence of all proteins with known knots and calculated the relative percentage of each amino acid composing the sequence. We plotted the average percentage of each amino acid, normalized to sequence length, along with the standard deviation. We computed the coefficient of variation, or percentage of the mean the standard deviation was.

Our data suggests that L, V and P are the most conserved amino acids found in knotted proteins, because they have the lowest degree of variability from their mean. These three amino acids are all hydrophobic, and thus would comprise the interior of a protein, shielded from water. A possible explanation for this conservation is that these amino acids make up the core foundation for the protein structure. Notably, cysteine has an incredibly high coefficient of variation compared to the rest of the residues despite having a lower mean, which follows from its role in protein structure: forming disulfide bonds, which are covalent bonds between two

It would be more informative to run this statistical analysis on local knot interactions, rather than the global sequence of the protein. This would indicate whether specific amino acids are more likely to be involved in the actual formation of knots, which may lend important insight into the mechanisms behind protein folding.

Though several computational methods for detecting knots in proteins have been proposed, there are still issues surrounding these methods. Nothing has been proposed that allows for automatic detection without direct intervention. Problems arise due to breaks in the chain or the absence of residues. Not the mention the computational complexity; our algorithm rapidly rises in computation time, on the scale of days once the number of residues exceeds 300 or so.

As mentioned earlier, the results of the algorithm must be met with healthy skepticism, and only after visual inspection of the output by the user should any conclusions be drawn, which becomes cumbersome at longer sequence lengths, due to high variability in output based on sequence length

or direction of input to the algorithm.

The issue of protein structure let alone detecting knots in proteins is inherently qualitative in nature. Describing a 3D structure precisely using any language, be it math or English, is no trivial task. These are both human inventions created in order to better communicate with one another; none can ever replace the incredible efficiency of our visual processing center. Research regarding computer vision is in high demand, and one day could surpass the capabilities of human eyesight. If a computer could be given the coordinates of amino acids in space, and just by "looking" at the protein, *i.e.* skipping the sequential flattening, this would drastically improve computation time and accuracy.

5 References

1. Faísca, Patrícia F.N. "Knotted Proteins: A Tangled Tale of Structural Biology." Computational and Structural Biotechnology Journal 13 (2015): 459–468. PMC. Web. 27 Feb. 2018.

2. Millet, Kenneth, Dobay, Akos and Stasiak. Andrzej. "Linear Random Knots and Their Scaling Behavior." Macromolecules 2005 38 (2), 601-606. DOI: 10.1021/ma048779a

3. S.E. Jackson, A. Suma, C. Micheletti. "How to fold intricately: using theory and experiments to unravel the properties of knotted proteins." Curr. Opin. Struct. Biol., 42 (2016), p. 6

4. Taylor, W. R. "A deeply knotted protein structure and how it might fold." Nature 406, 916–919 (2000).

5. R.C. Lua, A.Y. Grosberg. "Statistics of knots, geometry of conformations, and evolution of proteins."PLoS Comput Biol, 2 (2006), p. e45

6. Dabrowski-Tumanski, Pawel, Andrzej Stasiak, and Joanna I. Sulkowska. "In Search of Functional Advantages of Knots in Proteins." Ed. Eugene A. Permyakov. PLoS ONE 11.11 (2016): e0165986. PMC. Web. 27 Feb. 2018.

7. Nureki, O.; Shirouzu, M.; Hashimoto, K.; Ishitani, R.; Terada, T.; Tamakoshi, M.; Oshima, T.; Chijimatsu, M.; Takio, K.; Vassylyev, D.G.; et al. "An enzyme with a deep trefoil knot for the active-site architecture." Acta Cryst. Section D, Biol. Cryst. 2002, D58, 1129–1137.

8. N.P. King, E.O. Yeates, T.O. Yeates. "Identification of rare slipknots in proteins and their implications for stability and folding" J Mol Biol, 373 (2007), pp. 153-166

9. W.R. Taylor "Protein knots and fold complexity: some new twists." Comput Biol Chem, 31 (2007), pp. 151-162

10. Vitt UA, Hsu SY, Hsueh AJ. "Evolution and classification of cystine knot-containing hormones and related extracellular signaling molecules." Mol Endocrinol 2001; 15: 681-694.

11. "Point in Triangle Test." Point in Triangle Test, blackpawn.com/texts/pointinpoly/.

12. Hart WE, Istrail S: "Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than eighty-sex percent of optimal." J Comput Biol. 1997, 4 (3): 241-259.