CHAPTER 1

TYING DOWN OPEN KNOTS: A STATISTICAL METHOD FOR IDENTIFYING OPEN KNOTS WITH APPLICATIONS TO PROTEINS

Kenneth C. Millett
Department of Mathematics, University of California, Santa Barbara
Santa Barbara, CA 93016, USA
Email: millett@math.ucsb.edu
Homepage: http://www.math.ucsb.edu/˜millett/KM.html

Benjamin M. Sheldon
Department of Mathematics, University of California, Santa Barbara
Santa Barbara, CA 93016, USA
Email: bensheldon@umail.ucsb.edu

A mathematical knot is simply a closed curve in three-space. Classifying open knots, or knots that have not been closed, is a relatively unexplored area of knot theory. In this note, we report on our study of open random walks of varying length, creating a collection of open knots. Following the strategy of Millett, Dobay and Stasiak, an open knot is closed by connecting its two open endpoints to a third point, lying on a large sphere that encloses the random walk deeply within its interior. The resulting polygonal knot can be analyzed and its knot type determined, up to the indeterminacy of standard knot invariants, using the HOMFLY polynomial. With many closure points uniformly distributed on the large sphere, a statistical distribution of knot types is created for each open knot. We use this method to continue the exploration of the knottedness of linear random walks and apply it also to the study of several protein chains. One new feature of this work is the use of an Eckert IV planar projection, preserving area, of the knotting distribution on the sphere to characterise the spatial properties of the distribution.

1. Introduction

Proteins are linear polymeric chains of amino acids connected in a specific sequence nad folded into a specific spatial structure. While the linearity of proteins precludes them from forming real (closed) knots, their spatial structure has been shown to contain knotted conformations in a number of distinct methods developed for this purpose. Mansfield [?,?] used a method in which the protein structure is extended from the endpoints to two different points on a sphere that enclosed the protein at a relatively far distance. These two endpoints were then connected by a length of the generically unique great circle on the sphere and the knot type determined using knot invariants. Mansfield performed this operation 100 times for a protein in order to find a dominant knot in the resulting data. Taylor [?] described a continuous deformation of the protein structure in which the ends were held fixed and the rest of the protein is deformed around them until no further smoothing is possible. From the final configuration Taylor expects that the knot type can be identified. Applications of this method on a given protein can, however, yield different results depending upon order in which the deformations take place [?]. Both of these methods have something in common: in defining knots in proteins there is a degree of uncertainty.

The method we used, developed by Millett, Dobay and Stasiak [?] , is similar to Mansfields as it attempts to find the dominant knot type from multiple closures of the protein structure. In their method a large sphere, relative to the size of the open knot, centered at the center of the smallest sphere that encloses the protein (small-ball), is created. A point is then picked, at random, on this sphere and the two endpoints of the open knot are connected to this point, effectively closing the knot as shown in Figure 1.

Fig. 1. A random walk of length 50 is closed to a random point on an enclosing sphere, generating a unique knot type.

The resulting knot, now closed, is analyzed by computing the HOMFLY polynomial [?,?]. When multiple random points are chosen, a spherical distribution of knot types is created associated to the single open knot. These distributions provide the information, the knotting spectrum of the configuration, that is used to identify the dominate knot type of the configuration. In this study, we have first tested the method on 1000 random walks of varying lengths. This analysis provides a some insight into the structure of knotting. We wish, however, to have a deeper spatial sense of the distribution of knot types as they occur on the surface of the sphere containing the closure points. In order to capture the relative proportion and the spatial distribution, we selected the Eckert IV projection of the sphere on which we have coded the knot types by color.

The analysis of the data has provided some insight into the spatial structure of the random walks and how this structure is reflected in the knotting spectrum. In addition, we applied this method to eight proteins, those studied by Taylor, in order to test the extent of agreement between the two methods. In all cases, except the instance of 1kopA, the results are in agreement. In 1kopA [?], Taylor [?] finds a figure-eight knot while we find no dominate type but, rather, a balance between the unknot and the trefoil knot.

2. Random Walks and Knotting

Before analyzing the knottedness of protein chains we extended that initial work of Millett, Dobay and, Stasiak [?] by applying the method to 1000 classical random walks composed of uncorrelated equilateral segments with varying numbers of steps, 50, 100, 150 and, 200 steps. For each linear random walk, the smallest enclosing sphere was determined and 10,000 random points on the large enclosing sphere were chosen and analyzed via the computation HOMFLY knot polynomial invariant [?]. Typically, an open knot, will have different knot types depending upon the choice of closure point on the sphere. In some regions, this knot type will be the same as the end-to-end closure of the configuration. As the choice of closure point moves about the sphere, the two closing segments will pass through segments of the linear random walk. Such passages may or may not have the effect of changing the topological knot of the associated open knot closures. The first task is translate the 10,000 point knotting distribution, an approximation of the continuous distribution, into a histogram representing the knot spectrum of the knot. Of great interest is not only determining the dominant knot type in the spectrum, but identifying a standard by which a certain knot type should be considered as dominant and, therefore, considered as the knot type associated to the open walk. One standard could be to require its occurance in over 90% of the closures while, alternatively, one could require that its occurance be twice as large as its nearest competitor. Similarly, one could require that the knot type appear in 50% or more instances. In Figure 2 we show the spectrum associated to a single linear random walk having 50 steps. In Figure 3 we have collected 1,000 such spectra to give a sense of the presence of knots in such linear random walks. Figures 3 through 6 provide visual evidence of the increasing complexity of the spectra as the number steps in the random walk increases from 50 through 200 steps.

Fig. 2. The spectrum associated to a single linear random walk of 50 steps.

For random walks of length 50, 95.9% (959 out of 1000) showed a certain knot type for over 90% of its closures, see Figure 3. Of those, 32 had a probability of one, all were the unknot. The weaker criteria implies that in 99.1% of the cases there was a dominate knot type. In a walk of this relatively short length, the dominant knot type appears to be easy to recognize. For longer walks one expects a higher degree of uncertainty as to the dominant type. For a random walk of length 200, see Figure 5, only 84.9% (849 out of 1000) showed a certain knot type for over 90% of its closures. In this case, only 20 knots had a probability of one (7 unknot, 8 trefoil, 2 figure-eight, and 3 of more complicated knot types). Out of 1000 cases, 5 random walks of length 200. 0.5%, had no knot type appearing more than 50% of its closures. This provides concrete evidence that longer walks produce greater complexity in determining a dominant knot type. Figures 3 through 6 give the complete knot spectra for 1000 random walks of length 50, 100, 150 and 200.

Fig. 3. Knot spectra for 1000 random walks of length 50.

Fig. 4. Knot spectra for 1000 random walks of length 100.

Fig. 5. Knot spectra for 1000 random walks of length 150.

Fig. 6. Knot spectra for 1000 random walks of length 200.

We have explored the relationship between the spatial character of the linear random walks and the observed knot types. For example, one of these spatial properties is the end-to-end length, i.e. the distance between the start of the walk and its end. One knows that the average end-to-end length is proportional to the square root of the length of the walk. There are statistically very few instances of random walks with extreme end-to-end lengths: either very short or very long in relation to the number of steps in the random walk, see Figure 7. In either extreme, the distribution of knot types is more likely to have a dominant knot, for small end-to-end lengths this will be the knot type of the direct closure and for large end-to-end lengths one expects the trivial knot to dominate. Indeed, behaviour of this character is reflected in the data, for example consider the case of 50 step linear random walks, show in Figure 8. The relationship between the end-to-end length, d, of a random walk of length L and the complexity of the knotting, as measured by the number of knot types observed, N(d), seems to take the form of a power function: $N(d) = Constant * L/\sqrt{d}$ While this functional relationship is surprisingly simple, we believe that it warrants further investigation in view of data developed in this project. What it tells one is certainly quite intuitive: the shorter the end length, the more knot types will appear up to a certain threshold point at which time most of the knots will be of the type determined by the end-to-end closure.

Fig. 7. The probability distribution of end-to-end lengths for 1000 random walks of length 50

Fig. 8. The distribution of knot types as a function of the end-to-end distance for 1000 random walks of 150 steps. This function, the number of knot types, appears to be a simple function of the end-to-end distance, d, and the number of steps in the walk, L: $Constant * L/\sqrt{d}$

Another possibly interesting spatial characteristic is the compactness of the linear random walk. We calculated the ratio of the diameter of the smallest ball containing the walk and its end-to-end length distance. A smaller ratio signals a relatively compact walk. The general trend observed was that the more compact the random walk, the greater number of knots it would likely contain, see Figure 9.

We also looked at the distribution of the probabilities of the unknot knot. We observed that the distribution of the unknot in random walks

Fig. 9. A scatter plot of the number of knot types versus the ratio of the end-to-end length to the small ball radius for 1000 random walks of length 150.

was strongly bimodal. For a random walk, it is most likely that either almost all closures of a linear random walk are unknots, or almost none of the closures are unknots, for example see Figure 10.

Fig. 10. The bimodal distribution of trivial knots versus the ratio of the end-to-end length to the small ball radius for 1000 random walks of lenght 150.

3. Visualization of the Knotting Spectrum

The principal objective of this project is to develop a method to visualize the distribution of knotting that occurs with the spherical closure of the linear knot under investigation. To do so, we graphically reproduce the sphere and indicate the type of the knot closure at each point by means of a color assigned to each point on the sphere depending upon the type of knot created by closure from that point The resulting image provides a richer picture of the knotting spectrum of the random walk. This procedure determines, on the closure sphere, a collection of regions areas defined by the equally colored points within them. The regions of a given color may be disconnected or not simply connected, as a given knot types may show up in distant regions of the sphere.

Key to attaining our principal objective is to employ an area preserving planar representation of the sphere so as to link the statistics of the knot spectrum with the visual qualities of the image. The Eckert IV projection [?] of the closure sphere appears to be an excellent vehicle to accomplish this purpose. This is a pseudocylindrical projection that is non-conformal, but presents the area with minimal distortion, one of the critical properties, see Figure 11 **??**, [?]. Using this projection greatly simplifies the analysis of a sphere distribution and an equal area projection was chosen in order to visually reproduce the statistical prevalence of each knot type: a knot type occuring in 30% of closures will have an area equal to 30% of the projection of the sphere. In view of its finer presention of the knotting spectrum of the random walk, we call this the spectral sphere.

We believe that the local and global structure of the regions associated to the knot types presented by the spectral sphere are potentially reflections of the spatial properties of the random walk. From a topological perspective, generically, the borders between two regions on the sphere knot types represents a change of exactly one crossing between closures. If the cases observed in which two contiguous knot types can only be connected by a process involving more than one crossing change, this is evidence of an insufficient resolution of closures necessary to render the local properties.

4. Applications to the Identification of Knotting in Proteins

While we tested the spectral sphere method of analysis on linear random walk, the purpose for its development was its application to protein structure. The existence of classical knots in cyclic DNA, forming a closed path, is already well established. Knots in protein molecules have been studied but by taking into consideration the entire network of the protein [?], by simplifying the the structure [?], or by a closure method [?,?]. To test our spectral sphere method, we selected the 8 protein segments which had been studied by Taylor, [?], from the Protein Data Bank[?]. This data consists of coordinates of carbon atoms on the carbon backbone of the selected proteins. When the number of carbon atoms was too large for our HOMFLY programs, e.g. 1yve, 1znc, and 2btv, it was necessary to reduce the number of points by taking every other carbon atom in the backbone. In these cases, the data for both the full and reduced protein is shown when possible, Table 1 **??**. In the cases of 1yve 256 and 1yve 513 (where second number is the number of carbon atoms analyzed), we call attention to the fact that are significant differences in the results between the two cases. For that reason, we are unable to draw any conclusions from the spectral calculations for those examples. Due to the computational constraints arising in our HOMFLY programs, we are limited to under 1,000 crossings is the knot projectons. As a consequence, we were unable to apply our method to 2btvB 885 because this data produced too many crossings. Fortunately, taking every other carbon atom did produce useable data.

We note that, In the protein data, the expected termini behavior was not uniformly analogous to that observed in the linear random walks on which we tested our methods. To explore the consequences of this difference, we took the distance of the termini from the small-balls center and divided it by the small-balls radius. This provides one measure by which we can quantify the relative positon of the termini of the protein. A smaller ratio suggests that the termini maybe closer to the center of the protein mass. For the proteins investigated in this project (excluding 1yve) the average ratio was 0.751. Anticipating that the termini would lie close to the surface of the small ball inclosing the protein, this number is smaller than expected with 1cmxA having the lowest ratio of 0.479. The consequences this might have upon the observation of knotting encountered in the protein may be worthy of further investigation.

The types of knots found through our analysis were not surprising in view of previous analysis of the presence of knotting in these proteins [?,?,?]. The dominant knots are relatively simple ones: unknot, trefoils and figure-eight. Only one protein (1fugA) had a dominant knot type above 80%. 1kopA is especially interesting because it has two essentially equally likely knot types, the unknot and the right trefoil, whereas Taylor [?] found the right trefoil. In all other cases we find agreement with Taylor's results. We note, however, that in only two of the eight cases did the direct end-to-end closure result in the same knot type.

5. Conclusions and Speculations

This project principally consisted of an effort in numerical analysis, the development of a spectral sphere data visualization method, applied to linear random walks and, ultimately, to eight proteins. Our findings strengthen the proposal that certain properties of classical closed knots can be identified and quantified in open knots. The relationship between end-to-end length and the number of knots in a random walk was experimentally discovered and, we believe, should be explained by theoretical mathematical analysis. An observation, that is attractive for further investigation, is the bimodal nature of distributions of the unknot, and suggests the exploration of this property for other knot types. An analytical proof of this would provide a strong foundation on which to build a theoretical understanding of the knotting spectrum of an open chain.

Wesuggest that the geometry of spectral sphere is a rich arena for further research. The mechanics of changes in knot type and crossings as an open knot is closely related to the structure of the boundaries and connectivity of the knot regions on the spectral sphere. What is the fractal dimension of the boundary of the knotting regions on the sphere? How is the geometry of the knotting regions reflected in the spatial structure of the open knot or in the properties of the protein?

While it may appear that our explorations have raised more questions than it has answered, we dohope that it has provided new evidence that this method of analysis of the presence of knotting in open chains is a powerful and potentially important new approach to a quantative measure of knotting.

References

1. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Rersearch*, 28:235–242, 2000.
2. G. I. Evendem. Cartographic release procedures. *Rel. 4, Second Int. Report, USGS*, 2003.
3. Bruce Ewing and Kenneth C. Millett. Computational algorithms and the complexity of link polynomials. In *Progress in knot theory and related topics*, pages 51–68. Hermann, Paris, 1997.
4. P. Freyd, D. Yetter, J. Hoste, W. B. R. Lickorish, K. Millett, and A. Ocneanu. A new polynomial invariant of knots and links. *Bull. Amer. Math. Soc. (N.S.)*, 12(2):239–246, 1985.
5. C. Liang and K. Mislow. Topological chirality of proteins. *J. Am. Chem. Soc*, 116(1):3588–3592, 1994.
6. Marc L. Mansfield. Are their knots in proteins? *Structural Biology*, 41:213, 1994.
7. Marc L. Mansfield. Fit to be tied. *Structural Biology*, 43:166, 1997.
8. Kenneth C. Millett, Akos Dobay, and Andrzej Stasiak. Characterization of knots in linear random walks. *submitted*, page 12, 2004.
9. W. R. Taylor. A deeply knotted protein structure and how it might fold. *Nature*, 406:916–919, 2000.
10. Eric W. Weisstein. Eckert iv projection. *From MathWorld–A Wolfram Web Resource: http://mathworld.wolfram.com/EckertIVProjection.html*, 2004.

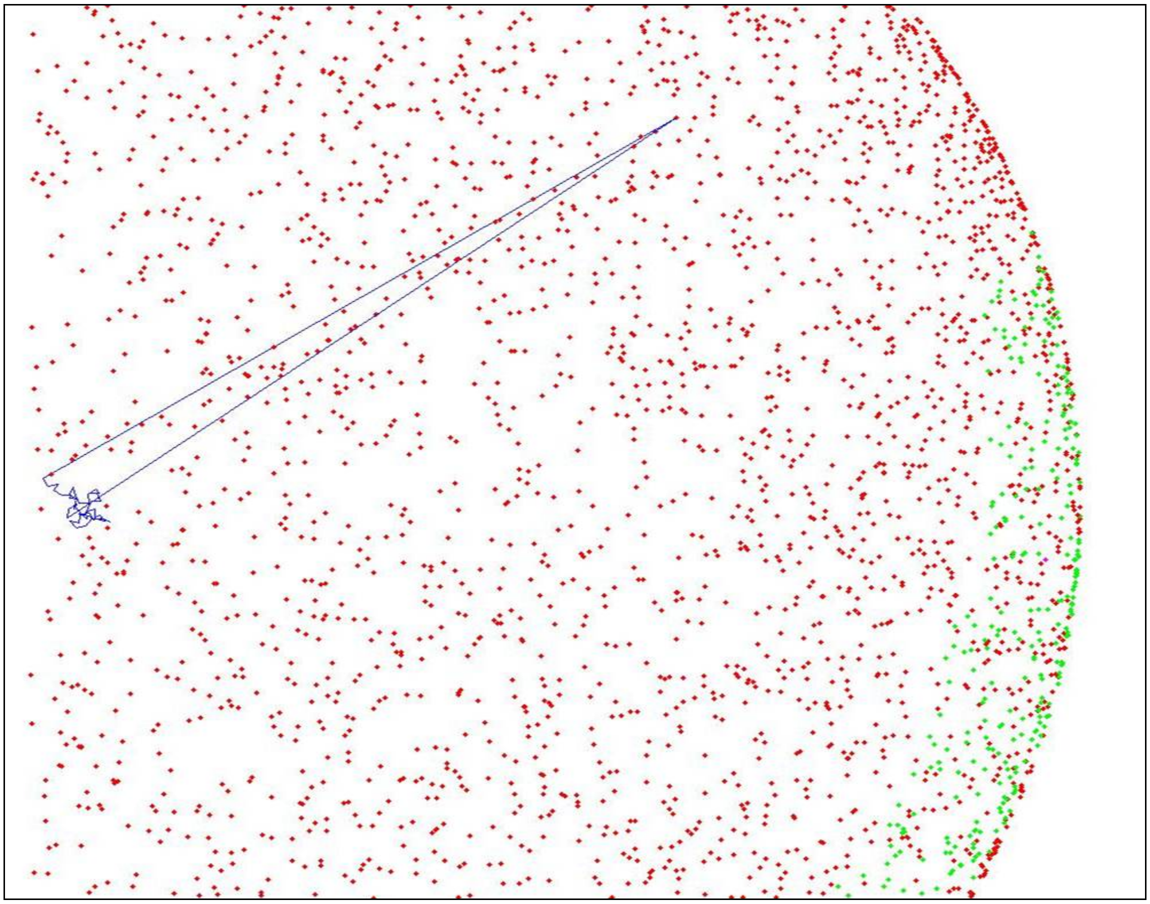


Figure 1

A random walk of length 50 is closed to a random point on an enclosing sphere, generating a unique knot type.

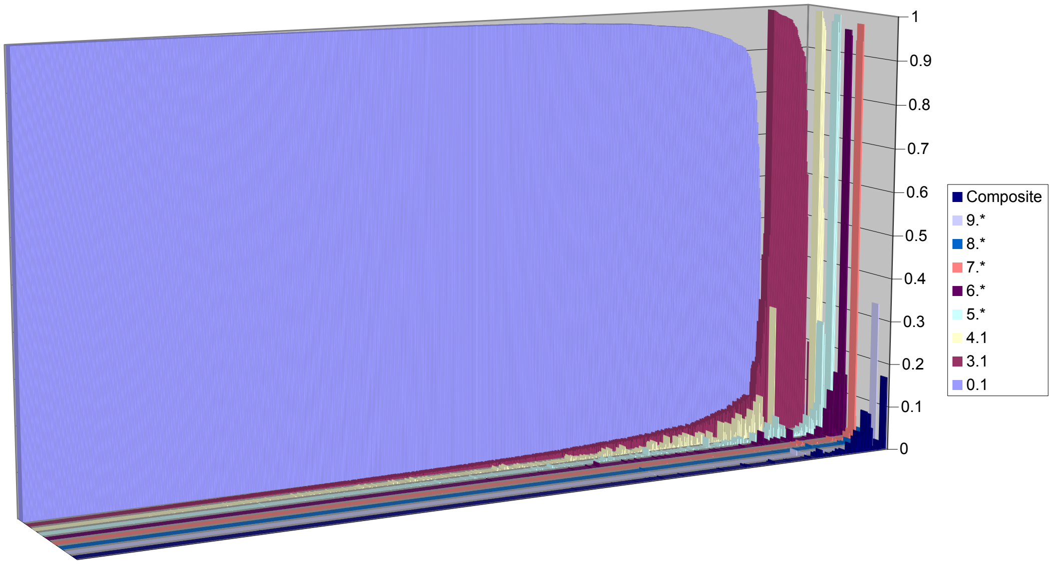


Figure 2

Knot distributions for random walks of length 50.

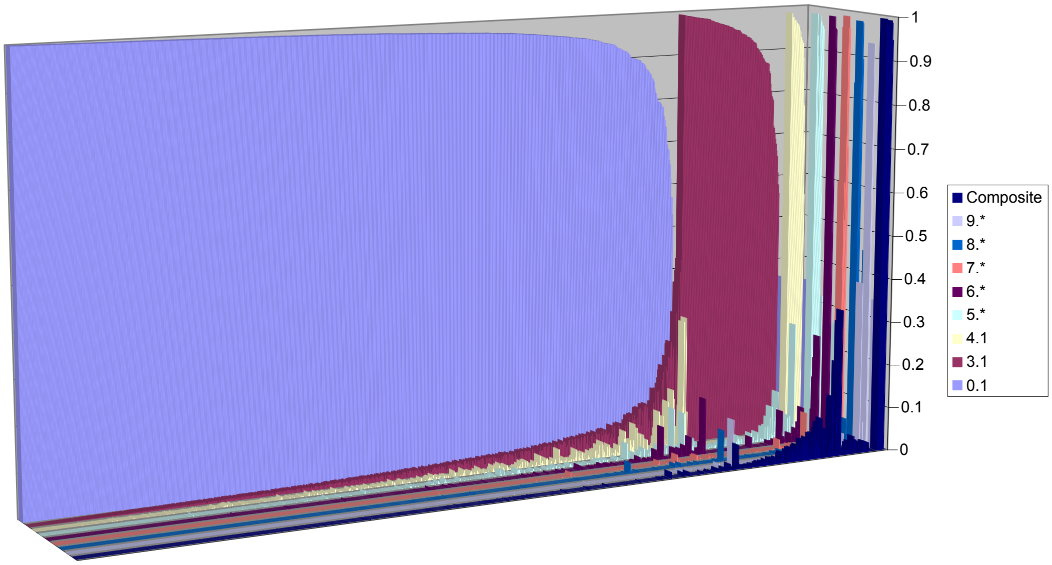


Figure 3

Knot distributions for random walks of length 100

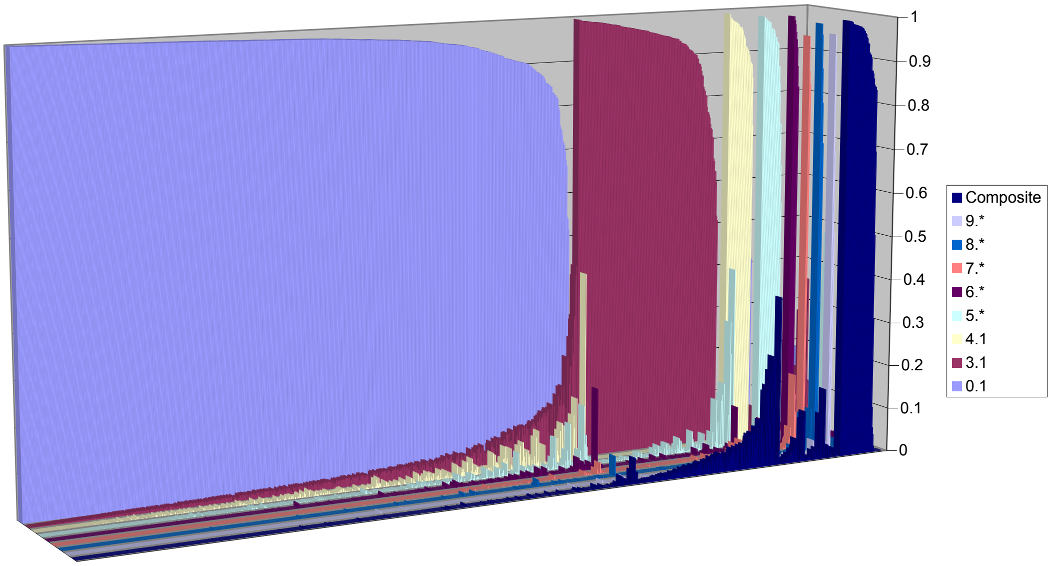


Figure 4

Knot distributions for random walks of length 150.

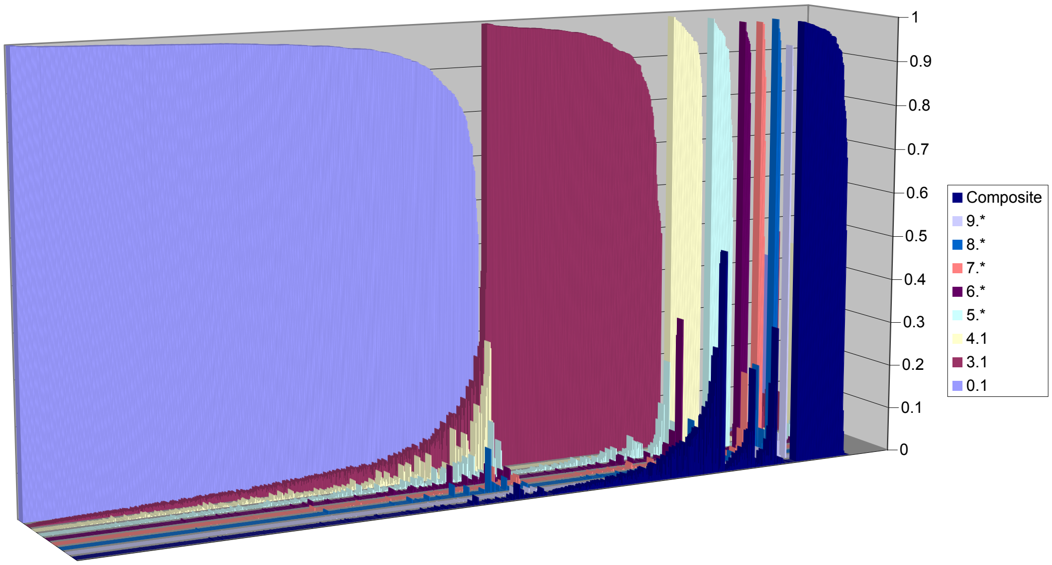


Figure 5

Knot distributions for random walks of length 200.

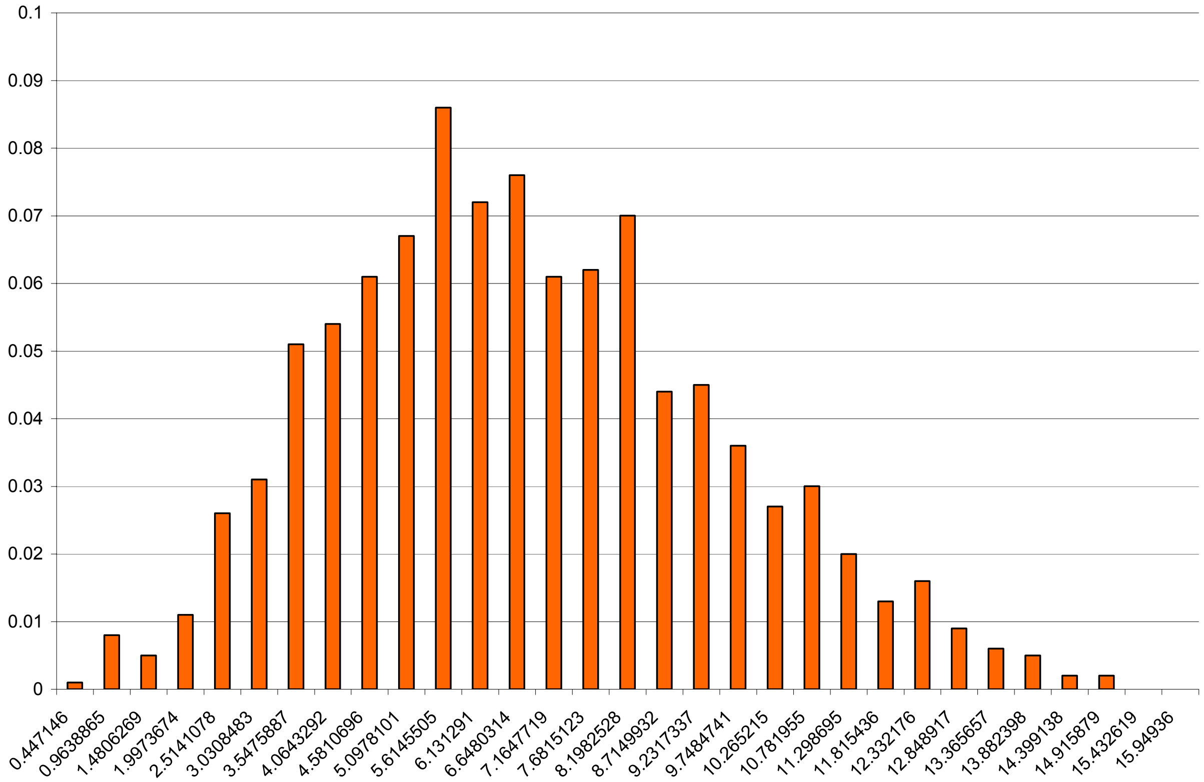


Figure 6

The probability distribution of end lengths for 1000 random walks of length 50.

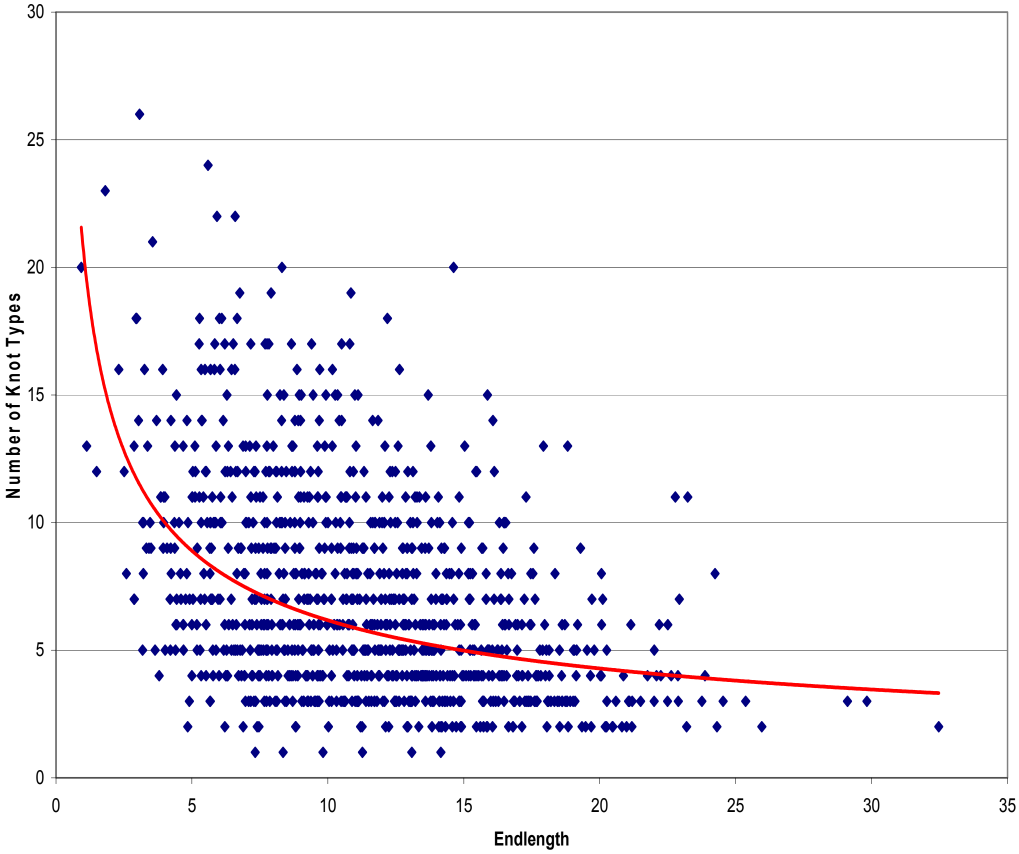


Figure 7

The distribution of knot types for a given end length for 1000 random walks of length 150. The trend line gives a power function that is a close fit to the average number of knot types:

$$Avg.\# KnotTypes = \frac{C_L \times LengthofRa\,ndomWalk}{\sqrt{Endlength}}$$

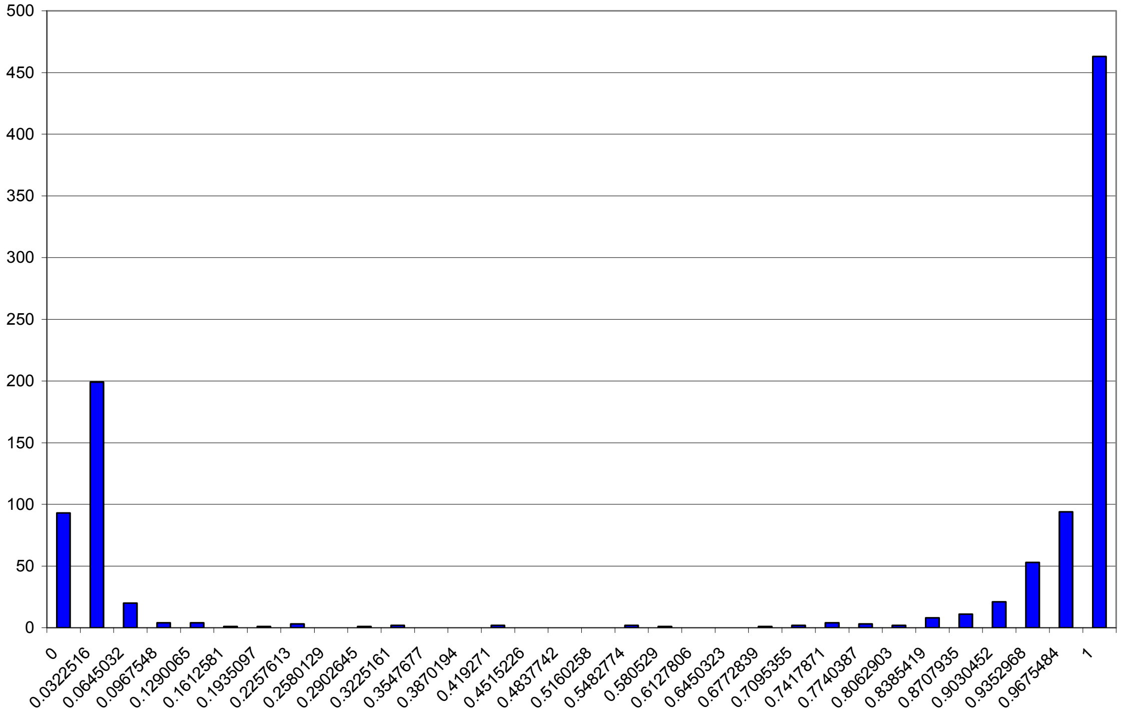


Figure 8

The bimodal nature of the probability of producing an unknot. This is taken from 995 random walks of length 150.

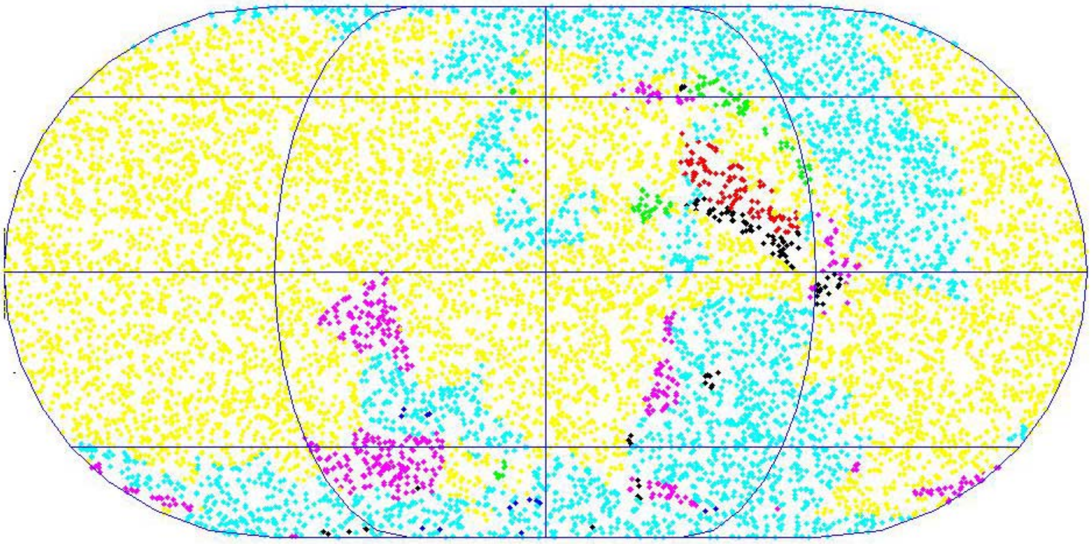
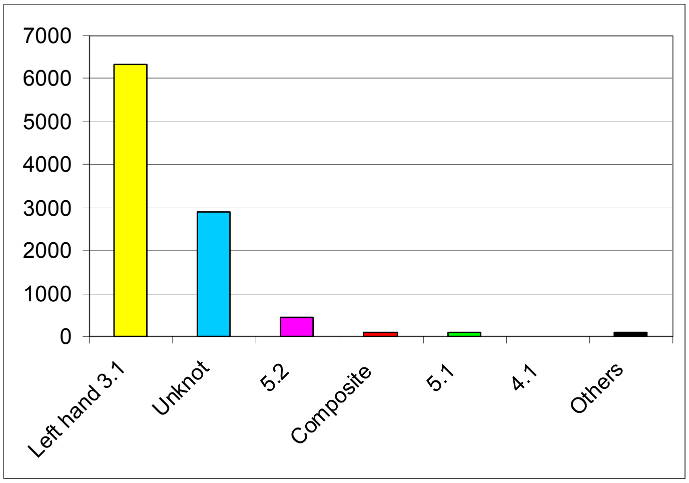


Figure 9

An Eckert IV projection of the protein 1cmxA. Color coding matches the chart below.



Distribution of knots in the protein 1cmxA

Protein Segment	smallball diameter	endlength	ratio	Termini Impactedness	Number of knot types	prob of 0.1	prob of 3.1	prob of 4.1	prob of 5.1	prob of 5.2	prob of 6+
1cmxA_214	74.5874	33.43906	0.448321	0.479127	13	0.2907	0.6424	0.0016	0.0102	0.0392	0.0159
direct						*1*	*0*	*0*	*0*	*0*	0
1dmxA_237	74.9052	30.13227	0.402272	0.679981	11	0.2481	0.6469	0.0185	0.0548	0.0121	0.0196
direct						*1*	*0*	*0*	*0*	*0*	0
1fugA_383	66.6972	29.29938	0.43929	0.909593	10	0.110791	0.851455	0.007414	0.01263	0.012493	0.005217
direct						*0*	*1*	*0*	*0*	*0*	0
1hcB_258	63.6452	41.36694	0.649962	0.696275	11	0.2782	0.6027	0.0152	0.0699	0.0165	0.0174
direct						*0*	*0*	*1*	*0*	*0*	0
1kopA_223	48.0467	36.9418	0.768873	0.816933	10	0.4534	0.4454	0.0137	0.0583	0.0155	0.0137
direct						*1*	*0*	*0*	*0*	*0*	0
1yvE_256	68.7558	43.23393	0.628804	0.642544	14	0.0897	0.0133	0.6607	0	0	0.2331
direct	*(obtained by removing last coord.)*					*0*	*0*	*1*	*0*	*0*	0
1yvE_513	70.0077	48.66591	0.695151	0.711969	1	1	0	0	0	0	0
direct						*1*	*0*	*0*	*0*	*0*	0
1zncA_131	54.4706	34.98487	0.642271	0.67267	11	0.3499	0.5024	0.0447	0.0672	0.0083	0.0275
direct	*(obtained by removing last coord.)*					*1*	*0*	*0*	*0*	*0*	0
1zncA_262	55.0504	35.86712	0.651532	0.652841	14	0.3538	0.5113	0.0243	0.0593	0.0155	0.0358
direct	*(obtained by removing last coord.)*					*1*	*0*	*0*	*0*	*0*	0
2btvB_442	260.721	29.89414	0.114659	0.92102	21	0.22	0.0176	0.6427	0.0001	0.001	0.1164
direct						0	0	1	0	0	0
2btvB_885	262.77	29.3935	0.11186	0.928338		Too many crossings					

Figure 10

Seven distinct protein segments analyzed. The last number in the protein segment name is the number of carbon atoms use to determine protein shape. The dominant knot type is colored in red, purple when dominance cannot be adequately identified. Yellow highlights show direct closure.