Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 6, Number 1 (2013) pp. 1-15

http://www.ripublication.com/jcib.htm

Novel Strategies to Classify Knot and Unknot Proteins

Lissy Anto P.

Department of Computer Science, St. Joseph's College,
Irinjalakuda, Kerala, India 680121
lissyantop@gmail.com

Abstract

Knots are relatively unexplained feature of proteins. The occurrences of knots in proteins play an important role in biological systems and processes. How the polypeptide chain is able to "knot" itself during the folding process to form these highly complicated protein topologies is not known and it is a new challenge to the research community. As of now, 278 known knotted proteins are identified in PDB. In the current investigation, features which characterize knot proteins are identified using computational methods. Using Fast Fourier Transforms (FFT), comparable spectral signatures were found for knot proteins and nonhub proteins (proteins which interact with less number of proteins) and its quantification led to the conclusion that knot proteins exhibit nonhub nature. Crosscorrelation of knot protein sequences with synthetic sequences revealed that a definite occurrence of hydrophobic domains exist in knot proteins. Using Chaos Game Representation (CGR) theory, we found that knot proteins exhibit special pattern in their CGR and these patterns were quantified by setting a ratio of CGR points clustering in the two opposite corners of their CGR (representing hydrophobic and hydrophilic amino acids). These characteristics can in turn be used as indicators for identifying knot proteins. The quantified parameters obtained from the above mentioned steps have been used as feature vectors to classify knot proteins and unknot proteins. The classification was done using Artificial Neural Network (ANN) and Support Vector Machines (SVM) for which the latter gave good results.

Keywords: Protein topology, folding rate, hydrophobicity, clustering, packing density.

1. Introduction

An increasing number of proteins are being discovered with a special feature, a knot in their native structures. Knots often appear in globular proteins. Knots and structures are studied traditionally in algebraic topology, a branch of mathematics. In algebraic topology, knots are defined as closed curves in three-dimensional Euclidean space $\Re^3$ and are categorized according to the minimal number of crossings in a projection onto a plane. Knots and structures found in algebraic topology are also found in DNA, RNA, and proteins. The function of knots in biological systems is in the realm of enzymatic activities. It has been found that knotting is happening during the folding process. The reason and factors that influence existence of knots in protein topologies is still unknown. Energy and hydrophobicity factors are known to influence formation of knotted structures.

Enzymatic activity is one of the major activities of a knot in protein. The knotted regions have been shown to be important in ligand binding too. Knotted structures are difficult to fold [1, 2], but once folded they maintain the conformation of the folded state which contribute to thermal stability [3]. The identification of knotted area of proteins is thus relevant in understanding protein functions.

Review of current literature suggests that hydrophobic interactions are the most important non-covalent forces that are responsible for structure stabilization of proteins. Hydrophobicity index is a determinant of protein-protein interactions too. Surface hydrophobicity can be used to identify regions of a protein's surface most likely to interact with a binding ligand. It is generally accepted that the hydrophobic effect is the main factor in stabilizing the folded structure of globular proteins [3, 4]. The recognition of residues forming hydrophobic microdomains is an important step in protein folding and it is determined by interactions between residues that are distant in sequences. About one third of the residues of the hydrophobic microdomains are knot residues. Each hydrophobic microdomain contains at least one knot residue. Most of the remaining knot residues are found to be adjacent to residues of the hydrophobic microdomains. Residues forming the hydrophobic microdomains are highly conserved in amino acid sequences. The localization of knot residues is achieved by finding these hydrophobic microdomains [5].

Knots are generally described by $C_{U,}$ where C is the ideal crossing number and U the unknotting number. The unknotting number of a knot is the minimal number of crossing changes required to convert a knot into unknot. Depending on the number of crossings, knots are named as trefoil knots (having 3 crossings denoted by 3_1), figure eight knot (having 4 crossings denoted by 4_1) and so on. As of now, 3_1, 4_1, 5_2, 6_1 knots are found to exist in proteins. Various algorithms have been developed to distinguish between different types of knots. Alexander polynomial algorithm and Homfly polynomial algorithm are two popular algorithms which are used to distinguish between different types of knots [6, 7]. All knotted proteins are enzymes and most of the knots in them are trefoil knots. The knotted proteins come from

the following classes: methyltransferase, transcarbamylase, carbonic anhydrase, ketol–acid reductosiomerase, ubiquitin hydrolase, methionine adenosyl transferase, the chromophore-binding domain of bacterial phytochrome and the inner core shell component protein of bluetongue virus [8]. Three different types of knots found in proteins are shown in Fig. 1.

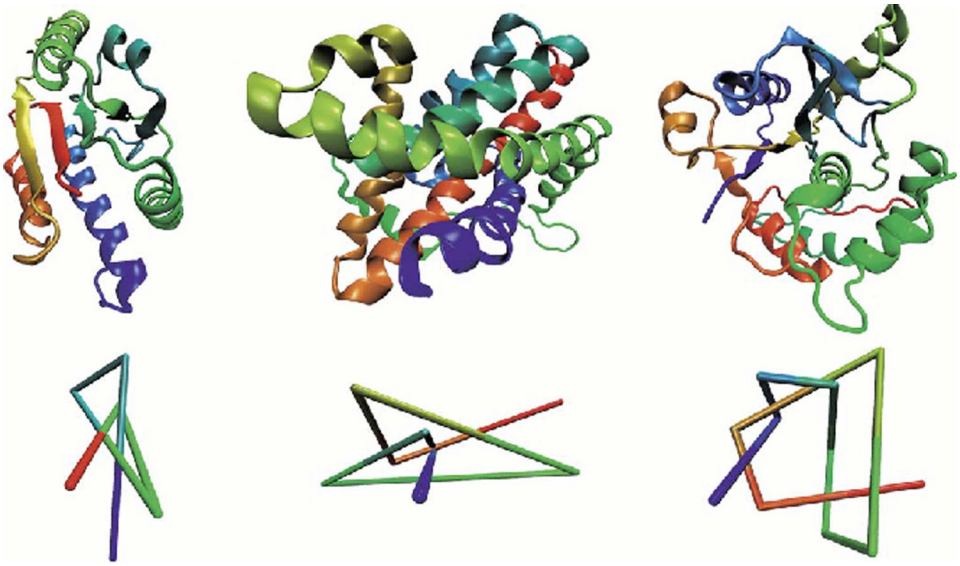

Figure 1. Examples of the three different types of knots found in proteins and their corresponding structures [9, 10]

In Figure 1, (Left) is a trefoil knot 3_1 in the YBEA methyltransferase from E. coli (pdb code 1ns5), (Middle) is a figure-eight knot 4_1 in the Class II ketol-acid reductoisomerase from Spinacia oleracea (pdb code 1yve) and (Right) is a 5_2 knot in ubiquitin hydrolase UCH-L3 (pdb code 1xd3). Pictures were generated with Visual Molecular Dynamics (http://www.ks.uiuc.edu/Research/vmd)

In the current investigation, features which characterize knot proteins are identified using computational methods. We were unable to spot work reporting computational analysis on knot proteins. The present investigation has three distinct parts: Firstly, knot protein sequences were mapped into discrete signals using hydrophobicity values and the mapped sequences were spectrally analysed using Fast Fourier Transforms (FFT). Comparable spectral signatures were found for knot proteins and nonhub proteins (proteins which interact with less number of proteins) in the analysis. Its quantification led to the conclusion that knot proteins exhibit nonhub nature. In the second part, Crosscorrelation of knot protein sequences was taken with discrete signals mapped from synthetic sequences of different hydrophobicity levels. This study revealed that a definite occurrence of hydrophobic domains exist in knot proteins. The third part of investigation involved mapping knot protein sequences into a graphical representation known as Chaos Game Representation (CGR). Knot proteins

exhibit special pattern in their CGR. These patterns were quantified by setting a ratio of CGR points clustering in the two opposite corners of their CGR (representing hydrophobic and hydrophilic amino acids). From web based tools, packing density, cavity and folding rate of knot proteins were computed. These features help to identify knot proteins as well as to classify them from unknot proteins using popular classification tools ANN and SVM.

2. Materials & Methods

A. Database

There are 278 entries in PDB known as knot proteins and are available at http://knots.mit.edu. It provides information on knots in structures of amino acid sequences. According to CATH (Class Architecture Topology Homology) classification of proteins, a majority of the knot proteins are of α/β structures. For classification, a negative dataset (unknot dataset) is also required. Presently there is no readily available unknot dataset. Hence the only available option was to choose a set of sequences belonging to a particular class, biologically known to be unknotted. Knots usually occur in hydrophobic microdomains. The possibility of hydrophobic microdomains in hemoglobin is very less [5] and hemoglobin proteins are so far considered as unknot proteins. Hence they were chosen as unknot dataset.

B. Spectral signatures of knot proteins

It is often useful to consider mapping of signals from one domain to another so that certain information hidden in the former domain may become explicit in the latter. One such transformation is the time domain to frequency domain transformation of signals and Fourier transformation is one such operation. That is, for a signal, it is possible to decompose it into a bunch of sine waves, each at a different frequency, amplitude and phase. Fourier analysis could be applied to a variety of signals- continuous, discrete, periodic, aperiodic or combinations thereof. Fourier Transform which converts information in time domain to frequency domain, is a signal analysis technique that have revolutionised the field of signal processing.

The Fourier Transform of a continuous aperiodic signal is defined by [11]

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \qquad (1)$$

where X(jω) is a complex quantity having magnitude and phase. If the signal is discrete and aperiodic, the Discrete Time Fourier Transform (DTFT) of the signal is defined in a similar way by

$$X(e^{j\omega}) = \sum_{n=-\infty}^{n=\infty} x[n]e^{-j\omega n} \qquad \text{……..(2)}$$

where $X(e^{j\omega})$ is a continuous function of ω and periodic with a period 2π.

To compute $X(e^{j\omega})$ for a finite length sequence x[n], (which is used in genomic signals), x[n] is restricted to N samples numbered from 0 to N-1. Also, instead of computing $X(e^{j\omega})$ for all frequencies in DTFT, it is only computed for a finite number of points in [0, 2π]. The fact that x[n] is a finite length sequence implies that the DTFT can be rewritten as

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n)\, e^{-j\omega n} \tag{3}$$

While computing DTFT only for a finite number of frequency points, this equation simplifies to

$$X(e^{j\omega_k}) = \sum_{n=0}^{N-1} x(n)\, e^{-j\omega_k n} \tag{4}$$

where $\omega_k = 2\pi k/N$ are the frequency samples. If there are N samples, then k takes values 0, 1, . . . , N−1. That is, $X(e^{j\omega})$ is evaluated only at the frequency values $\omega_k = 2\pi k/N$ for k = 0, 1, . . . , N −1 The resulting expression is

$$X(e^{j\omega_k}) = \sum_{n=0}^{N-1} x(n)\, e^{-j\frac{2\pi kn}{N}} \tag{5}$$

The N-point discrete Fourier transform (DFT), X(k), of an N-point discrete-time sequence x[n], is defined by

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi kn/N} \tag{6}$$

for k = 0, 1, . . . , N − 1. The DFT, X[k], is just a sampled version of the DTFT, $X(e^{j\omega})$, at $\omega = 2\pi k/N$ [12].

The DFT power spectrum of a signal at frequency k is defined as

$$S(k)=|X(k)|^2,\ k = 0,1,2\ldots\ N\text{-}1 \tag{7}$$

where X(k) is the k^{th} DFT coefficient.

For applying Fourier Transform in biological sequences, they are converted into numeric bio-signals using suitable physicochemical parameters. For eg., Let x[n] = [A R N M Y P L N M D A C] be an amino acid sequence. Replacing the sequence characters by a physico chemical parameter hydrophobicity in Table 1, the string becomes a discrete signal

i.e x[n] = [1.8 -4.5 -3.5 1.9 -1.3 -1.6 3.8 -3.5 1.9 -3.5 1.8 2.5]

Table 1. Kyte Doolittle Hydrophobicity Values [13]

Amino Acid	HydrophobicityValue
Ala(A)	1.8
CyS(C)	2.5
Asp(D)	-3.5
Glu(E)	-3.5
Phe(F)	2.8
Gly(G)	-0.4
His(H)	-3.2
Ile(I)	4.5
Lys(K)	-3.9
Leu(L)	3.8
Met(M)	1.9
Asn(N)	-3.5
Pro(P)	-1.6
Gln(Q)	-3.5
Arg(R)	-4.5
Ser(S)	-0.8
Thr(T)	-0.7
Val(V)	4.2
Try(W)	-0.9
Tyr(Y)	-1.3

Fourier spectrum of this discrete signal is found out using the above equations (6) and (7). Fourier Transform method has proved its relevance to find special characteristics in biological sequences. For example, Fourier power spectrum of the DNA sequence is used to identify the gene coding area in a DNA sequence of eukaryotes. Like DNA sequences, for a given collection of proteins which have a common function, it is possible to identify this commonality by analyzing the amino acid sequence by Fourier transform techniques.

The area under the spectral curve of a signal is a measure of the total energy of the signal [14]. The energy spectrum is capable in discerning where the principal energy for a given signal is located. For hub and non hub proteins, it has been found that the average energy content of hub sequences has a great contrast with that of non-hub sequences. Results show that the proportion of multi-domain proteins in hubs is larger than the corresponding fraction in non-hubs [15]. Also, repeated domains are clearly overrepresented in hub proteins [15]. The domain repeats in hub proteins cause comparatively high frequency spectrum. In nonhubs, the possibility of multiple domains is

very less and this leads to a low frequency spectrum. So nonhubs have less spectral area compared to that of hub proteins. In the current investigation, the spectral signatures of knot proteins are found to be similar with nonhub proteins. The spectral area for 278 sequences were calculated for knot and unknot dataset. Each set of values is then clustered and the clustering cenroids are shown in Table 2. It could be inferred that similar to nonhub proteins, knot proteins too exhibit less multiple domains.

Table 2. Spectral area clusters for knot & unknot proteins

Knot proteins	Unknot proteins
0.4226	2.2077
1.5447	8.388
3.3668	22.655
6.1855	38.933
17.267	66.459

C. Crosscorrelation of knot proteins

Crosscorrelation is a powerful tool of signal processing that can be applied for detecting signal similarities, pattern recognition and signal detection. Correlation uses two signals to produce a third signal, which is called the cross-correlation of the two input signals. Cross-correlation function for two sequences x(n)and y(n) is defined as:

$$r_{x,y}(l) = \Sigma\ x(n)*y(n-l)\ \ l = 0, \pm1, \pm2,\ldots\ \ldots.. \quad (8)$$

or, equivalently, as

$$r_{x,y}(l) = \Sigma\ x(n+l)*y(n)\ \ l = 0, \pm1, \pm2,\ldots\ \ldots.. \quad (9)$$

When the two signals are similar in shape and unshifted with respect to each other, their product is all positive. This is like constructive interference, where the peaks add and the troughs subtract to emphasise each other. The area under this curve gives the value of the correlation function at point zero, and this is a large value. As one signal is shifted with respect to the other, the signals go out of phase, the peaks no longer coincide, so the product can have negative going parts. This is a bit like destructive interference, where the troughs cancel the peaks. The area under this curve gives the value of the correlation function at the value of the shift. The negative going parts of the curve now cancel some of the positive going parts, so the correlation function is smaller. The breadth of the correlation function where it has significant value shows for how long the signals remain similar [16].

Traditional method of using crosscorrelation has been extended to compare biological sequences [16]. The hydrophobic nature of knot proteins has been analysed using crosscorrelation method in this investigation. In order to study

hydrophobic nature of knot protein sequences, these sequences were correlated with synthetic sequences of varying hydrophobicities. To apply crosscorrelation on knot protein sequences, they are mapped into numeric sequences by replacing amino acids with hydrophobicity values in Table 1. For the construction of synthetic sequences the software package DAMBE was used [A18]. DAMBE (Data Analysis in Molecular Biology and Evolution) is an integrated software package for retrieving, organizing, manipulating aligning and analyzing molecular sequence data. The input to DAMBE tool is percentage of amino acids in a sequence. The hydrophobicity values in Table 1 are clustered to get 4 groups namely highly hydrophobic (I, L, V), hydrophobic (C, A, F, M), neutral (P, Y, T, G, S, W), and not hydrophobic (R, K, N, Q, E, H, D) amino acids. Six different groups were chosen with combinations of decreasing hydrophobic contents. The symbols used for these six combinations are as follows:

- HH – highly hydrophobic
- HY – highly hydrophobic + hydrophobic
- HN – highly hydrophobic + hydrophobic + neutral
- HU – highly hydrophobic + hydrophobic + neutral + not hydrophobic
- NU – neutral
- NH – not hydrophobic

Equal percentages for each amino acid in all these six cases were used which to create synthetic sequences by DAMBE. These synthetic sequences are also converted into numeric sequences, thereby creating signals with the help of hydrophobicity values as given in Table 1. Among different hydrophobicity measures, Kyte-Doolittle hydrophobicity was chosen as it is more related to the structure of proteins [13].

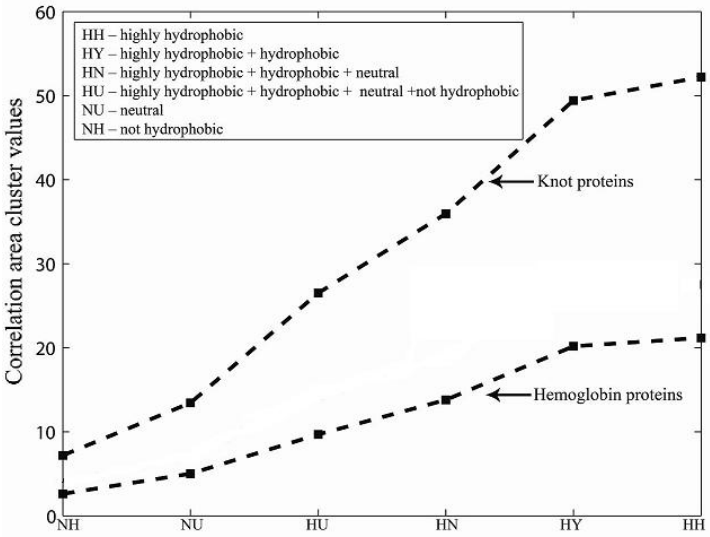

Figure 2. Correlation area cluster values for knot proteins and hemoglobin proteins

Highly hydrophobic (HH), combination of hydrophobic and highly hydrophobic (HY) sequences when correlated with knot proteins produced stronger signals. But the other four cases - 1) highly hydrophobic + hydrophobic + neutral (HN), 2) highly hydrophobic + hydrophobic + neutral + not hydrophobic (HU), 3) neutral (NU), 4) not hydrophobic (NH) showed relatively weaker signals. To quantify crosscorrelation of sequences, the area under the crosscorrelation signals are considered. These are designated as the correlation area CA values. For each knot protein, there are six different CA values. The CA values for the whole knot data set is found out. Each set of values under the same hydrophobicity level is clustered using k-means clustering to find a common value about which most of the data is spread around. So there are 6 cluster centroid values to represent the whole knot dataset. The same process was repeated for a group of unknot proteins. The six cluster points for knot proteins and hemoglobin proteins are shown in Figure 2. The variation in CA values give special insights for the investigation. The hydrophobic domains in knot sequences have lead to higher CA values than that of unknot proteins. Hemoglobin proteins which can never be knotted [5]. have less CA values than that of knot proteins. Hence it is concluded that correlation method can be used to detect knot proteins from unknot proteins [18].

D. CGR of DNA sequences

Chaos Game Representation (CGR) can recognize patterns in the sequences and recognition of patterns helps visual identification of the sequences. The scope of CGRs as useful signature images of bio-sequences such as DNA has been investigated since early 1990s. CGR exhibits striking pattern characteristics of the genome and so is helpful to identify genome fragments or to detect gene. [19]. To derive a chaos game representation of a DNA sequence, a square is first drawn to any desired scale and corners marked A, T, G and C. The first point is plotted halfway between the center of the square and the corner corresponding to the first nucleotide of the sequence, and successive points are plotted halfway between the previous point, and the corner corresponding to the base of each successive nucleotide. Mathematically, coordinates of the successive points in the chaos game representation of a DNA sequence is described by an iterated function system defined by

$$X_i = 0.5(X_{i-1} + g_{ix})$$

$$Y_i = 0.5(Y_{i-1} + g_{iy})$$

g_{ix} and g_{iy} are the X and Y co-ordinates respectively of the corners corresponding to the nucleotide at position i in the sequence [19, 20].

CGRs of amino acid sequences present a very difierent challenge, as CGRs can be constructed using a 20 sided regular polygon or alternatively using smaller polygons, by assigning groups of amino acids to the corners in a variety of ways. In this investigation, CGRs are drawn as follows: The 20

amino acids are divided into four classes depending on their hydrophobicity values. The six residues F, I, W, L, M, V designate the highly hydrophobic class; the three residues A, C, Y designate the hydrophobic class; the six residues T, H, G, S, Q designate the neutral class; and the remaining six residues R, K, N, E, P, D designate the hydrophilic class. For a given protein sequence $s = s_1s_2\ldots s_{20}$ where s_i is one of the twenty amino acids, define

a_i = 1, if si is highly hydrophobic

= 2, if si is hydrophobic

= 3, if si is hydrophilic

= 4, if si is neutral

The sequence s is subsequently mapped into another sequence $a = a_1a_2a_3a_4$ and the CGR is constructed using the new sequence a. The CGR is plotted on a square with vertices 1, 2, 3 and 4 where vertex 1 represents the highly hydrophobic amino acids, 2 represents hydrophobic amino acids, 3 represents hydrophilic and 4 represents amino acids that are neutral.

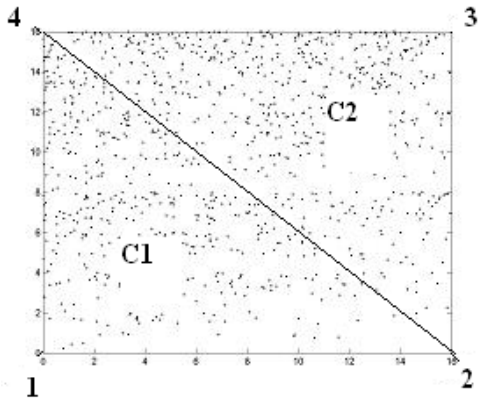

Figure 3. Division of points in a CGR image

E. Evaluation of packing density and folding rate

Packing density and folding rate are related characteristics to the hydrophobic nature of protein sequences which in turn influence knottedness in proteins. Residues in the interior of protein structures are closely packed with mean volumes which are a little smaller than those in the primary sequences. This rate of interior packing of residues is known as packing density of a protein [21]. Amino acids interact with each other and fold to produce a well-defined three dimensional structure, the native state. The three dimensional structure is essential for a protein to function. The rate of folding is the rate of the physical process by which a polypeptide folds into its characteristic and functional three dimensional structure. It has been proved that the α/β class of proteins have the tightest packing [22]. As the set of knot proteins are of α/β

class, this implies that knot proteins too have tightest packing. The packing density of the proteins was computed by a software Voronoia [23] which is available at http://bioinformatics.charite.de/. This gives packing density for input given as PDB structures of proteins. Along with packing density, this tool provides another feature named cavity, which measures a protein's capacity to interact with other proteins.

The folding rate values of proteins were calculated by the prediction server fold-rate [24]. This tool computes folding rate of proteins on the basis of 49 diverse amino acid properties. Length of a polymer chain is a parameter which determines folding rate of proteins [25, 26]. The folding rate of knot proteins were six times less than that of unknot proteins.

F. Classification by Artificial neural network & Support vector machine

Good features simplify the process of a classier whereas features with little discriminating power can hardly be compensated with any classier [27]. The discriminative nature of features characterize the items to be classified. The best feature set is a representative of the underlying data set and it can contribute to high classification rate. A list of 11 features has been investigated, both sequence level and structural level. Sequence level features are derived from Fourier analysis (1 feature), Crosscorrelation method (6 features) and CGR theory (1 feature). Structural level features were limited to secondary structural motifs as the tertiary and quaternary structures are not available for all proteins. From available online tools folding rate, packing density, cavity has also added to the feature set. These 11 features performed well and were selected for classification.

The three essential features of ANN are basic computing elements referred to as neurons, the network architecture describing the connections between the neurons and the training algorithm used to find values of the network features for performing a particular task. Various configurations of ANN were initially tested in order to confirm the best architecture for the classification problem. They are Feed Forward Network (FFN), Radial Basis Function (RBF) and Probabilistic Neural Network (PNN, which is a variant of Radial basis network) of which Radial Basis Function (RBF) was selected. In Back propagation ANN, 11 input nodes were used as the feature vector representing the input data and one output node for output. The input data was tested with different number of hidden layers and layer neurons. The trained algorithm used was Levenberg-Marquardt training algorithm [28], which is widely used. Configuration which gave the best accuracy was the one with 2 hidden layers and 6 & 4 neurons in the first and second layers respectively.

SVM is a nonlinear classification algorithm based on kernel methods. There are a number of kernels that can be used in Support Vector Machines models. These include linear, polynomial and radial basis function (RBF). The RBF is the most popular choice of kernel types used in Support Vector Machines[29]. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

3. Results & Discussion

The computational analysis carried out in this work is comprised of signal processing tools and pattern recognition method. The packing density, cavity and folding rate were also analysed in order to explore how these factors vary in the case of knot and unknot sets of sequences.

Fourier analysis of knot proteins revealed the following. In knot proteins, interactions within the chain are predominantly local and it is probable that their interaction capability is less with other proteins. So the possibility of multiple domains is very less in knot proteins and this leads to a low frequency spectrum. That is, spectral content of knot proteins are very less compared to that of unknot proteins.

The correlation analysis of knot proteins resulted in unveiling the significant relationship of knot proteins and hydrophobic domains. Synthetic sequences of varying hydrophobic contents were correlated with knot and unknot proteins. Correlation of hydrophobic and highly hydrophobic synthetic sequences with knot proteins resulted in stronger signals whereas correlation of the lesser hydrophobic sequences with knot proteins produced relatively weaker signals. The cross correlation results were quantified by considering the area under the cross-correlation curves defined as correlation area (CA) values. The CA values for all the knot proteins were high whereas CA values for unknot proteins were low. This is due to the hydrophobic domains in knot proteins.

Quantification of these CGR image patterns can be used to differentiate knot and unknot proteins. A new measure, Hydrophobic CGR Ratio (HCGRR) is set as follows: The CGR square image can be divided in two ways - either by the diagonal 1 joining highly hydrophobic vertex and hydrophilic vertex or by the diagonal 2 joining hydrophobic vertex and neutral vertex. Diagonal 2 was chosen as it gave a discriminative measure in the count of points. Taking the ratio of the number of points at the two sides of the diagonal 2 of the CGR square image. HCGRR is defined by, HCGRR = C1/C2, where C1 is the number of points in the lower triangular portion of the CGR square and C2 is the number of points in the upper triangular portion of the square. HCGRR for knot proteins are high compared to that of unknot proteins.

The soft computing algorithms ANN & SVM have been used for classification of knot and unknot proteins. With the 11 features selected as feature vector the classifiers performed well to achieve the following results as shown in Table 3.

Table 3. Results of Classification

Performance measure	ANN (%)	SVM (%)
Sensitivity	90.64	89.20
Specificity	89.20	93.52
Accuracy	89.92	91.36

4. Conclusions

Knots in the backbone of proteins are significant structural motifs that appear at different levels of protein complexity and might offer new insight in the understanding of protein folding mechanisms. Knots in protein are related to enzymatic activity and ligand binding of proteins. Hence the characteristics of this special type of proteins are explored in this study. The three-dimensional structure of a protein is uniquely dictated by its primary sequence. Hence sequence based experiments on knot proteins have relevance in bringing out their distinct features. Computational approach performed in this study is the first step in analyzing their special features. In this work, different features triggering knottedness were investigated to distinguish knot proteins. Our results proved that knot proteins are of nonhub nature. The presence of hydrophobic domains is obvious in knot proteins. The newly developed quantitative measure HCGRR derived from the visual pattern of knot proteins is a discriminative measure for these type of proteins. Also knot proteins have high packing density, cavity and less folding rate. With these special features classification of knot and unknot proteins by popular classification tools ANN & SVM brought the result 89.92%, 91.36% respectively.

References

[1] W. R. Taylor, 2000, “A deeply knotted protein and how it may fold,” Nature, Vol. 406, pp. 916-919.

[2] F. Khatib, M.T. Weirauch, C.A. Rohl, 2006, “Rapid knot detection and application to protein structure prediction,” Bioinformatics, Vol. 22, pp. 252–259.

[3] K. Wagschal, B. Tripet, P. Lavigne, C. Mant, R.S. Hodge, 1999, “The role of position in determining the stability and oligomerization state of α-helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins,” Protein Sci., Vol. 8, pp. 2312-2329.

[4] B. Y. Zhu, N.E. Zhou, C.M. Kay, R.S. Hodges, 1993, “Packing and hydrophobicity effects on protein folding and stability,” Protein Sci., Vol. 2, pp. 383-394.

[5] R. B. Gregory, 1994, “Protein-solvent interactions,” American Chemical Society, New York.

[6] Alan L. R. et al., 2005, “Sequence Alignment by Cross-Correlation,” Journal of Biomolecular Techniques, Vol. 16, pp. 453-458.

[7] P. Virnau, 2007, “Knots in Macromolecular Systems:Concepts and Challenges, Computational Biophysics to Systems Biology (CBSB07),” Proceedings of the NIC Workshop 2007, NIC Series, Vol. 36, pp. 287-289.

[8] Y. L. Lai, S.C. Yen, S.H. Yu, J.K. Hwang, 2007, “pKNOT: the proteinKNOT web server,” Nucleic Acids Research, Vol. 35, pp. 420-424.

[9] P. Virnau, L.A. Mirny, M. Kardar, 2006, "Intricate Knots in Proteins:Function and Evolution," PLoS Comput Biol., Vol. 2, pp. 1074-1079.
[10] W. Humphrey, A. Dalke, K. Schulten, 1996, "VMD—Visual molecular dynamics," J. Mol. Graphics, Vol. 14, pp. 33–38.
[11] Simon H. and Barry V. V., 1999, "Signals and Systems," John Wiley & Sons, Chichester, England.
[12] Cohen L., 1995, "Time-frequency Analysis," Prentice Hall, Englewood Cliffs, USA.
[13] Kyte, J. and Doolittle, R., 1982, "A simple method for displaying the hydropathic character of a protein," J. Mol. Biol., Vol. 157, pp. 105-132.
[14] Dennis J. W., Mario E. C. and Jennifer K., 2001, "Waters Teaching time-series analysis - Finite Fourier analysis of ocean waves," American Journal of Physics, Vol. 69.
[15] Diana E., Sara L., Sa K. B. and Arne E., 2006, "What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?,", Genome Biology, Vol. 7, pp. 45.1-45.13.
[16] T. A. Wren, K. P. Do, S. A. Rethlefsen, B. Healy, 2005, "Cross-correlation as a method for comparing dynamic electromyography signals during gait," J. Biomech, Vol. 39, pp. 2714-2718.
[17] Xia, X., and Xie. Z., 2001, "DAMBE: Data analysis in molecular biology and evolution," Journal of Heredity, Vol. 92, pp. 371-373.
[18] Lissy Anto P. and Achuthsankar S. Nair, 2009, "Hydrophobic tint of knot proteins," Proceedings of the International Symposium of Bio Computing, NITC Calicut, India.
[19] Jeffrey H. J., 1992, "Chaos game visualization of sequences," Computers and Graphics, Vol. 16, pp. 25-33.
[20] Wang Y., Hill K., Singh S. and Kari L., 2005, "The spectrum of genomic signatures: from dinucleotides to chaos game representation," Gene, Vol. 346, pp. 173-185.
[21] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, 1999, "The Packing Density in Proteins: Standard Radii and Volumes," J. Mol. Biol., Vol. 290, pp. 253-266.
[22] M. M. Gromiha, A. M. Thangakani, S. Selvaraj, 2006, "FOLD-RATE: prediction of protein folding rates from amino acid sequence," Nucleic Acids Res., Vol. 34, pp. 70-74.
[23] K. Rother, P.W. Hildebrand, A. Goede, B. Gruening, R. Preissner, 2009, "Voronoia: analyzing packing in protein structures," Nucleic Acids Res., Vol. 37, pp. 393-395.
[24] M.Gromiha, " A Statistical Model for Predicting Protein Folding Rates from Amino Acid Sequence with Structural Class Information," J.Chem.Inf.Model, Vol. 45,pp. 494-501.
[25] Y. Zhu, X. Fu, T. Wang, A. Tamura, S. Takada, J.G. Saven, F. Gai, 2004, "Guiding the search for a protein's maximum rate of folding," Chem. Phys., Vol. 307, pp. 99-109.

[26] J. T. Huang, J.P. Cheng, H. Chen, 2007, "Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics," Proteins, Vol. 67, pp. 12-17.
[27] Arbib A.M., 1995, "The Handbook of Brain Theory and Neural Networks," MA:MIT Press, Cambridge.
[28] Ozgur K., 2004, "Multi-layer perceptrons with Levenberg-Marquardt training algorithm for suspended sediment concentration prediction and estimation", Hydrological Sciences-Journal, Vol. 49, pp. 1025-1040.
[29] Vapnik V., 1995, "The Nature of Statistical Learning Theory," Springer, New York.