Chapter 6

Conclusions and Future work

6.1 Opening Remarks

This chapter concludes the thesis with a recap of what was aimed at, what has been achieved and what open questions remain.

6.2 Objectives of the research work - restated

For the sake of easy reference, we restate here the objectives of the present research work, as mentioned in section 1.2: The objectives of the present research work are, in broad terms:

1. To assess the present status of knowledge related to class of proteins known as knot proteins, with specific reference to computational analysis and recognition

2. To collate the known physico chemical parameters which characterise knottedness in proteins

3. To conduct in-silico investigations to bring out computationally derived features from sequence data, which have potential in serving as characteristic signatures

4. To design signature feature vector incorporating known physico chemical parameters and also computationally derived features to design soft computing tools that can predict knottedness
5. To compile dataset of knot and unknot proteins and process them into feature vectors so as to train the soft computing predictor tools

6. To evaluate the performance of the soft computing predictor

7. To compile unanswered questions that arose from the current research work as a pointer for future research

In the next section, we summarise and conclude the work in the backdrop of the original objectives.

### 6.3 Summary and Conclusions

The present research began with a review of work done currently on knottedness in protein. The major observation in this regard is that the sequence based studies on knottedness has not happened in critical volumes. To compound this, the number of sequences confirmed as knot proteins is also small and only growing very slowly. Hydrophobicity, folding rate and packing density were features that have been investigated in relation to knottedness. This study has taken up these features and analysed them computationally and derived signature features out of them.

New signature features were derived using computational methods that have been popularly used by the Kerala University Computational biology research group. Spectral analysis of signals mapped from biosequences, cross-correlation with synthetic sequences of designed hydrophobicity patterns, quantitative features derived out of the chaos game representation images were the ones identified as having discriminative power. All the identified features have been compiled into a 11-member feature vector. Existing knot data and carefully chosen unknot dataset were preprocessed and used to train and test soft computing based knot prediction. The accuracy levels of 86-87% show considerable success in this experiment.

The successful results demonstrate yet again that a carefully researched pattern recognition tool can be used to provide a powerful hypothesis to the
biologists in their wet lab investigation.

The investigation also highlights some biologically relevant conclusions which require further study by biologists to integrate it into the existing knowledgescape about proteins. The following conclusions of biological relevance are of special significance:

1. Fourier analysis of knot proteins revealed that knot proteins are of nonhub nature. Spectral content of knot proteins is very less compared to that of unknot proteins.

2. Connectivity of almost all knot proteins is less than 12.

3. Crosscorrelation of knot proteins with synthetic sequences with different levels of hydrophobic content proved the existence of hydrophobic domains in knot proteins.

4. CGRs of these special type of proteins (drawn with 4 sided polygons by different order of hydrophobic percentages of amino acids) exhibited signature patterns.

5. Folding rate of knot proteins are very less compared to that of unknot proteins.

6. Packing density is high for knot proteins.

7. Signature feature of knots are organism specific.

### 6.4 Limitations

In addition to common limitation that arise from the availability of time, resources, computing power etc. (which were not of grave nature), there were some severe limitations that has constrained the present work. Any soft computing tool will be rugged only if trained with sufficient data which is representative of the problem space. The number of knot sequences available in the Protein Data Bank was limited to 278 and there was no confirmed unknot dataset. The investigator had to do some reasoning to compile a
unknot dataset and to balance it with the knot dataset. Organism-specific experiments also could not be explored fully due to lack of sufficient data for organism other than Homosapiens. This limitation can be overcome only when a broader dataset emerges from life science researchers. The present tool could be retrained on such an occasion. As the percentage of knot proteins in PDB rises from 0.8% (that is, 278 PDB entries having knots from 32853 PDB entries) to higher values, it will strengthen the present research.

The accuracy of the tool is reported based on correct prediction in the test dataset which is created out of this 278 sequences already confirmed as knots. Ideally the tool should have been used to suggest new knot protein and confirmed them in the wet lab. As this investigator is purely computational scientist by training, this could not be undertaken.

6.5 Suggestions for further study

Why a protein would have such a complicated knot is still a mystery to be unveiled. Some of the pioneers in this new topic of research Virnau, Mirny, and Kardar suggested that “If knotted proteins are in fact more difficult to degrade, it might also be disadvantageous for most proteins to be knotted in the first place”. Human ubiquitin hydrolase protein contains the five-crossing knot. This protein rescues proteins otherwise marked for destruction. Normally, a protein destined for destruction within a cell is labeled with another protein called ubiquitin. Marked proteins are then shuttled to a cell structure called a proteasome, which takes in the protein and chops it up. Ubiquitin hydrolase, however, intercedes in this process and removes the offending ubiquitin, saving the condemned protein. A proteasome starts its job of degrading a protein by unfolding it—a task accomplished by threading the protein through a narrow pore. Researchers suggest that rescue protein’s own entangled structure makes it resistant to such unfolding and degradation. So investigations on knot proteins will throw light on various function and folding mechanisms of proteins.
Knots are examples of topologically nontrivial structures in proteins with potentially very high biomedical relevance. They may even have some substantial biomedical significance in relation to illnesses such as Parkinson’s disease. Ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) (EC 3.1.2.15) is a deubiquitinating enzyme. UCHL1 is a member of a gene family whose products hydrolyze small C-terminal adducts of ubiquitin to generate the ubiquitin monomer. Expression of UCHL1 is highly specific to neurons and to cells of the diffuse neuroendocrine system and their tumors. It is present in all neuron. A point mutation (I93M) in this gene encoding this protein is implicated as the cause of Parkinson’s disease [137]. Furthermore, a polymorphism (S18Y) in this gene has been found to be associated with a reduced risk for Parkinson’s disease. The gene is also associated with the Alzheimer's disease, and required for normal synaptic and cognitive function [137]. Hence further studies in knot proteins may provide relevant insightful methods in drug discovery for such diseases.

For proteins, knot theory is likely to be useful, because the geometry and motions of protein backbones may be modeled using techniques from knot theory. Two proteins or subsegments of proteins are similar if there is a motion that transforms one into the other while avoiding backbone self-collisions. Knot invariants help to assess the similarity of proteins [86]. This invites the need to develop study about knot invariants and it could be taken up as a future work since they play an important role in finding similarity of proteins.

A significant sequence homology is found among a sizeable group of knotted and unknotted proteins. In this family, knotted members occupy a primary sub-branch of the phylogenetic tree and differ from unknotted ones only by additional loop segments. These “knot-promoting” loops, whose virtual bridging eliminates the knot, are found in various types of knotted proteins. Valuable insight into how knots form, or are encoded, in proteins could be obtained by targeting these regions in future computational studies or excision experiments [109]. Also as these steps are developed in structural level, sequential level studies and experiments will fasten this phylogenetic study
of knot proteins.

The type $5_2$ knot, which was recently discovered in the structure of human ubiquitin hydrolase, slows down translocation by about two orders of magnitude, as compared to the unknotted chain. In contrast to the unknotted chain case, the translocation mechanism of knotted chains involves multiple slippage events. Further studies in this area will show how the presence of a deep knot affects threading of a polypeptide chain [80]. Analyses of folding and stability in knotted proteins have thus far suffered from a lack of unknotted controls. In order to pinpoint the effects of knotting, it would be desirable to compare a knotted protein to a control protein having a similar core structure, but lacking the knot. In the analysis of the knotted RNA methyltransferase YibK, Lim et al. noted that the knot could be resolved (i.e. removed) by altering the connectivity of the protein backbone at two points [21]. A wide variety of corresponding knotted and unknotted protein pairs could be generated. Such pairs of proteins could be valuable in both experimental and computational studies. In addition, if an increase in stability frequently accompanies knotting, then synthetic knotting could become a new method for engineering novel proteins with enhanced stabilities [134].

6.6 Closing remarks

As with every research work, this work also raises more questions than it answered. Knot protein being an area with not much of work done especially in computational analysis, remains a fertile field for further investigations. I hope that, the humble clouds of knowledge that have been raised in this research will cause some rain in the knowledgescape of bioinformatics, though some might just drift away into non-consequence. I enjoyed this knowledge advancing journey and remain automatically rewarded.