Construction of a Database of Secondary Structure Segments and Short Regions of Disorder and Analysis of Their Properties

Yizhi Zhang

Submitted to the faculty of the School of Informatics

in partial fulfillment of the requirements

for the degree

Master of Science in Bioinformatics

in the School of Informatics,

Indiana University

August, 2004

Accepted by the Faculty, Indiana University School of Informatics, in partial fulfillment of the requirements for the degree of Master of Science.

_____

Dr. Keith Dunker, Chair.

_____

Narayanan Perumal, Ph.D.

_____

Snehasis Mukhopadhyay, Ph.D.

Acknowledgments

Most of all I would like to thank my research advisor, **Professor Keith Dunker** of Biochemistry and Molecular Biology, Director of Center for Computational Biology at School of Medicine, IUPUI. He is not only a great scientist with deep vision but also and most importantly a kind person. His trust and scientific excitement inspired me in the most important moments of making right decisions and I am glad to have had the opportunity to work with him.

Interdisciplinary knowledge obtained on my classes in the School of Informatics allowed me to accomplish this project. I want to thank all great people who taught these classes: **Dr. Narayanan Perumal, Dr. Snehasis Mukhopadhyay, Dr. Wei-min Liu**, **Dr. Jeffery Huang** and others.

I also want to thank **Dr. Predrag Radivojac** for his patience and professionalism that were instrumental in this project.

My special thanks to **Dr. Marc Cortese** for being kind and patient while editing this text and political discussions which helped me to relax from writing it.

As always it is impossible to mention everybody who had an impact on this work. Some others are: Dr. Jack Yang, Mr. Andrew Campen, Ms. Mary K O'Neill, etc….

However there are those whose spiritual support is even more important. I thank my Wife Meihua Luo and my son Zhixiang Zhang.

Table of Contents

Abstract

Prediction of the secondary structure of a protein from its amino acid sequence remains an important task. Not only did the growth of database holding only protein sequences outpace that of solved protein structures, but successful predictions can provide a starting point for direct tertiary structure modeling [1],[2], and they can also significantly improve sequence analysis and sequence-structure threading [3],[4] for aiding in structure and function determination. Previous works on predicting secondary structures of proteins have yielded the best percent accuracy ranging from 63% to 71% [5]. These numbers, however, should be taken with caution since performance of a method based on a training set may vary when trained on a different training set. In order to improve predictions of secondary structure, there are three challenges. The first challenge is establishing an appropriate database. The next challenge is to represent the protein sequence appropriately. The third challenge is finding an appropriate method of classification. So, two of three challenges are related to an appropriate database and characteristic features. Here, we report the development of a database of non-identical segments of secondary structure elements and fragments with missing electron densities (disordered fragments) extracted from Protein Data Bank and categorized into groups of equal lengths, from 6 to 40. The number of residues corresponding to the above-mentioned categories is: 219,788 for α-helices, 82,070 for β-sheets, 179,388 for coils, and 74,724 for disorder. The total number of fragments in the database is 49,544; 17,794 of which are α-helices, 10,216 β-sheets, 16,318 coils, and 5,216 disordered regions. Across the

whole range of lengths, α-helices were found to be enriched in L, A, E, I, and R, β-sheets were enriched in V, I, F, Y, and L, coils were enriched in P, G, N, D, and S, while disordered regions were enriched in S, G, P, H, and D. In addition to the amino acid sequence, for each fragment of every structural type, we calculated the distance between the residues immediately flanking its termini. The observed distances have ranges between 3 and 30Å. We found that for the three secondary structure types the average distance between the bookending residues linearly increases with sequence length, while distances were more constant for disorder. For each length between 6 and 40, we compared amino acid compositions of all four structural types and found a strong compositional dependence on length only for the β-sheet fragments, while the other three types showed virtually no change with length. Using the Kullback-Leibler (KL) distance between amino acid compositions, we quantified the differences between the four categories. We found that the closest pair in terms of the KL-distance were coil and disorder (d_{KL} = 0.06 bits), then α-helix and β-sheet (d_{KL} = 0.14 bits), while all other pairs we almost equidistant from one another ($d_{KL} \approx$ 0.25 bits). With the increasing segment length we found a decreasing KL-distance between sheet and coil, sheet and disorder, and disorder and helix. Analyzing hierarchical clustering of length from 6 to 18 for sheet, coil, disorder, and helix, we found that the group coil had the closet proximity among lengths from 6 to 18. The next closest were helix and disorder. The sheet has the most difference among its length from 6 to 18. In group sheet and coil, fragments of length 17 had the longest distance while fragments of length 6 had the longest distance in group disorder and helix.

Introduction

A. Introduction of subject

Proteins are macromolecules (heteropolymers) consisting of 20 different L–α–amino acids, also referred to as residues. Usually a heteropolymer with less about 40 residues is called a peptide. A certain number of residues are necessary to perform particular biochemical functions, Protein sizes range up to several hundred residues in multi-functional proteins. Very large macromolecules can be formed from protein subunits, for example several thousand actin molecules assemble into an actin filament. Large protein complexes with RNA are found in ribosome particles, which are in fact 'ribozymes'.

Proteins are not linear molecules as suggested when we write out a "string" of amino acid sequence. Rather, usually this "string" folds into an intricate three-dimensional structure. It is this three-dimensional structure that allows proteins to function. Thus in order to understand the details of protein function, one must understand protein structure. Protein structure is broken down into four levels. Primary structure refers to the "linear" sequence of amino acids. Secondary structure is "local" ordered structure brought about via hydrogen bonding mainly within the peptide backbone. The two most common secondary structure arrangements are the right-handed **α-helix** and the **β-sheet**, which can be connected into a larger tertiary structure (or fold) by turns and loops of a variety of types. These two secondary structure elements satisfy a strong hydrogen bond network within the geometric constraints of the bond angles ω, φ and ϕ. . Not all

amino acids favor α-helix formation due to rigid constraints of the amino acid side chains. Amino acids such as A, D, E, I, L and M favor the formation of α-helices, whereas, G and P tend to disrupt helicies. This is particularly true for P since it is a pyrrolidine based imino acid (HN=) whose structure significantly restricts movement about the peptide bond in which it is present, thereby interfering with extension of the helix. The disruption of the helix is important as it introduces additional folding of the polypeptide backbone allowing the formation of globular proteins. Sheets can be formed by parallel or, more common, antiparallel arrangement of individual **β-sheets**. β -sheets are composed of two or more different stretches of at least 5-10 amino acids. Folding and alignment of stretches of the polypeptide backbone beside one another form β-sheets which are stabilized by H-bonding between amide nitrogens and carbonyl carbons. However, the H-bonding residues are present in adjacently opposed stretches of the polypetide backbone as opposed to a linearly contiguous region of the backbone in the α-helix. Tertiary structure is the "global" folding of a single polypeptide chain. A major driving force in determining the tertiary structure of globular proteins is the hydrophobic effect. The polypeptide chain folds such that the side chains of the nonpolar amino acids are "hidden" within the structure and the side chains of the polar residues are exposed on the outer surface. Hydrogen bonding involving groups from both the peptide backbone and the side chains are important in stabilizing tertiary structure. The tertiary structure of some proteins is stabilized by disulfide bonds between cysteine residues. Quartenary structure involves the association of two or more polypeptide chains into a multi-subunit

structure. Quartenary structure is the stable association of multiple polypeptide chains resulting in an active unit. Not all proteins exhibit quartenary structure. Usually, each polypeptide within a multisubunit protein folds more-or-less independently into a stable tertiary structure and the folded subunits then associate with each other to form the final structure. Quartenary structures are stabilized mainly by noncovalent interactions; all types of noncolvalent interactions: hydrogen bonding, van der Walls interactions and ionic bonding are involved in the interactions between subunits. In rare instances, disulfide bonds between cysteine residues in different polypeptide chains are involved in stabilizing quartenary structure.

Based on the theory that function of protein is determined by its structure which is thought be encoded by its primary amino sequence, much effort has been made in the area of predicting structures from the amino sequences. One form of predicting the protein structure from the amino acid sequence is the secondary structure prediction. Instead of predicting the full 3-D coordinates of the structure, the task is to predict a sequence of secondary structure based on the amino acid sequence alone. Present work uses the set of secondary structure labels whose size is three (i.e. Helix, Coil, and Sheet).

B. Importance of subject

Prediction of the secondary structure of a protein from its amino acid sequence remains an important task. Not only did the growth of databases holding only protein sequences outpace that of solved protein structures, but successful secondary structure predictions can provide a starting point for direct tertiary

structure modeling [1],[2], and they can also significantly improve sequence analysis and sequence-structure threading [3],[4] which aid structure and function determination.

C. Knowledge gap

Previous works on predicting secondary structures of proteins have achieved percentage accuracies ranging from 63% to 71% [5]. These numbers, however, should be taken with caution since performance of a method based on a training set may vary when trained on a different training set. In order to improve predictions of secondary structure, there are three challenges. The first challenge is establishing an appropriate database. For example, disordered proteins, existing protein structural databases are strongly biased against disorder. As a result, in the previous work [25] just 32 proteins with disorder longer then 40 amino acids were available. Later, about 110 more disordered proteins were added. The next challenge is to represent the protein sequence appropriately. The third challenge is finding an appropriate method of classification. So, two of the three challenges are related to an appropriate database and characteristic features.

Background

A. Related research

The effort of predicting protein secondary structure began even before the structure of first protein was solved by x-ray crystallography. The size of the database collecting those structures is a testament to the fact that there exists recurring shapes representing various parts of the protein. The geometries of these domains are guided by the composition of the amino acid sequence [26]. Initially these recurring shapes were given secondary structure labels by the experts in the area. But this method of labeling introduced subjectivity. In 1983, Kabsch et al. introduced the DSSP program that consistently assigned secondary structure labels to the solved structures. This program bases its method on the hydrogen bonding patterns found in the solved structure. According to DSSP, 8 types of protein secondary structure elements were classified and denoted by letters: H (α-helix), E (extended β -strand), G (3/10 helix), I (5-helix), B (isolated β-strand), T (turn), S (bend) and "_" (coil). The 8 classes are usually simplified to three states, helix (H), sheet (E), and coil (C) by different reduction methods[6]. Thus, the secondary structure prediction can be analyzed as a typical three-state pattern recognition or classification problem, where the secondary structure class of a given amino acid residue in a protein is predicted based on its sequence features. Since the 1970s, many methods have been developed for predicting protein secondary structures. Early works usually relied on the single-residue statistics of various secondary structural elements, for example, the Chou–Fasman method[7] and the Garnier–Osguthorpe–Robson (GOR I) method[8]. Nearly 20 years later, a

significant improvement was made in the PHD method[9], which is a three-level neural network including some machine learning techniques. After the PHD method, many further neural networks and machine learning refinements were developed[10],[11],[12, 13]. Several machine learning approaches have successfully predicted protein secondary structures, and prediction accuracies were further improved. There have been many previous efforts to predict disorder. Perhaps the earliest are methods based on regions of low-complexity. Although many such regions are structurally disordered, the correlation is far from perfect between regions of low sequence complexity and disordered segments (and vice versa) [13]. Likely the strongest evidence for this correlation comes from the fact that low-complexity regions are rarely seen in protein 3D structures [14]. Methods to predict low complexity, like SEG [15] and CAST [16], are thus often used for this purpose. Methods using hydrophobicity can also give hints as to disordered regions, as they are typically exposed and rarely hydrophobic. Regions without regular secondary structure can be predicted by the NORSp (NO Regular Structure) server[17] , however as the authors indicate that such regions are not necessarily disordered. For examples structures such as the Kringle domain (PDB: 1KRN) are almost entirely without regular secondary structure in their native state but they still have tertiary structure wherein the basic building block are coils.

B. Current understanding

The fundamental elements of protein secondary structure are α -helices, β- sheets, coils, and turns. Some methods have been developed for defining various protein secondary structure elements from the atomic coordinates in the Protein Data

Bank (PDB), such as DSSP[18], STRIDE[19], and DEFINE[20]. Recently, there are many facts showing that many functionally important protein segments that appear to adopt regular structure only upon binding to substrates or other proteins [21] [22] ; these segments don't have rigid second structure, they are referred to as floppy, natively disordered, natively unfolded, or loopy[23],[24],[25] .

The current understanding of disorder is that disordered proteins are flexible to allow for more interaction partners and modification sites[21]. It has also been thought that disordered proteins exist to provide a simple solution to having large intermolecular interfaces in a smaller protein. Usage of smaller proteins would also reduce the required cell and genome sizes [26]. It has been demonstrated that having several relatively low affinity linear interaction sites allows for a flexible, subtle regulation and can accounts for specificity with fewer linear motifs types[27]. It has also been noted that protein disorder plays an important role in biology and in diseases mediated by protein misfolding and aggregation [28], [29], [30]. There is no commonly agreed definition about protein disorder. The thermodynamic definition of disorder in a polypeptide chain is the "random coil" structural state. The random coil state can best be understood as the structural ensemble spanned by a given polypeptide in which all degrees of freedom are used within the conformational space. However, even under extremely denaturing solution conditions, such as 8M urea, this theoretical state is not observed in solvated proteins [31], [32], [33]. Proteins in solution thus seem to always keep a certain amount of residual structure.

There is a database of disorder protein and the first tool named PONDR (Predictor Of Naturally Disordered Regions, http://www.pondr.com) designed specifically for prediction of protein disorder.

C. Hypothesis or research question

Based on the assumption that amino acid sequence determines structure, it was proposed that sequence also determines intrinsic disorder as well. Some predictors of order and disorder have been developed. There are two aspects of secondary structure prediction. In the ab initio or single sequence prediction, the test sequence does not exhibit significant similarity to any of the training sequences at the sequence level. This is a limiting factor for the prediction accuracy. On the other hand, if there are closely related sequences, this generally implies their structural similarity, and the predictions are improved by considering an appropriate database and a characteristic feature. In this paper, we address the problem of establishing a database containing helix (H), sheet (E), coil (C), and disorder (D) sequence segments with lengths from 6 to 40, and analyze their amino acid composition and similarity.

D. Intended research project

We intended to construct a database containing helix (H), sheet (E), coil (C), and disorder (D) sequence segments by using a perl program. Then, we will compute the amino acid composition of these four structures with a MATLAB program. After that, we will calculated the Kullback-Leibler (KL) [34] distance between each pair of distributions in different data sets. Finally, we attempted to cluster

length segments of each structural type in order to find optimal groupings and improve prediction accuracy of short disorder regions.

Methods

A. Materials and instruments

Disorder (D) segments with lengths from 6 to 40 residues were extracted from DisProt ((http://divac.ist.temple.edu/disprot/database.php). The helix (H), sheet (E), coil (C) segment data set was constructed based on DSSP (Kabsch and Sander, 1983) secondary structure assignments as described in Linding et al. 2003 (http://www.cmbi.kun.nl/gv/dssp/). We grouped H (α-helix), G (3/10 helix) into H (α-helix); grouped E (extended β -strand), B (isolated β -strand) into E (strand), and T (turn), S (bend), I (5-helix), and "_" (coil) into C (coil) giving. This data set only contains chains from each PDB_ID according to DisProt database. In each length sub-data set we removed all identical segments. All segments of this database were internal.

B. Statistical analysis

1. we calculated the Kullback-Leibler (KL) [34] distance between each pair of distributions $p1$ and $p2$ as

$$D_{kl}(S1,S2) = \sum_{i=1}^{20} p_1(i) \bullet \log_2 \frac{p_1(i)}{p_2(i)}$$

where $p1(i)$ and $p2(i)$ represent relative frequencies of amino acid i in samples $S1$ and $S2$. In all cases, KL distance of s1 to s2 is half KL distance of s1 to s2 plus half KL distance of s2 to s1.

2. KL-distance was also used as a test statistic to evaluate the significance of the differences between the pairs of underlying sample distributions. Using bootstrapping, we tested the null hypothesis that each pair of samples was generated from the same distribution. Estimates of *P*-values were calculated using 5,000 bootstrap iterations.

3. We calculated the standard deviation of each amino acid composition as

$$Sd = \sqrt{\frac{reference * (1 - reference)}{datasize}}$$

where reference is a percent of the amino acid frequency and datasize is all the residue number of data.

4. We calculated the distance of each fragment between bookending residues as

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Where x1, y1, and z1 represent the beginning bookending residue coordination in PDB, x2, y2, and z2 represent the end bookending residue coordination.

5. Hierarchical Clustering

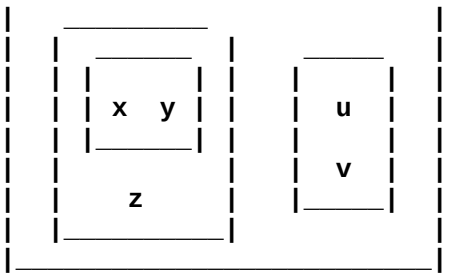
A hierarchical data clustering algorithm yields a multi-level dendrogram. The agglomerative hierarchical clustering (HAC) algorithms operate by maintaining a sorted list of inter-cluster distances. Initially, each data instance forms a cluster. The clustering algorithm repetitively merges the two clusters with the minimum inter-cluster distance. Upon merging two clusters, the clustering algorithm

computes the distances between the newly-formed cluster and the remaining clusters and maintains the sorted list of inter-cluster distances accordingly. There are a number of ways to define the inter-cluster distance: minimum distance (single-link), maximum distance (complete-link), average distance (average-link)

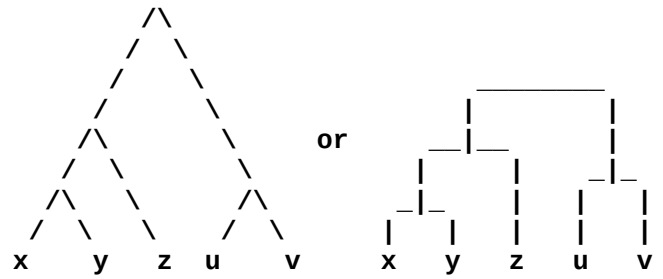
mean distance.

Hierarchical clustering refers to the formation of a recursive clustering of the data points: a partition into two clusters, each of which is itself hierarchically clustered.

One way to draw this is some kind of system of nested subsets, maximal in the sense that one can't identify any additional subsets without violating the nesting:



Alternatively, one can draw a "dendrogram", that is, a binary tree with a distinguished root, which has all the data items at its leaves:



Conventionally, all the leaves are shown at the same level of the drawing. The ordering of the leaves is arbitrary, as is their horizontal position. The heights of the internal nodes are usually related to the metric information used to form the clustering.

The tree is not a single set of clusters, but rather a multi-level hierarchy, where clusters at one level are joined as clusters at the next higher level. This allows you to decide what level or scale of clustering is most appropriate in your application.

We use Matlab hierarchical clustering function and KL-distance matrix to build hierarchical tree. The matlab hierarchical clustering function takes the KL-distance information and links pairs of objects that are close together into binary clusters (clusters made up of two objects). The Matlab hierarchical clustering function then links these newly formed clusters to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree.

Results

1. Comparing sheet, coil, disorder, and helix fragments

Table 1. The data size of four fragments.

Type	Number of fragments	Number of residues
Coil	16,318	179,388
Disorder	5,216	74,724
Sheet	10,216	82,070
Helix	17,794	219,788

In this study, we developed a database of non-identical segments of secondary structure elements and fragments with missing electron densities (disordered fragments) extracted from Protein Data Bank and categorized into groups of equal lengths, from 6 to 40. The number of residues corresponding to the above-mentioned categories was: 219,788 for α-helices, 82,070 for β-sheets, 179,388 for coils, and 74,724 for disorder. The total number of fragments in the database is 49,544; 17,794 of which are α-helices, 10,216 are β-sheets, 16,318 coils, and 5,216 are disordered regions (Table1). Across the whole range of lengths (figure.1), α-helices were found to be enriched in L, A, E, I, and R, β-sheets were enriched in V, I, F, Y, and L, coils were enriched in P, G, N, D, and S, while disordered regions were enriched in S, G, P, H, and D.

Figure.2 shows the amino acid compositions of coil compared in fragment length from 6 to 18. The cysteine and tryptophan are the most depletion in each fragment group, the glycine and proline are the most enriched. With the increase of

fragment length, the amino acid leucine and proline exhibit tendency of more enriched, the amino acid glycine exhibits tendency of more depletion.

The amino acid compositions of disorder are compared in fragment length from 6 to 18 are showed in figure 3. The cysteine and tryptophan are the most depletion in each fragment length, the glycine, Serine and glutamic acid are the most enriched. With the increase of fragment length, the amino acid leucine and isoleusine exhibit tendency of more enriched. The amino acid histidine exhibits tendency of more enriched with fragment length increasing from 6 to 9 but shows tendency of more depletion from length 10 to 18.

Figure.4 shows the amino acid compositions of sheet compared in fragment length from 6 to 18. All fragments are depleted in cysteine and tryptophan, while enriched in valine and leucine. With the increase of fragment length, threonine, glutamine, serine, asparagines, proline, aspartic acid, glutamic acid, and lysine exhibit tendency of being more prevalent; valine, leucine, isoleucine, and cycteine exhibit tendency of being less prevalent.

Figure.5 shows the amino acid compositions of helix compared in fragment length from 6 to 18. Cysteine and tryptophan are the least frequent for each fragment length; leucine and alanine are the most frequent. Cysteine and methionine exhibit tendency of being more prevalent, while proline, and glutamic acid exhibit tendency of being less prevalent with the increase of fragment length.

The thirteen distributions of various lengthy of four structures can also be compared using a more rigorous statistical approach. Because there is little higher order Markov dependence in proteins (Nevill-Manning and Witten 1999), all

segments from each group can be concatenated to form four distinct samples, *Sk* (*k* _ 1. . .4). Each sample *Sk* can be considered a realization of an independent and identically distributed random process that emits symbols from an alphabet of 20 amino-acid codes. To compare the thirteen amino-acid frequency distributions, we calculated the Kullback-Leibler (KL) distance between each pair of distributions *p*1 and *p*2 as described in the method Figure 6 presents the KL-distances. The KL –distance suggest that the two most similar sets are coil and disorder (d_{KL} = 0.06 bits), then α-helix and β-sheet (d_{KL} = 0.14 bits), while all other pairs show equidistant from one another ($d_{KL} \approx 0.25$ bits). With the increase of segment length, there is a tendency of decreasing KL-distance between sheet and coil, sheet and disorder, and disorder and helix.

2. Comparing distance of each length fragment between bookending residues

The twenty distances of each length fragment between bookending residues of the four structures are also compared. The observed distances of all four structure of various length range between 3 and 30Å. In all lengths from 6 to 25 (Figure.7), Sheet had the greatest distance, then helix. Coil had a longer distance than disorder except in segment length of 14, 15, 20, and 25. With increasing to length from 6 to 25, the distance of sheet, helix and coil exhibits tendency to increase. But disorder has the shortest distance in length 18. We also compared distance of disorder between coil, sheet and helix (Figure. 8), the coil and disorder are the closest, then helix and disorder, finally sheet and disorder.

3. Hierarchical clustering of length 6 to 18 for sheet, coil, disorder, and helix.

Hierarchical cluster analysis is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. In the figures of hierarchical clustering, the numbers along the horizontal axis represent the indices of the objects in the original data set, for example L18 is the set of sequences with length of 18. The links between objects are represented as upside down U-shaped lines. The height of the U indicates the distance between the objects taking into account the length of both vertical lines.

Figure 9 shows hierarchical clustering of length from 6 to 18 for sheet. The range of distances is between 2.4×10^{-3} to 44×10^{-3}. L8, L9, L10, L6, L7, L11, L12, L13, and L15 form a closely related cluster with quite close distances between all members. The fragments of length 17 are the most further from the nearest neighbor.

Figure 10 shows hierarchical clustering of length from 6 to 18 for coil. The range of distances is between 0.5×10^{-3} to 4.3×10^{-3}. Fragments of length 9 and 10 have the closest proximity. Sequence of length 7, 8, 11, 13, and 14 have almost the same proximity. The fragment of length 17 is the longest distance from its nearest neighbor.

Figure 11 shows hierarchical clustering of length from 6 to 18 for disorder. The range of distance is between 4.2×10^{-3} and 13.9×10^{-3}. Fragments of length 9 and 10 have the closest proximity. Sequence of length 7, 8, 11, 12, 13, 14, 15, 16, 17 and 18 have almost the same proximity. The fragment of length 6 is the longest distance from its nearest neighbor.

Figure 12 shows hierarchical clustering of length from 6 to 18 for helix. The range of distance is between 0.5×10^{-3} and 4.5×10^{-3}. Fragments of length 13 and 14 have the closest proximity. Sequence of length 7, 8, 9, 10, 11, 12, 15, 16, 17 and 18 have almost the same proximity. The fragment of length 6 is the longest distance from its nearest neighbor.

Analyzing the hierarchical clustering of length from 6 to 18 for sheet, coil, disorder, and helix, we found that the group coil has the closet proximity among its lengths from 6 to 18. The next closest were helix and disorder. The sheet had the most difference among its length from 6 to 18. In group sheet and coil, fragments of length 17 had the longest distance while the fragments of length 6 had the longest distance in group disorder and helix.

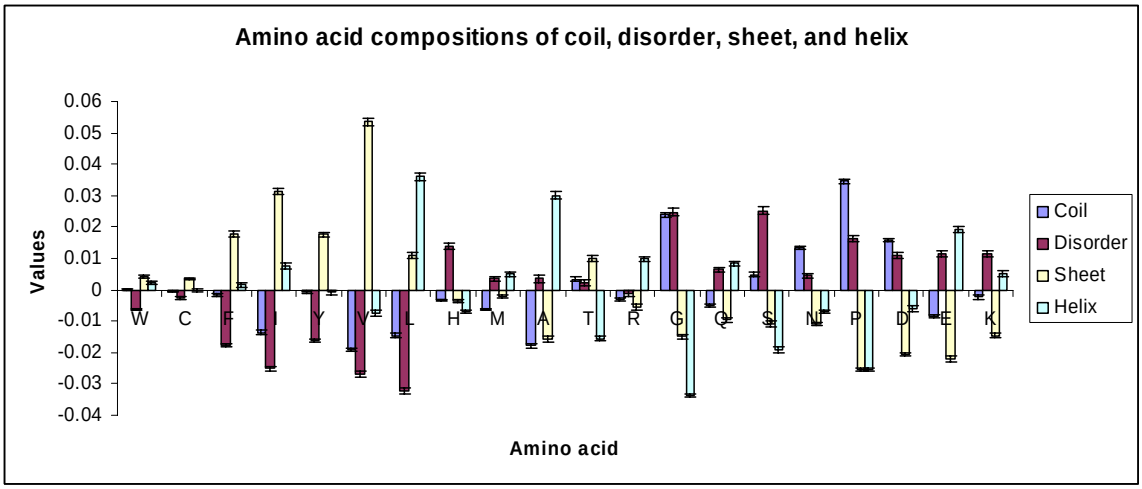


Figure1: Amino acid composition of various data sets. The composition of each amino acid was subtracted from the average composition of the four sets described herein; thus, negative peaks indicate depletions compared to the average of the four reference sets, and positive peaks represent enrichments. The order of the amino acids along the *x*-axis (as in figure 2) is from the most buried (*left*) to the most exposed (*right*) in typical globular proteins.

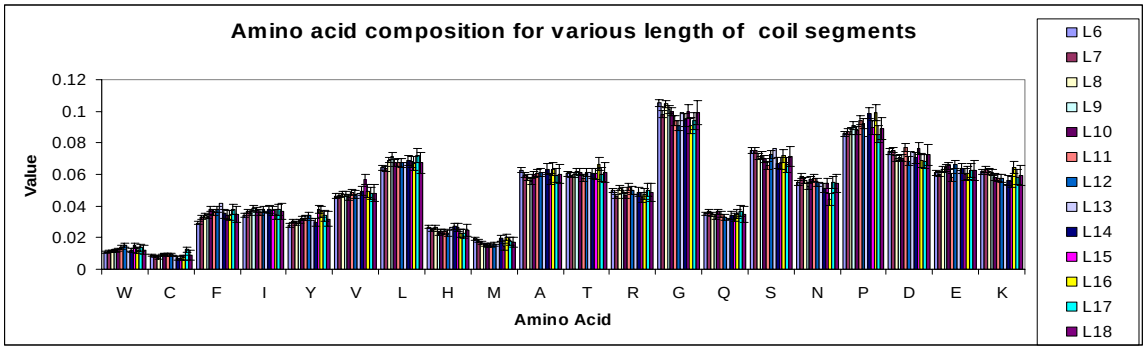


Figure 2. Amino acid compositions of coil for fragment lengths from 6 to 18.

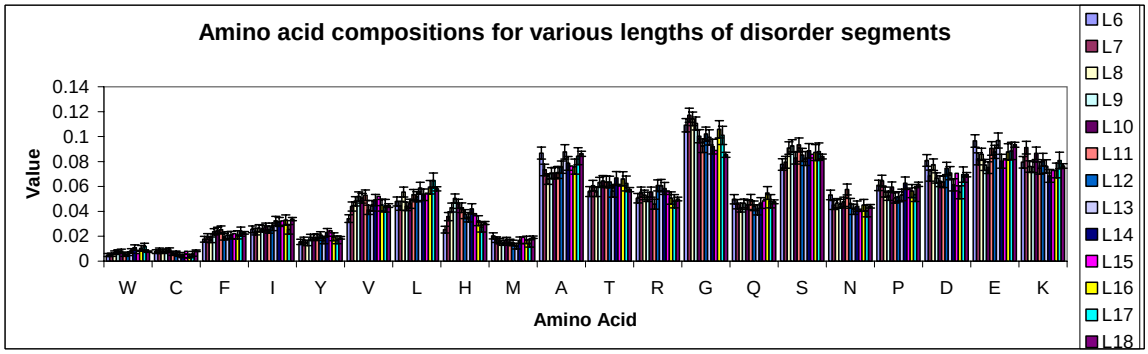


Figure 3. Amino acid compositions of disorder for fragment lengths from 6 to 18.

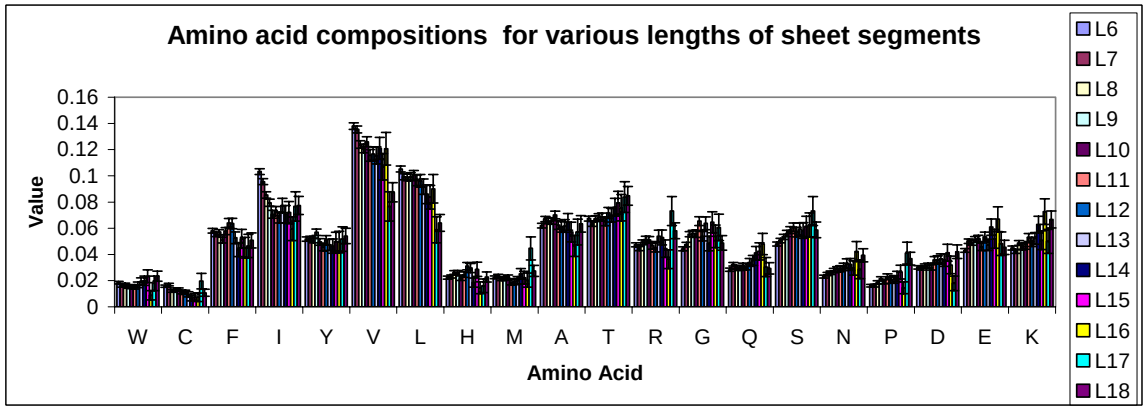


Figure 4. Amino acid compositions of sheet for fragment lengths from 6 to18.

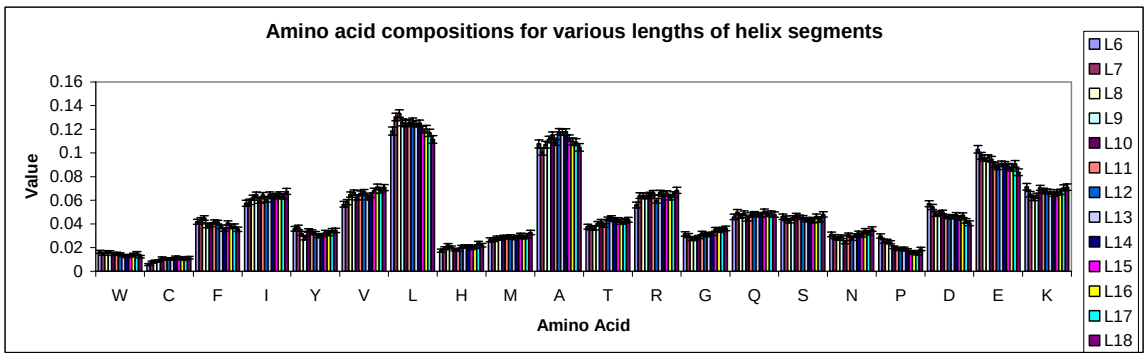


Figure 5. Amino acid compositions of helix for fragment lengths from 6 to 18.

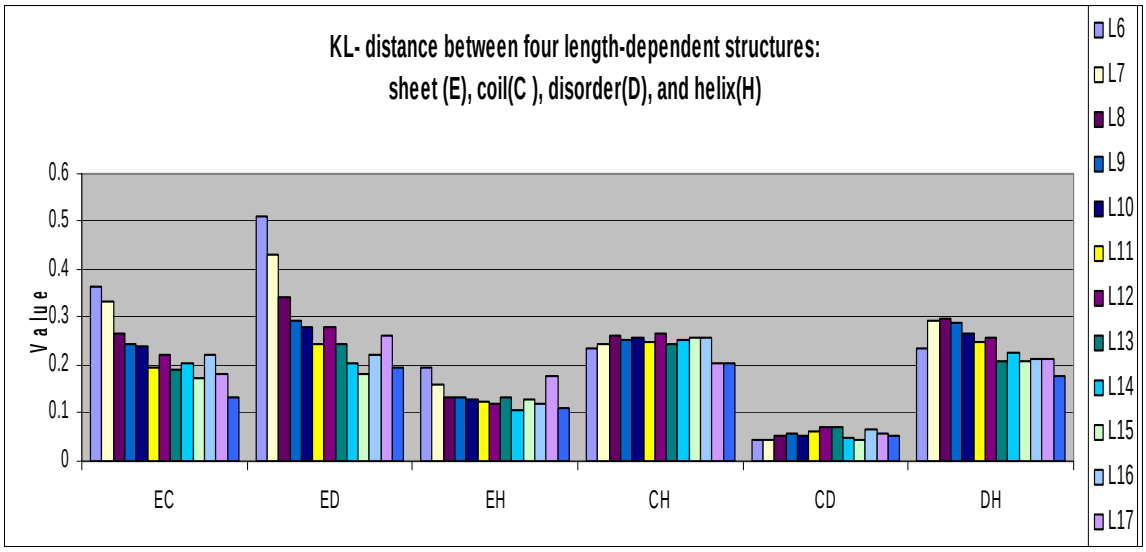


Figure 6. KL distance of sheet, coil, disorder, and helix for fragment length from 6 to 18

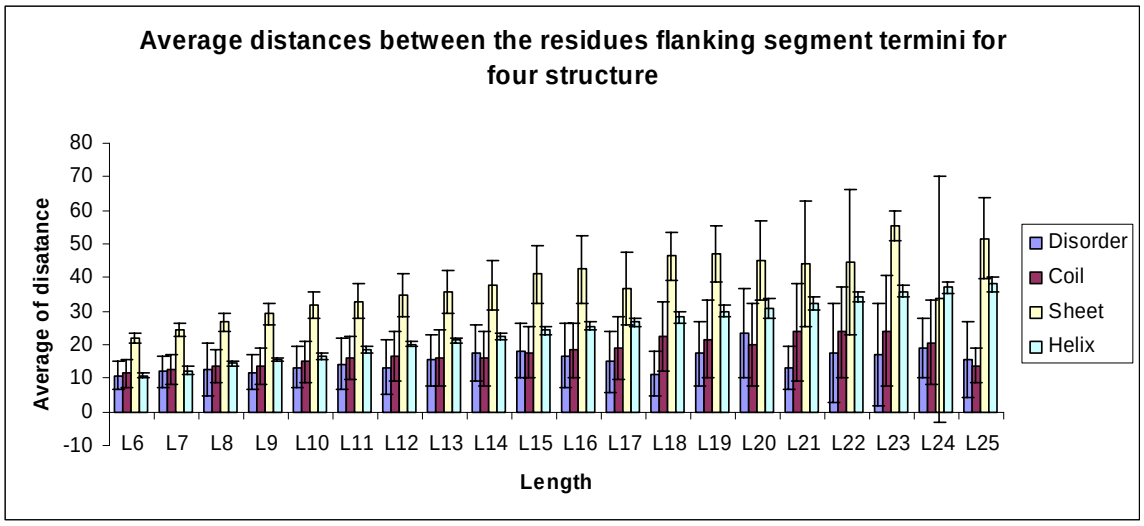


Figure7. Average distance between residues flanking segment termini for disorder, coil, sheet, and helix.

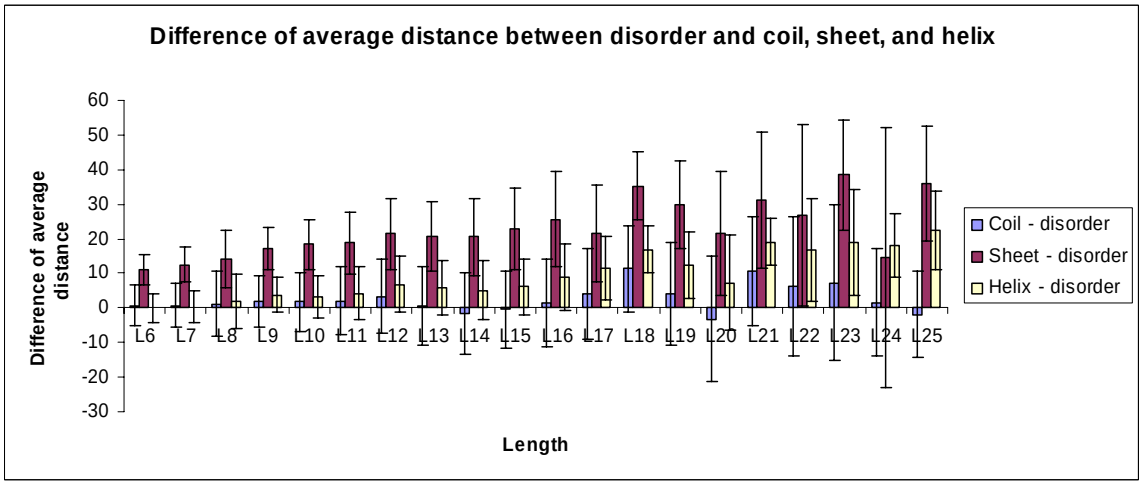


Figure 8. Difference of average bookending residue distances between disorder and coil, sheet and helix for fragment length 6 to 25.

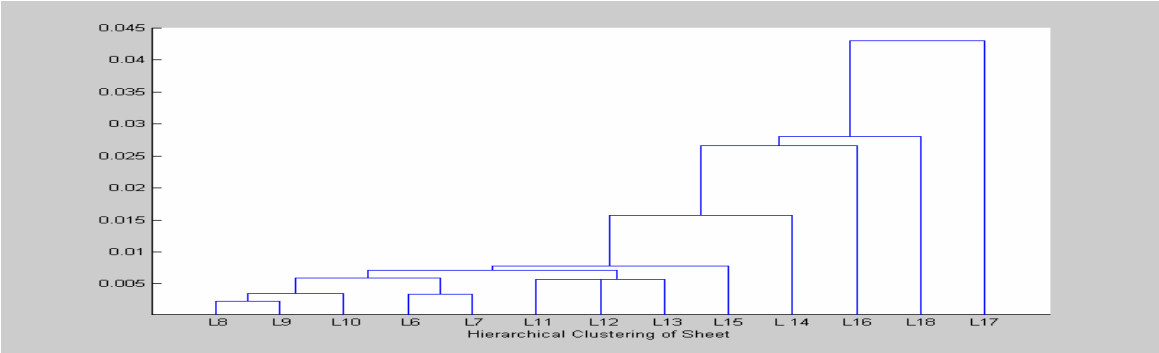


Figure9. Hierarchical clustering of length from 6 to 18 for sheet.

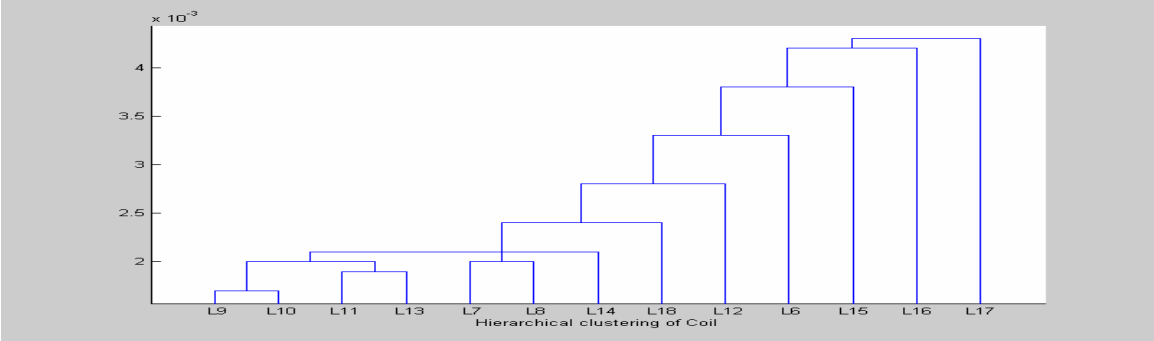


Figure10. Hierarchical clustering of length from 6 to 18 for coil

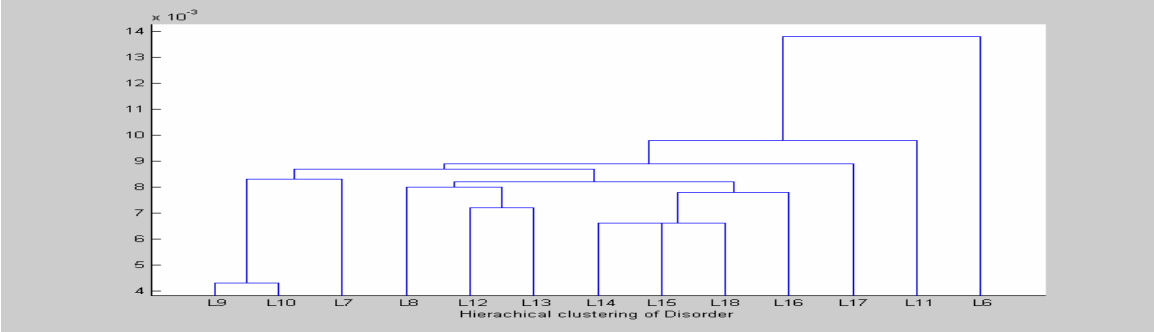


Figure11. Hierarchical clustering of length from 6 to 18 for disorder.

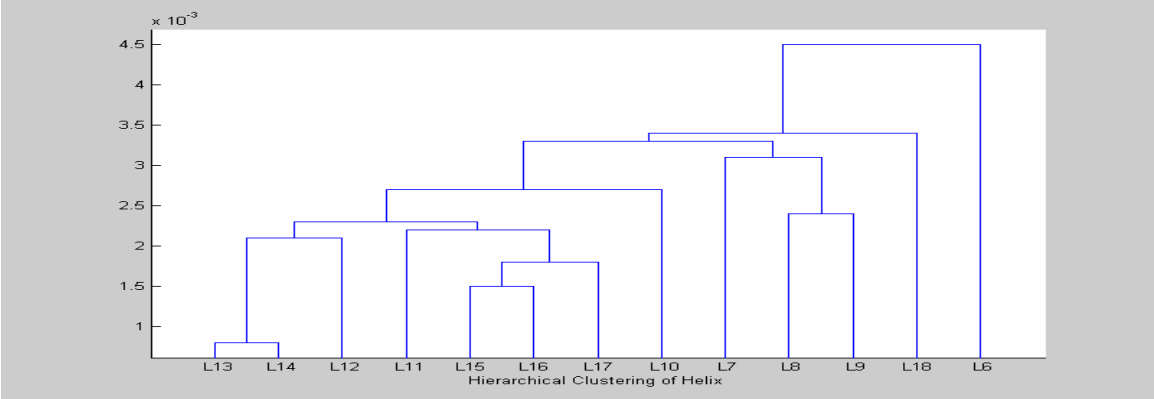


Figure12. Hierarchical clustering of length from 6 to 18 for helix.

Conclusion

Across the whole range of lengths, α-helices were found to be enriched in L, A, E, I, and R, β-sheets were enriched in V, I, F, Y, and L, coils were enriched in P, G, N, D, and S, while disordered regions were enriched in S, G, P, H, and D. The cysteine and tryptophan are the least frequent in all fragments of all four structural types. The helix and sheet have more buried amino acid in typical globular proteins than coil and disorder, while the coil and disorder have more of exposed amino acids. For each length between 6 and 18, we compared amino acid compositions of all four structural types and found a strong compositional dependence on length only for the β-sheet fragments, while the other three types showed virtually no change with length. Using the Kullback-Leibler (KL) distance between amino acid compositions, we quantified the differences between the four categories. We found that the closest pair in terms of the KL-distance were coil and disorder (dKL = 0.06 bits), then α-helix and β-sheet (dKL = 0.14 bits), while all other pairs were almost equidistant from one another ($dKL \approx 0.25$ bits). With the increase of segment length we found a decreasing KL-distance between sheet and coil, sheet and disorder, and disorder and helix.

The observed distances between the residues immediately flanking ranged between 3 and 30Å. We found that for the three secondary structure types the average distance between the bookending residues linearly increases with sequence length from 6 to 25, while it is more constant for disorder. Based on distances between the residues immediately flanking segment termini, we also found the coil and disorders are the closest, then helix and disorder, finally sheet

and disorder. Analyzing hierarchical clustering of length from 6 to 18 for sheet, coil, disorder, and helix, we found that the group coil has the closet proximity among its lengths from 6 to 18. The following were helix and disorder. The sheet had the most difference among lengths from 6 to 18. In group sheet and coil, Fragments of length 17 had the longest distance while fragments of length 6 had the longest distance in group disorder and helix.

Discussion

Recently, it is becoming increasingly clear that many functionally important protein segments appear to adopt regular structure only upon binding to substrates or other proteins[21]; they are referred to as floppy, natively disordered, natively unfolded, or loopy[23],[24],[25]. More than 100 such proteins are found including Tau, Prions, Bcl-2, p53, 4E-BP1 and eIF1A [22],[35]. It seems that these disorder regions are important for function. They are assumed to become ordered only when bound to another molecule (e.g. CREB-CBP complex [36]) or owing to changes in the biochemical environment[37] . Because of their flexibility, the disorder proteins play an important role in the process of molecular recognition, assembly/disassembly, highly-entropy chairs, protein modification.

Protein disorder can be studied by a variety of experimental methods, such as X-ray crystallography, NMR, CD-spectroscopy and hydrodynamic measurements [38]. For example, one class of 'natively disordered' regions was defined as regions missing coordination in X-ray diffraction, presumably since the flexibility keeps them from crystallizing into well-ordered structures. These regions are sometimes associated with regions with 'compositional bias' or 'low sequence complexity' [39], [13], [40]. Another class is characterized by proteins that appear unfolded by CD measurements [41]. In vivo studies of disorder are possible with NMR spectroscopy on living cells (e.g. anti-sigma factor FlgM [42]). Each one of these methods detects different aspects of disorder resulting in several operational definitions of protein disorder.

Based on the theory that function of protein is determined by its structure which is thought be encoded by its primary amino sequence, much effort has been made in the area of predicting structures from the amino sequences. The first tool designed specifically for prediction of protein disorder was PONDR (Predictor Of Naturally Disordered Regions, http://www.pondr.com)[43],[44]. It is based on artificial neural networks. An alternative method is GlobPlot (http://globplot.embl.de) that instead relies on a novel propensity based disorder prediction algorithm [45].

Here, we report the development of a database of non-identical segments of secondary structure elements and fragments with missing electron densities (disordered fragments) extracted from Protein Data Bank and categorized into groups of equal lengths, from 6 to 40. The number of residues corresponding to the above-mentioned categories is: 219,788 for α-helices, 82,070 for β-sheets, 179,388 for coils, and 74,724 for disorder. The total number of fragments in the database is 49,544; 17,794 of which are α-helices, 10,216 β-sheets, 16,318 coils, and 5,216 disordered regions. Across the whole range of lengths, α-helices were found to be enriched in L, A, E, I, and R, β-sheets were enriched in V, I, F, Y, and L, coils were enriched in P, G, N, D, and S, while disordered regions were enriched in S, G, P, H, and D. In addition to the amino acid sequence, for each fragment of every structural type, we calculated the distance between the residues immediately flanking its termini. The observed distances have ranges between 3 and 30Å. We found that for the three secondary structure types the average distance between the bookending residues linearly increases with sequence length,

while distances were more constant for disorder. For each length between 6 and 25, we compared amino acid compositions of all four structural types and found a strong compositional dependence on length only for the β-sheet fragments, while the other three types showed virtually no change with length. Using the Kullback-Leibler (KL) distance between amino acid compositions, we quantified the differences between the four categories. We found that the closest pair in terms of the KL-distance were coil and disorder (d_{KL} = 0.06 bits), then α-helix and β-sheet (d_{KL} = 0.14 bits), while all other pairs we almost equidistant from one another (d_{KL} ≈ 0.25 bits). With the increasing segment length we found a decreasing KL-distance between sheet and coil, sheet and disorder, and disorder and helix. Analyzing hierarchical clustering of length from 6 to 18 for sheet, coil, disorder, and helix, we found that the group coil had the closet proximity among lengths from 6 to 18. The next closest were helix and disorder. The sheet has the most difference among its length from 6 to 18. In group sheet and coil, fragments of length 17 had the longest distance while fragments of length 6 had the longest distance in group disorder and helix.

References

1. Friesner, R.A. and J.R. Gunn, *Computational studies of protein folding.* Annu Rev Biophys Biomol Struct, 1996. **25**: p. 315-42.
2. Rost, B., P. Fariselli, and R. Casadio, *Topology prediction for helical transmembrane proteins at 86% accuracy.* Protein Sci, 1996. **5**(8): p. 1704-18.
3. Fischer, D. and D. Eisenberg, *Protein fold recognition using sequence-derived predictions.* Protein Sci, 1996. **5**(5): p. 947-55.
4. Russell, R.B., R.R. Copley, and G.J. Barton, *Protein fold recognition by mapping predicted secondary structures.* J Mol Biol, 1996. **259**(3): p. 349-65.
5. Salzberg, S. and S. Cost, *Predicting protein secondary structure with a nearest-neighbor algorithm.* J Mol Biol, 1992. **227**(2): p. 371-4.
6. Cuff, J.A. and G.J. Barton, *Evaluation and improvement of multiple sequence methods for protein secondary structure prediction.* Proteins, 1999. **34**(4): p. 508-19.
7. Chou, P.Y. and G.D. Fasman, *Prediction of protein conformation.* Biochemistry, 1974. **13**(2): p. 222-45.
8. Garnier, J., D.J. Osguthorpe, and B. Robson, *Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.* J Mol Biol, 1978. **120**(1): p. 97-120.
9. Rost, B. and C. Sander, *Prediction of protein secondary structure at better than 70% accuracy.* J Mol Biol, 1993. **232**(2): p. 584-99.
10. Riis, S.K. and A. Krogh, *Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments.* J Comput Biol, 1996. **3**(1): p. 163-83.
11. Baldi, P., et al., *Exploiting the past and the future in protein secondary structure prediction.* Bioinformatics, 1999. **15**(11): p. 937-46.
12. Chandonia, J.M. and M. Karplus, *New methods for accurate prediction of protein secondary structure.* Proteins, 1999. **35**(3): p. 293-306.
13. Dunker, A.K., et al., *Protein disorder and the evolution of molecular recognition: theory, predictions and observations.* Pac Symp Biocomput, 1998: p. 473-84.
14. Saqi, M.A. and M.J. Sternberg, *Identification of sequence motifs from a set of proteins with related function.* Protein Eng, 1994. **7**(2): p. 165-71.
15. Wootton, J.C., *Non-globular domains in protein sequences: automated segmentation using complexity measures.* Comput Chem, 1994. **18**(3): p. 269-85.
16. Promponas, V.J., et al., *CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts.* Bioinformatics, 2000. **16**(10): p. 915-22.
17. Liu, J., H. Tan, and B. Rost, *Loopy proteins appear conserved in evolution.* J Mol Biol, 2002. **322**(1): p. 53-64.
18. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
19. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment.* Proteins, 1995. **23**(4): p. 566-79.

20. Richards, F.M. and C.E. Kundrot, *Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure.* Proteins, 1988. **3**(2): p. 71-84.
21. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J Mol Biol, 1999. **293**(2): p. 321-31.
22. Tompa, P., *Intrinsically unstructured proteins.* Trends Biochem Sci, 2002. **27**(10): p. 527-33.
23. Dunker, A.K. and Z. Obradovic, *The protein trinity--linking function and disorder.* Nat Biotechnol, 2001. **19**(9): p. 805-6.
24. Namba, K., *Roles of partly unfolded conformations in macromolecular self-assembly.* Genes Cells, 2001. **6**(1): p. 1-12.
25. Zetina, C.R., *A conserved helix-unfolding motif in the naturally unfolded proteins.* Proteins, 2001. **44**(4): p. 479-83.
26. Gunasekaran, K., et al., *Extended disordered proteins: targeting function with less scaffold.* Trends Biochem Sci, 2003. **28**(2): p. 81-5.
27. Evans, P.R. and D.J. Owen, *Endocytosis and vesicle trafficking.* Curr Opin Struct Biol, 2002. **12**(6): p. 814-21.
28. Schweers, O., et al., *Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure.* J Biol Chem, 1994. **269**(39): p. 24290-7.
29. Kaplan, B., V. Ratner, and E. Haas, *Alpha-synuclein: its biological function and role in neurodegenerative diseases.* J Mol Neurosci, 2003. **20**(2): p. 83-92.
30. Bates, G., *Huntingtin aggregation and toxicity in Huntington's disease.* Lancet, 2003. **361**(9369): p. 1642-4.
31. Shortle, D. and M.S. Ackerman, *Persistence of native-like topology in a denatured protein in 8 M urea.* Science, 2001. **293**(5529): p. 487-9.
32. Ackerman, M.S. and D. Shortle, *Robustness of the long-range structure in denatured staphylococcal nuclease to changes in amino acid sequence.* Biochemistry, 2002. **41**(46): p. 13791-7.
33. Klein-Seetharaman, J., et al., *Long-range interactions within a nonnative protein.* Science, 2002. **295**(5560): p. 1719-22.
34. Radivojac, P., et al., *Protein flexibility and intrinsic disorder.* Protein Sci, 2004. **13**(1): p. 71-80.
35. Uversky, V.N., *Natively unfolded proteins: a point where biology waits for physics.* Protein Sci, 2002. **11**(4): p. 739-56.
36. Radhakrishnan, I., et al., *Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions.* Cell, 1997. **91**(6): p. 741-52.
37. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
38. Smyth, E., et al., *Solution structure of native proteins with irregular folds from Raman optical activity.* Biopolymers, 2001. **58**(2): p. 138-51.
39. Wootton, J.C. and S. Federhen, *Analysis of compositionally biased regions in sequence databases.* Methods Enzymol, 1996. **266**: p. 554-71.
40. Dunker, A.K., et al., *Intrinsically disordered protein.* J Mol Graph Model, 2001. **19**(1): p. 26-59.

41. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-27.
42. Dedmon, M.M., et al., *FlgM gains structure in living cells.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12681-4.
43. Garner, E., et al., *Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization.* Genome Inform Ser Workshop Genome Inform, 1998. **9**: p. 201-213.
44. Garner, E., et al., *Predicting Binding Regions within Disordered Proteins.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 41-50.
45. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder.* Nucleic Acids Res, 2003. **31**(13): p. 3701-8.