

A Robust Grouping Algorithm for Clustering of Similar Protein Folding Units

¹Li, Z., ¹Brener, N.E., ^{1*}Iyengar, S.S.,
¹Seetharaman, G., ¹Dua, S., ²Ramakumar, S.,
²Manikandan, K., and ³Barhen, J.

¹Department of Computer Science, Louisiana State
University, Baton Rouge, LA 70803, USA
Phone: (225) 578-1495

²Department of Physics and Bioinformatics Centre,
Indian Institute of Science, Bangalore-560012, INDIA

³CESAR Laboratory, Oak Ridge National Laboratory,
Oak Ridge, TN 37831, USA

*Correspondence should be addressed to:
S.S. Iyengar at iyengar@bit.csc.lsu.edu

ABSTRACT

The properties of a protein depend on its sequence of amino acids and its three-dimensional structure which consists of multiple folds of the peptide chain. If some of the properties depend primarily on the folding structure, then proteins with certain folding units may exhibit properties specific to those units. In that case, a classification of proteins based on folding units would facilitate the selection of proteins with certain desired properties. With this in mind, we propose an efficient clustering algorithm that can be used to classify proteins according to common folding units. Our algorithm has the following steps:

- Represent the protein structure as a series of conformational angles.
- Partition the proteins into fragments (folding units) of a specified size.
- Cluster the fragments into groups.

The use of overlapped substrings makes our unique demographic clustering technique not susceptible to noise and outliers. Preliminary implementation of this algorithm indicates that it has the capability to discover secondary structural elements (folding units) in proteins and can be generalized to large protein data banks. The algorithm has been applied to a set of 20 randomly selected proteins from the Protein Data Bank and a set of 12 non-homologous α/β protein structures from the PDBSELECT. The algorithm not only identifies the secondary structural elements such as α -helices and β -strands, but also uncovers different turn types which link extended and helical structures.

CATEGORY

- Protein Structural Analysis and Prediction

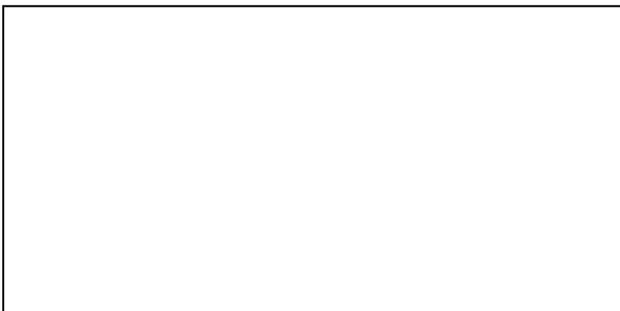
Keywords: Grouping algorithm
Protein folding units

1. INTRODUCTION

A protein is a sequence of amino acids joined by a backbone structure called a peptide chain. In addition to the peptide chain, proteins have a three-dimensional structure which consists of multiple folds of the chain [15,17]. The specific properties of the protein depend on both the amino acid sequence and the folding structure. If some of the properties of a protein depend primarily on the folding structure, then proteins with certain folding units may exhibit properties specific to those units. In that case, a classification of proteins based on folding units would facilitate the selection of proteins with certain desired properties.

The library of protein fragments (referred to as folding units in our study) derived from the experimentally solved proteins structures is shown to be useful in the process of the *ab initio* prediction of the 3D structure of proteins from the primary sequence [3,6,11]. A number of clustering methods have been proposed in the past to identify such representative fragments [8,14].

Currently there is a large quantity of protein structure data available in protein databases [22], and the amount of data is steadily increasing [10]. In order to facilitate the search for common folding units in large protein data banks, we propose a new efficient grouping algorithm derived from demographic clustering techniques used in data mining applications [2]. This algorithm, which is described in detail below, is used to perform case studies on a set of 20 randomly selected proteins from the Protein Data Bank and a set of 12 non-homologous α/β protein structures from the PDBSELECT and the identified clusters are discussed.



2. DATA STRUCTURE REPRESENTATION SCHEME

As mentioned above, a large number of protein 3D structures are now stored in databases, and the number of structure submissions is steadily increasing. Basically, the protein data banks store the protein's atomic coordinates, as derived from crystallographic studies. Although these coordinates contain the structure information precisely, they are not the best representation for detecting similar folds.

A common way of reducing the number of parameters needed to describe the conformation of a protein backbone is to take advantage of the fact that the backbone contains planar units which are connected at C_α atoms, with six atoms per planar unit. Two adjacent planar units, $(C_{\alpha,i-1}, C_{i-1}, O, N_i, H, C_{\alpha i})$ and $(C_{\alpha i}, C_i, O, N_{i+1}, H, C_{\alpha,i+1})$, are shown in Figure 1. Each C_α atom belongs to two of these planar units. The two adjacent planar units which meet at a C_α atom are free to rotate about the C_α -N or C_α -C bond at the junction. This leads to a wide range of three-dimensional configurations for the protein.

There are a number of ways that the protein backbone can be represented [for example: 5, 16, 21], including the following:

1. Express the backbone as a series of C_α points in 3D space, with 3 coordinates for each point. This is a very precise way to describe the backbone. A large amount of work has been done based on this scheme [12,20]. But this approach demands too much computation to search for common folding units and hence is not applicable for large databases.
2. Classify the conformations that an amino acid can take into several categories and represent them as symbols [23], and implement string alignment to search for similarity between folding units. One of these approaches is to divide the Ramachandran map [18] into domains [19]; another attempt is to divide the whole conformation space directly into subspaces [13]. Then based on string comparison, we can search for similar folds. These representations greatly decrease the computational tasks by simplifying a 3-D problem to a 1-D problem. But there is a contradiction in this scheme: if the number of subspaces is large, then it is not easy to find similar structures; or, if there are only a few subspaces, the comparison will be too inaccurate.

3. Express the backbone as a series of conformational angles ϕ and ψ , where ϕ is the rotation angle of the planar unit about the bond between the C_α atom and the nitrogen atom, i.e., the C_α -N bond, and ψ is the rotation angle of the planar unit about the bond between the two carbon atoms, i.e., the C_α -C bond, as shown in Figure 1. When comparing the similarity of two folding units, we simply compute the difference between each pair of ϕ angles and each pair of ψ angles on the same position in their respective folding units, and then sum up these differences. In this way, we simplify the 3-D problem to a 2-D one while preserving all of the conformational information. This data representation scheme enables the efficient detection of folding similarities and hence will be used in the present study. An added advantage is that, from the (ϕ, ψ) s of a cluster of fragments, it is easy to directly identify different secondary structural elements and turn types [9] represented by that cluster.

The Protein Data Bank (PDB) [1] and the PDBSELECT [7] are archives of experimentally determined three-dimensional structures of proteins. The archives contain the coordinates of each atom in the proteins. We will extract the atomic coordinates of the backbone atoms and use these to compute the dihedral (conformational) angle pairs (ϕ, ψ) .

3. GROUPING ALGORITHM

Once the dihedral angle pairs have been computed, we will use them to search for similar folding units in proteins. The technique we propose to use is based on dividing the protein into fragments of a specified size. For the first study described in this paper, we have selected a fragment length of 8, that is, 8 pairs of dihedral angles. For each protein to be included in the search, we first compute the following series of dihedral angles:

$$\{ (\phi, \psi)_1 (\phi, \psi)_2 (\phi, \psi)_3 (\phi, \psi)_4 (\phi, \psi)_5 \dots (\phi, \psi)_{n-1} \}$$

where n is the number of amino acids used to obtain the fragments and the range of the dihedral angles is -180° to 180° . The peptide chain is then decomposed into a series of overlapping fragments of length 8:

Fragment 1:

$$[(\phi, \psi)_1 (\phi, \psi)_2 (\phi, \psi)_3 (\phi, \psi)_4 (\phi, \psi)_5 (\phi, \psi)_6 (\phi, \psi)_7 (\phi, \psi)_8]$$

Fragment 2:

$$[(\phi, \psi)_2 (\phi, \psi)_3 (\phi, \psi)_4 (\phi, \psi)_5 (\phi, \psi)_6 (\phi, \psi)_7 (\phi, \psi)_8 (\phi, \psi)_9]$$

Fragment 3:

$$[(\phi, \psi)_3 (\phi, \psi)_4 (\phi, \psi)_5 (\phi, \psi)_6 (\phi, \psi)_7 (\phi, \psi)_8 (\phi, \psi)_9 (\phi, \psi)_{10}]$$

....

Then we apply a grouping algorithm, which is based on the demographic clustering technique of data mining [2]. In the following, we treat the fragments as points in a 16-dimensional space. We define the distance between two points A_i and A_j , $DIST(A_i, A_j)$, as

$$DIST(A_i, A_j) = ((\phi_{i1}-\phi_{j1})^2 + (\psi_{i1}-\psi_{j1})^2 + (\phi_{i2}-\phi_{j2})^2 + (\psi_{i2}-\psi_{j2})^2 + \dots + (\phi_{i8}-\phi_{j8})^2 + (\psi_{i8}-\psi_{j8})^2)^{1/2}$$

where

$$A_i = [(\phi_{i1}, \psi_{i1}), (\phi_{i2}, \psi_{i2}), \dots, (\phi_{i8}, \psi_{i8})]$$

$$A_j = [(\phi_{j1}, \psi_{j1}), (\phi_{j2}, \psi_{j2}), \dots, (\phi_{j8}, \psi_{j8})]$$

For every $(\psi_{im}-\psi_{jm})$, if $|\psi_{im}-\psi_{jm}| > 180$, then we will use $360-|\psi_{im}-\psi_{jm}|$, and similarly for $(\phi_{im}-\phi_{jm})$. Let j be the index that labels the groups. We define the center of group j , C_j , as

$$C_j = [(\phi_{j1}, \psi_{j1}), (\phi_{j2}, \psi_{j2}), \dots, (\phi_{j8}, \psi_{j8})]$$

where

$$\phi_{jm} = \sum \phi_{im} / N_j$$

$$\psi_{jm} = \sum \psi_{im} / N_j \quad (i = 1, 2, \dots, N_j; m = 1, 2, \dots, 8),$$

N_j is the number of points in the group, and the sum is over i . Such groups are regarded as folding units in our current work.

Algorithm

Input: A set of points in 16-dimensional space and a distance measure R .

Output: A set of groups into which the points have been divided, where every point in a group is within the distance R of the group center.

Begin:

- I. Start a stack with all of the points in it.
- II. Do an operation "pop up" of a point A_1 , create group 1, with center C_1 equal to A_1 , set N_1 to 1.
- III. While (stack is not empty)
 - {
 - a. Do an operation "pop up" of a point A_p .
 - b. Compute the distances between A_p and each existing group center C_j (suppose we have k groups now, then $1 \leq j \leq k$).
 - c. Suppose when $j = j_{min}$, the distance is a minimum. If $DIST(C_{j_{min}}, A_p) > R$, then create a new group $k+1$, with center C_{k+1} equal to A_p , set N_{k+1} to 1. Else
 1. Insert A_p into group j_{min} , add 1 to $N_{j_{min}}$.
 2. Compute the new center $C'_{j_{min}}$ of group j_{min} .
 3. For $i = 1, 2, \dots, N_{j_{min}}$
 - }

- i. Re-compute the distance $DIST(A_{j_{min}, i}, C'_{j_{min}})$ between the point $A_{j_{min}, i}$ in group j_{min} and the new group center $C'_{j_{min}}$.
- ii. If $DIST(A_{j_{min}, i}, C'_{j_{min}}) > R$, push $A_{j_{min}, i}$ into the stack, subtract 1 from $N_{j_{min}}$, go to step 2.

- IV. For each group, re-calculate the distances between the contained points and all of the group centers. If there is any point that has a shorter distance with another group center than with its own group center, move it to the other group where the distance is shorter. If there are no such points, go to END.
 - V. Re-compute all the group centers. If any point is no longer within distance R of the center of its group, push it into the stack. If there are points in the stack, go back to step III. If there are no points in the stack, go back to step IV.
- END**

4. CASE STUDIES

In this section we show how our grouping algorithm can be applied to a set of proteins. The software programs have been implemented in the C language on a high-performance computer system. To test our algorithm, we conducted the following two case studies:

Case Study A: 20 Randomly Selected Proteins from the PDB

In this study, we randomly selected 20 proteins from the PDB. These proteins have different numbers of amino acids and are from different protein families. Only the amino acids with good resolutions are chosen for computing the fragments. Table 1 shows the 20 proteins that were selected and the number of points (fragments) derived from each one.

In our test, we set R (the maximum allowed distance from the center of a group) to 240° . Although R seems to be large in this case, it will be significantly decreased if we include more proteins in the study. We obtained a total of 3083 points from these 20 proteins and used our algorithm to group them into 1734 groups. The group center is the average of the coordinates of all the points in the group and thus is usually not an actual fragment from one of the proteins. Therefore, in order to represent each group more reasonably, we choose the fragment that is closest to the group center. Table 2 gives the five largest groups, labeled A, B, C, D, E, and Figure 2 shows the fragments (folding units) that have the minimum distance from the centers of these groups. The table shows that for

each group, the fragments are from different proteins, which means that our algorithm is capable of efficiently detecting common folding units in a set of proteins.

Case Study B: 12 Non-Homologous α/β Proteins from the PDBSELECT

A small set of 12 non-homologous α/β protein structures was selected from the PDBSELECT April 2003 list [7]. For a residue to be part of a fragment, the torsion angle defining atoms (N, C $_{\alpha}$ and C) of the residue should have the B-factor of less than 60 Å² so that the atoms are well defined in the electron density maps. Any missing residues or atoms are considered as a discontinuity in the polypeptide chain. Accordingly, an input set of 3636 fragments has been derived from the selected 12 proteins (Table 3).

In this test, we set R (the maximum allowed distance of a fragment from its group center) to 240°. In order to ensure that the deviations are more uniformly distributed along the fragment, the maximum allowed deviation for any main-chain dihedral angle from the corresponding angle in the cluster centroid is taken to be 60°. The deviation can be adjusted to do fine clustering (say 30°) or coarse clustering (say 90°) depending upon the interest of the study. With the R value of 240° and the maximum residue level deviation of 60°, our algorithm grouped the 3636 points into 1858 clusters which include single member clusters. Table 4 gives the five largest clusters, labeled A, B, C, D and E and Figure 3 shows the corresponding fragments (folding units) that have the least distance from the centers of these groups. The top ten clusters identified for varying fragment lengths (6 – 9) and their secondary structure descriptions are given in Table 5. Upon testing on various R values (160°, 240° and 320°) for fragment length 8 (FL8), it is noticed that as expected, the number of clusters decreases as R increases. The nomenclature used to designate the conformation of a given residue is according to Efimov [4].

Focusing on clusters with FL8 for detailed discussion (Table 4 and column FL8 of Table 5), it may be seen that all the clusters represent either the regular secondary structural elements or combination of them. In general, the conformation of helical residues is well defined even though deviations are observed for the N- and C- termini of helices compared to the body of the helix. The deviations are more pronounced at the C-terminal region suggesting fraying of the C-terminus [21]). The drift of (ϕ, ψ) from the helical region at the C-terminus presumably maximizes capping interactions. An analysis

such as the one reported here would help not only in modeling helical regions but also the helix termini.

There are two clusters which correspond to two different kinds of termination of α -helices (clusters 4 and 7 of column FL8 of Table 5). In cluster 4, the residue in the α_L conformation is predominantly Glycine and in many cases the residues $\gamma\alpha_L$ interlink a β -strand with the preceding α -helix.

Focusing now on clusters with fragment length 6 (column FL6 of Table 5), it is seen that here only the β -strand emerges as an independent cluster among the first ten, presumably because β -strands in general are shorter secondary structural elements compared to α -helices. Also the standard deviation in (ϕ, ψ) associated with β -strands (β_6) is approximately 19° whereas the corresponding value for α_6 is 6° suggesting that an α -helix is conformationally more rigid compared to a β -strand. This may have some implications for the main-chain conformational entropies associated with different types of secondary structural elements in proteins.

5. CONCLUSION

This paper proposes a unique demographic clustering algorithm that can be used to classify proteins according to similar folding units. Such a classification has the potential to facilitate the selection of proteins with specific desired properties. Preliminary implementation of this algorithm indicates that it has the capability to discover secondary structural elements (folding units) in proteins and can be generalized to large protein data banks.

This novel clustering technique is likely to be useful in generating different sizes of libraries of protein fragments which may be helpful in the design of peptides with required 3D structures. The algorithm may also be used to find preferred conformers either within a structural class or across structural classes of proteins.

6. ACKNOWLEDGMENTS

This work was supported in part by the Materials Science and Engineering Program, DOE Office of Basic Energy Sciences. Additional support was provided by NSF. ORNL is operated by UT-Battelle LLC, under contract DE-AC05-00OR22725.

KM thanks CSIR (India) for a fellowship. Facilities at the Bioinformatics center funded by DBT (India) were used and are gratefully acknowledged.

7. REFERENCES

- [1] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures, *J. Mol. Biol.*, **112**, 535-542; www.pdb.bnl.gov.
- [2] Cabena, P., Hadjnia, Stadler, Verhees and Zanasi (1997) Discovering Data Mining – From Concept to Implementation. Prentice Hall PTR.
- [3] de Brevern, A. G. and Hazout, S. (2003) Hybrid protein model for optimally defining 3D protein structure fragments, *Bioinformatics*, **19**, 345 – 353.
- [4] Efimov, A. V. (1991) Structure of α - α -hairpins with short connections, *Protein Engg*, **4**, 245 – 250.
- [5] Flocco, M. M. and Mowbray, S. L. (1995) C-alpha based torsion angles: A simple tool to analyze protein conformational changes, *Protein Sci*, **4**, 2118 – 2122.
- [6] Haspel, N., Tsai, C. J., Wolfson, H. and Nussinov, R. (2003) Reducing the computational complexity of protein folding via fragment folding and assembly, *Protein Sci*, **12**, 1177 - 1187.
- [7] Hobohm, U. and Sander, C. (1994) Enlarged representative set of protein structures, *Protein Sci*, **3**, 522 - 524.
- [8] Hunter, C. G. and Subramaniam, S. (2003) Protein fragment clustering and canonical local shapes, *Proteins*, **50**, 580 – 588.
- [9] Hutchinson, E. G. and Thornton, J. M. (1996) PROMOTIF--a program to identify and analyze structural motifs in proteins, *Protein Sci*, **5**, 212 - 220.
- [10] Iyengar, S. S. (1998) Computer modeling and simulations of complex biological systems. Iyengar, S. S. (editor). CRC Press, Boca Raton, Florida.
- [11] Kolodny, R., Koehi, P., Guibas, L. and Levitt, M. (2002) Small libraries of protein fragments model native protein structures accurately, *J. Mol. Biol*, **23**, 297 – 307.
- [12] Levitt, M. and Chothia, C. (1976) Structural patterns in globular proteins, *Nature*, **261**, 552 - 558.
- [13] Matsuda, H., Taniguchi, F. and Hashimoto, A. (1997) An approach to detection of protein structural motifs using an encoding scheme of backbone conformations, *Pac. Symp. Biocomput1997*, 280 – 291.
- [14] Micheletti, C., Seno, F. and Maritan, A. (2000) Recurrent oligomers in proteins: An optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies, *Proteins*, **40**, 662 – 674.
- [15] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*, **247**, 536-540.
- [16] Oldfield, T. J. and Hubbard, R. E. (1994) Analysis of C-alpha geometry in protein structures, *Proteins*, **18**, 324 – 337.
- [17] Orengo, C. (1994) Classification of protein folds, *Curr. Opin. Struc. Biol*, **4**, 429 -440.
- [18] Ramachandran, G. N. and Sasisekharan, V. (1968) Conformation of polypeptides and proteins, *Advan. Protein Chem*, **23**, 283 - 437.
- [19] Rومان, M. J., Kocher, J. A. and Wodak, S. J. (1991) Prediction of protein backbone conformation based on seven structure assignments, *J. Mol. Biol.*, **221**, 961 - 979.
- [20] Russel, R. B. and Barton, G. J. (1993) Multiple protein sequence alignment from tertiary structure comparisons: Assignments of global and residue level confidences, *Proteins*, **14**, 309 - 323.
- [21] Soman, K. V., Karimi, A. and Case, D. A. (1991) Unfolding of an alpha-helix in water, *Biopolymers*, **31**, 1351 – 1361.
- [22] Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W. F., Weissig, H., Greer, D. S., Bourne, P. E. and Berman, H. M. (2002), The Protein Data Bank: Unifying the archive, *Nucleic Acids Research*, **30**, 245 – 248.
- [23] Wintjens, R, Wodak, S. J. and Rومان, M. (1998) Typical interaction patterns in $\alpha\beta$ and $\beta\alpha$ turn motifs, *Protein Engg*, **11**, 505 – 522.

8. TABLES AND FIGURES

Table 1. A short list of proteins that were randomly selected for the demographic clustering of dihedral angles in the peptide chain. For each protein, the table shows the amino acids that were selected and the number of points that

PDB Entry	Name of the Protein	Amino Acids Selected	Points Derived
1ash	HEMOGLOBIN (DOMAIN ONE)	1 – 146	138
1bsr	RIBONUCLEASE(BOVINE, SEMINAL) (CHAIN A)	1 – 124	115
1cca	CYTOCHROME C PEROXIDASE	4 – 294	282
1cew	CYSTATIN	9- 116	99
1clm	CALMODULIN (PARAMECIUM TETRAURELIA)	4 – 147	135
1crn	CRAMBIN	1 – 46	37
1ctt	CYTIDINE DEAMINASE	4 – 294	285
1erb	RETINOL BINDING PROTEIN COMPLEX WITH N-ETHYL RETINAMIDE 2	2 – 174	164
1fut	RIBONUCLEASE F1	1 – 107	98
1hng	CD2 (RAT) (CHAIN B)	2 – 176	166
1hoe	ALPHA-*AMYLASE INHIBITOR HOE-467*A	1- 74	65
1lbu	HYDROLASE METALLO (ZN) DD-PEPTIDASE	1 – 213	204
1mka	BETA-HYDROXYDECANOYL THIOL ESTER DEHYDRASE (CHAIN A)	1- 171	162
1mng	MANGANESE SUPEROXIDE DISMUTASE (CHAIN A)	1 – 203	194
1pkp	RIBOSOMAL PROTEIN S5	4 – 148	137
1udi	URACIL-DNA GLYCOSYLASE	18 – 244	218
1utg	UTEROGLOBIN(OXIDIZED)	1 – 70	61
1yal	CARICA PAPAYA CHYMOPAPAIN	1 – 218	209
2vab	MHC CLASS I H-2KB HEAVY CHAIN	1 – 274	265
5pti	TRYPSIN INHIBITOR	1 – 58	49

Table 2. The top 5 groups detected by our grouping algorithm. For each group, the table gives the coordinates of the group center, the number of points in the group, the point nearest the group center, and the number of points derived from the various proteins.

Group Name	A	B	C	D	E
ϕ_1	-67.8	-118.0	-105.2	-81.8	-80.7
ψ_1	-39.1	139.9	127.0	132.4	-36.7
ϕ_2	-67.0	-117.6	-120.8	-64.6	-107.8
ψ_2	-37.3	139.6	141.4	48.4	106.2
ϕ_3	-67.2	-120.3	-119.2	-63.2	-106.6
ψ_3	-38.6	140.4	126.9	-29.4	130.9
ϕ_4	-67.3	-118.3	-120.8	-72.1	-114.4
ψ_4	-38.1	139.2	138.6	-35.5	130.7
ϕ_5	-68.1	-113.8	-115.5	-71.7	-102.6
ψ_5	-36.6	137.6	143.2	-34.6	119.9
ϕ_6	-65.8	-111.5	-113.8	-66.7	-105.7
ψ_6	-36.1	134.9	132.3	-32.5	121.1
ϕ_7	-68.1	-113.7	-85.4	-69.5	-104.5
ψ_7	-35.2	128.4	132.0	-32.4	112.3
ϕ_8	-70.8	-112.4	-15.5	-72.9	-105.5
ψ_8	-31.6	141.0	-29.6	-30.5	127.1
Points in the Group	202	109	42	40	38
The Nearest Points	1mka81-90	1cew92-101	1hng75-84	1udi133-142	1mka121-130
Sources of Points	1ash: 1 1bsr: 11 1cew: 15 1mka: 16 1mng: 60 1udi: 43 2vab: 56	1bsr: 16 1cew: 13 1hng: 27 1mka: 11 1mng: 1 1udi: 2 2vab: 39	1bsr: 2 1cew: 2 1hng: 11 1mka: 7 1mng: 3 1udi: 3 2vab: 14	1bsr: 6 1cca: 1 1cew: 3 1mka: 4 1mng: 9 1udi: 13 2vab: 4	1bsr: 8 1cew: 4 1hng: 7 1mka: 5 1udi: 5 2vab: 9

Table 3. A list of non-homologous α/β proteins used for the case study and the number of 8 residue fragments derived from each protein.

PDB code	Name of the protein	# fragments
1byi_	Dethiobiotin Synthase	208
1g66A	Acetyl xylan esterase II	191
1ga6A	Serine-carboxyl proteinase	353
1gci_	Subtilisin	205
1i1wA	Endo-1, 4-beta-xylanase	286
1ixh_	Phosphate binding protein	305
1muwA	Xylose isomerase	370
1mxtA	Cholesterol oxidase	482
1n55A	Triosephosphate isomerase	233
1o7jA	L-asparaginase	309
1ug6A	Beta-glycosidase	410
7a3hA	Endoglucanase	284

Table 4. The group centers of the top 5 groups detected by our grouping algorithm for a fragment length of 8 (FL8) and the nearest points to the group centers in each group. The root mean square deviations of each position (ϕ , ψ) and the overall deviation are also given.

Group Name	A		B		C		D		E	
ϕ_1	-62.7	6.7	-66.2	10.1	-61.2	5.4	-61.8	4.0	-74.7	14.5
ψ_1	-41.5	7.7	-40.7	9.6	-41.7	6.5	-41.7	4.9	136.6	22.1
ϕ_2	-63.2	5.7	-63.0	4.7	-63.1	6.8	-64.2	4.8	-64.1	12.5
ψ_2	-42.3	6.7	-42.1	5.8	-41.1	5.7	-42.9	5.5	-33.7	12.8
ϕ_3	-62.8	5.3	-63.3	6.3	-63.9	4.6	-61.9	5.0	-63.3	4.7
ψ_3	-42.9	6.2	-41.2	5.1	-42.7	5.5	-43.9	5.0	-37.7	9.4
ϕ_4	-62.7	5.8	-63.5	3.9	-62.2	4.5	-64.1	6.9	-69.8	15.6
ψ_4	-42.8	6.2	-43.2	5.1	-43.1	5.2	-40.2	7.4	-35.9	17.0
ϕ_5	-62.7	4.3	-62.8	4.9	-64.0	6.5	-67.2	8.8	-61.8	8.9
ψ_5	-42.7	5.2	-42.4	5.6	-40.9	7.0	-28.4	8.7	-44.3	10.3
ϕ_6	-62.9	4.3	-63.9	8.6	-66.7	8.3	-91.6	13.2	-61.5	4.8
ψ_6	-42.8	5.1	-39.0	9.9	-28.8	9.1	-0.1	11.5	-41.5	4.9
ϕ_7	-62.6	4.2	-69.4	11.1	-90.3	14.3	79.8	13.3	-64.3	5.3
ψ_7	-42.7	5.4	-28.9	10.1	-0.9	13.2	19.3	13.0	-42.1	7.1
ϕ_8	-63.3	4.9	-94.4	16.1	78.1	13.1	-84.7	18.1	-65.8	10.5
ψ_8	-40.8	7.3	-6.52	14.7	20.3	13.4	139.2	17.5	-40.5	8.1
Fragments in the Group	443	5.8	87	9.0	52	8.7	39	10.3	38	11.6
The Nearest Fragment	1n55A 110A – 117A	7a3hA 85A – 92A	1ug6A 63A – 70A	1o7jA 264A – 271A	1i1wA 244A – 251A					
Description of the fragments	An α - helix [α_8]	An α - helix with type I β turn at the C terminal [α_7 - β I]	An α - helix with type I β turn at the C terminal followed by an α_L residue [α_6 - β I- α_L]	[α_5 - β I- α_L - β]	[β - α_7]					

β I – refers to type I β turn

Table 5. The first 10 ranked clusters identified with the fragment length (FL) varying from 6 to 9 and their corresponding secondary structure combination are listed. Values in parentheses refer to the number of fragments in each cluster.

Cluster Number	FL 9	FL 8	FL 7	FL 6
1	α_9 (367)	α_8 (443)	α_7 (522)	α_6 (626)
2	$\alpha_8 \gamma$ (81)	$\alpha_7 \gamma$ (87)	$\alpha_6 \gamma$ (97)	$\alpha_5 \gamma$ (107)
3	$\alpha_7 \gamma \alpha_L$ (45)	$\alpha_6 \gamma \alpha_L$ (52)	$\alpha_5 \gamma \alpha_L$ (55)	$\beta \alpha_5$ (59)
4	$\alpha_6 \gamma \alpha_L \beta$ (39)	$\alpha_5 \gamma \alpha_L \beta$ (39)	$\beta \alpha_6$ (53)	$\alpha_4 \gamma \alpha_L$ (55)
5	$\beta \alpha_8$ (35)	$\beta \alpha_7$ (38)	$\alpha_4 \gamma \alpha_L \beta$ (39)	β_6 (50)
6	$\alpha_7 \gamma \beta$ (32)	$\alpha_6 \gamma \beta$ (33)	$\beta_6 \alpha$ (32)	β_6 (41)
7	$\beta_2 \alpha_7$ (30)	$\alpha_5 \gamma \beta \gamma$ (26)	$\alpha_5 \gamma \beta$ (30)	$\alpha_3 \gamma \alpha_L \beta$ (39)
8	$\alpha_5 \gamma \alpha_L \beta_2$ (26)	$\beta_2 \alpha_6$ (26)	$\alpha_4 \gamma \beta \gamma$ (29)	$\alpha_4 \gamma \beta$ (39)
9	$\alpha_5 \gamma \alpha_3$ (25)	$\alpha_4 \gamma \alpha_3$ (26)	$\beta_2 \alpha_5$ (28)	$\alpha \gamma \alpha_L \beta_3$ (34)
10	$\alpha_2 \gamma \alpha_6$ (24)	$\alpha_2 \gamma \alpha_5$ (25)	$\alpha \gamma \alpha \beta_4$ (28)	$\beta_2 \alpha_4$ (33)

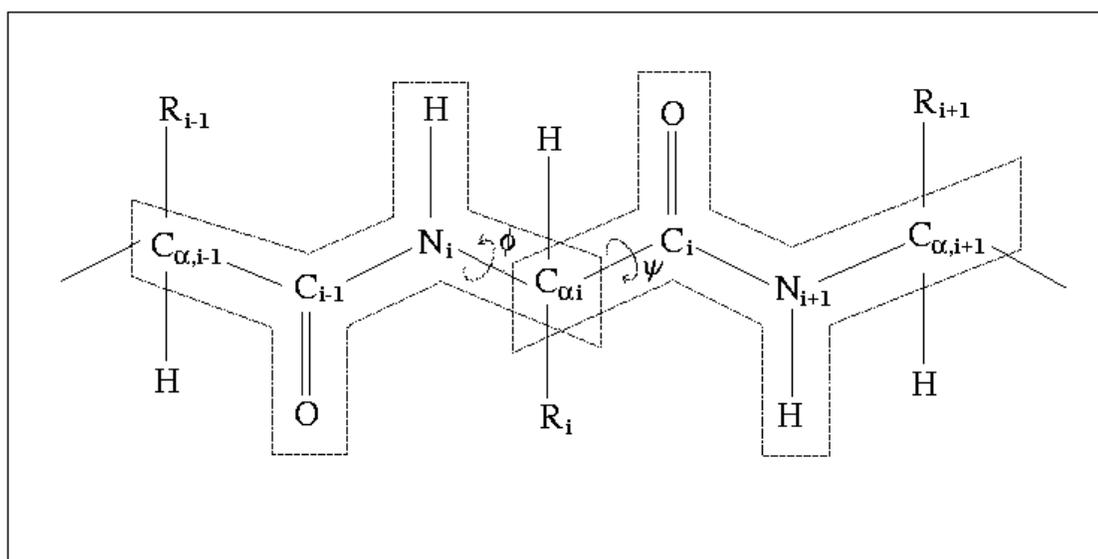
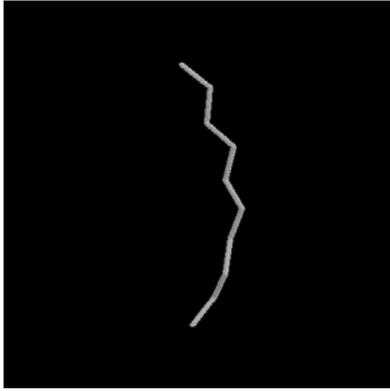
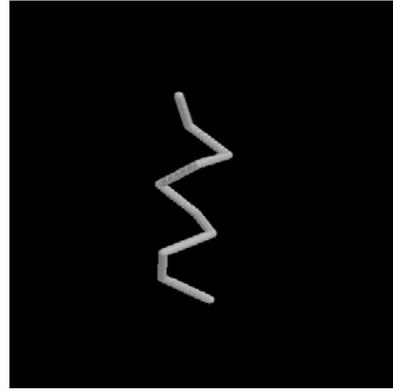


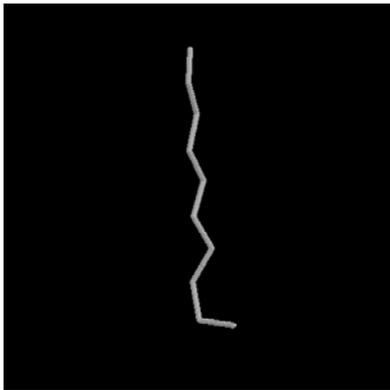
Figure 1. The two rotation angles ϕ and ψ characterize the three-dimensional nature of the protein molecule.



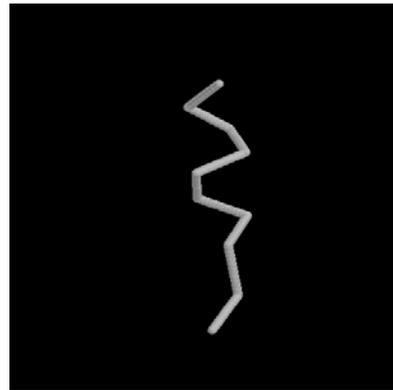
A



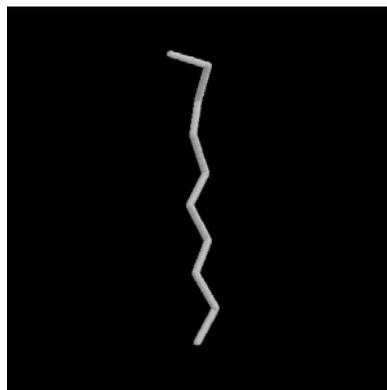
B



C



D



E

Figure 2. Examples of Folding Units produced by the grouping algorithm. Figures A, B, C, D, E show the conformation of the point nearest the center in groups A, B, C, D, E .

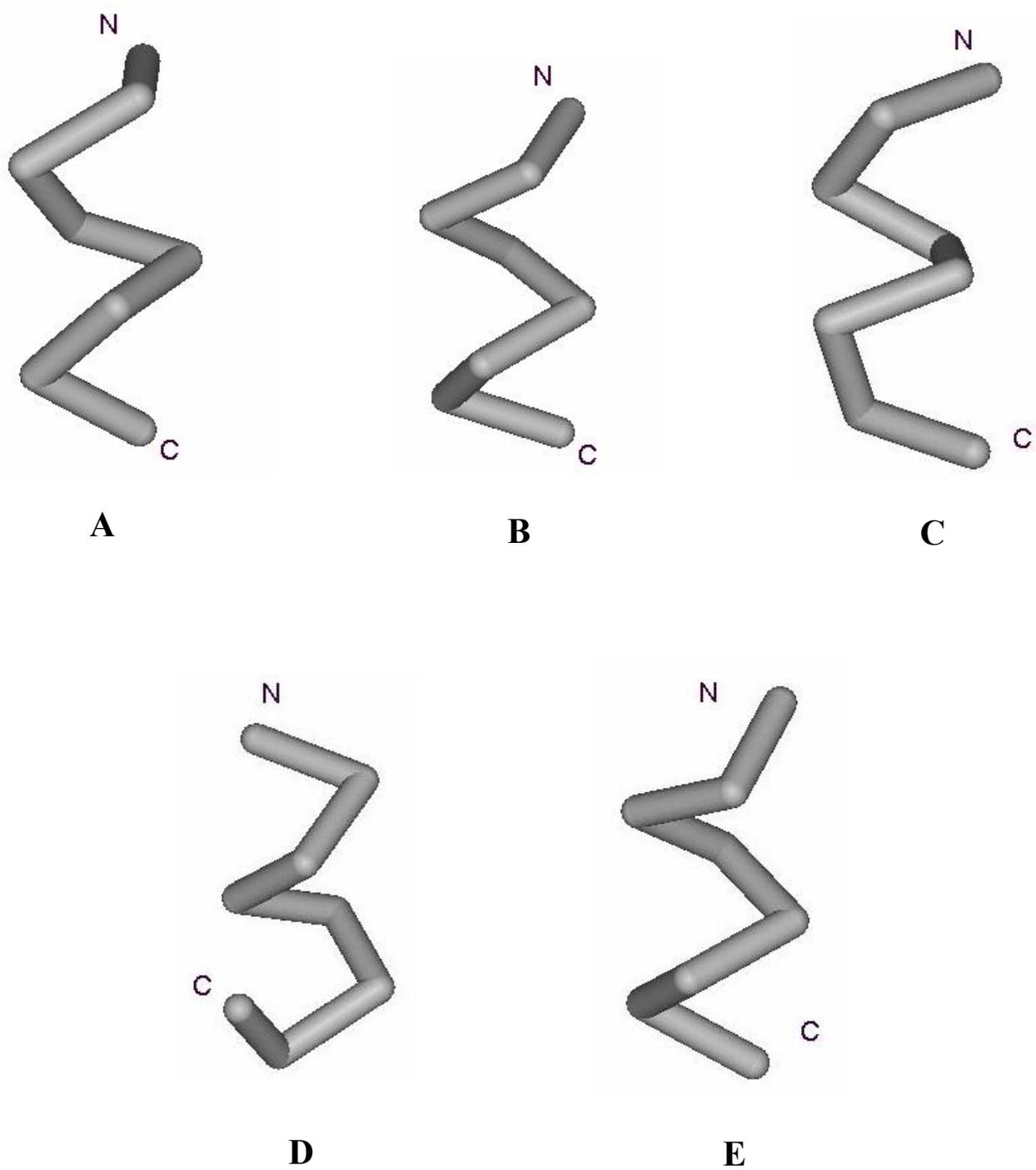


Figure 3. C α traces of nearest fragments for the first five clusters listed in Table 4. The amino and carboxyl terminal ends of the fragments are denoted as N and C respectively.