A Central-3-Residues-Based Clustering Approach for Studying the Effect of Hydrophobicity on Protein Backbone Angles

Ibrahim.M.Al-Henawy
Computer Science department, Faculty of Computers and Informatics, Zagazig University
henawy2000@yahoo.com
Ahmed H.Kamal
Computer Science department, Faculty of Computers and Information, Cairo University
a.kamal@fci-cu.edu.eg
Hisham Al-Shishiny
Research and Development department, IBM Egypt
shishiny@eg.ibm.com
Haitham Gamal
Computer Science department, Faculty of Computers and Informatics, Zagazig University
haitham.gamal@yahoo.com

Abstract
This paper aims to study the effect of the hydrophobicity of protein subsequences on the backbone structure. In this work backbone structure is represented by the sequence of angles between each three consecutive C_α atoms in space. The main idea is based on clustering a large set of protein subsequences – taken from a big number of proteins – using their hydrophobicity patterns as a similarity measurement. The standard probability density function that best fit the observed angle measurements of each of the resulting clusters is determined through a Kolmogorov-Smirnov (KS) test. The resulting fits are then studied in terms of their KS-statistics and the number of rejected critical values in order to determine the relationship between subsequence length, hydrophobicity and backbone angle measurements. It is found that the longer the protein subsequence the higher the possibility of getting a good fit. Also, it is observed that clustered data achieve better fits than unclustered data.

Keywords: K-means clustering, subsequences, Kolmogorov-Smirnov, Protein Folding

1. Introduction

Protein structure prediction has been one of the most challenging problems facing researchers over the few last decades. Exact prediction was found to be too far from today's state of the art. Even using simplified forms such as the Hydrophobic-Polar (HP) model is found to be NP-complete [1]. Several approaches have been proposed to simplify the prediction process. These approaches are either *ab initio* that assumes no prior knowledge about the protein under study or homology methods that use sequence similarity with already known structures to guide the search. Hybrid approaches are also available. Most of the hybrid approaches tend to use double-staged prediction algorithms [16]. In this type, the output of the first stage is taken as an input into the second stage. In many cases the first stage is used to approximately predict the secondary structure of a protein [9] while the second stage continues to approximately predict the tertiary structure [16]. Due to the difficulty of the folding problem, researchers use approximated and simpler protein representations like using lattice structures such as HP models [9] and face-centered-cubic lattice [8,16] and/or using heuristic techniques in order to simplify the calculations [2,7,17]. This study tries to introduce an alternative approach to be used in the first stage through building a probabilistic database of protein backbone angles based on the hydrophobicity of its constituent amino-acids.

The use of subsequence structural information as a step towards the ultimate goal of complete prediction is widely used in literature [4,6,10,14,15]. Another approach is to deal with the protein as a whole [7,11,12,13,16,17] trying to find the optimum conformation with minimum free energy [11]. Statistical analysis of protein subsequences has appeared in literature too. Rong She et al. used two types of subsequence classifiers to identify outer membrane proteins of Gram-negative bacteria [14]. Eran Segale and Daphne Koller introduced a general probabilistic framework for clustering biological data into a hierarchy [5]. Eli Hershkovitz et al. used torsion angles to search for clusters in RNA conformational space [4]. Estimating the probability density function was used by Diego Rother et al. with the notion of ensembles [3]. Marcio Dorn and Osmar Norberto used the ϕ and Ψ angles of the central residue of a subsequence along with a secondary structure prediction method to cluster subsequences (fragments) [10]. The approach proposed here tries to find the best probability distributions fitting the angle measurements of subsequences clustered based on their hydrophobicity. These fits are then analyzed statistically to determine the effect of hydrophobicity and subsequence length on backbone angles. The paper is organized as follows; section 1 explains

the representation used throughout the rest of the paper. Section 2 discusses the proposed approach in steps. Section 3 discusses the experimental results and finally section 4 concludes the study.

2. Representation

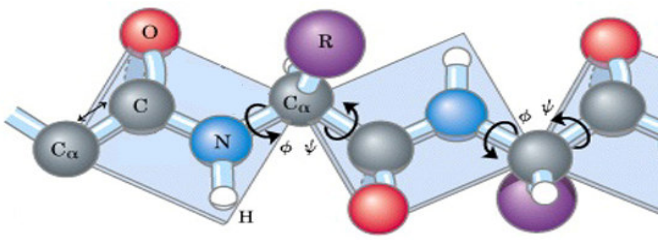

Fig.1 ϕ and Ψ torsion angles

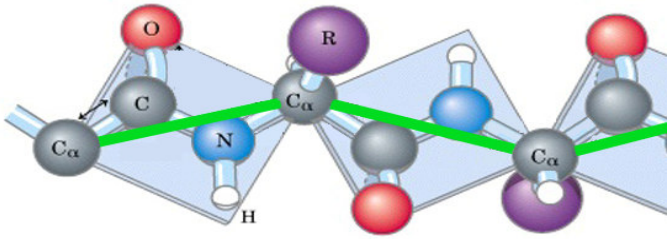

Fig.2 Θ angles

A subsequence of residues is represented by a vector *(v)*. Each residue contains three main consecutive atoms; a central Carbon atom (C_α) surrounded by another Carbon atom *(C)* and a Nitrogen atom *(N)*. A side chain is connected to the central C_α atom. This side chain is what differentiates residues from each other. Each amino acid contains two torsion angles; ϕ and Ψ as shown in fig.1. This study is not concerned with these torsion angles however the main concern is the angle Θ which is the angle between the three consecutive C_α atoms of the three central residues of the subsequence. Θ is the angle between each two lines connecting C_α atoms in fig.2. Thus a subsequence *S* is represented by a vector *v* and an angle Θ:

$$S = (v, \Theta: v = <aa_p, aa_{p+1} \dots aa_{p+n-1}>) \quad (1)$$

Where *p* is the starting position of the subsequence and aa_p represents the amino-acid at position *p*. Notice that the angle taken here is neither the ϕ nor the Ψ angles of the central amino-acid, alternatively one angle is taken to represent the relative positions of every three consecutive amino-acids. A centroid is represented by a simple vector of *n* hydrophobicity values:

$$C = <h_0, h_1, \dots h_{n-1}> \quad (2)$$

3. Proposed approach

In this study we used a sample of 1089 proteins randomly selected from the SCOP protein database. Subsequences of length 3, 5 and 7 are extracted from each protein. K-means clustering is performed on the three groups of subsequences according to their hydrophobicity. Both before and after clustering a KS test is used to fit the observed angles measurements into one of 66 standard continuous probability distributions, which are: Beta, Burr, Burr (4P)[1], Cauchy, Chi-Squared, Chi-Squared (2P), Dagum, Dagum (4P), Erlang, Erlang (3P), Error, Error Function, Exponential, Exponential (2P), Fatigue Life, Fatigue Life (3P), Frechet, Frechet (3P), Gamma, Gamma (3P), Gen. Extreme Value, Gen. Gamma, Gen. Gamma (4P), Gen. Logistic, Gen. Pareto, Gumbel Max, Gumbel Min, Hypersecant, Inv. Gaussian, Inv. Gaussian (3P), Johnson SB, Johnson SU, Kumaraswamy, Laplace, Levy, Levy (2P), Log-Gamma, Log-Logistic, Log-Logistic (3P), Log-Pearson 3, Logistic, Lognormal, Lognormal (3P), Nakagami, Normal, Pareto, Pareto 2, Pearson 5, Pearson 5 (3P), Pearson 6, Pearson 6 (4P), Pert, Phased Bi-Exponential, Phased Bi-Weibull, Power Function, Rayleigh, Rayleigh (2P), Reciprocal, Rice, Student's t,

[1] *2P*, *3P* and *4P* refer to two, three and four parameters distributions respectively. A typical distribution is said to be *(n-1)P* if its location parameter is set to 1 and *(n)P* otherwise.

Triangular, Uniform, Wakeby, Weibull and Weibull (3P). Parameters are estimated using Maximum Likelihood Estimation (MLE). The following points explain the steps of the approach:

1. Initially all proteins are divided into subsequences of residues i.e. amino acids of length *n*. A protein of length *L* is divided into *L-n+1* subsequences starting at (aa_0, … aa_{n-1}) and ending at (aa_{L-n}, … aa_{L-1}). Therefore the total number of subsequences is $\sum_{i=0}^{N}(Li - n + 1)$ where *N* is the total number of proteins. The angle Θ is the angle between the three residues in the center of the subsequence ($aa_{(p+n)/2}$, $aa_{(p+n+2)/2}$, $aa_{(p+n+4)/2}$), where *p* is the start position of the subsequence in the whole protein sequence. The subsequences are overlapping i.e. every two consecutive subsequences of length *n* shares *n-1* residues. The value of Θ is calculated using the coordinates of C_α atoms of these residues in the SCOP database. Obviously the number of residues in a subsequence must be odd so that the number of residues on both sides of the angle is the same. Typical values of *n* used in this study are 3, 5 and 7. Higher values of *n* are possible but they are computationally intensive.

2. Before the clustering algorithm starts, a KS test is performed to fit all the observed – unclustered – measurements into one of the previously listed standard continuous probability distributions.

3. The sets of subsequences generated from step 1 are then fed to the k-means clustering algorithm each at a time. This algorithm uses residues hydrophobicity as a similarity measurement (discussed later).

4. Centroids are pre-known and are based on the value of *n*. When creating the initial centroids each position in the subsequence is assumed to be either hydrophobic[2] (*H*) or hydrophilic[3] (*P*)[4]. Taking the two extremes of hydrophobicity into consideration, namely *Isoleucine* (*+4.5*) and *Arginine* (*-4.5*), leaves us with only two choices for each residue position. Calculating all the permutations of a subsequence of length *n* results in a total of 2^n centroids.

5. Similarity function: Let the hydrophobicity of a residue (*aa*) be (*aa.h*). The similarity function measures how a subsequence *S* is similar to some centroid *C* in terms of hydrophobicity. The function simply calculates the average of differences in hydrophobicity between the residues of *S* and the corresponding hydrophobicity values in *C*.

$$(\sum_{i=0}^{n}(aa_i.h - h_i))/n \quad (3)$$

6. After clustering is completed, another KS test is performed on the observed clustered measurements of each centroid against the same 66 standard continuous probability distributions.

7. The KS statistic as well as the number of average number of rejected critical values (out of five values) is recorded before and after clustering.

4. Experiments and results

Table 1 summarizes the best fitting distributions for all the three values of n. The goodness of these fits will be discussed later in this section.

Table 1 best fitting distribution of each centroid of the three values of *n*

continuous distribution	centroids that the continuous distribution best fits
n = 3	
Burr	1, 4
Burr(4p)	7
Gen. Extreme Value	6
Gen. Pareto	2, 3, 5
Johnson SB	0
n = 5	
Dagum(4p)	0, 5, 7, 19
Gumbel Min.	1, 2, 3, 17, 20
Gen. Extreme Value	4, 32
Burr(4p)	6, 8, 10, 11, 14, 18, 21, 22, 23, 24, 27, 30, 31
Weibull(3p)	9, 12, 13, 15, 16, 25, 26, 28, 29

[2] Having a strong aversion for water
[3] Having a strong affinity for water
[4] *P* here stands for *"polar"* which has the same meaning as *"hydrophilic"*.

n = 7	
Weibull(3p)	3, 21, 79
Burr(4p)	20, 32, 40, 60, 67, 71, 74, 75, 83, 85, 105
Dagum	4, 80
Dagum(4p)	41, 90
Gen. Gamma(4p)	69, 84, 106
Gen. Logistic	2, 6, 7, 9, 12, 14, 15, 19, 33, 34, 35, 36, 37, 45, 46, 47, 49, 79, 87, 89, 94, 95, 107, 117, 125
Gumbel Min.	66
Log-Logistic	42, 116, 118
Wakeby	1, 5, 8, 10, 11, 13, 16, 17, 18, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 38, 39, 43, 44, 48, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 68, 70, 72, 73, 76, 77, 78, 81, 82, 86, 88, 91, 92, 93, 96, 98, 99, 100, 101, 102, 103, 104, 108, 109, 110, 111, 112, 113, 114, 115, 119, 120, 121, 122, 123, 124, 126, 127

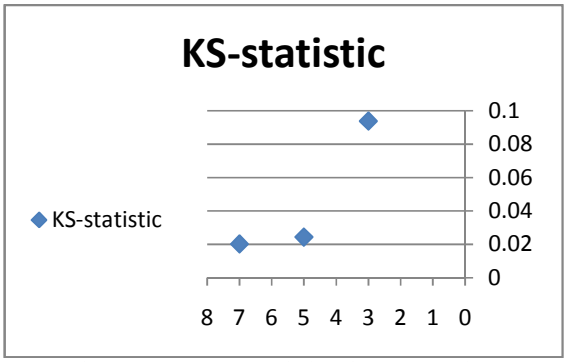

Fig.3 average KS-statistic of clustered data

From fig.3 it is quite obvious that the longer the sequence the smaller the KS-statistic. The values of the statistic are 0.0937, 0.0243 and 0.0202 for subsequences of length 3, 5 and 7 respectively. From fig.4 it is apparent that the number of rejected critical values greatly decreases with longer subsequences. For subsequences of length 3 all the five critical points are rejected for all the centroids. Thus subsequences of length 3 have no reliable fit among the tested distributions. 5 residues centroids have better results in terms of the number of rejected points. An average of 2.94 critical points is rejected among all the centroids. Finally centroids of length 7 achieves an average of zero rejected critical point, i.e. all the critical point for all the centroids of length 7 are accepted. Clearly, the length of the subsequence is effective in terms of the KS-statistic and the number of rejected critical values.

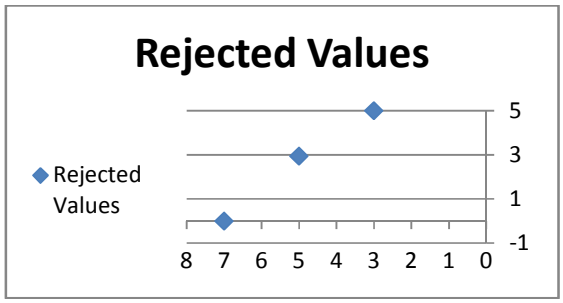

Fig. 4 average number of rejected critical values out of 5

The same KS-test was performed on unclustered data for n=3, n=5 and n=7. For the three values of *n* the best fitting distribution was the *Wakeby* distribution. The value of the test statistics for the three was found to be 0.09041, 0.012 and 0.013 respectively. However these results are not as interesting as they seem to be. Actually all the 5 critical values were rejected for all the values of n for the unclustered data.

5. Conclusion

From the previous results it is now clear that there exists a direct relationship between the hydrophobicity of the residues of a subsequence and the measurements of the backbone angles. Classifying a subsequence into one of the available clusters will give a good insight of the angle measurements and consequently the structure of the

subsequence. Also the length of the subsequence is an effective factor in the angle prediction process. Longer subsequences achieve better fits in one of the standard continuous probability distributions. It is found that unclustered subsequences have unreliable results compared to clustered subsequences.

6. Future work

These results can be used to guide the search process in a complete protein structure prediction algorithm. Using these results will greatly reduce the search space which can increase both the efficiency and the effectiveness of the search process. This angle-hydrophobicity relationship can be used combined with heuristic techniques like genetic algorithm to restrict the initial population to statistically familiar conformation. In this case it is better to apply these guiding rules to only a portion of the initial population in order to leave a chance to the new unfamiliar conformations. Approximations of our results can be applied to crystalline lattices protein models like cube octahedron lattice model which allows the use of several possible angles 60", 90", 120" and 180". Applying the results to this algorithm will allow the predictor to use the most statistically realistic angle of the available alternatives based on its neighboring residues. Also, it is possible to investigate applying the same approach on subsequences of length more than 7 residues and minimize the processing time.

References

[1] Bonnie Berger, T. L., "Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete." Proceedings of the second annual international conference on Computational molecular biology, vol. 5, issue 1, 1998.

[2] Clayton Matthew Johnson, A. K., "A genetic algorithm with backtracking for protein structure prediction." Proceedings of the 8th annual conference on Genetic and evolutionary computation, 2006.

[3] Diego Rother, G. S., Vijay Pande, "Statistical Characterization of Protein Ensembles." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 5, issue 1, 2008.

[4] Eli Hershkovitz, G. S., Allen Tannenbaum, Loren Dean Williams, "Statistical Analysis of RNA Backbone." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 3, issue 1, 2006.

[5] Eran Segal, D. K., "Probabilistic hierarchical clustering for biological data." Proceedings of the sixth annual international conference on Computational biology, 2002.

[6] Hardik A. Sheth, S. K., "Motif discovery for proteins using subsequence clustering." Proceedings of the 5th international workshop on Bioinformatics, 2005.

[7] Hoos, A. S. a. H. H., "An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem." BMC Bioinformatics, vol. 6, issue 30, 2005.

[8] Jernigan, G. R. a. R. L., "Ideal architecture of residue packing and its observation in protein structures." Protein Science, vol. 6, issue 10, 1997.

[9] Karplus, J. M. C. a. M., "Neural networks for secondary structure and structural class prediction." Protein Science, 1995.

[10] Márcio Dorn, O. N. d. S., "CReF: a central-residue-fragment-based method for predicting approximate 3-D polypeptides structures." Proceedings of the 2008 ACM symposium on applied computing, 2008.

[11] Nancy M. Amato, K. A. D., Guang Song, "Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures." Proceedings of the sixth annual international conference on Computational biology, 2002.

[12] Neal Lesh, M. M., Sue Whitesides, "A complete and effective move set for simplified protein folding." Proceedings of the seventh annual international conference on Research in computational molecular biology, 2003.

[13] Richa Agarwala, S. B., Vlado Dančík, Scott E. Decatur, Martin Farach, Sridhar Hannenhalli, S. Muthukrishnan, Steven Skiena, "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model." Proceedings of the first annual international conference on Computational molecular biology, 1997.

[14] Rong She, F. C., Ke Wang, Martin Ester, Jennifer L. Gardy and Fiona S. L. Brinkman, "Frequent-subsequence-based prediction of outer membrane proteins." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.

[15] Saravanan Dayalan, S. B., Heiko Schroder, "Dihedral angle database of short sub-sequences for protein structure prediction." Proceedings of the second conference on Asia-Pacific bioinformatics 29, 2004

[16] Sergio Raul Duarte Torres, D. C. B. R., Luis Fernando Nino Vasquez, Yoan Jose Pinzon Ardila, "A novel ab-initio genetic-based approach for protein folding prediction." Proceedings of the 9th annual conference on Genetic and evolutionary computation, 2007.

[17] Thang N. Bui, G. S., "An efficient genetic algorithm for predicting protein tertiary structures in the 2D HP model." Proceedings of the 2005 conference on Genetic and evolutionary computation, 2005.