Gaussian Mixture Modeling of Helix Subclasses: Structure and Sequence Variations

Ashish V. Tendulkar, Babatunde Ogunnaike, and Pramod P. Wangikar

Pacific Symposium on Biocomputing 11:291-302(2006)

GAUSSIAN MIXTURE MODELING OF α-HELIX SUBCLASSES: STRUCTURE AND SEQUENCE VARIATIONS

ASHISH. V. TENDULKAR*
Kanwal Rekhi School of Information Technology,
Indian Institute of Technology Bombay,
Powai, Mumbai-400 076, India.
E-mail: ashish@it.iitb.ac.in

BABATUNDE OGUNNAIKE
Department of Chemical Engg.,
University of Delaware,
Newark, DE 19716.
Email: ogunnaik@che.udel.edu

PRAMOD P. WANGIKAR†
Department of Chemical Engg.,
Indian Institute of Technology Bombay,
Powai, Mumbai-400 076, India.
E-mail: pramodw@iitb.ac.in

Classification of helical structures and identification of class specific sequence features is of interest for protein structure modeling. We use geometric invariant based method to first select helix-like local conformations. These conformations are mapped in a principal component space and subjected to Gaussian mixture modeling. The largest Gaussian corresponds to the regular α-helix. Kinked helix and curved helix appear as a separate gaussians. Class conditional, position specific amino acid propensity analysis reveals striking difference among the three classes. In regular helix, proline propensity is significant only in the beginning and low in the rest of the region regardless of length of the helix. In kinked helix, the proline propensity has a sharp peak at the helix center, while in the curved helix, the proline propensity has a broad peak in the middle region.

*presenting author
†corresponding author

1. Introduction

α-Helices are the most important structural elements in globular proteins. Based on handedness of helix turn, they have been classified as right handed α-helices and left handed α-helices[1]. The right handed helices are most commonly occurring helices in globular proteins. Although, α-helices are certainly the most regular structural building blocks, they show significant imperfections[1,2,3]. The perturbation is caused by a variety of reasons such as occurrence of proline residue[4] in the middle producing a kink. Based on structural and geometric features, the helices are categorized as linear, curved and kinked α-helices[4,5].

It is well known that the structural and geometric differences give rise to different subclasses of α-helix. Moreover, the structural features of a helix is encoded in its sequence composition. Propensities of different amino acids for formation of a specific secondary structure is a basis of many secondary structure prediction methods[6]. Kumar and Bansal[5] have reported a strong correlation between propensities of individual amino acids and helix length, geometry and location on protein globe. Doig and co-workers[7] have reported amino acid propensities at N and C terminus of α-helices. Engel and DeGrado[8] have reported very strong position dependent propensities of different amino acids throughout the length of a helix. These methods first extract helices from protein dataset based on secondary structure assignment[9] or steriochemical punctuation marks[10]. These methods have limitations in terms of accuracy in determining beginning and endpoints of the helix.

In our earlier work, we have reported classification of overlapping octapeptide substructures in globular proteins[11]. To obtain more fine-grained classification of α-helices, we have combined all the octapeptides classified as helices and modeled the structure space as a mixture of Gaussian. The resulting clusters represents various helix subclasses. The overlapping octapeptides in a particular subclass are merged to form longer length peptides. Thus, we construct our dataset of helices of varying length. Thus, our method provides structure based unbiased way of extracting helices. The analysis of amino acid propensities for different subclasses of helices reveals that the amino acid propensities in helices are strongly dependent on the subclass and position in the helix structure.

2. Method

2.1. *Selection of Helices*

We had performed k-means clustering(k=150) on overlapping octapeptides local conformations drawn from ASTRAL_95 dataset, version(1.67)[12]. The octapeptides were described using a set of 29 non-redundant geometric invariants[13,2] such as edge, perimeter, volume, area of triangle etc.[11]. The geometric invariants were directly computed from x, y, z coordinates of C_α atom. Thus, we have approximated backbone geometry with C_α geometry[14]. Consensus secondary structure was calculated for the clusters resulting from K-means application[11]. To obtain a more finer level classification of α-helices, we have combined all the octapeptide local conformations, which have been classified in the clusters, which have consensus secondary structure as HHHHHHHH.

2.2. *Gaussian Mixture Modeling of α-helix Structure Space*

Each geometric invariant was normalized to mean-centric, unity standard deviation values[15]. Principal component analysis[15] was performed on the standardized geometric invariants of octapeptides helical structures. We have chosen first s principal components to represent an octapeptide helix structure. Thus, n octapeptide helical structures are represented as a vector in space spanned by the first s principal components.

Let $y = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ be the set of n helix octapeptides. We model the data as a mixture of k Gaussian,

$$f(y) = \sum_{i=1}^{k} \phi_i f_i(y, \theta_i) \tag{1}$$

where, k is the number of components in the mixture, ϕ_i is the probability that a given helix octapeptides will come from ith component, also called as ith mixing proportion, θ_i is the vector of parameters describing the ith component. Since we're using Gaussian mixture model, θ_i consists of the mean vector, μ_i and covariance matrix, Σ_i. Thus ith component in the mixture is characterized by ϕ_i, μ_i and Σ_i. To estimate the parameters of the Gaussian mixture model, we used Expectation Maximization(EM) algorithm, fastmix[16], which automatically determines optimal number of components k required to maximize the expectation based on Bayesian information content[16].

Each helix octapeptide $\vec{x}$ is scored against each Gaussian i in the mixture using the following formula:

$$ln\ p(i|\vec{x}) = (-\frac{1}{2}\ ln\ |\Sigma_i| - \frac{1}{2}(\vec{x} - \vec{\mu_i})'\ {\Sigma_i}^{-1}\ (\vec{x} - \vec{\mu_i}) + ln\ \phi_i) \qquad (2)$$

The score signifies amount of influence each Gaussian exerts on the helix octapeptide. The octapeptides is assigned to the highest scoring Gaussian.

2.3. *Visualization of Clusters*

The clusters are visualized in form of bivariate distribution of the first two principal components conditioned on third and fourth principal component and marginal on the subsequent principal components. The detail procedure for obtaining such conditional bivariate distribution is given in [11]. The third and fourth principal components are divided into 4x4 equidensity grid. For each cell in the grid, we calculate bivariate density for the first two principal components. The bivariate distribution shows multiple peaks. Each peak has one to one corresponds to a mixture components and hence with clusters. Each peak is labeled with the cluster number and cartoon representing the structure of the helix octapeptide closest to its mean.

2.4. *Concatenation of octapeptides to form longer helices*

The octapeptide helices are assigned to different subclasses based on Gaussian mixture modeling. We carried out analysis of structural subclass assigned to the neighboring octapeptides i and $i+1$, which share an overlap of consecutive seven residues at the end of i and at the start of $i+1$. Suppose that the length of ith octapeptides is l. The neighboring octapeptides, assigned to the identical subclass, are merged to from helix of length $l+1$. Such kind of neighboring helices can be combined to form longer helices as they share similar geometric and structural properties.

2.5. *Amino Acid Propensity Analysis*

We analyzed position specific propensity of different amino acids for different subclasses. The position specific propensity is defined as [8],

$$P_{ij} = \frac{f_{ij}}{f_i} = \frac{n_{ij}/\sum_i n_{ij}}{N_i/\sum_i N_i} \qquad (3)$$

where, f_{ij} is the fraction of ith amino acid at jth helix position, n_{ij} is the number of ith amino acid at jth helix position, f_i is the fraction of ith

amino acid over entire protein structure dataset and N_i is the number of ith amino acid over entire protein structure dataset.

3. Results

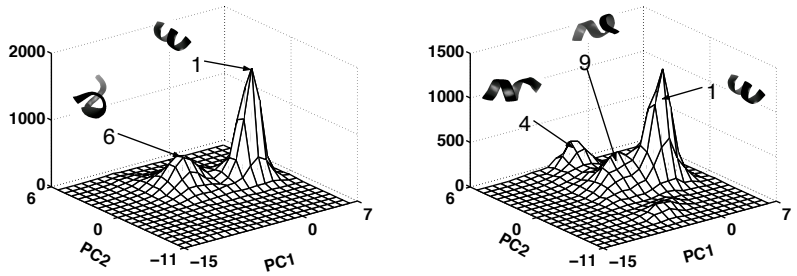
Approximately 0.4 million helices were selected from 1.7 million overlapping octapeptides drawn from ASTRAL_95 dataset(version 1.67)[12] based on the criteria defined in subsection 2.1. Principal component analysis of the dataset reveals that the first six principal components explain about 80% of variance in the data. The details about interpretation of principal components based on contributions from individual geometric invariants are documented in our earlier work[11]

3.1. *Visualization of Clusters*

Due to space constraints, we have shown the two most interesting cells from 4x4 equidensity grid in Figure 1. The fig. 1a shows two distinct and well separated peaks corresponding to gaussians 1 and 6. The cartoons of representative structures suggest that the Gaussian 1 is regular α-helix, whereas Gaussian 6 is kinked helix. The peak for regular α-helix is sharp and tall, whereas the peak corresponding to kinked helix is broad and short in height. The sharp nature of regular α-helix peak signifies highly regular nature of the class with lesser tolerance towards structural variations. The tall height of the peak signifies that the regular α-helix class is the most dominant class among helix subclasses. The broad nature of kinked helix peak suggests tolerance for structural variations. The short height of the peak denotes lesser probability of occurrence of kinked helices regular α-helices

The fig. 1b shows three distinct peaks corresponding to gaussians 1, 4 and 9. The cartoons denotes the structural differences between the classes corresponding to the peak. It further suggests that the peak for Gaussian 1 represents regular α-helix class, the peak for Gaussian 4 denotes a extended helix and the peak for Gaussian 9 denotes a helix class with distortion in the middle. The height of the corresponding peaks suggests that the regular α-helices have the highest probability of occurrence than the remaining two subclasses.

It is interesting to note that the helix structure space is sparse as shown by a lot of open area in both the bivariate plots in fig. 1a and fig. 1b.



(a) (b)

Figure 1. Representative conditional bivariate probability distribution of local helix conformations(A, B). Each panel shows a bivariate distribution on the first two principal component values, conditional on a specific range of the third and fourth principal components and marginal on the subsequent principal components values. (A) A bivariate distribution on the first two principal component values, conditional on the following values of the third and fourth principal components, $-\infty \leq PC_3 < -0.84$ and $-\infty \leq PC_4 < -0.70$. (B) A bivariate distribution on the first two principal component values, conditional on the following values of the third and fourth principal components, $+0.68 < PC_3 \leq +\infty$ and $+0.72 < PC_4 \leq +\infty$

3.2. *Properties of Gaussian Mixtures*

The summary of mixture parameters estimated by Expectation maximization algorithm[16] are provided in Table 1. Total of eleven gaussians have been detected with skewed mixing proportions. The gaussians are arranged in their descending values of mixing proportions. The most heavily represented Gaussian has mixing proportion of 76%. The lowest mixing proportion is 1%. The helix octapeptides are assigned to appropriate gaussians.

The first Gaussian represent regular right handed α helix subclass. Its centroid occupies positive value on the first principal component, a small positive value on fourth one, and small negative values on second, third, fifth and sixth principal components(Table 1). These values are in agreement with the interpretation of contributions of individual geometric invariants to the principal components. The covariance matrix of the first cluster is the tightest among all the clusters, suggesting highly regular nature of the helix subclass(Table 1).

The Gaussian number 6 and 10 have substantial negative values for their means along the first principal component. The Gaussian number 6 corresponds to kinked helix and number 10 corresponds to curved helix. The curved helix shows smaller length between the first and the last residue of helix octapeptide(Table 2). The decrease in length is accompanied by in-

crease in the area of triangle formed by first, fifth and the last residue(Table 2). The kinked helices have mixing proportion of 2%, whereas the curved helix has a mixing proportion of 1%.

Comparison of covariance matrices of regular, kinked and curved helix reveals that regular helix has the tightest covariance matrix, kinked helix has moderate covariance matrix, whereas the curved helix had the largest covariance matrix.

Table 1. Gaussian mixture characteristics of the helix subclasses

Characteristics	Regular α-helix	Kinked Helix	Curved Helix
ϕ_i	0.76	0.02	0.01
μ_1	1.12	-5.94	-7.71
μ_2	-0.06	2.02	-5.33
μ_3	-0.14	-3.03	-4.27
μ_4	0.04	-4.82	2.51
μ_5	-0.03	-1.29	1.39
μ_6	-0.04	-2.24	-0.41
σ_{11}	1.20	4.28	7.23
σ_{22}	0.75	4.93	7.70
σ_{33}	0.93	2.94	6.33
σ_{44}	0.85	3.91	5.22
σ_{55}	0.73	2.82	7.04
σ_{66}	0.63	2.16	9.63

Note: a. $\sum_i^k \phi_i = 1$. ϕ_i represent probability of a helix octapeptides belonging to mixture i. The total number of helix subclasses, $k = 11$. The three most important helix subclasses are described here.
b. μ_{ij} represents mean of mixture i on jth principal component.
c. σ_{ii} represents variance of mixture i along ith principal component. The covariance matrix for each mixture contains non-zero values only along its diagonal.

3.3. *Finer structural differences between distinct subclasses*

The analysis reveals three distinct helix subclasses: regular, kinked and curved. We analyzed structural differences between these distinct subclasses based on structural descriptors used for representing the helix octapeptides. The structural differences between these subclasses have been summarized in Table 2.

The structural descriptors characterize differences between various subclasses. The distance between i and $i + 3$ residues($d_{i,i+3}$) characterizes regularity of helix. The regular α-helix has the least $d_{i,i+3}$ with the least

Table 2. Mean and standard deviation of $d_{i,i+3}$, d_{18}, $Vol_{i,i+1,i+2,i+3}$, and $Area_{158}$ for vastly differing subclasses.

Structure Descriptors	Regular α-helix	Kinked Helix	Curved Helix
$d_{i,i+3}$	$5.15 +/- 0.20$	$5.51 +/- 0.45$	$5.64 +/- 0.57$
d_{18}	$10.64 +/- 0.33$	$11.38 +/- 0.74$	$8.64 +/- 1.59$
$vol_{i,i+1,i+2,i+3}$	$6.99 +/- 0.57$	$5.56 +/- 2.61$	$5.31 +/- 3.00$
$area_{158}$	$10.40 +/- 1.49$	$18.53 +/- 4.16$	$11.38 +/- 5.93$

Note: (i) $d_{i,i+3}$ denotes distance between i and $i+3$ residues.
(ii) d_{18} denotes distance between first and the last residue.
(iii) $vol_{i,i+1,i+2,i+3}$ denotes volume of tetrahedron formed by $i, i+1, i+2$and $i+3$ residues.
(iv) $area_{158}$ denotes area of triangle formed by first, fifth and eighth residue.

standard deviation signifying regular nature of helices. The kinked helix show longer average $d_{i,i+3}$ with more deviation. The longest average $d_{i,i+3}$ is assumed by the curved helix. The larger values of mean and standard deviation of $d_{i,i+3}$ denotes significant departure from the regularity.

The distance between end to end residues d_{18} of helices characterizes extended structure. The kinked helix is the most extended structures among the distinct subclasses. The regular α-helix has moderate end to end length. The decrease in the end to end distance either denotes shrink or a curve in the helix octapeptide. The area of triangle($Area_{158}$) formed by first, fifth and the last residue differentiate between helices with a bend , kinked and regular ones when coupled with d_{18}. With reference to regular α-helix, the decrease in d_{18} and increase in $Area_{158}$ denotes a curve in the middle of the octapeptide. The increase in both d_{18} and $Area_{158}$ denotes a kinked helix.

The sign of $Vol_{i,i+1,i+2,i+3}$ differentiate between left handed and right handed systems[11]. The positive values of volumes in all three distinct subclasses means that all the helix subclasses make a right handed system.

3.4. *Concatenation of octapeptides to form longer helices*

For the regular α-helix subclass, the majority of neighboring octapeptides lies in the same subclass endorsing the strong regular nature of the subclass. This leads to formation of variable length helices. The distribution of length of regular helices shows a wide spectrum of values ranging from 8 to 82, with the maximum number(6909) of regular α-helix of length 8, while a singleton α-helix of length 82. The distribution of length of regular α-helix is shown in Fig. 2. The distribution shows roughly exponential trend. The

maximum helix length in kinked and curved helix is 9.

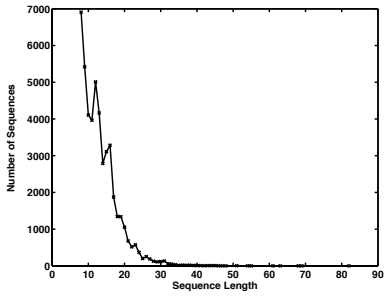


Figure 2. Distribution of length of regular α-helices in cluster 1 ($8 \leq length \leq 82$)

3.5. *Analysis of amino acid propensities in helix subclasses*

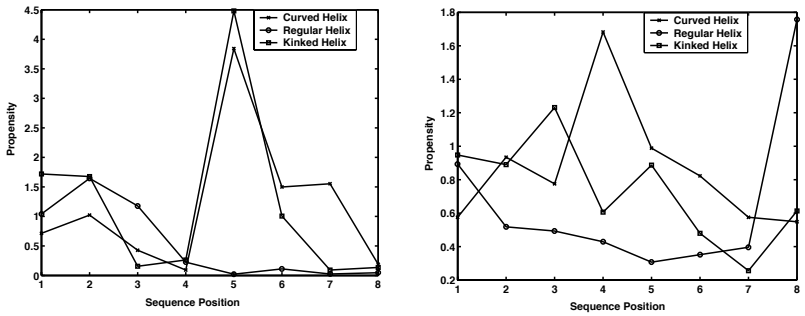


Figure 3. Propensity of (a) Proline and (b) Glycine in regular α-helix, kinked helix and curved helix.

The analysis of amino acid propensities at different positions in different helix subclasses was carried out on the helices having length 8. The propensities of an individual amino acids at various positions in different subclasses were plotted. We have shown propensity graphs for two representative amino acids, glycine and proline, in fig 3.

Proline is considered as a prominent helix breaker[4]. The propensity graph of proline (Fig. 3a) shows different propensity numbers at different positions of different helix classes. The proline has significant propensity

up to third position in regular α-helix. The highest proline propensity was observed in kinked helix at fifth position, where the kink occur. The kinked helix also shows significant proline propensity at second position. The curved helix has highest proline propensity at fifth position. It also shows significant proline propensities at sixth and seventh position also. Thus, proline propensity is higher in the curving region of curved helix.

The propensity graph of glycine (Fig. 3b) shows that the propensity of glycine shows different trends based on position and helix subclass. The regular α-helices shows significant glycine propensity only at the end position of the helix. The other position shows moderate glycine propensities. Kinked helix shows significant glycine propensity at third position and moderate propensities at the remaining positions. The curved helix shows significant glycine propensities at second, fourth, fifth and sixth position position with highest propensity at fourth position.

3.6. *Analysis of lengthwise amino acid propensities for regular α helix*

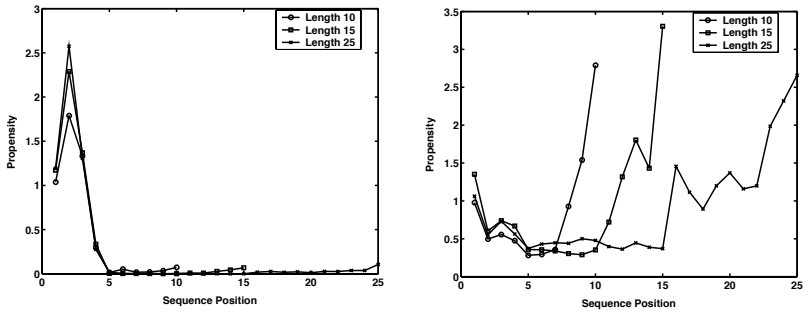


Figure 4. Propensity of (a) Proline and (b) Glycine in regular α-helix of varying lengths

The lengthwise propensity analysis was done for the regular α-helix class. We have selected three groups of lengths: 10, 15 and 25 for analysis. All the sequences having length ≤ 10 were merged in first group, the sequences having length between 11 and 15 were merged in secondary group, while the sequences having length between 16 and 25 are merged in the third group. The lengthwise propensity analysis is carried out on all the three groups for all 20 amino acids. Here, We have included two represent

propensity graphs of proline and glycine(Fig 4).

The propensity of proline is shown in fig 4a. and the propensity of glycine is shown in fig 4b. Regardless of the total length of helix, we have observed almost similar trends for both proline and glycine. Proline is having significant propensity up to first three positions in all three groups. For the rest of the positions the propensity is almost zero. Glycine is having significant propensity in the beginning of helix, then low propensity in the middle region and then high propensity at towards the end. The same trend is observed in all three groups. Since we have combined helices having length between 11 to 15 in group of 15, glycine propensity starts upward trend from 11th position for group of length 15, the same is true for the other two groups.

4. Discussion

We have reported a novel method for fine grain classification of helical substructures into its subclasses using geometric invariants and Gaussian mixture modeling. We also provide detailed explanation about roles of various structure descriptors in differentiating between various helix subclasses. It is interesting to note that the individual geometric invariants are capable of differentiating one or other feature of helix geometries(Table 2). The linear combination of these individual descriptors differentiates between various subclasses more efficiently.

Gaussian mixture modeling of helix local conformation space provides a formal framework for analyzing geometry of theoretical helix structures or newly formed helix structure. The modeling of this sort provides a formal method for detecting outliers in the data as well as subclass of a particular helix structure. The Bayesian information content based expectation maximization algorithm[16] ensures that the right number of mixture components are selected to model the Gaussian mixture accurately. The mixing proportion ϕ_i assigned to different subclasses matches well with literature reported mixing proportions(Table 1). The regular α-helix subclass has been assigned a mixing proportion of 76%, which is in accordance with 74% mixing proportion reported by Kumar and Bansal[5]. The visualization of helix local conformation space in form of conditional bivariate distribution plots(Fig 1) helps in getting quick idea about separation between various subclasses and differences in their geometry.

The merging of neighboring octapeptides having identical subclass provides more accurate and unbiased approach for extracting helices from pro-

tein structures. This is a fundamental shift from the literature reported methods, which depends on secondary structure assignment[5] or helix breaking signals in protein sequence[8]. The method provides structure based method for extracting helices more accurately.

Analysis of propensities of amino acids for different classes of helix reveals different position specific trends. It implies that the propensity of amino acid is also dependent on the class of helix. The analysis of propensity for different length of helices reveals similar trends regardless of length of helix.

Class conditional position specific analysis of amino acid propensities provides vital clues in better understanding sequence-structure relationship in various subclasses of α-helices, leading to better prediction of helix subclass in protein structure prediction. The results presented in the paper are also useful for designing artificial helices from a specific subclass.

References

1. J. Richardson, *Adv. Prot. Chem.* **34**, 167 (1981).
2. A. Tendulkar, A. Joshi, M. Sohono, P. Wagikar, *J. Mol. Biol.* **338**, 611 (2004).
3. E. Emberly, R. Mukhopadhyay, N. Wingrren, C. Tang, *J. Mol. Biol.* **327**, 229 (2003).
4. D. Barlow, J. Thornton, *J. Mol. Biol.* **201**, 601 (1988).
5. S. Kumar, J. Bansal, *Proteins: Struct. Funct. Genet.* **31**, 460 (1998).
6. S. Dasgupta, J. Bell, *Int. J. Pept. Protein Res.* **41**, 499 (1993).
7. A. Doig, R. Baldwin, *Protein Sci.* **4**, 1325 (1995).
8. D. Engel, W. DeGrado, *J. Mol. Biol.* **337**, 1195 (2004).
9. W. Kabsch, C. Sander, *Biopolymers.* **22**, 2577 (1983).
10. K. Gunasekaran, H. Nagrajaram, C. Ramkrishnan, P. Balaram, *J. Mol. Biol.* **275**, 917 (1998).
11. A. Tendulkar, M. Sohono, B. Ogunnaike, P. Wagikar, *Bioinformatics.* **21**, 18 (2005).
12. S. Brenner, P. Koehl, *Nucleic Acid Research.* **28**, 254 (2000).
13. A. Tendulkar, V. Samant, C. Mone, M. Sohono, P. Wagikar, *J. Mol. Biol.* **334**, 157 (2003).
14. T. Oldfield, R. Hubbard, *Proteins: Struct. Funct. Genet.* **18**, 324 (1994).
15. R. Johson, D. Wichern, *Prentice Hall of India.* (2003).
16. A. Moore, *Adv. Neural Information Processing Systems.* **11** (1999).