Optimization of the Sliding Window Size for Protein Structure Prediction

Ke Chen*[1], Lukasz Kurgan[1] and Jishou Ruan[2]

[1]University of Alberta, Department of Electrical and Computer Engineering, Edmonton, CANADA, T6G 2V4
[2]Chern Institute of Mathematics, College of Mathematical Science & LPMC, Nankai University, Tianjin, PRC 300071
*kchen1@ece.ualberta.ca (corresponding author)

Abstract - **Sliding window based methods are relatively often applied in prediction of various aspects related to protein structure. Despite their wide spread use, researchers did not establish a standard related to the size of the window, i.e., window sizes ranging between 7 and 17 residues were used in the past. To this end, this paper performs a computational study based on a probabilistic approach that aims at finding an optimal sliding window size. The results shows that formation of helical structure can be affected by amino acids (AAs) that are up to 9 positions away in the sequence, while the formation of coils and strands can be affected by AAs that are up to 3 and 6 positions away, respectively. Overall, our results suggest that a sliding window with 19 residues is optimal for secondary structure prediction, while for a specific prediction tasks, such as prediction of β-strands, a smaller window size is sufficient. Finally, the 20 AAs are categorized into five groups based on their influence of formation of the secondary structure. The finding related to the optimal window size was confirmed based on an independent experimental study related to the prediction of secondary protein structure.**

Keywords – **Protein Structure, Protein Structure Prediction, Sliding Window, Secondary Protein Structure**

I. INTRODUCTION

Protein folding is a complex process that involves all amino acids (AAs) of the corresponding protein sequence. However, to improve computational efficiency and due to the limited number of structure samples, many structure prediction methods use a sliding window that covers fragments of the sequence, rather than the whole sequence, as the input. In fact, sliding window is adopted in many areas such as prediction of cis/trans isomerization in proteins [1], optimization of hydrophobicity tables [2], prediction of flexibility and rigidity of proteins [3], prediction of solvent accessibility of proteins [4], and in numerous methods for the secondary [5-10] and the tertiary structure prediction [11,12]. The sheer number and breath of these applications provide strong motivation for the research presented in this paper.

When predicting or analyzing some characteristics of an amino acid A_i, researchers relatively often use a window of $2n+1$ AAs that is centered at Ai. In other words, a segment composed of $A_{i-n}\ A_{i-n+1}\ldots A_{i-1}\ A_i\ A_{i+1}\ldots A_{i+n}$, AAs is used since a given characteristic of the central AA is determined not only by the AA itself, but also by the adjacent AAs. At the same time, different studies apply different window sizes. The sizes range between 7 and 17 residues. In local, secondary structure prediction, the 7-residue window is adopted [9]. In the prediction of flexibility and rigidity of proteins, researchers applied 9-residue window [3]. The 11-residue window was adopted in the cis/trans isomerization prediction [1] and the 13-residue window was selected in the prediction of helices in trans-membrane proteins [7]. Finally, several secondary structure prediction methods use the 15-residue window [5,6], and at least one of them uses an even larger, 17-residue window [8]. It is clear that scientists have not been able come up with a standard with respect to the length of the window that provides the optimal results. A recent study shows that significant majority of identical sequence segments that consist of 10-20 residues fold into similar structures in different proteins [13], which suggests that their structure is conserved and that the other regions of the protein sequence may have limited impact on their structure. This supports the premise that the structure prediction tasks that were investigated by the above mentioned researchers can be performed based on sequence segments of limited length, which in turn provides validation for the sliding window based methods.

However, it is obvious that the strength of the impact that one AA has on another AA's secondary structure conformation will on average decay as their corresponding positions in the sequence is farther apart. We emphasize that this statement is true "on average", as the actual spatial packing of the sequence brings some of the distant, in terms of the position in the sequence, AA close together. At the same time, the secondary structure arrangements are mostly local (except the long range interactions in β-sheets), and thus they motivate the above statement.

This work estimates the distance-impact relation between an AA at position *i* and the adjacent, in terms of the position in the sequence, AAs. In other words, we investigate the impact of an AA at position *i*+*k* on the formation of the secondary structure of the AA at position *i*. Although intuitively the secondary structure of a given AA at position *i* is strongly affected by the type of the immediately adjacent AAs, it is not so obvious how

1-4244-0623-4/06/$20.00 ©2006 IEEE

this impact changes as the value of *k* increases. This impact can be quantified based on conditional probabilities that the AA at position *i* assumes helical, strand or coil structures given the particular types of AAs at the positions *i* and *i*+*k*. The paper performs a carefully designed analysis of these probabilities for a large set of protein sequence that are characterized by low homology. The latter assumption allows us to draw unbiased, in terms of the sequence homology distribution, conclusions with respect to the optimal selection of the sliding window length. The analysis of these conditional probabilities with respect to individual secondary structures allows us also to investigate which AAs provide the strongest influence on the formation of specific secondary structures. Finally, the above findings are verified based on an experimental study that performs a window based prediction of the secondary protein structure.

II. Methods and Datasets

A. *Dataset*

The low homology dataset used to compute the probabilities was generated using the PICSEC protein sequence culling server, which uses a combination of structural and sequence alignments to limit the sequence homology. We used the cullpdb25 set, which is characterized by the maximum of 25% sequence identity and includes proteins that were measured at 3.0 resolution and with R-factor equal 1.0 [14]. The original set includes 4127 protein sequences, which were further filtered to remove: 1) sequences with less than 20 AAs; 2) sequences, for which secondary structure information was incomplete in the PDB [16]; 3) sequence that included non-standard AAs; and 4) sequences, for which side chain coordinates are not provided in PDB. The missing side chain information may results in erroneous secondary structure assignments, as performed by the DSSP that requires the coordinates of the side chains [15]. The remaining 1743 sequences constitute the dataset that was used to compute the probabilities. The secondary structure of these 1743 sequences was assigned by DSSP.

B. *Unconditional Probabilities of AAs Being in the Helical, Strand and Coil Secondary Structure States*

A conditional probability that a given AA belongs to one of the three major secondary structures is defined as

$$P_{class}(AA) = \frac{Class(AA)}{N(AA)} \qquad (1)$$

where *class* is one of the secondary structures, i.e., helix (H), strand (E) and coil (C), *AA* denotes one of the 20 AAs, i.e., *AA= A, C, D…, V, W, Y*, *N*(*AA*) denotes the frequency of *AA* in the input dataset and *Class*(*AA*) is the frequency of a given *AA* among residues that belong to secondary structure *class* in the dataset. The corresponding 60 probabilities are denoted as $P_H(A)$, $P_E(A)$, $P_C(A)$, $P_H(C)$, $P_E(C)$, $P_C(C)$,……, $P_H(Y)$, $P_E(Y)$, $P_C(Y)$. We note that $P_H(AA)+P_E(AA)+P_C(AA)=1$.

C. *Conditional Probabilities*

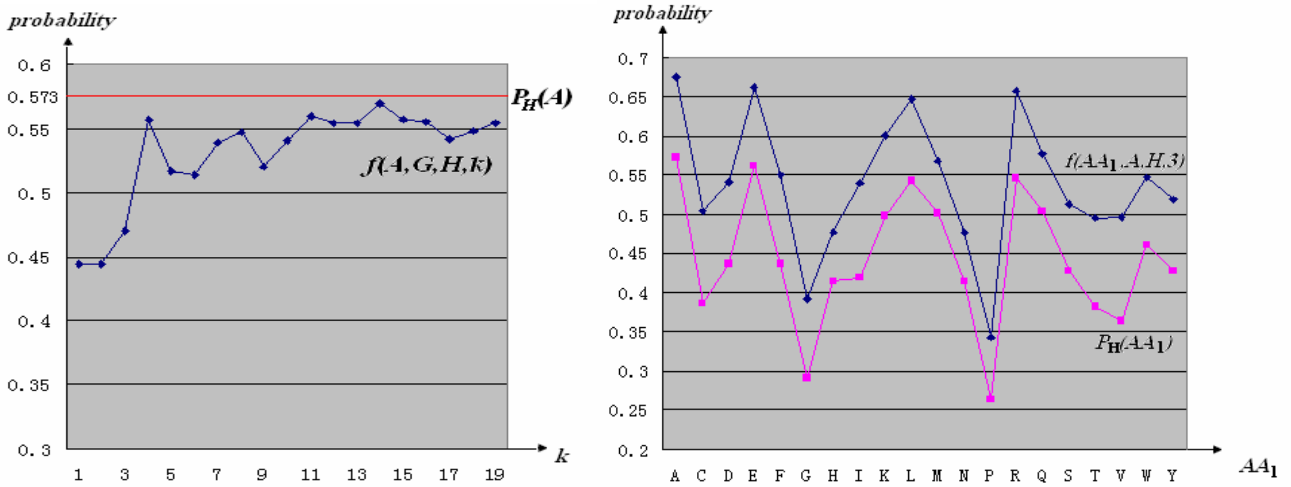
Segment $A_{i-n}\, A_{i-n+1} \ldots A_{i-1}\, A_i\, A_{i+1} \ldots A_{i+n}$ defines a window of *2n+1* AAs that is centered at AA A_i. The impact of an AA at position *k*, i.e., A_{i+k} and *k = -1, -2, …, -n, 1, 2, …, n*, on the secondary structure of central AA at position *i* can be estimated based on conditional probability, which is defined as

$$P(SS(A_i) = Class \mid A_{i+k} = AA_2, A_i = AA_1)$$
$$= \frac{(P(SS(A_i) = Class, A_{i+k} = AA_2, A_i = AA_1)}{P(A_{i+k} = AA_2, A_i = AA_1)} \qquad (2)$$

where AA_1 and AA_2 are any two out of the 20 AAs (they could be the same), and *SS*(A_i) denotes the secondary structure of A_i.

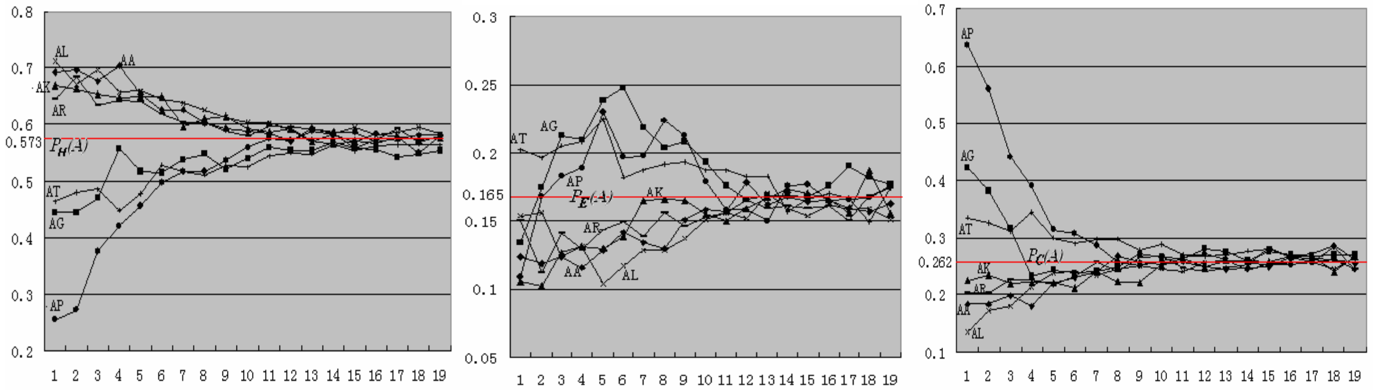
For example, assuming that $SS(A_i)=H$, $A_{i+k}=G$, and $A_i=A$, $P(SS(A_i)=H \mid A_{i+k}=G, A_i=A)$ corresponds to the conditional probability that an AA at i^{th} position is a helix given that $(i+k)^{th}$ AA in the sequence is *C* and the i^{th} AA is *A*.

If an AA at position (*i*+*k)* has an impact on the structure of the AA at position *i*, then the probability of A_i being a helix, strand or coil, is expected to be different for some A_{i+k} when compared with the corresponding unconditional probabilities $P_H(A_i)$, $P_E(A_i)$, and $P_C(A_i)$.



a) b)

Figure 1. a) The function *f*(*A,G,H, k*) (blue, lower line) and the corresponding unconditional probability $P_H(A)$ (red, upper line). The graph shows how AA *G* affects the secondary structure of the central AA *A* with the increasing distance *k* between them; b) The function $f(AA_1,A,H,3)$ (blue, upper line) and the corresponding unconditional probability $P_H(AA_1)$ (red, lower line). The graph shows how AA *A* affects the probability of different central AAs to form a helix when they are three residues away in sequence. The x axis corresponds to the 20 different central AAs.



a) b) c)

Figure 2. a) The function $f(A, AA_2,H, k)$ (black lines) and the corresponding unconditional probability $P_H(A)$ (red, straight line); b) The function $f(A, AA_2,E, k)$ (black lines) and the corresponding unconditional probability $P_E(A)$ (red, straight line); c) The function $f(A, AA_2,C, k)$ (black lines) and the corresponding unconditional probability $P_C(A)$ (red, straight line). The x-axis corresponds to k and the y-axis corresponds to the conditional probabilities.

Table 1. The standard deviation of $f(A, AA_2,H, k)$, $f(A, AA_2, E, k)$, $f(A, AA_2, C, k)$ with respect to the unconditional probability of A being helix, strand and coil, i.e., $P_H(A)$, $P_E(A)$ and $P_C(A)$, in the function of k.

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
H	0.11	0.11	0.08	0.08	0.06	0.05	0.04	0.03	0.04	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02
E	0.06	0.05	0.03	0.03	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02
C	0.12	0.09	0.06	0.05	0.03	0.04	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01

The conditional probabilities are expressed using a function of four variables, i.e., $f(AA_1, AA_2, class, k) = P(SS(A_i)= class \mid A_{i+k} = AA_2, A_i = AA_1)$. Based on this definition two relevant cases can be considered:

1. When assuming k as a variable and fixing the AA_1, AA_2, and *class* we can investigate the differences in the impact of a given AA on a given secondary structure of the central AA with respect to different distance between the residues. Consequently, this allows us to find the optimal sliding window size.

   For example, when we assume that $AA_1 =A$, $AA_2 =G$, and $class = H$, then $f(AA_1, AA_2, class, k)$ is simplified to $f(A,G,H, k)$. This function describes how AA G affects the probability of AA A to be a helix with respect to k that describes how far apart these two AAs are in the sequence. We expect that $f(A,G,H, k)\approx P_H(A)$ when k is large enough, since G would be too far away from A to have any significant influence on the structure of A. Figure 1(a), which is computed using the dataset of 1743 sequences, illustrates this relation. The straight line corresponds to $P_H(A) = 0.573$. The Figure confirms that $f(A,G,H, k)$ is approaching $P_H(A)$ with the increasing value of k.

2. When assuming AA_1 as variable and the values of the other three variables, i.e., AA_2, *class* and k, as fixed, we can study how a particular AA_2, which is k residues away from AA_1, affects the latter AA to form a specific secondary structure, which is denoted by *class*.

   For example, when we assume that $AA_2 =A$, $class = H$, and $k = 3$, then $f(AA_1, AA_2, class, k)$ is simplified to $f(AA_1,A,H, 3)$. This function describes how AA A, which is 3-residue away from AA_1 in sequence, affects the probability of AA_1 to be helix. The corresponding relation is shown and contrasted with the unconditional probability $P_H(AA_1)$ in Figure 1(b). In this case, it is evident that AA A increases the probability of any central AA to be a helix.

III. Results and Discussion

A. *Optimization of the Sliding Window Size for Protein Structure Prediction*

The optimal window size for prediction of protein structure, which is centered on AA A_i, should include all AAs that impact the structure of A_i while discarding those AAs that are too far away to have the impact. This can be equivalently described by finding the minimal value of k such that

$$f(AA_1, AA_2, class, k) \approx P_{class}(AA_1) \qquad (3)$$

for any AA_1, AA_2, and *class*. The equation (3) describes a situation, in which no matter which AA will be selected as AA_2, it will have no impact on the secondary structure of the central AA given the interval of $(k\text{-}1)$ AAs. Given that (3) is true for any AA_1, AA_2 and *class,* we can conclude that the AA_2 has no impact on the structure of AA_1.

Due to limited space, only a sample subset of results obtained for AA A is shown. The three functions, $f(A, AA_2,H, k)$, $f(A, AA_2, E, k)$, $f(A, AA_2, C, k)$, show how the probabilities of A to be helix, strand and coil, respectively, are influenced by the adjacent AAs. The corresponding graphs that include 7 out of the 20 functions for the $AA_2 = \{A, G, K, L, P, R, T\}$ are shown in Figure2. These curves give an insight of how these 7 AAs impact the secondary structure of A for different values of k. The graphs show that when k increases to about 10, the

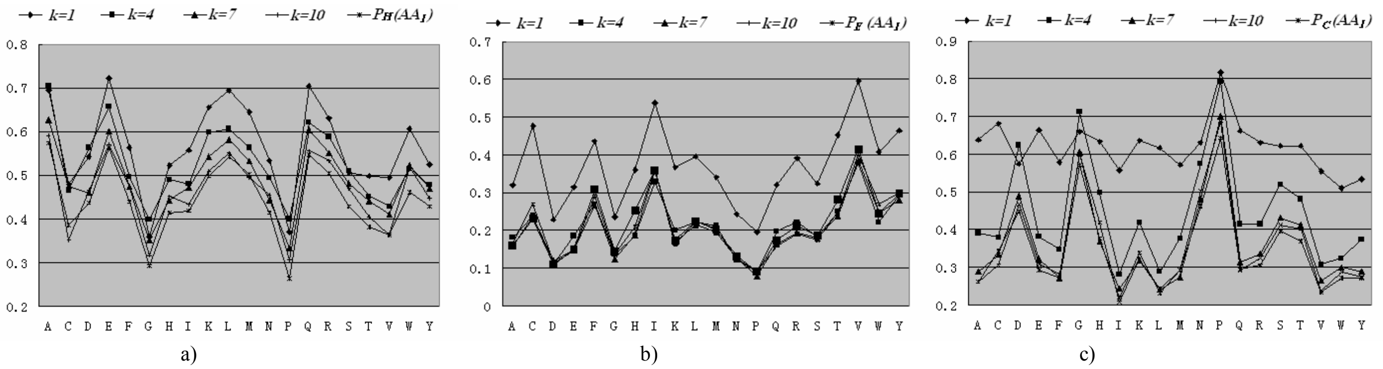


Figure 3. The influence of selected three AAs, i.e., A, I, and P on the formation of the helical, strand and coil structures, respectively: a) function $f(AA_1,A,H,k)$ for k = 1, 4, 7, 10, contrasted with $P_H(A)$; b) function $f(AA_1,I,E,k)$ for k = 1, 4, 7, 10, contrasted with $P_E(A)$. c) function $f(AA_1,P,C,k)$ for k = 1, 4, 7, 10, contrasted with $P_C(A)$.

Table 2. The influence of the 20 AAs on formation of the helical, strand and coil secondary structure of the immediately adjacent AAs (k = 1). For example, for AA A, the three values: 0.11 for H, -0.03 for E, and -0.08 for C, represent influence on the adjacent AA to form the corresponding structures, i.e., the influence on AA_1 to be a helix is positive and equals 0.11, while the influence on formation of a strand and coil is negative and equals -0.03 and -0.08, respectively.

AA secondary structure	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
H	0.11	-0.03	-0.05	0.09	0.01	-0.08	-0.02	-0.02	0.06	0.09	0.07	-0.04	-0.17	0.10	0.06	-0.05	-0.08	-0.06	0.04	-0.01
E	-0.03	0.08	-0.06	-0.06	0.07	-0.06	0	0.12	-0.06	0.03	0.02	-0.05	-0.10	-0.05	-0.03	-0.02	0.02	0.15	0.03	0.07
C	-0.08	-0.05	0.11	-0.03	-0.08	0.14	0.02	-0.10	0	-0.12	-0.09	0.09	0.27	-0.05	-0.03	0.07	0.06	-0.09	-0.07	-0.06

seven curves cluster together at $P_H(A)$, $P_E(A)$ and $P_C(A)$, respectively. This means that when AA_2 is about 10 residues away from A, it has virtually no impact on the secondary structure of A.

The impact of all 20 AA_2 is estimated based on the corresponding standard deviations with respect to the unconditional probability of A, which are defined as

$$deviation(Class,k)=\sqrt{\frac{\sum_{AA2}(f(A,AA_2,Class,k)-P_{Class}(A))^2}{20}} \quad (4)$$

The standard deviation is based on the average difference between $f(A, AA_2, Class, k)$ and $P_{Class}(A)$. Larger standard deviation corresponds to bigger impact of AA_2 on the secondary structure of AA A.

The example standard deviations for AA A are shown in Table 1. The Table shows that the impact decreases when AA_2 is farther apart from A in the sequence. The standard deviation saturates at k = 10. The values for k ≥ 10, which range between 0 and 0.02, can be considered as noise. We conclude that AAs that are more than 9 residues away from the central AA A have no impact to the secondary structure of A. Experiments performed for the remaining 19 central AAs show the same results, i.e., the secondary structure of the central AA is affected by the AAs that are 9 or less residues away. This suggests that prediction/analysis of protein structure of the central AA should be performed based on a 19 residues long window, i.e., 9 residues on the left and the right side of the central AA. We again emphasize that this window size assures that all local (with respect to sequence neighborhood) secondary structure interactions for the central AA are taken into account.

B. *Influence of Individual AA on the Formation of the Secondary Structure.*

Each of the 20 AAs is quite different in terms of its physiochemical properties such as hydrophobicity, charge, weight, etc. This section analyzes the differences with respect to their influence to form the secondary structure. The influence of AA_2 on AA_1 to form a certain secondary structure is measured using the following criteria

$$impact(AA_2,Class,k)=\frac{\sum_{AA_1}(f(AA_1,AA_2,k,Class)-P_{Class}(AA_1))}{20} \quad (5)$$

If AA_2 is helpful in establishing a given secondary structure, then the value of *impact* will be positive. The negative value of *impact* indicates that AA_2 diminished the probability of the central AA to assume this structure. By the definition, the influence of a given AA_2 is averaged over the 20 different AA_1.

Based on the computation of the *impact* values, the 20 AAs can be divided into five categories:

1. AAs that are strongly related to formation of helices, which include *A*, *E*, *K*, *L*, *M*, *Q* and *R*,
2. AAs that are strongly related to formation of strands, which include *C*, *F*, *I*, *V*, and *Y*,
3. AAs that are strongly related to formation of coils, which include *D*, *G*, *N*, *P*, *S*, and *T*,
4. AAs that are strongly related with formation of both helices and strands, which include *W*,
5. and finally AAs which have no significant impact on formation of the secondary structure of the adjacent AAs, which include *H*.

Due to space limitations, we discuss only representative results for the first three categories. AA *A*, *I* and *P* are selected as the typical cases that influence formation of the helix, strand and coil structures, respectively. We note that based on the below discussion the reader can easily reconstruct the reasoning behind the presented results for all 20 AAs.

Figure 3(a) shows that AA *A* increases the probability of an adjacent AA to form a helix. The five curves correspond to different values of *k*, which equals to the distance between the AA *A* and the central AA. The curve for $k = 1$ corresponds to the probability of AA_1 forming a helix when *A* is immediately adjacent to it in the sequence, i.e., $f(AA_1,A,H, 1)$. Analogously, curves for $k = 4$, $k = 7$, and $k = 10$ represent functions $f(AA_1, A, H, 4)$, $f(AA_1, A, H, 7)$, and $f(AA_1, A, H, 10)$, respectively. To analyze the influence of AA *A* on the secondary structure of the central AA, the 4 curves are compared with $P_H(AA1)$. The latter curve is the lowest among the 5 curves for the left most point of the graph, which corresponds to AA *A*. We observe that the smaller the value of *k*, the bigger the corresponding probability, which indicates that the influence of the AA *A* on the helical conformation of the central AA increases as the two AAs are closer to each other in the sequence.

Figure 3(b) presents results with respect to the influence of each of the 20 AAs on the strand conformation of the central AA. In this case, we focus our analysis on AA *I*. The corresponding probability values for $k = 4$, $k = 7$, $k = 10$ and $P_E(AA_1)$ are overlapping, which means that if AA *I* is at least 4 residues away from AA_1, then it does not influence the formation of a strand structure of the central AA. At the same time, *I* provides a strong influence on formation of a strand structure for the immediately adjacent AA since the value for $k = 1$ is significantly bigger than the value of $P_E(AA_1)$. This is in contrast with results in Figure 3(a), where *A* is shown to influence formation of the helix for AA_1 even being 9 residues apart in the sequence. Therefore, we conclude that the AA *I* is characterized by a short range, i.e., up to 3 residues away, influence on the formation of strands. Similar pattern can be observed for the remaining AAs, and thus we conclude that in case of the prediction efforts related directly to strand structure, the window size could be reduced to length of 7, i.e., 3 residues on each side. At the same time, we note that strands are characterized by long range interactions between residues that form β-sheets, which cannot be successfully addressed by the sliding window.

Figure 3(c) indicates existence of a medium range influence on the formation of coils, i.e., the points for $k = 7$, $k = 10$ and $P_C(AA_1)$ are overlapping, while the points for $k = 1$ and $k = 4$ are different than the base line curve $P_C(AA_1)$. Therefore, we conclude that AA *P* (as well as AAs *D*, *G*, *N*, *S*, and *T*) influence the coil conformation of the central residue from the distance of up to 6 residues. The size of the corresponding sliding window that specifically targets coil structure should be set to 13 residues.

C. Experimental Verification of the Proposed Optimal Window Size

The results presented in sections III.A and III.B demonstrate that the optimal window size should be 19 resides long, i.e., $k = 9$. This section provides independent experimental evaluation of this result by performing a *K*-nearest neighbor based prediction of the protein secondary structure.

More specifically, for a given AA in a protein sequence, a window that consists of $2n+1$ residues, i.e., *n* residues are to the left and to the right of the central AA, is used for the prediction of the secondary structure of the central AA. In case when either side of the window stretches outside of the sequence, the corresponding positions are filled with blanks. The prediction method applies $K = 25$, which gives highest prediction accuracy, and considers $k = 2, 3,\ldots, 11, 12$. The input sequence dataset that consists of 1743 sequences was randomly divided into two subsets, i.e., 1243 chains were used as the training set and the remaining 500 sequences were used as the test set. The methods predicts the secondary structure of the central residue by finding the 25 nearest neighbors from the training set based on the sequence similarity. Next, the predicted secondary structure is set as the most frequent secondary structure of the central residues of the neighbors. The resulting three state secondary structure prediction accuracies when using sliding windows of varying length are shown in Figure 4.

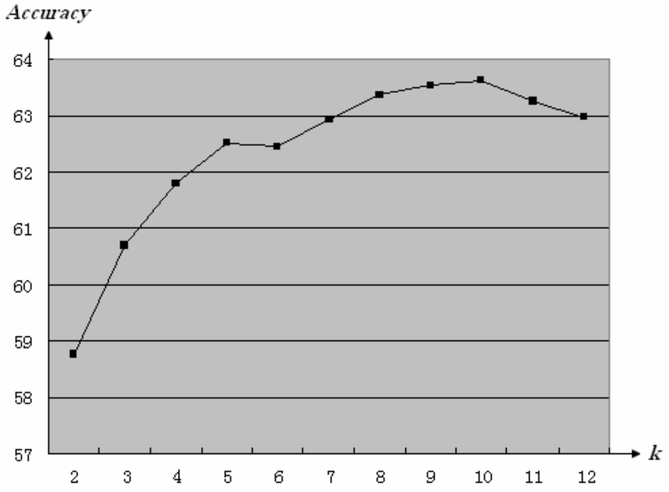


Figure 4. The accuracy of the three state secondary structure prediction performed using *K*-nearest neighbor method with sliding windows of varying length.

The prediction method achieves the peak accuracy of 63.62% for $k = 10$. At the same time, the accuracy for $k = 9$ is 63.55%, which is just 0.07% less than the best result. Therefore, we conclude that there is no significant difference between window sizes of 19 and 21 residues. For $k > 10$, the accuracy drops, illustrating that a wider window may result in a worse performance. We speculate that the main reason for the decrease in the accuracy is noise that comes from the information located on the window's edges. This result provides validation for our findings.

Although majority of secondary structure prediction methods use multiple sequence alignment profiles and position-specific scoring matrixes, and achieve accuracy of about 80%, a recent study showed that accuracy of these methods, (i.e., PSIPRED, Errsig, YASPIN, PHDpsi) drops significantly to 65%-67.5% when training and test sets are composed of 25% homology sequences [5]. The main reason for these relatively low results was that the prediction methods were based on sequence alignment, which in case of such low homology often produces sequence pairs that have different structure [17]. This suggests that alternative methods, which are not affected by the negative impact of alignment for low homology sequences, should be introduced to cope with the predictions for future targets.

The achieved secondary structure prediction results are compared with results of the above mentioned study, which considered sequences characterized by the same homology threshold and a single test set composed of 500 sequences [5]. We note that our simple method used sequences and structure information based on the 1243 chains in the training set, while in [5], the prediction was based on PSI-BLAST profiles that use sequence information from hundreds of thousands of known proteins, and structural information based on a much larger training set of 3553 sequences. Since the amount of training information gives an advantage to methods tested in [5], we conclude that the prediction accuracy achieved by our simple method (63.5%) is relatively high. This demonstrates feasibility of using a sliding window based protein structure prediction methods for the low homology sequences.

IV. CONCLUSIONS

Sliding windows based approaches are relatively popular in prediction of various aspects related to protein structure. Our study shows that selection of an optimal window length can increase the prediction accuracy. Based on a probabilistic approach, we estimated that the optimal window length should be 19 residues, which includes the central AA and the 9 adjacent AAs on both of its sides. Such window includes information required to predict and analyze folding of local structures. The study also shows that helical structure is characterized by the long range interactions, which include AAs that are up to 9 residues away from the central AA. At the same time, strands are characterized by the short range interactions (up to 3 residues away) and coils by the medium range interactions (up to 6 residues away). Therefore, a smaller window size may be sufficient for some prediction tasks, such as prediction of beta strands. Finally, our results indicate that the 20 AAs can be categorized into five groups according to their influence on formation of different secondary structures, which may give new understanding of relations between the AAs and new insights for the protein primary sequence analysis.

An independent, experimental evaluation of the proposed results concerning the window size have shown that the optimal results for the sliding window based secondary structure prediction are achieved for windows of 19 or 21 residues. The prediction accuracy drops when either narrower or wider windows are used, which confirms our findings.

This study provides helpful input for numerous protein structure prediction studies, including prediction of cis/trans isomerization, prediction of flexibility and rigidity, prediction of solvent accessibility, and especially for prediction of the secondary structure.

At the same time, we note that further verification of the results of this research, which would encompass application of the windows of the proposed size to the above prediction tasks to check their impact on the quality of the prediction, is beyond the scope of this paper and will be addresses in our future work.

ACKNOWLEDGMENT

KC and LAK gratefully acknowledge support from NSERC Canada under the Discovery program and MITACS Canada under the Internship program. JR was supported by Liuhui Center for Applied Mathematics and China-Canada exchange program administered by MITACS and NSFC.

REFERENCES

[1] Song J, Burrage K, Yuan Z, Huber T. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information, *BMC Bioinformatics*, 2006 Mar 9; 7:124.

[2] Zviling M, Leonov H, Arkin IT. Genetic algorithm-based optimization of hydrophobicity tables, *Bioinformatics*, 2005 Jun 1; 21(11):2651-6. Epub 2005 Mar 29.

[3] Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence, *Proteins*, 2005 Oct 1; 61(1):115-26.

[4] Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines, *Proteins*, 2002 Aug 15; 48(3):566-70.

[5] Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks, *Bioinformatics*, 2005 Jan 15; 21(2):152-9, Epub 2004 Sep 17.

[6] Jones,D.T., Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, 1999; 292:195-202.

[7] Rost B, Casadio R, Fariselli P, Sander C. Trans-membrane helices predicted at 95% accuracy, *Protein Science*, 1995 Mar; 4(3):521-33.

[8] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology*, 1978 Mar 25; 120(1):97-120.

[9] Boden M, Yuan Z, Bailey TL. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures, *BMC Bioinformatics*, 2006 Feb 14; 7:68.

[10] Sadeghi M, Parto S, Arab S, Ranjbar B. Prediction of protein secondary structure based on residue pair types and conformational states using dynamic programming algorithm, *FEBS Letters*, 2005 Jun 20; 579 (16): 3397-400.

[11] Sander O, Sommer I, Lengauer T, Local protein structure prediction using discriminative models, *BMC Bioinformatics*, 2006 Jan 11; 7:14.

[12] Benros C, de Brevern AG, Etchebest C, Hazout S. Assessing a novel approach for predicting local 3D protein structures from sequence, *Proteins*, 2006 Mar 1; 62(4):865-80.

[13] Ruan, J., Chen, K, Tuszynski, J., and Kurgan, L., Quantitative Analysis of the Conservation of the Tertiary Structure of Protein Segments, *Protein Journal*, 2006; accepted

[14] Wang G, Dunbrack RL Jr. PISCES: recent improvements to a PDB sequence culling server, *Nucleic Acids Research*, 2005 Jul 1; 33(Web Server issue):W94-8.

[15] Kabsch W, Sander C, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 1983 Dec; 22(12):2577-637.

[16] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P., Protein Data Bank, *Nucleic Acids Research*, 2000; 28:235-242.

[17] Rost B., Twilight Zone of Protein Sequence Alignments, *Protein Engineering*, 1999; 12:85-94.