

# Statistical Analysis of Long-Range Interactions in Proteins

Jinmiao Chen  
School of Computer Engineering  
Nanyang Technological University  
Singapore, 639798  
pg05205549@ntu.edu.sg

Narendra S. Chaudhari  
School of Computer Engineering  
Nanyang Technological University  
Singapore, 639798  
asnarendra@ntu.edu.sg

## Abstract

*We carry out a comprehensive study of long-range interactions on a large data set of non-homologous proteins. Our study reveals that the long-range interactions between amino acids far apart are common in protein folding, and play an important role on the formation of secondary structure. Using residue-wise contact order(RWCO) to describe long-range interactions, we further evaluate the effect of long-range interactions on secondary structure prediction. We select six most popular prediction methods and collect their prediction results on the same set of proteins. All the six prediction methods show a significant negative correlation between prediction accuracy and RWCO.*

**Key words:** Long-range interactions, protein secondary structure, mutual information, residue-wise contact order, correlation coefficient.

## 1. Introduction

One of the most important open problems in computational biology concerns the computational prediction of the secondary structure of a protein given only the underlying amino acid sequence. During the last few decades, much effort has been made toward solving this problem, with various approaches including GOR (Garnier, Osguthorpe, and Robson) techniques [8], PHD(Profile network from HeiDelberg) method [4], nearest-neighbor methods [24] and support vector machines (SVMs) [11]. These methods are all based on a fixed-width window around a particular amino acid of interest, and thus usually do not explicitly consider long-range interactions between distant amino acids. This weakness of local window approaches is often cited as the current limitation to accurate secondary structure prediction.

In this paper, a comprehensive study is carried out on a large data set of non-homologous proteins. We statistically analyze the long-range interactions in proteins and study their effect on protein secondary structure prediction.

## 2. Data acquisition

### 2.1. EVA set

EVA(EVAluation of Automatic protein structure prediction) is a web-based server that provides a continuous, fully automated, and statistically significant analysis of structure prediction servers. Everyday, EVA downloads the newest protein structures from Protein Data Bank (PDB) [2]. The structures are added to MySQL databases, sequences are extracted for every protein chain, and are sent to each prediction server by META-PredictProtein [16]. META-PP collects the results and sends them to EVA. The central EVA site at Columbia is available at <http://cubic.bioc.columbia.edu/eva/>.

The EVA set, the largest sequence unique subset of PDB, fulfills the criteria that no pair in subset has more than 33% identical residues over more than 100 residues aligned. We extracted the EVA set from PDB based on the current EVA list with 3449 chains([http://cubic.bioc.columbia.edu/eva/res/unique\\_list.html](http://cubic.bioc.columbia.edu/eva/res/unique_list.html)). All proteins in the EVA set were sequence unique, i.e. they do NOT have *significant sequence similarity* to proteins of known structure (at the date of the submission of the new protein). The term *significant sequence similarity* refers to the following operational definition: Two proteins A and B are considered to be significantly similar if we can safely predict that B has the same structure as A by simply comparing their sequences.

To extract 3D coordinates of  $C_{\alpha}$  atoms, together with secondary structure, we run the DSSP program [15] on all the PDB files of the EVA proteins, excluding those

for which DSSP crashes due, for instance, to missing entries or format errors. The DSSP program was designed by Wolfgang Kabsch and Chris Sander to standardize secondary structure assignment. The DSSP is also a database of secondary structure assignments for all protein entries in the Protein Data Bank. In the current EVA list, there are some theoretical model structures and outdated PDB IDs(each structure in the PDB is represented by a 4 character identifier). After removing the redundant entries and those for which DSSP could not produce an output, we obtained the final set consisting of 3374 protein chains, which forms the source for our present study.

## 2.2. Collection of secondary structure predictions

Twenty-one secondary structure prediction methods participate in the EVA evaluation. They are apssp, phdpsi, jpred, phd, psipred, prof\_king, profsec, sspro2, jufo, phdprof, prof, prof0, prof1, prospect, pssp, sable, samt99\_sec, scratch, sspro, sspro1 and apssp2(available online at [http://cubic.bioc.columbia.edu/eva/doc/explain\\_methods.html](http://cubic.bioc.columbia.edu/eva/doc/explain_methods.html)). For each protein in the EVA unique list, EVA evaluated these secondary structure prediction methods at the deposition date of that protein. The prediction results of each prediction sever are downloadable at <http://cubic.bioc.columbia.edu/eva/doc/ftp.html>. Out of these 21 methods, we selected apssp [22], phdpsi [21], jpred [6], phd [3], psipred [14] and sspro2 [20] to evaluate the effect of long-range interactions on prediction accuracies.

## 3. Importance of long-range interactions in protein folding

The folding of a polypeptide chain into a compact, unique 3D structure is directed and stabilized by intra-molecular interactions between the constituent amino acid residues along the chain. The residues in a protein molecule are represented by their  $C_\alpha$  atoms. The complicated atomic structure of a protein can be represented as a chain trace, that is, the ordered succession of residue centers ( $C_\alpha$  atoms) described by their X, Y, Z coordinates. We used the three-dimensional  $C_\alpha$  coordinates to calculate residue-residue ( $C_\alpha$ - $C_\alpha$ ) distance matrices:

$$r_{i,j} = \text{sqrt}[(X_{C_\alpha}^i - X_{C_\alpha}^j)^2 + (Y_{C_\alpha}^i - Y_{C_\alpha}^j)^2 + (Z_{C_\alpha}^i - Z_{C_\alpha}^j)^2] \quad (1)$$

Two residues are considered to be in contact if any pair of  $C_\alpha$  atoms from each residue locate closer than a threshold value. Different threshold values have been used to define inter-residue contact, for example, 4.5Å, 6Å, 8Å, 10Å and 12Å.

Gromiha and Selvaraj [23, 9, 10] defined long-range interactions to be those residue-contacts for which the se-

quence separation of contacting residues are larger than 4 residues, and analyzed the role of long-range interactions in the folding of globular,  $(\alpha/\beta)_8$  barrel and membrane proteins. Their work revealed that long-range interactions play an active role in the stability of protein molecules. For globular proteins, they studied the influence of long-range interactions in different structural classes of proteins and found that 85% of residues are involved in long-range contacts. The all- $\alpha$  class proteins have more long-range contacts in the 4-10 range and the all- $\beta$  class proteins have more long-range contacts in the 11-20 range. The range 4-10 is favored by the  $\alpha + \beta$  class of proteins. In the  $\alpha/\beta$  class, the  $\alpha$ -helices and  $\beta$ -strands occur alternatively and some residue distances are necessary to form  $\beta$ -strand and barrel, which leads to having contacts in the 21-30 range. A similar trend was also observed in their study of  $(\alpha/\beta)_8$  barrel proteins. In the case of membrane proteins, the long-range interactions play an important role in the stabilization of helix-helix interactions in transmembrane helical(TM<sub>H</sub>) proteins and in the close packing of  $\beta$ -strands in transmembrane strand(TMS) proteins. Specifically, the TMS proteins prefer the 11-20 range, and the higher long-range contacts are influenced up to the range 21-30.

The above observation was obtained from a relative small data set(150 globular proteins and 49 membrane proteins). In our study, a more comprehensive survey is conducted on a large data set. We calculated the numbers of residue-contacts and the percentage of long-range contacts in the EVA protein set. Different partitioned distances are used to define residue contact and different threshold values of sequence separation are used for long-range contact. The results for different combinations of parameters are recorded in Table 1.

As observed from Table 1, long-range interactions are important in protein folding, and the fraction of long-range interactions is smallest when 6Å is selected as the threshold value for defining residue contact. Therefore, in the rest of this paper, we fixed the threshold value for residue contact to be 6Å.

## 4. Influence of long-range interactions on SS formation

In this section, we address the effect of long-range interactions on the formation of secondary structures of proteins.

There are several concrete examples of secondary structures whose formation is influenced by long-range interactions between amino acids [17, 18]. Minor and Kim [17] designed an 11-amino-acid sequence that folds as an alpha-helix when in one position but as a beta-sheet when in another position of the primary sequence of the IgG-binding domain of protein G. Their experiment demonstrated that non-local interactions

Table 1: The percentage of residues that are involved in long-range contacts. *c\_num* refers to the number of residue-contacts, and each column represents the percentage of long-range contacts defined by different threshold values.

cutoff	4.5Å	6 Å	8Å	10Å	12Å
<i>c_num</i>	52424	900798	2236370	5231096	9455424
> 8	76.80	48.09	57.04	62.40	69.86
> 9	74.42	46.80	55.66	60.94	68.25
> 10	71.56	45.56	54.33	59.64	66.85
> 11	69.17	44.41	53.06	58.41	65.60
> 12	67.65	43.26	51.86	57.27	64.42
> 13	65.69	42.20	50.75	56.20	63.29
> 14	63.60	41.18	49.68	55.18	62.18
> 15	61.52	40.17	48.64	54.17	61.10
> 16	60.17	39.21	47.60	53.16	60.05
> 17	58.61	38.27	46.60	52.17	58.99
> 18	56.99	37.33	45.63	51.19	57.94
> 19	55.49	36.46	44.66	50.21	56.89
> 20	54.15	35.60	43.70	49.22	55.83

can determine the secondary structure of peptide sequences of substantial length. It was also observed that small fragments of the same sequence are found in different secondary structures [19, 13, 25, 12]. Most recently, Crooks and Brenner showed that local inter-sequence information is insufficient to determine secondary structure, implying indirectly that nonlocal interactions is important for secondary structure formation [5]. In particular, they found that correlations between neighboring amino acids are essentially uninformative and only one-fourth of the total information needed to determine the secondary structure is available from local inter-sequence correlations. As they estimated, the entropy density of secondary structure sequences is 0.60 bits per residue, while the local mutual information between primary and secondary structure is only 0.16 bits per residue. The scarcity of local sequence information places severe constraints on any prediction algorithm that purports extract secondary structure information from local sequence correlations.

On the other hand, some other researchers denied the significant influence of long-range interactions on secondary structure specification. For example, Pan, X. M. et al [19] analyzed the relationship between the fraction of  $n$ -mers with a unique central secondary structural state and the length of  $n$ -mers in a low-homologous database. They computed the fraction of determinable  $n$ -mers as below:

$$F_{dt} = N_{dt}/(N_{dt} + N_{undt}) \quad (2)$$

where  $N_{dt}$  is the sum of appearance frequencies of determinable  $n$ -mers, and  $N_{undt}$  is the sum of appearance frequencies of undeterminable  $n$ -mers. It is found that the

fraction of determinable  $n$ -mers increases with increasing subsequence length. Based on their observation, they concluded that the minimal length of subsequence required to determine the central secondary structural state is about 14-17 residues. However, when  $n > 11$ ,  $N_{dt}$  and  $N_{undt}$  were too small to supply statistically relevant information for analysis, due to the limited size of the database used. For  $n = 14$  and 17, the value of  $F_{dt}$  is not presented in the original paper.

Is protein secondary structure primarily determined by local interactions between residues closely spaced along the amino acid backbone or by non-local tertiary interactions? As we know, quantitative measurement of sequence-structure correlations can elucidate the relative importance of different interactions to protein structure and facilitate the rational design of structure prediction algorithms. To answer the above question, we examined the strength and relative importance of correlations among amino acid identity and secondary structure, both for contacting residues closely located along the amino acid chain and for residues proximate in space but distantly separated along the chain.

The average strength of correlations between two discrete random variables is the mutual information, an information-theoretic measure of the knowledge that each variable carries about the other.

$$\begin{aligned} MI(X, Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= H(X) + H(Y) - H(X, Y) \quad (3) \end{aligned}$$

where  $H(X)$  is the entropy of  $X$ . When calculated with base 2 logarithms, mutual information is expressed in units of bits. A high mutual information value is a result of strong correlation, whereas a mutual information value of zero indicates uncorrelated variables.

Robson's studies show that the effect of one residue type on the conformation of residues up to eight residues distant plays a predominating role, while choosing shorter separation distance neglects significant information. The GOR techniques take into account only the local subsequence around the residue of interest by using an approximation:

$$I(S_i; R_1, R_2, \dots, R_{last}) \simeq I(S_i; R_{i-8}, \dots, R_{i+8}) \quad (4)$$

in which interactions between residues separated by more than 8 positions in the amino acid sequence are neglected. In our work, we define two residues to be in long-range contact if they are separated by at least 9 residues in the chain, and the distance between the  $\alpha$  carbon atoms is less than the threshold distance of 6 Å.

In table 2, we quantify the protein sequence-sequence, sequence-structure and structure-structure correlations using mutual information. All contacts have a  $C_\alpha$  distance of 6 Å or less, and are separated in the polypeptide chain by 2 or more residues;  $R_i$  is the amino acid at position  $i$

Table 2: Summary of mutual information in inter-residue contacts.

	$ i - j  \geq 2$	$2 \leq  i - j  \leq 8$	$ i - j  > 8$
$MI(R_i, R_j)$	0.007289	0.006261	0.017497
$MI(R_i, S_j)$	0.024069	0.014110	0.033919
$MI(S_i, S_j)$	0.459712	0.265390	0.262924

and  $S_i$  is the corresponding secondary structure. As we observed, the mutual information between residues in long-range contact is comparable with the mutual information between residues in local contact.

## 5. Negative effect of long-range interaction on prediction accuracy

Most of the current secondary structure prediction methods assign a secondary structure to the center of a local segment and thus usually do not explicitly consider long-range interactions between amino acids. There have been several experimental examples where a local segment alone is not sufficient to determine its secondary structure. So far, there are not many systematic studies on the effect of long-range interactions on prediction accuracy. Interestingly, some earlier studies suggested that the current insufficient accuracy of secondary structure prediction does not result from the neglect of long-range interactions. Fiser et al. [7] compared the accuracy of secondary structure prediction for residues involved in significant long-range interactions and for the other residues, and concluded that the role of long-range interactions in defining the secondary structures is overestimated.

In the present study, we address the effect of long-range interaction on the prediction accuracy of secondary structures and come to a contrary conclusion that the accuracy gets worse on average for residues with long-range interactions. We use Residue Contact Order(RCO) and Residue-Wise Contact Order(RWCO) to describe the extent of long-range interactions and examine the relationship between the RCO/RWCO and the prediction accuracy on the EVA set.

### 5.1. Quantity of long-range interactions

We need to specify numerically long-range interaction in proteins before we study its effect on prediction accuracy.

**5.1.1. Contact order** The contact order is the average sequence separation between contacting residues in the native state of a protein. This simple index has been used as

a measure of the complexity of protein topology.

$$CO = \frac{1}{N} \sum_{i=1}^{L-3} \sum_{j=i+3}^L |i - j| \delta_{ij} \quad (5)$$

where  $N$  is the total number of contacts in the protein;  $L$  is the length of the protein;  $\delta_{ij}$  ( $1 \leq i, j \leq L$ ) represent the contact map of the protein, i.e.  $\delta_{ij} = 1$  if residue  $i$  and  $j$  are in contact, and 0 otherwise.

**5.1.2. Relative contact order** Relative contact order reflects the relative importance of local and non-local contacts to a protein’s native structure. Relative contact order is the average sequence distance between all pairs of contacting residues normalized by the total sequence length.

$$CO_{rel} = \frac{1}{N \times L} \sum_{i=1}^{L-3} \sum_{j=i+3}^L |i - j| \delta_{ij} \quad (6)$$

The contact order and relative contact order were originally introduced to quantify the complexity of the native topology of proteins and investigate the correlation between the native structure and its folding rate. As such, both the contact order and the relative contact order are per-protein quantity. In the below sections, we extend the definition of contact order to make it a per-residue quantity.

**5.1.3. Residue contact order** Kihara, D. [16] introduced for the first time the residue contact order to evaluate the long-range interactions for each residue. The RCO indicates the average sequence separation of contacting residues to a residue of interest.

$$RCO_i = \frac{1}{n} \sum_{j:|j-i|>2}^L |i - j| \delta_{ij} \quad (7)$$

where  $n$  is the number of contacts between residue  $i$  and the others. We don’t take into account residue-residue pairs for which  $|i - j|$  is smaller than 3. In this way, we exclude trivial contacts between nearest- and next-nearest residues along the sequence.

**5.1.4. Residue-wise contact order** Residue-wise contact order was originally introduced by Nishikawa and Kinjo as a new kind of 1D protein structure, which represents the extent of long-range contacts. The residue-wise contact order (RWCO) of the  $i$ -th residue of a protein is defined by

$$o_i = \frac{1}{L} \sum_{j:|j-i|>2}^L |i - j| C_{i,j} \quad (8)$$

where  $L$  is the length of the amino acid sequence of the protein and the contact between two residues is defined by a sigmoid function:

$$C_{i,j} = \frac{1}{1 + \exp(r_{i,j} - d_c)} \quad (9)$$

where  $r_{i,j}$  is the distance between the  $C_\alpha$  atoms of the  $i$ -th and  $j$ -th residues,  $d_c$  is the cut-off distance for the contact definition. The RWCO is a generalization of the residue contact order, and is also a per-residue quantity.

Kihara, D. [16] have examined the effect of global features of proteins on the accuracy of protein secondary structure prediction. He found that the prediction accuracy is not affected by the length, the contact order, and relative contact order of proteins. In the next section, we will investigate the influence of long-range interactions on the prediction accuracy by using the RCO and RWCO as measures of long-range interaction.

## 5.2. $\alpha$ -helices are better predicted than $\beta$ -strands

Existing approaches can predict  $\alpha$ -helices with accuracies between 70% and 80%. However, the problem of predicting  $\beta$ -sheet regions has not been treated at a comparable level. The overall prediction accuracies of apssp, phd, phdpsi, jpred, psipred and sspro2 for the EVA proteins are summarized in Table 3. The EVA set most probably includes sequences that were used to train these programs, which would inflate the accuracy beyond that shown in the original publications.

Table 3: Performance of several prediction methods

method	num.P	num.R	Q <sub>3</sub>	Q <sub>C</sub>	Q <sub>H</sub>	Q <sub>E</sub>
apssp	1264	247762	77.05	82.70	77.58	65.38
phdpsi	1780	359152	74.11	76.97	77.18	63.17
jpred	1301	258007	74.23	84.61	71.34	59.01
phd	1883	380026	70.62	72.56	74.12	60.94
psipred	1667	337333	78.21	79.08	81.97	70.18
sspro2	1320	270399	77.28	79.82	81.84	64.66

In table 3, num.P and num.R represent the number of proteins and number of residues respectively;  $Q_3$  is the overall three-state prediction percentage defined as the ratio of correctly predicted residues to the total number of residues;  $Q_H$ ,  $Q_E$  and  $Q_C$  are the percentage of correctly predicted residues observed in class H( $\alpha$ -helices), E( $\beta$ -sheets) and C(others) respectively. It is shown that  $\alpha$ -helices are better predicted than  $\beta$ -strands by each prediction method.

$\beta$ -sheet structures typically range over several discontinuous sections in an amino acid sequence, whereas the configurations of  $\alpha$ -helix are continuous and their dependency patterns are more regular. We calculated the RWCO and RCO values for residues in different classes of secondary structures. As indicated by the results in Table 4 5, the RWCO and RCO values of residues in class E are larger than those of residues in class H. That is, the formation of  $\beta$ -strands is more affected by long-range interactions than that of  $\alpha$ -helices.

Table 4: RWCO of residues in three secondary structure classes (threshold=6Å).

class	num	max	min	average
H	144182	5.38045	0.00104984	0.280893
E	85319	5.5523	0.000118382	0.718816
C	164782	6.12381	0	0.360048

Table 5: RCO of residues in three secondary structure classes ( threshold=6Å).

class	num	max	min	average
H	144182	838.5	0	11.6377
E	85319	595	0	41.9357
C	164782	1001	0	26.1306

## 5.3. Relationship between RCO and prediction accuracy

Residues with RCO values larger than 8 are considered to have long-range contacts with other residues located beyond the local window of most window-based predictors.

Table 6: The prediction accuracy for residues with local interactions and long-range interactions.

method	RCO $\leq$ 8		RCO $>$ 8	
	num.R	accuracy	num.R	accuracy
apssp	141951	78.92%	105811	74.54%
phdpsi	204357	76.50%	154795	70.95%
jpred	148020	76.55%	109987	71.12%
phd	214573	73.18%	165453	67.31%
psipred	190802	80.00%	146531	75.88%
prof_king	167460	76.09%	125468	73.10%
profsec	187740	78.41%	140746	73.82%
sspro2	153471	80.01%	116928	73.70%

Table 6 records the prediction accuracies for residues with RCO $\leq$ 8 and those with RCO $>$ 8. Prediction accuracies of all the predictors drop if the residue has contacts with distant residues located outside the local input window.

## 5.4. Prediction accuracy with respect to RWCO

The correlation of prediction accuracy and RWCO is illustrated in Figure 1- 6. Six Angstroms is used as the threshold value for residue contact. The residues of the test proteins are split into subsets with a bin size of around 5000. We observed that the secondary structure prediction accuracy drops as the RWCO value becomes higher. The significant negative correlation coefficient in Table 7 verifies our observations. Note that all the prediction methods used show this negative correlation.

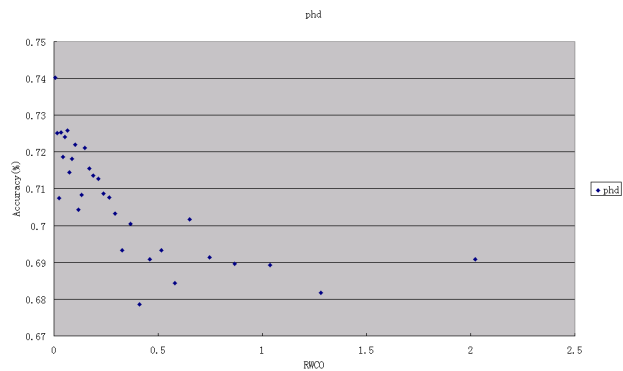


Figure 1: The correlation between the  $Q_3$  prediction accuracy of phd and RWCO

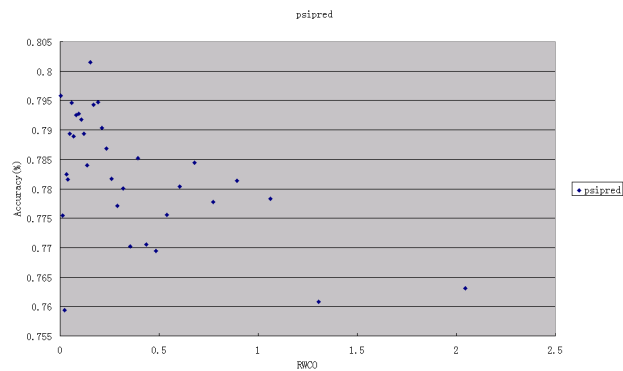


Figure 3: The correlation between the  $Q_3$  prediction accuracy of psipred and RWCO

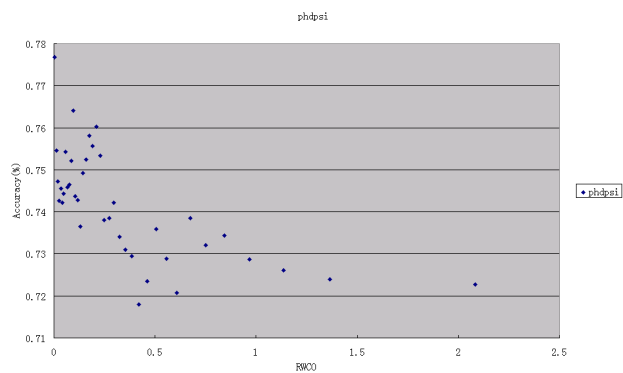


Figure 2: The correlation between the  $Q_3$  prediction accuracy of phdpsi and RWCO

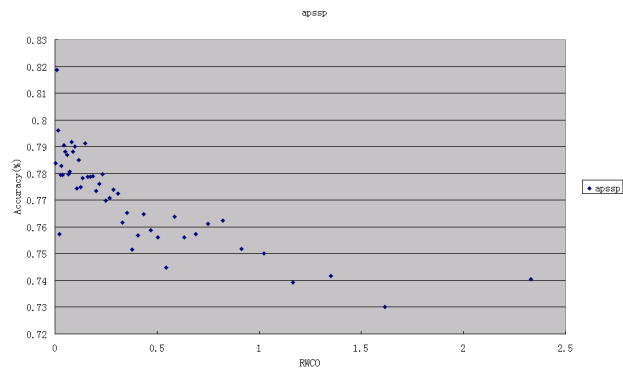


Figure 4: The correlation between the  $Q_3$  prediction accuracy of apssp and RWCO

We compute the Pearson's Correlation Coefficient between RWCO and prediction accuracy using the data points in Figure 1- 6. The formula for Pearson's correlation takes on many forms. A commonly used formula is shown below:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (10)$$

Table 7: The correlation coefficient of the residue-wise contact order and the prediction accuracy.

method	correlation coefficient
apssp	-0.793196
jpred	-0.834448
phd	-0.694418
phdpsi	-0.646376
psipred	-0.563313
sspro2	-0.763202

## 6. Discussion

We have addressed the negative effect of long-range interactions on protein secondary structure prediction accuracy. It is suggested that to improve protein secondary structure prediction, one should use distant information, that is not contained in local input windows. Substantially increasing the size of input window does not improve the performance because of the problem of over-fitting. Gianluca Pollastri [1] developed a bidirectional recurrent neural network(BRNN) to overcome the drawbacks of local fixed-window approaches. However, recurrent neural networks have difficulties in learning long-term dependencies. Moreover, it is hard to detect the sparse and weak long-range signals while ignore the additional noise found over large distances. The BRNN system achieves an accuracy close to 76%, comparable to the current window-based approaches. There is still a margin of 10%-15% left for further improvement to reach the upper limit of the prediction accuracy of 90%, especially for the prediction of  $\beta$ -strands.

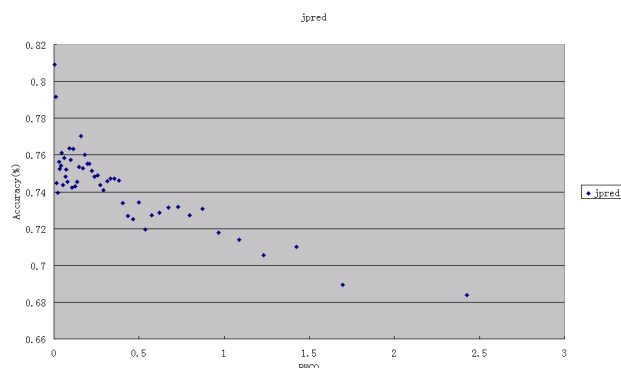


Figure 5: The correlation between the  $Q_3$  accuracy of jpred prediction and RWCO.

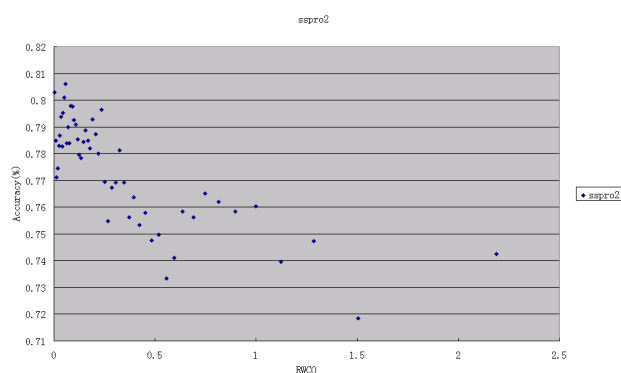


Figure 6: The correlation between the  $Q_3$  accuracy of sppro2 prediction and RWCO.

## References

- [1] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda, *Bidirectional iohmms and recurrent neural networks for protein secondary structure prediction*, CLUEB, Bologna, Italy, 2000.
- [2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The protein data bank*, *Nucleic Acids Res.* **28** (2000), 235–242.
- [3] B. Rost, *Phd: predicting one-dimensional protein structure by profile based neural networks*, *Meth. Enzymol.* **266** (1996), 525–539.
- [4] B. Rost and C. Sander, *Prediction of protein secondary structure at better than 70% accuracy*, *J. Mol. Biol.* **232** (1993), 584–599.
- [5] G. E. Crooks and S. E. Brenner, *Protein secondary structure: Entropy, correlations and prediction*, *Bioinformatics* **20** (2004), no. 10, 1603–1611.
- [6] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, *Jpred: A consensus secondary structure prediction server*, *Bioinformatics* **14** (1998), 892–893.
- [7] A. Fiser, Z. Dosztányi, and I. Simon, *The role of long-range interactions in defining the secondary structure of proteins is overestimated*, *Comput. Appl. Biosci.* **13** (1997), no. 3, 297–301.
- [8] J. F. Gibrat, J. Garnier, and B. Robson, *Further developments of protein secondary structure prediction using information theory*, *Journal of Molecular Biology* **198** (1987), no. 3, 425–443.
- [9] M. M. Gromiha and S. Selvaraj, *Importance of long-range interactions in protein folding*, *Biophysical Chemistry* **77** (1999), 49–68.
- [10] ———, *Role of medium- and long-range interactions in discriminating globular and membrane proteins*, *International Journal of Biological Macromolecules* **29** (2001), 25–34.
- [11] S. Hua and Z. Sun, *A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach*, *J. Mol. Biol.* **308** (2001), 397–407.
- [12] K. Ikeda and J. Higo, *Free-energy landscape of a chameleon sequence in explicit water and its inherent alpha/beta bifacial property*, *Protein Sci.* **12** (2003), no. 11, 2542–2548.
- [13] I. Jacoboni, P. L. Martelli, P. Fariselli, M. Compiani, and R. Casadio, *Predictions of protein segments with the same amino acid sequence and different secondary structure: a benchmark for predictive methods*, *Proteins* **41** (2000), no. 4, 535–544.
- [14] D. T. Jones, *Protein secondary structure prediction based on position-specific scoring matrices*, *Journal of Molecular Biology* **292** (1999), no. 2, 195–202.
- [15] W. Kabsch and C. Sander, *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features*, *Biopolymers* **22** (1989), 2577–2637.
- [16] D. Kihara, *The effect of long-range interactions on the secondary structure formation of proteins*, *Protein Science* **14** (2005), 1955–1963.
- [17] D. L. Minor and P. S. Kim, *Context-dependent secondary structure formation of a designed protein sequence*, *Nature* **380** (1996), 730–734.
- [18] V. Munoz, P. Cronet, E. Lopez-Hernandez, and L. Serrano, *Analysis of the effect of local interactions on protein stability*, *Folding and Design* **1** (1996), 167–178.
- [19] X. M. Pan, W. D. Niu, and Z. X. Wang, *What is the minimum number of residues to determine the secondary structural state*, *J. Protein Chem.* **18** (1999), no. 5, 579–584.
- [20] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, *Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles*, *Proteins* **47** (2002), 228–235.
- [21] D. Przybylski and B. Rost, *Alignments grow, secondary structure prediction improves*, *Proteins* **46** (2002), 197–205.
- [22] G. P. S. Raghava, *Protein secondary structure prediction using nearest neighbor and neural network approach*, *CASP4* (2000), 75–76.
- [23] S. Selvaraj and M. M. Gromiha, *Importance of long-range interactions in  $(\alpha/\beta)_8$* .
- [24] Tau-Mu Yi and Eric S. Lander, *Protein secondary structure prediction using nearest-neighbor methods*, *J. Mol. Biol.* **232** (1993), 1117–1129.
- [25] X. Zhou, F. Alber, G. Folkers, G. H. Gonnet, and G. Chelvanayagam, *An analysis of the helix-to-strand transition between peptides with identical sequence*, *Proteins* **41** (2000), no. 2, 248–256.