The effect of long-range interactions on the secondary structure formation of proteins

DAISUKE KIHARA
Department of Biological Sciences/Computer Science, Markey Center for Structural Biology, The Bindley Bioscience Center, Purdue University, West Lafayette, Indiana 47907, USA

(RECEIVED March 25, 2005; FINAL REVISION April 19, 2005; ACCEPTED April 19, 2005)

Abstract

The influence of long-range residue interactions on defining secondary structure in a protein has long been discussed and is often cited as the current limitation to accurate secondary structure prediction. There are several experimental examples where a local sequence alone is not sufficient to determine its secondary structure, but a comprehensive survey on a large data set has not yet been done. Interestingly, some earlier studies denied the negative effect of long-range interactions on secondary structure prediction accuracy. Here, we have introduced the residue contact order (RCO), which directly indicates the separation of contacting residues in terms of the position in the sequence, and examined the relationship between the RCO and the prediction accuracy. A large data set of 2777 nonhomologous proteins was used in our analysis. Unlike previous studies, we do find that prediction accuracy drops as residues have contacts with more distant residues. Moreover, this negative correlation between the RCO and the prediction accuracy was found not only for β-strands, but also for α-helices. The prediction accuracy of β-strands is lower if residues have a high RCO or a low RCO, which corresponds to the situation that a β-sheet is formed by β-strands from different chains in a protein complex. The reason why the current study draws the opposite conclusion from the previous studies is examined. The implication for protein folding is also discussed.

Keywords: secondary structure prediction; long-range interaction; residue contact order; β-strand formation

Predicting the secondary structure of proteins from their amino acid sequences has one of the oldest histories in bioinformatics research (Rost and Sander 2000). Earlier works mostly relied on propensities of amino acids for three states of the secondary structure, namely, α-helices, β-strands, or others (often referred to as coils). Essentially in these methods, a local region with a high density of amino acids with a high propensity to a certain type of secondary structure is predicted to form that particular secondary structure type (Lim 1974; Chou and Fasman 1978a,b; Garnier et al. 1978). The current generation of prediction methods employ machine learning techniques such as neural networks (Rost and Sander 1994; Jones 1999; Petersen et al. 2000), hidden Markov models (Karplus et al. 1998; Lin et al. 2005), and support vector machines (Hua and Sun 2001; Ward et al. 2003; Guo et al. 2004), which try to capture local sequence patterns of known examples of secondary structures in an input multiple sequence alignment. Recently some prediction methods have extended their prediction capability from the conventional three-state prediction to more states, including 3_{10} helices, β-bulges, and turns (Karchin et al. 2003; Kuang et al. 2004).

Current best methods achieve a three-state per residue-based accuracy (the Q3 measure) of 75%–80% (Rost 2001; Rost and Eyrich 2001; McGuffin and Jones 2003). In spite of gradual improvements made by

Reprint requests to: Daisuke Kihara, Department of Biological Sciences/Computer Science, Markey Center for Structural Biology, The Bindley Bioscience Center, Purdue University, Lilly Hall, West Lafayette, IN 47907, USA; e-mail: dkihara@purdue.edu; fax: (765) 496-1189.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051479505.

modern methods including those mentioned above, there is still a margin of 10%–15% left for further improvement to reach the upper limit of the prediction accuracy of ∼90%. This upper limit of the prediction accuracy was estimated based on two observations: first, 5%–15% of secondary structure points can differ between different X-ray structures and NMR models of the same protein; second, there is inconsistency of secondary structure assignments by different methods, e.g., DSSP (Kabsch and Sander 1983) and STRIDE (Frishman and Argos 1995), and also of their parameters (Levin 1997; Rost 2001).

It has been discussed that one of the main reasons for the limitation comes from long-range amino acid interactions, which may overwrite local sequence propensity of secondary structures, since most of the current methods assign a secondary structure to a window of a local segment and thus usually do not explicitly consider long-range interactions of amino acids. Indeed, we can easily find several concrete examples of secondary structures whose formation is influenced by long-range interactions (Minor and Kim 1996; Munoz et al. 1996). An interesting experiment was conducted by Minor and Kim (1996), where an 11-residue-long sequence changed its secondary structure according to its position in the global fold of protein G. It was also observed that small fragments of the same sequence are found in different secondary structures (Pan et al. 1999; Jacoboni et al. 2000; Zhou et al. 2000; Ikeda and Higo 2003).

In spite of the general consensus and discussion, so far there are not many systematic studies on the influence of long-range interactions on the prediction accuracy of secondary structures. Fiser et al. (1997) compared the accuracy of secondary structure prediction for residues with many long-range contacts and for the other residues and concluded that the role of long-range interactions in defining the secondary structures is overestimated. Pan et al. (1999) concluded that the current insufficient accuracy of secondary structure prediction may result from the limitation of the available database size. Therefore, interestingly, both of them concluded that the long-range interaction does not have a strong effect on the prediction accuracy of secondary structures. But in our opinion, their statements come from indirect observation of long-range effects and do not approach analysis of long-range interactions well enough. Crooks and Brenner (2004) showed that local sequence information is insufficient to determine secondary structure, implying indirectly that nonlocal interaction is important for secondary structure formation. There are some other benchmark reports of secondary structure prediction methods, but none of them mention the effect of long-range interactions (Levin 1997; Przybylski and Rost 2002; McGuffin and Jones 2003).

In the current study, we directly address the effect of long-range interaction on the accuracy of current secondary structure prediction methods and come to a different conclusion. We introduce the residue contact order (RCO), which describes the separation of contacting residues in terms of the position in the sequence, and examine the relationship between the RCO and the prediction accuracy on a large, nonhomologous data set of 2777 proteins. Unlike previous studies, we do find a negative correlation between high RCO and the prediction accuracy. Typically, mispredicted residues with a high RCO are those that interact with other residues in a different domain. Interestingly, the negative correlation was found not only for β-strands, but also for α-helices. For β-strands, the prediction accuracy is relatively low when residues have a high RCO or a low RCO, indicating that there are many cases when formation of β-strands is affected by long-range interactions.

Results

Overall prediction accuracy

The overall prediction accuracy of PSIPRED, Jnet, and PREDATOR for the entire 2777 benchmark proteins is summarized in Table 1, although comparison of their performance is not the main interest. It is shown that α-helices can be better predicted than β-strands (e.g., cf. Q3a and Q3b), which is consistent with other studies (Rost 2001; Rost and Eyrich 2001; McGuffin and Jones 2003). Keep in mind that the benchmark proteins most probably include sequences that were used to train these programs, which would inflate the accuracy beyond that shown in the original publications. It is notable that ∼60% of the proteins have at least one residue whose secondary structure is oppositely predicted (BADp). Comparing BADa and BADb, more residues in β-strands tend to be predicted oppositely than those in α-helices. This may indicate that some β-strands have a high propensity for α-helices, which were captured by the prediction methods.

Generally, the accuracy (Q3 and SOV) is not affected by the size of proteins (Fig. 1). The large standard deviation for smaller proteins is simply because the variation of the number of correctly predicted residues (i.e., the numerator) has a greater effect on the final accuracy because of smaller length (i.e., the denominator). This plot is the same as Figure 3 of Levin's report (Levin 1997). Figure 2 shows the Q3 accuracy relative to the relative contact order (CO_{rel}) for all the benchmark proteins. The accuracy is not affected by the CO_{rel} of proteins. The relation between the contact order (CO) of proteins and the accuracy is very similar to Figure 1, because the CO largely reflects the length of proteins (data not shown). Results of Jnet and PREDATOR are

Table 1. *Summary of the prediction accuracy*

	Sov[a]	Q3[b]	Q3ab[c]	Q3a[d]	Q3b[e]	Q3pab[f]	CorrH[g]	CorrE[h]	BADp[i]	BADa[j]	BADb[k]
PSIPRED	83.4	79.8	78.6	82.2	69.9	80.6	0.70	0.64	58.0	1.8	4.3
Jnet	81.6	76.3	71.7	75.8	64.9	78.4	0.65	0.59	67.0	3.3	5.6
PREDATOR	75.9	69.0	58.8	68.3	47.2	72.2	0.54	0.45	78.2	5.4	11.5

[a] The modified Sov segment-based measure (Zemla et al. 1999).

[b] The average Q3 measure over the data set. The Q3 measure is the percentage of the correctly predicted residues in a test protein (Schulz and Shirmer 1979).

[c] The average Q3 measure is calculated only for residues that form α-helices or β-strands.

[d] The average Q3 measure calculated only for residues in α-helices.

[e] The average Q3 measure calculated only for β-strand residues.

[f] The average Q3 measure is calculated for residues predicted to be either in α-helices or β-strands.

[g] Matthews' correlation coefficient (Matthews 1975) for α-helices. The value is averaged over all the test proteins that have α-helices.

[h] Matthews' correlation coefficient for β-strands.

[i] Percentage of test proteins in the data set that have helical residues predicted as β-strand or β-strand residues predicted as α-helix.

[j] Percentage of helical residues predicted as β-strand.

[k] Percentage of β-strand residues predicted as α-helix.

not shown here, because they are essentially the same as in Figures 1 and 2. In summary, the accuracy of secondary structure prediction is not affected by the length, CO, or CO_{rel} of proteins.

The residue contact order (RCO) and the number of contacts

In the previous section, the effect of global features of proteins on the accuracy of secondary structure prediction, namely, the length and the topology (CO and CO_{rel}) was examined. Next, before investigating the influence of long-range interactions on the prediction accuracy by using the RCO, we will discuss the number of residue contacts, which indicates the packing density of residues (Nishikawa and Ooi 1980). The RCO is weakly related to the number of contacts, because a sufficient protein size is required to have a high number of contacts and a high RCO (Fig. 3B). Figure 3A shows the cumulative fraction of the RCO for different secondary structures. Around 80% of residues have an RCO of <50. On average, residues in β-strands have slightly higher RCO values.

Figure 4 shows the prediction accuracy with respect to the number of contacts. Looking at the range of the number of contacts (X-axis) where there is a sufficient amount of data, say ~5–12, α-helices and β-strands are better predicted when they are well packed (i.e., higher number of contacts). Residues in the other conformations (turns, loops) tend to be better predicted when they have a rather smaller number of contacts. One of the reasons for this observation may be that the sequence patterns of α-helices and β-strands in the core and loops or turns in the surface of proteins are better learned due to the abundance of training data.

Prediction accuracy with respect to RCO

In our study, we used RCO to describe long-range interactions. Unlike previous works, which mention the effect of long-range interactions (Fiser et al. 1997; Pan et al. 1999), the RCO directly indicates the average sequence separation of contacting residues to a residue of interest. Figure 5A illustrates a strong negative correlation between RCO values and prediction accuracy by PSIPRED, clearly showing that the long-range interactions do affect the accuracy of secondary structure prediction. In this plot, a contact threshold value of 6 Å and a window size of three

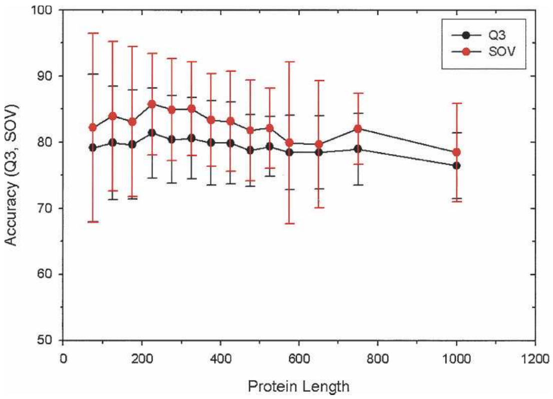

Figure 1. The accuracy of the secondary structure prediction by PSIPRED. The data set is split into subsets by length, with a bin size of 50 for proteins up to 600 residues long (50–100, 100–150, etc.) and a bin size of 100 for proteins 600–800 residues long. Proteins of 800 residues or longer are put together. The Q3 and the Sov accuracy measures are computed for each protein and averaged over each subset. The dots show the average and the error bars show the standard deviation.

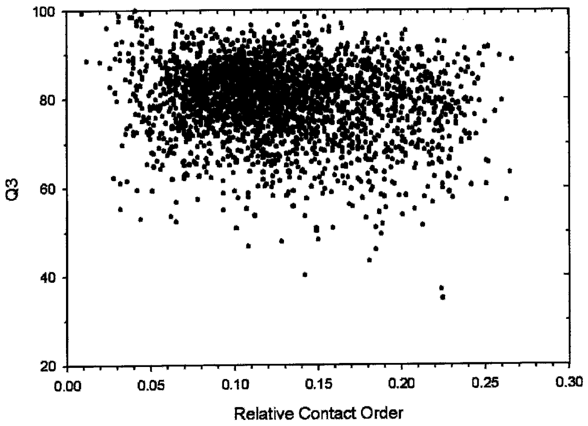

Figure 2. The Q3 accuracy measure with respect to the relative contact order of the proteins. PSIPRED is used for the prediction. Any two residues are considered to be in contact if any pair of the heavy atoms from each residue locate within 6 Å.

are used for illustration because this combination of the parameters showed the strongest negative correlation among those combinations tested, although almost all the other combinations of parameters also showed significant negative correlation coefficients (Table 2). The negative correlation becomes even larger when only residues with an RCO of up to 150 are used (residues with an RCO of up to 150 account for > 99% of the data points). Note that all the prediction methods used show this negative correlation. It was observed that the negative correlation decays quickly in the case of PREDATOR. This might be caused by the interesting feature of PREDATOR that seeks possible hydrogen-bonded residue pairs in interacting β-strands, thus taking long-range interaction into account. We have also tried a slightly different definition of the RCO, where adjacent residues are not counted in the contacting residues, resulting in essentially the same negative correlation with prediction accuracy.

In Figure 5B, the correlation is shown separately for the three states of secondary structures. The prediction accuracy of not only β-strands, but also α-helices, is strongly influenced by high RCO interactions, which shows a clear contrast with residues in coils. To examine the statistical significance of this observation, we performed a χ^2 test using four categories of residues, both with a lower (50–150) and a higher (150–250) RCO, and for each of them, its secondary structure is either correctly or wrongly predicted. The null hypothesis that the two valuables are independent is clearly rejected for residues in α-helices and β-strands, with the χ^2 value of 59.0 and 80.8 for α-helices and β-strands, respectively. As for the coil regions, the null hypothesis still holds with the χ^2 value of 2.6 using the significance level of 5%.

Intriguingly, from Figure 5B we can also see that β-strands with a very low RCO (say, < 20) are also not well predicted. A very low RCO indicates that the residue has no interactions with others or has only very local interactions.

Several examples of wrongly predicted secondary structure segments are shown in Figure 6. The first three figures, A–C, show high RCO residues whose secondary structures are wrongly predicted. The first protein, 1d3gA, forms a TIM-barrel fold (Fig. 6A). Eighteen residues at the interface of closing up the barrel have high RCO values, and the secondary structures of 44.4% of them are not correctly predicted. In the second example of 1a0i (Fig. 6B), the N-terminal tail wraps around the middle part of the protein, with a β-strand in the tail forming two β-sheets (shown in yellow), and β-strands in the middle of the protein.

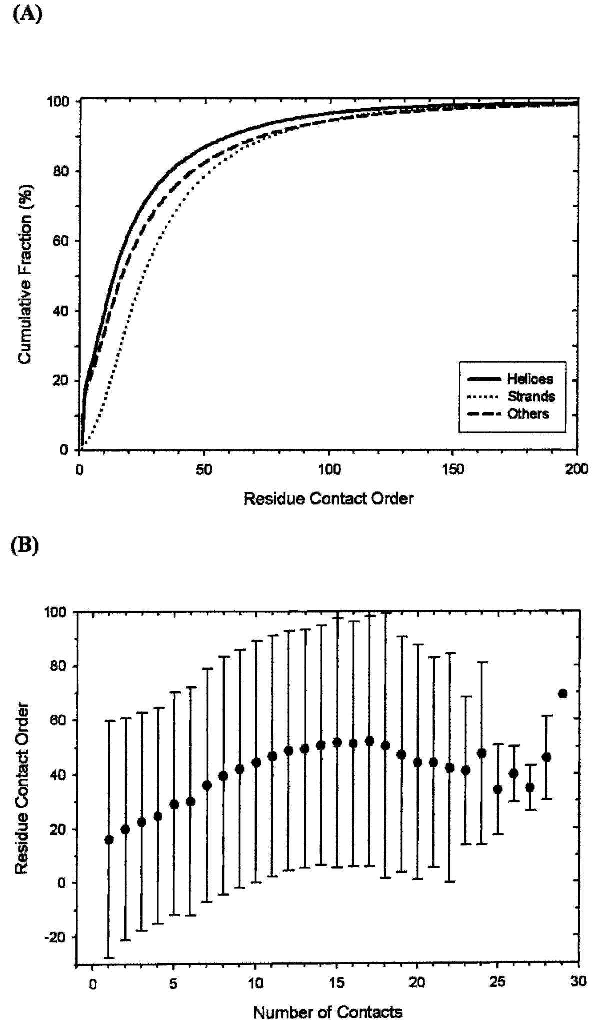

Figure 3. (*A*) Cumulative fraction of the RCO for each secondary structure. (*B*) The RCO as a function of the number of contacts of residues. The threshold value for residue contact is 6 Å. Error bars show the standard deviation.

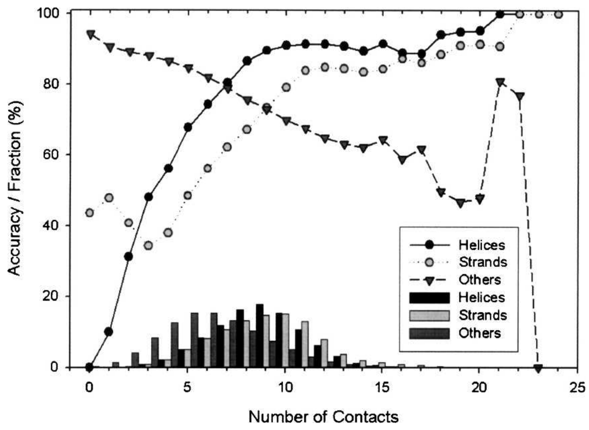

Figure 4. The Q3 accuracy with respect to the number of contacts. The line plots show the Q3 accuracy of PSIPRED for residues with different numbers of contacts in α-helices (black dots), β-strands (gray dots), and others (triangles). The bars at the *bottom* of the chart show the amount (fraction) of residues with each number of contacts classified by the secondary structure: black, α-helices; pale gray, β-strands; dark gray, others.

Nine out of 11 residues in the β-sheets are wrongly predicted. 1aofA (Fig. 6C) forms an eight-propeller fold (the bottom domain in the figure), and again residues at the interface of closing up the propeller have wrong secondary structure prediction. They include residues in an α-helix and a β-strand at the C terminus (shown in red). Figure 6D is an example of a mispredicted β-strand with a high RCO value: The highlighted strand of residues 146–152 in 1pmi is sandwiched between two β-strands that consist of distant residues, 266–271 and 285–297, respectively; thus, its strand formation is speculated to be assisted by hydrogen bonds from the two adjacent strands. In contrast to Figure 6D, the last two examples, Figure 6, E and F, show mispredicted strands with a low RCO value. The C-terminal β-strand in 1k3bC forms a β-sheet with another strand in a different chain, 1k3bB. Since the C-terminal strand is isolated from the core body of 1k3bC, its RCO value calculated within the chain C is very low (Fig. 6E). Similarly, the highlighted strand in Figure 6F is stabilized by forming a β-sheet with a strand from a different chain. Another type of observed mispredicted strand with a low RCO is a strand that forms a two-stranded β-sheet with adjacent strands.

What we observed in Figure 5 is that the secondary structure prediction accuracy drops as the RCO value becomes higher. But note that high RCO here means an RCO of >100–250, not just 20 or 30. Recalling that neural network-based approaches (PSIPRED and Jnet) use a window of a short length (15 and 17, respectively), it would be easy to understand if the prediction accuracy drops when the RCO of a residue goes beyond that window size. But the observation is different than that; the accuracy keeps dropping as the RCO grows. Why is the average accuracy of residues with an RCO of 20 and those with an RCO of 200 so different, although both of them have inter-residue contacts beyond the window size? Residues with a very high RCO of 150–250 are

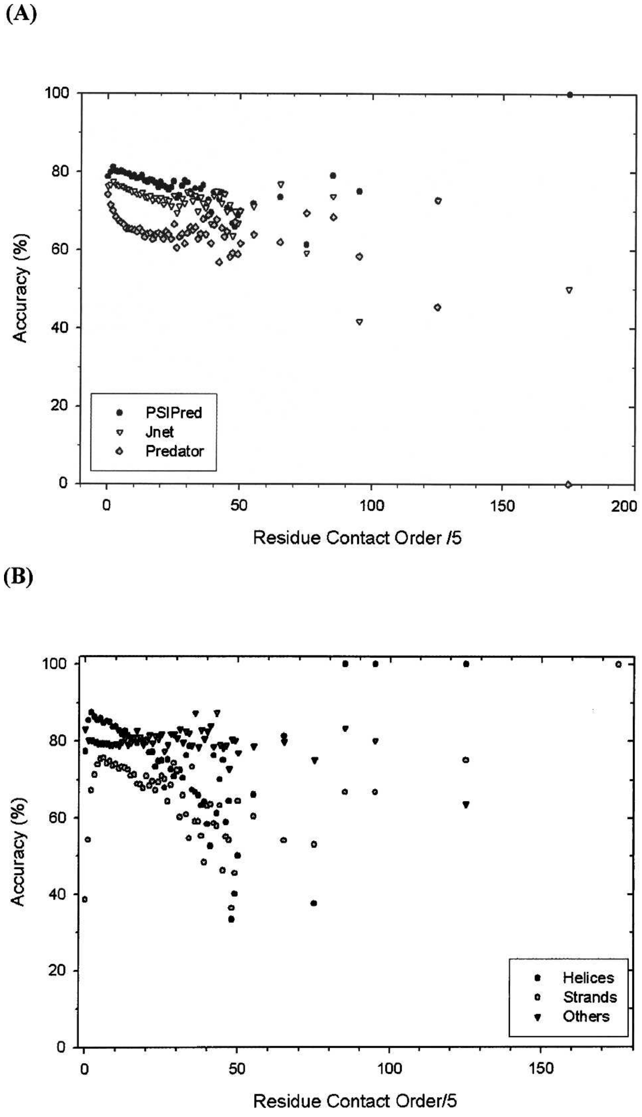

Figure 5. The correlation between the Q3 prediction accuracy and RCO. Six Angstroms is used as the threshold value for residue contacts. Note that the RCO is divided by five on the *X*-axis. (*A*) The RCO for each residue is averaged over a window size of three. The residues in the test proteins are split into subsets by the RCO with a bin size of 5–250 (0–4, 5–9 . . . 266–270, 271–275) and with a bin size of 50 for the residues with an RCO 276–624 residues long. Then the rest of the residues with an RCO of 625 or higher are put together. Predictions by the three methods are plotted: black dots, PSIPRED; triangles, Jnet; diamonds, PREDATOR. (*B*) The three different secondary structure conformations are separately plotted: black dots, α-helices; gray dots, β-strands; triangles, others. PSIPRED is used.

Table 2. *The correlation coefficient of the residue contact order and the prediction accuracy*

Threshold value[a]	4.5 Å						6 Å					
Window size[b]	3	5	9	15	21	25	3	5	9	15	21	25
PSIPRED	−0.720	−0.729	−0.703	−0.531	−0.236	−0.360	−0.845	−0.772	−0.673	−0.795	−0.818	−0.657
Jnet	−0.518	−0.589	−0.553	−0.510	−0.147	−0.141	−0.571	−0.587	−0.517	−0.483	−0.703	−0.563
PREDATOR	−0.324	−0.499	−0.586	−0.586	−0.411	−0.436	−0.490	−0.489	−0.572	−0.790	−0.694	−0.496

The correlation coefficient is calculated using data points of residue contact order (RCO) with > 100 residues.

[a] The threshold value used to define inter-residue contacts.

[b] The RCOs were averaged using a window of this size.

usually those that are caught at an interface of two domains, or that locate in a long terminal tail that wraps around a different domain of the protein. What our results imply is that the secondary structure of these kinds of residues is affected by interactions with distant domains, but the secondary structure of residues with an RCO of just beyond the window size, and interacting just with residues in the same domain (thus well packed, Fig. 4), is already well learned by the neural network, because there are plenty of examples of sequence patterns for secondary structure in a core of proteins. Practically, regardless of the existence of long-range interactions, current machine learning techniques are good enough to learn sequence patterns for secondary structures, but since the examples of residues sandwiched by two domains or in a wrapping tail whose secondary structure is severely affected by long-range interactions are still not abundant compared with the residues in a core of proteins, the prediction methods tested failed to predict their secondary structure. In other words, sequence patterns of the wrongly predicted secondary structures in domain interfaces or wrapping tails are different from the patterns of the sequences of the same secondary structure in the core of proteins, and even current machine learning techniques could not capture them. This indicates that long-range interactions indeed override the secondary structure propensity of local sequences.

Discussion

We have addressed the effect of long-range interactions on the formation of secondary structures of proteins, a topic that has been long discussed and has accumulated much anecdotal experimental evidence. If it is not rare that long-range interaction overrides the secondary structure propensity of local sequences, it is natural to speculate that the accuracy of secondary structure prediction gets worse on average for residues with such long-range interactions. Contrary to this speculation, previous studies did not see a clear negative correlation between the long-range interaction and the accuracy of prediction (Fiser et al. 1997; Pan et al. 1999). Here we have introduced the residue contact order to evaluate the long-range interactions for each residue and have examined its effect on the accuracy of secondary structure prediction, using a large data set of nonhomologous protein sequences. Indeed, we did find that the accuracy of secondary structure prediction decays as the RCO of residues grows (Fig. 5A). The dependency of the prediction accuracy on the RCO is evident and statistically significant for residues in α-helices and β-strands, which shows a clear difference from residues in coils (Fig. 5B). The prediction accuracy does not depend on the length (Fig. 1), the contact order, and relative contact order of proteins (Fig. 2). The prediction accuracy also does not show a similar negative correlation with the number of residue contacts, which indicates the packing density of residues (Fig. 4). Actually, the effect of packing density seems to be rather opposite; on average, the prediction accuracy improves when a residue is more packed in the protein structure. In summary, only the RCO showed the negative correlation with the secondary structure prediction accuracy among characteristics we have examined and this observation is independent from the prediction methods and the threshold value used for defining inter-residue contacts used in the analyses (Table 2).

At this juncture, it would be appropriate to closely examine the different results between ours and Fiser's report (Fiser et al. 1997). In their study, Fiser et al. prepared two types of data sets, "highly interacting residues" and "stabilization center residues." The former residues are those that have a higher number of long-range interactions (defined as interactions between residues separated by at least 10 residues), and the latter residues are defined as pairs of distant residues in contact, with some of their flanking residues also in contact. Therefore, Fiser et al. are looking at the residues that are well packed in the core of proteins, which corresponds to our results of the correlation of the number of contacts in residues and the prediction

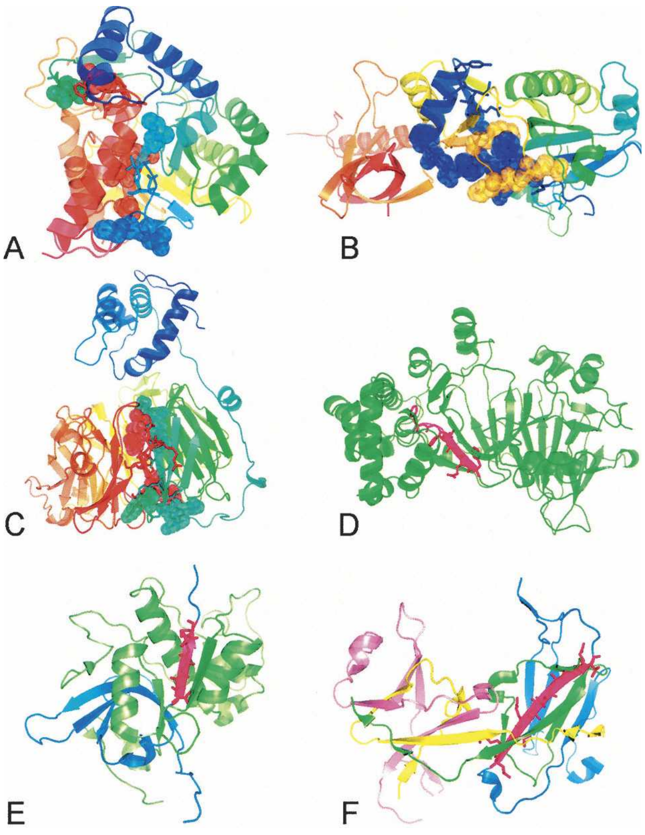

Figure 6. Examples of wrongly predicted secondary structures with a large or small RCO. In *A–C*, chains are colored in blue to red from the N to the C terminus. Residues that have an RCO in the range of 100–250 (referred to as high RCO residues) are shown with side chains, and among them, a sphere representation is used for those residues with a wrongly predicted secondary structure. (*A*) In 1d3gA (360 residues long), there are 18 high RCO residues, and eight residues (three in helices, four in strands, one in coil) among them (44.4%) have wrong secondary structure predictions. (*B*) In the case of 1a0i (332 residues long), among 18 high RCO residues, secondary structures of 12 residues (66.7%, nine of them are in strand conformation) are wrongly predicted. (*C*) 1aofA (532 residues long) has 29 high RCO residues including 14 residues (48.3%) that have wrong predictions (10 in helices, three in strands, one in coil). In *D* and *E*, wrongly predicted secondary structure segments are colored in pink with side chain atoms. (*D*) A seven-residue-long strand (residues 146–152, average RCO: 95.5) in 1pmi (440 residues long). (*E*) The C-terminal strand (residues 431–435, average RCO: 1.4) of 1k3bC (69 residues long, the chain in blue) forms a β-sheet with a strand in a different chain, 1k3bB (the green chain). (*F*) The C-terminal 10-residue-long strand (average RCO: 4.9) in 1fi8C (70 residues long, the chain in blue). 1fi8C forms a complex with other chains, 1fi8D (green), 1fi8E (pale pink), and 1fi8F (yellow). The strand has hydrogen bonds with a strand from 1fi8D. All figures are made by PyMOL (Delano 2002).

accuracy (Fig. 4). We would also point out that their data set size at that time was much smaller (80 proteins) compared with the current study.

In contrast, the RCO that we have introduced differentiates the sequential distance of contacts, i.e., unlike the case in Fiser's report, a residue contact between residues 10 and 25 is categorized differently from that of a contact between residues 10 and 200. In addition, our results show that residues of a very high RCO value, say >150, tend to have a low prediction accuracy on average, which are residues that interact with other residues in a different domain. So in this sense, we could say that our results do not contradict Fiser's report; rather, we extended the analyses in a different way.

All the secondary prediction methods are parameterized (trained) with known examples of sequences for each secondary structure (this is especially true for neural network-based approaches). What our analyses imply is that the secondary structure in the core of proteins is already well learned by the methods due to the abundance of examples, but some secondary structures in domain interfaces do not share sequence patterns with those in cores, which is the reason that prediction fails. The formation of these secondary structures with a high RCO is assisted by long-range interactions, which override the local propensity to a different secondary structure. It is also interesting to see that some of the isolated β-strands that form a β-sheet with another β-strand in a different chain (and thus have a very low RCO when the single chain is considered) are also difficult to predict. In the same way, the formation of these β-strands with a low RCO is also assisted and stabilized by β-strands from another chain.

Having observed that there are secondary structures influenced by long-range interactions, how can we then overcome this limitation of the accuracy of secondary structure prediction? Probably one of the most possible and practical ways is to still use machine learning techniques, such as neural network or support vector machine, but to separately train a neural network with those sequences of secondary structures with a high RCO (and β-strands with a low RCO) that were thus previously mispredicted. Another possible approach is to consider tertiary structures of proteins in secondary structure prediction. Actually there are some attempts along this line that show some improvements in the accuracy (Ito et al. 1997; Meiler and Baker 2003). Since secondary structure prediction is a crucial step for threading (Skolnick and Kihara 2001; McGuffin and Jones 2003; Skolnick et al. 2004) and ab initio-type protein tertiary structure prediction methods (Kihara et al. 2001), it will be worthwhile to revisit secondary structure prediction methods, taking advantage of the recent exponential increase of both sequence and structure data of proteins.

Interestingly, since secondary structure prediction methods can capture the intrinsic secondary structure propensity in a sequence, they can sometimes predict a more dynamic aspect of protein structures. For example, regions of proteins that undergo conformational switches

(Young et al. 1999) and secondary structures in folding intermediate conformations of a protein (Shiraki et al. 1995) can be predicted by secondary structure prediction methods. Along this line, we would also like to mention recent results that show that nonnative secondary structures in folding intermediates are captured by tertiary structure prediction methods: Formation of nonnative secondary structure is suggested for some proteins in protein folding simulation using a coarse-grained protein model by Liwo et al. (2005; Skolnick 2005). Another example is a recent folding simulation of the SH3 domain using a fragment assembly-based protein structure prediction method (Chikenji et al. 2004), which showed formation of nonnative α-helices consistent with a folding experiment at subzero temperatures (H. Kihara, pers. comm.). Of course molecular dynamics (MD) would be a natural tool to observe conformational transition. An example is provided by Ikeda and Higo (2003), who employed a simulation of the "chameleon" sequence in the MATα2/MCM1/DNA complex that showed two local minima in its free energy landscape, which correspond to α-helical and β-strand conformations. Thus, not only rigorous MD simulations, but also other computational methods, are becoming capable of investigating folding intermediates and conformational changes of proteins. To conclude, we would like to emphasize that this will also be applied in bioinformatics-type approaches, which can take advantage of the growing number of available sequence or structure data.

Materials and methods

Data set

A total of 2777 nonredundant protein sequences were selected from the PDB database (Berman et al. 2000) with a sequence identity threshold value of 30%. Sequences were taken from the ATOM field of the PDB files. Sequences < 50 residues long and those that had gaps were discarded. We used the secondary structure definition of the DSSP program (Kabsch and Sander 1983): Residues in H (α-helix), G (3/10 helix), and I (pi helix) are considered to be in helical conformation, and those in E (extended strand) and B (β-bridge) are in β-strand conformation.

Secondary structure prediction methods

Three secondary structure prediction methods, PSIPRED (Jones 1999), Jnet (Cuff and Barton 2000), and PREDATOR (Frishman and Argos 1996), were used in this study. PSIPRED uses a neural network that takes a multiple sequence alignment of a query sequence generated by PSI-BLAST as input (Altschul et al. 1997). It assigns a secondary structure (α-helix, β-strand, or coil) to a residue at the center of a size 15 sliding window. Jnet is another neural network-based approach, which uses a size 17 sliding window. It uses three different forms of multiple sequence alignments generated with homologous sequences retrieved by PSI-BLAST. PREDATOR uses FASTA (Pearson and Lipman 1988) for collecting sequences. An interesting feature of PREDATOR is that it tries to recognize potentially hydrogen-bonded residues in amino acid sequences using database-derived statistics on residue-type occurrences in different classes of β-bridges to delineate interacting β-strands. For all three methods, homologous sequences for multiple sequence alignments are collected from a sequence database that consists of SWISS-PROT, trEMBL (Boeckmann et al. 2003), and KEGG genes database (Kanehisa et al. 2004).

Residue contact order

The contact order (CO) is the average sequence separation between contacting residues in the native state of a protein, which is defined by Equation 1, below. This simple index for describing the complexity of protein topology is often used in the context of the protein folding rate (Plaxco et al. 1998; Zhou and Zhou 2002; Ivankov et al. 2003):

$$CO = \frac{1}{N}\sum_{i=1}^{L-1}\sum_{j=2}^{L}|i-j|\delta_{ij} \qquad (1)$$

Here $\delta_{ij} = 1$ when residue i and j are in contact, and 0 otherwise. Two residues are considered to be in contact if any pair of heavy atoms from each residue locate closer than a threshold value. We used 4.5 and 6 Å as the threshold values. L is the length of the protein; N is the total number of contacts in the protein.

The relative contact order is a normalized CO by the length (L) of the protein:

$$CO_{rel} = \frac{1}{N \cdot L}\sum_{i=1}^{L-1}\sum_{j=2}^{L}|i-j|\delta_{ij} \qquad (2)$$

Similarly, we define the residue contact order for residue i as the average contact order for the residue:

$$RCO_i = \frac{1}{n}\sum_{j,i\neq j}^{L}|i-j|\delta_{ij} \qquad (3)$$

Here n is the number of contacts between the ith residue and the others. Recently a similar value named "residue-wise" contact order (RWCO) was introduced (Kinjo and Nishikawa 2005), which is just the sum of the sequence separation of contacting residues (i.e., $RWCO_i = n \times RCO_i$). Hence, RWCO has a higher correlation to the number of contacts than RCO does (Fig. 3B). After RCO for individual residues are calculated, the average RCO values in a smoothing window is assigned to the center residue of the window.

Acknowledgments

The author thanks Troy Hawkins for proofreading this manuscript. The discussion with Hiroshi Kihara (Kansai Medical University, Japan) is also much appreciated.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31:** 365–370.

Chikenji, G., Fujitsuka, W., and Takada, S. 2004. Protein folding mechanisms and energy landscape of src SH3 domain studied by a structure prediction toolbox. *Chem. Phys.* **307:** 157–162.

Chou, P.Y. and Fasman, G.D. 1978a. Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47:** 251–276.

———. 1978b. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47:** 45–148.

Crooks, G.E. and Brenner, S.E. 2004. Protein secondary structure: Entropy, correlations and prediction. *Bioinformatics* **20:** 1603–1611.

Cuff, J.A. and Barton, G.J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40:** 502–511.

Delano, W.L. 2002. The PyMOL Molecular Graphics System. http://www.pymol.org.

Fiser, A., Dosztanyi, Z., and Simon, I. 1997. The role of long-range interactions in defining the secondary structure of proteins is overestimated. *Comput. Appl. Biosci.* **13:** 297–301.

Frishman, D. and Argos, P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23:** 566–579.

———. 1996. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9:** 133–142.

Garnier, J., Osguthorpe, D.J., and Robson, B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120:** 97–120.

Guo, J., Chen, H., Sun, Z., and Lin, Y. 2004. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins* **54:** 738–743.

Hua, S. and Sun, Z. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J. Mol. Biol.* **308:** 397–407.

Ikeda, K. and Higo, J. 2003. Free-energy landscape of a chameleon sequence in explicit water and its inherent α/β bifacial property. *Protein Sci.* **12:** 2542–2548.

Ito, M., Matsuo, Y., and Nishikawa, K. 1997. Prediction of protein secondary structure using the 3D-1D compatibility algorithm. *Comput. Appl. Biosci.* **13:** 415–424.

Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., and Finkelstein, A.V. 2003. Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* **12:** 2057–2062.

Jacoboni, I., Martelli, P.L., Fariselli, P., Compiani, M., and Casadio, R. 2000. Predictions of protein segments with the same amino acid sequence and different secondary structure: A benchmark for predictive methods. *Proteins* **41:** 535–544.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292:** 195–202.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32 Database issue:** D277–D280.

Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K. 2003. Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry. *Proteins* **51:** 504–514.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14:** 846–856.

Kihara, D., Lu, H., Kolinski, A., and Skolnick, J. 2001. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci.* **98:** 10125–10130.

Kinjo, A.R. and Nishikawa, K. 2005. Recoverable one-dimensional encoding of three-dimensional protein structures. *Bioinformatics* **21:** 2167–2170.

Kuang, R., Leslie, C.S., and Yang, A.S. 2004. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* **20:** 1612–1621.

Levin, J.M. 1997. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.* **10:** 771–776.

Lim, V.I. 1974. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88:** 857–872.

Lin, K., Simossis, V.A., Taylor, W.R., and Heringa, J. 2005. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **21:** 152–159.

Liwo, A., Khalili, M., and Scheraga, H.A. 2005. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci.* **102:** 2362–2367.

Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405:** 442–451.

McGuffin, L.J. and Jones, D.T. 2003. Benchmarking secondary structure prediction for fold recognition. *Proteins* **52:** 166–175.

Meiler, J. and Baker, D. 2003. Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci.* **100:** 12105–12110.

Minor Jr., D.L. and Kim, P.S. 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380:** 730–734.

Munoz, V., Cronet, P., Lopez-Hernandez, E., and Serrano, L. 1996. Analysis of the effect of local interactions on protein stability. *Fold. Des.* **1:** 167–178.

Nishikawa, K. and Ooi, T. 1980. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Pept. Protein Res.* **16:** 19–32.

Pan, X.M., Niu, W.D., and Wang, Z.X. 1999. What is the minimum number of residues to determine the secondary structural state? *J. Protein Chem.* **18:** 579–584.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P., and Lund, O. 2000. Prediction of protein secondary structure at 80% accuracy. *Proteins* **41:** 17–20.

Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277:** 985–994.

Przybylski, D. and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins* **46:** 197–205.

Rost, B. 2001. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134:** 204–218.

Rost, B. and Eyrich, V.A. 2001. EVA: Large-scale analysis of secondary structure prediction. *Proteins* **(Suppl.) 5:** 192–199.

Rost, B. and Sander, C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:** 55–72.

———. 2000. Third generation prediction of secondary structure. In *Protein structure prediction* (ed. B. Webster), pp 71–95. Humana Press, Clifton, NJ.

Schulz, G.E. and Shirmer, R.H. 1979. *Principles of protein structure*. Springer-Verlag, New York.

Shiraki, K., Nishikawa, K., and Goto, Y. 1995. Trifluoroethanol-induced stabilization of the α-helical structure of β-lactoglobulin: Implication for non-hierarchical protein folding. *J. Mol. Biol.* **245:** 180–194.

Skolnick, J. 2005. Putting the pathway back into protein folding. *Proc. Natl. Acad. Sci.* **102:** 2265–2266.

Skolnick, J. and Kihara, D. 2001. Defrosting the frozen approximation: PROSPECTOR—A new approach to threading. *Proteins* **42:** 319–331.

Skolnick, J., Kihara, D., and Zhang, Y. 2004. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* **56:** 502–518.

Ward, J.J., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2003. Secondary structure prediction with support vector machines. *Bioinformatics* **19:** 1650–1655.

Young, M., Kirshenbaum, K., Dill, K.A., and Highsmith, S. 1999. Predicting conformational switches in proteins. *Protein Sci.* **8:** 1752–1764.

Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. 1999. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* **34:** 220–223.

Zhou, H. and Zhou, Y. 2002. Folding rate prediction using total contact distance. *Biophys. J.* **82:** 458–463.

Zhou, X., Alber, F., Folkers, G., Gonnet, G.H., and Chelvanayagam, G. 2000. An analysis of the helix-to-strand transition between peptides with identical sequence. *Proteins* **41:** 248–256.