ARTICLE

Correlations Between Amino Acids at Different Sites in Local Sequences of Protein Fragments with Given Structural Patterns

Wen Lu, Hai-yan Liu*

Hefei National Laboratory for Physical Sciences at Microscale, Key Laboratory of Structural Biology, School of Life Sciences, University of Science and Technology of China, Hefei 230026, China

(Dated: Received on April 30, 2006; Accepted on May 18, 2006)

Ample evidence suggests that the local structures of peptide fragments in native proteins are to some extent encoded by their local sequences. Detecting such local correlations is important but it is still an open question what would be the most appropriate method. This is partly because conventional sequence analyses treat amino acid preferences at each site of a protein sequence independently, while it is often the inter-site interactions that bring about local sequence-structure correlations. Here a new scheme is introduced to capture the correlation between amino acid preferences at different sites for different local structure types. A library of nine-residue fragments is constructed, and the fragments are divided into clusters based on their local structures. For each local structure cluster or type, chi-square tests are used to identify correlated preferences of amino acid combinations at pairs of sites. A score function is constructed including both the single site amino acid preferences and the dual-site amino acid combination preferences, which can be used to identify whether a sequence fragment would have a strong tendency to form a particular local structure in native proteins. The results show that, given a local structure pattern, dual-site amino acid combinations contain different information from single site amino acid preferences. Representative examples show that many of the statistically identified correlations agree with previously-proposed heuristic rules about local sequence-structure correlations, or are consistent with physical-chemical interactions required to stabilize particular local structures. Results also show that such dual-site correlations in the score function significantly improves the Z-score matching a sequence fragment to its native local structure relative to non-native local structures, and certain local structure types are highly predictable from the local sequence alone if inter-site correlations are considered.

Key words: Protein, Local structure pattern, Sequence-structure correlation, Hydrogen bonding

I. INTRODUCTION

In recent years much progress has been made in protein structure predictions by incorporating information about local structure patterns [1,2]. It has been shown that besides *ab initio* predictions, fold recognitions and remote homology detections can be improved by the integration of local sequence-structure correlations with global evaluations of sequence-structure compatibility [3]. These achievements indicated that further studies on local sequence-structure relationships in peptide fragments of native proteins could be very valuable. One important question is, given the local sequence of a protein fragment, how much can we tell about its likely local structure in a native fold? It is expected that in many cases knowledge of the local sequence alone is insufficient to specify structure. Such peptide fragments can be called chameleons because their structures are determined by long-range interactions [4]. However, there are also other cases in which local sequences allow for a unique or just a few types of local structures. The identification of such conserved cases from the large body of protein structure and sequence data can enable the prediction of local structures in respective regions of the protein [5]. Various secondary structure prediction approaches are the most widely known examples [6-9].

Theoretically, the local structure should result from the subtle balance between local and long-range interactions [10,11]. Because local structure patterns can transcend protein family boundaries, it is challenging to exploit the structural information contained in local sequences thoroughly [12-15]. In this work, we use cluster analysis to identify recurring structural patterns of fragments from different protein families. Then sequences of fragments which share the same structure pattern are analyzed. Important features of this work are that we focus mainly on boundaries of secondary structures, and for each structure pattern we extract not only the information of amino acid preferences at individual sites, but also correlation between amino acid preferences at different sites. The existence of such correlations is unambiguous from our results, and most of the time it can be explained by physical-chemical interactions specifically to stabilize the respective local structure pattern.

*Author to whom correspondence should be addressed. E-mail: hyliu@ustc.edu.cn, Tel.: +86-551-3607451, Fax: +86-551-3607451

II. DATA AND METHODS

A. Data

A total of 2447 protein structures were chosen from PDBSELECT(2004) [16], all of them with resolutions higher than 0.2 nm and *R*-factors<0.2. The sequence identity between any pair of selected proteins was below 25%. These proteins were processed, resulting in a total of 350768 peptide segments of nine residues, for which their sequences and structures in native proteins are known. It is known that sequences of longer fragments may provide more information about local structure, but a library of longer fragments would be much larger. For example, a seven-residue library is 100 times larger than a five-residue one [17]. Reference [18] illustrates that the nine-residues sequence length is a good choice.

B. Methods

1. Cluster peptide fragments based on their structures

Secondary structure types and Ca atom positions were used as local structural features. The secondary structure type for each residue was calculated by STRIDE [19]. Strand, a-helix, pi-helix, turn, and coil were selected as secondary structure types, and the fragments were sorted so that all fragments with the same composition and order of secondary structure types belong to one class. Each class was further divided into different clusters using ISODATA [20], requiring that each cluster should contain more than 40 fragments, and the RMS Ca atom position deviations between each member of a cluster and the representative fragment of the same cluster should be less than 0.15 nm. This step assigned 198189 fragments into different structure clusters. The remaining fragments were clustered in the next step, in which only the Ca atom positional RMSDs were employed as distance measures. As described above, it was required that the average RMSD from cluster members to the cluster center (represented by one member) should be less than 0.15 nm. This resulted in 2126 clusters in total, and each represents a local structure pattern observed in native protein structures.

2. Amino acid preferences at different sites of clustered fragments

Each structure pattern obtained above contains nine sites. At some sites, a certain type or types of amino acids may be preferred, and such preferences provide a basis for analyzing sequence-structure correlations.

First, a conventional method was used to analyze amino acid preferences at each site of each structure pattern. The amino acid preference of one site was considered as independent from the other sites. This preference was quantified by a simple evaluation of the frequency of occurrence of each amino acid at each site by using a Bayesian probability, from which a sequence profile for each structure pattern can be constructed. The Kullback-Leibler divergence measure was used to judge the amount of information provided by each type of amino acid at each site. In the formula below, p stands for the frequency of certain type of residue at a given site, q stands for the frequency of the residue type in all sites.

$$K(p,q) = \sum_i p_i \ln\left(\frac{p_i}{q_i}\right) \tag{1}$$

The assumption that amino-acid preferences at different sites are independent from each other is very crude. Obviously, the amount of data in this study does not allow for a full correlation analysis involving all 9 sites in a structure pattern simultaneously, so it considers only correlations between two sites.

The chi-square test is used to determine whether amino-acid preferences at two sites of a given structure pattern are correlated [21]. Given the residue type at one site, the null hypothesis is that this would not influence the amino-acid preferences of the other site. When the chi-square value is so large that there is less than a 0.05 chance for the null hypothesis to be true, the null hypothesis is rejected and the amino-acid preferences at the two sites are considered to be correlated. Each of these correlated pairs defines a sequence pattern for the structure pattern, noted as (a_{i1}, a_{i2}), where $i1$, $i2$ are indices of the correlated sites and a_{i1}, a_{i2} refers to the preferred amino acid combinations at these two sites .

Based on the above analyses, a score function was constructed for each structure pattern, which can be applied to score the likelihood that a fragment of given sequence will have the structure pattern in a native protein,

$$P = \prod_{i=1}^{9} P_i(a_1, a_2, ..., a_9) \tag{2}$$

if the template sequence pattern contains the correlated pattern (a_{i1}, a_{i2}), then

$$p_{i1}(a_1, a_2, ..., a_9) = p_{i2}(a_1, a_2, ..., a_9) = 1 \tag{3}$$

for other sites,

$$P_i(a_1, a_2, ..., a_9) = P_{\text{Bayes}}(a_i) \tag{4}$$

where $P_{\text{Bayes}}(a_i)$ is the Bayesian probability of observing amino acid type a_i at site i computed using the independent site assumption.

For each structure pattern, the amino acid preferences given by the Bayesian probability for each site and the preferred amino acid combinations for correlated sites can be predetermined using the cluster library of fragments. The function defined by Eq.(2) can be applied to score any 9-residue peptide fragment

ISSN 1003-7713/DOI:10.1360/cjcp2007.20(1).71.7

(a query) with sequence $(a_1, a_2, ..., a_9)$ against each of the structure patterns (templates). The highest scoring template can be employed as a prediction of the structure of the query peptide. In the mean while, a Z-score of the prediction can be obtained, which can be used as an indicator or threshold for reliable predictions.

III. RESULTS

A. Sequence variances and distributions of inter-site correlations at secondary structure boundaries

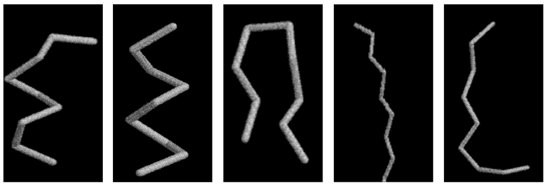
We first consider the sequence variances at different sites of given local structure classes. These variances quantify how much structural information is contained in the amino acid type at a given site. We use two quantities to indicate this variance, one is the Kullback-Leibler divergence (defined by Eq.(1)) at the site, and the other is the number of preferred amino acid combinations involving the site. The first quantity reflects how much the amino acid preferences at a site differ from background preferences (zero corresponds to no difference). The second reflects how much correlation exists between the amino acid preferences at the given site and other sites. (Note: the numbers are not normalized by sizes of clusters.)

As examples, three common secondary structure boundaries are considered, i.e., the N-terminal cap of helices, the C terminus of a strand connected to a turn, and the C terminus of a strand connected to a coil. These are chosen because each of the clusters contains a large number of members, and they represent secondary structure boundaries important for structure prediction. Figure 1 shows the sequence variances at each site of each of the three such structure clusters measured by the Kullback-Leibler divergences (Fig.1(1a)-(3a)) and by the number of preferred amino acid combinations involving the site (Fig.1(1b)-(3b)). Figure 1(1a) shows that for the N-terminal cap cluster, the third site (the first residue in the helix) shows the largest deviation from the background amino-acid preferences, while the first residue is most variable. However, if we consider correlations between sites, Figure 1(1b) shows that each site is represented by similar number of members showing inter-site correlations. More interestingly, the first and second sites, which show less amino-acid preferences when considered as independent sites, show more correlation. Figure 1(2a) shows that the 7th site of the strand-turn cluster, or the last residue of the strand, has the strongest amino acid preference, and the 5th site shows little amino acid preference. However, when inter-site correlations are considered, the 5th site shows the strongest inter-site correlations. In general, Figure 1 (3a) and (3b) indicate that when amino-acid preferences are considered for each site independently, different sites show significantly different amino-acid variability. When inter-site correlations are considered, different sites show generally similar degrees of correlations with other sites. Also, the strengths of single site amino-acid preferences are not correlated with probabilities of observing inter-site correlations.

B. Predicting structure clusters by the single-site profile scores and combined single-site and inter-site correlation scores

For the five clusters shown in Table I, each fragment from each cluster was taken as a query sequence. For each query sequence, the query sequence from its cluster was excluded, the sequence profile of the cluster was recalculated, the chi-square-based correlation analysis for this cluster was carried out again and the score functions for this cluster were reconstructed. Then the query sequence was matched to the independent-site profiles and the combined score functions for different clusters. From the resulting distribution of scores, the Z-score matching the query sequence to its native structure pattern was computed. Table I shows that including inter-site correlation in the score function significantly improved the Z-scores.

TABLE I Structure pattern prediction

S [a]					
Z1 [b]	2.91	2.10	3.18	1.77	2.08
Z2 [c]	7.13	9.55	5.80	6.73	4.04
R [d]	0.57	0.90	0.29	0.53	0.17

[a] The respective structures of selected clusters.
[b] The averaged Z-scores for scoring each member sequence against the sequence profiles of the native cluster and all clusters.
[c] The Z-scores with the inter-site correlations included in the score function.
[d] The success rates for structure pattern prediction (see text for definition).

If we use the criterion that the score for the native cluster is among the top five scoring clusters for a successful prediction, we observe that predictions on different local structures have different successful. This rate reflects the predictability of the structure pattern based on local sequences. Certain local structure patterns are highly predictable from local sequences (after considering inter-site correlations), including all the helical fragments, the N-terminal cap of helices, and the extended strand. Interestingly, in Table II, we see that the largest improvement after including inter-site correlations in the score function is for the extended strand.

ISSN 1003-7713/DOI:10.1360/cjcp2007.20(1).71.7

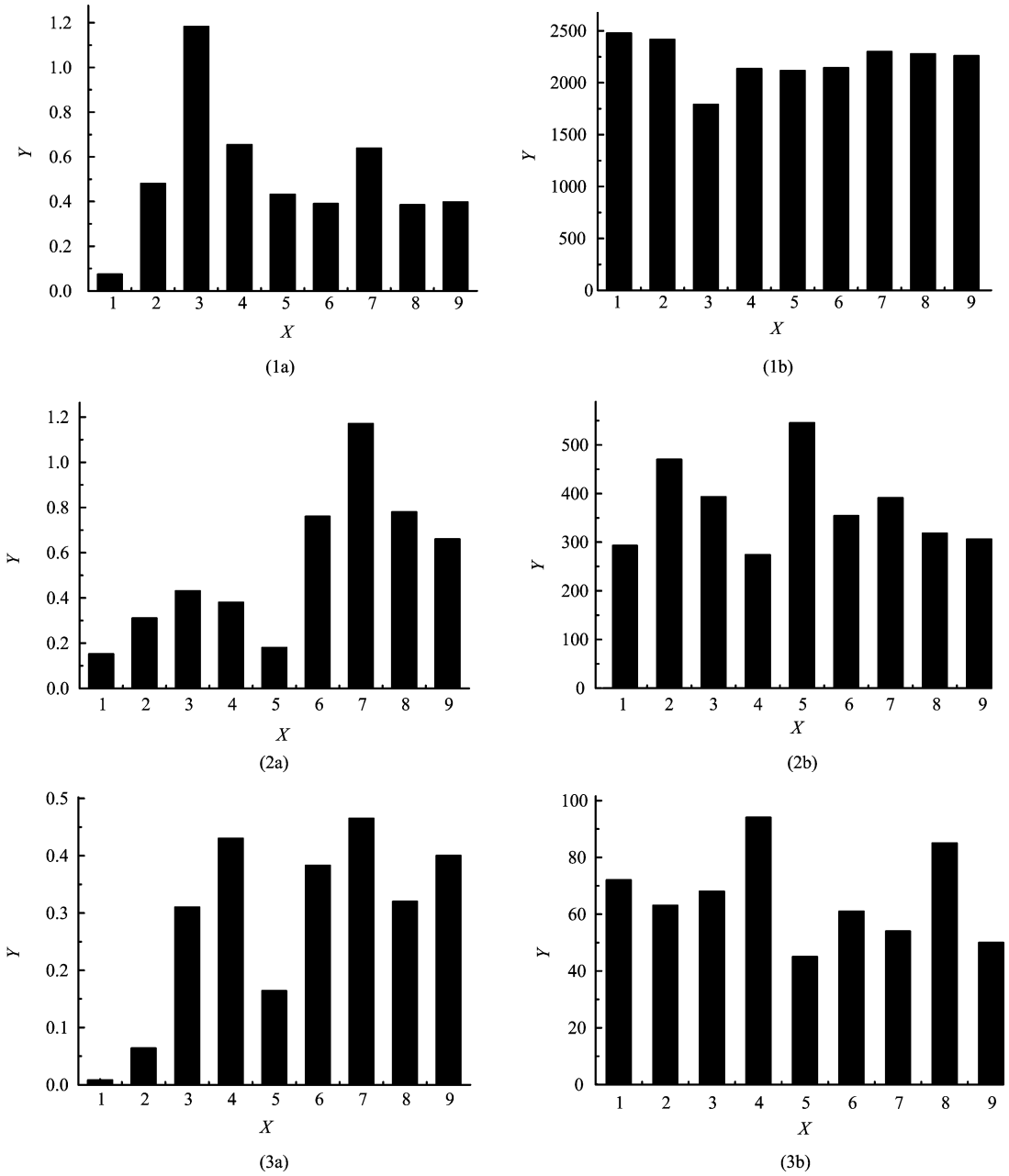


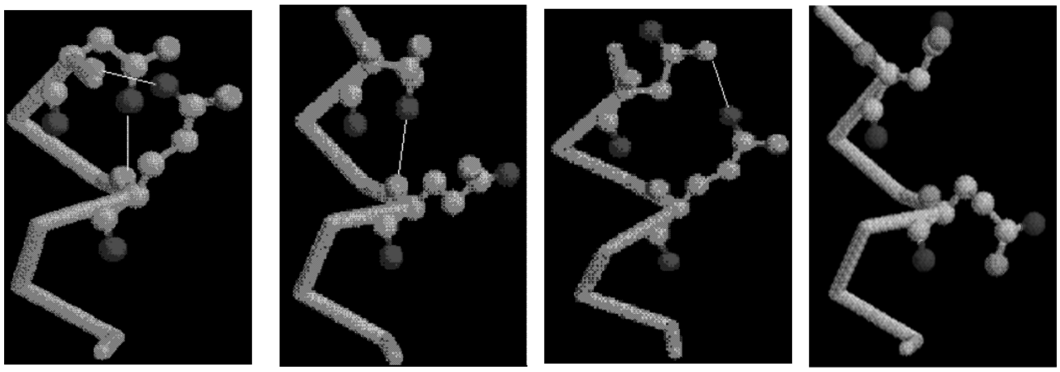
FIG. 1 (a) The Kullback-Leibler divergence measure of amino acid preferences at individual sites; (b) For each site, the number of cluster members satisfying the inter-site correlation rule involving the site is plotted. The clusters shown are a cluster with the secondary structure composition CCAAAAAAA, SSSSSSSTT and SSSSSSCCC for (1a)/(1b), (2a)/(2b) and (3a)/(3b), respectively. A stands for a-helix, C for coil, S for strand, and T for turn.

C. Inter-site correlations and physical chemical interactions stabilizing local structures: comparing with N-terminal capping boxes of helices

The physical-chemical nature of molecular interactions stabilizing protein 3D structures has been well understood. Such interactions include hydrogen-bonding, hydrophobic packing, electrostatic interactions (e.g., salt bridges) and van der Waals packing. Many of the above statistically identified preferred amino-acid combinations at different sites of a given structure pattern can be attributed to specific physical-chemical interactions stabilizing the corresponding local structures.

Some such interactions have been identified based on heuristic observations. For example, reference [22] describes three patterns of capping at helix N-termini. Each pattern has its own sequence characteristics, involving interactions between amino-acid residues at different sites. The proposed nomenclature uses the following definition-N‴-N″-N′-Ncap-N1-N2-N3-N4-to re-

ISSN 1003-7713/DOI:10.1360/cjcp2007.20(1).71.7



(a) (b) (c) (d)

FIG. 2 Different specific physical chemical interactions between ASN at Ncap site and GLN at N3 site for members in a cluster of the secondary structure composition CCAAAAAAA.

TABLE II Comparison of amino-acid combination preferences in a CCAAAAAAA cluster with the normal N-terminal capping box type

Site	Ncap–N3	N′–N3	N′–N4
A	218	333	244
B	47	299	244

A: The number of observed members satisfying the site-correlation criteria for the site combination.
B: The number of observed members satisfying the respective rules defining the box type.

fer to different sites. Three types of "capping boxes" have been identified: normal boxes containing a hydrogen bond between the side chain of Ncap and the backbone of N3, another hydrogen bond between the side chain of N3 and the backbone of Ncap, and hydrophobic interactions between residues N′ and N4; big boxes containing a hydrogen bond between the side chain of Ncap and the backbone of N3, and hydrophobic interactions between the residues N″ and N3 and also between N″ and N4; and β-boxes containing a hydrogen bond between the backbone of Ncap and the backbone of N‴, and hydrophobic interaction between residues N‴ and N3 and also between N‴ and N4.

From Tables II-IV we can see for each box type a corresponding structure cluster. These list the number of members demonstrating amino-acid combination preferences at each pair of sites involved in the box type definition, and the number of members demonstrating amino-acid combination preferences satisfying the box type definition. For example, in Table II, a cluster with the secondary structure composition CCAAAAAAA contains 218 members satisfying the amino-acid combination preference rules obtained for site pair Ncap-N3 from the statistical analyses here. Among them, 47 members satisfy the rule defining a normal box type, i.e. reciprocal hydrogen bonds between Ncap and N3.

From Tables II-V, we observe that while many of the observed amino-acid combination preferences have been identified before, a large number of them have not been covered by heuristic analysis. That is, the inter-site correlations obtained from statistical analyses have not been included in the heuristic rules defining the structure patterns.

D. Preferred amino-acid combinations and physical-chemical interactions stabilizing local structures

As examples, we look at some of the specific interactions present in one of the preferred Ncap-N3 combinations. The 218 members showing preferred amino-acid combinations at the Ncap and N3 sites of the cluster considered in Table II-IV have ASN at the Ncap site and GLN at the N3 site. Among the 34 members, 7 members contain the reciprocal hydrogen bonds between Ncap and N3 (Fig.2(a)), 16 members contain the hydrogen bond between the side chain of Ncap and the backbone of N3 (Fig.2(b)); 4 members contain the hydrogen bond between the side chain of Ncap and the side chain of N3 (Fig.2(c)); 7 members contain no hydrogen bond between the residue pair (Fig.2(d)). Thus the observed preferences can be traced to specific physical chemical interactions which preferentially stabilize

TABLE III Comparison of amino-acid combination preferences in a CCCAAAAAAA cluster with the big N-terminal capping box type

Site	Ncap–N3	N3-N′	N″–N3	N″–N4
A	110	123	167	102
B	50	33	116	91

TABLE IV Comparing amino-acid combination preferences in a CCCCAAAAA cluster with the N-terminal capping β-box type

Site	Ncap–N‴	N‴–N3	N‴–N4
A	25	14	24
B	8	3	18

ISSN 1003-7713/DOI:10.1360/cjcp2007.20(1).71.7

TABLE V The Nearest neighbor preferences for a helix N-terminal cap cluster.

Sites[a]	Preferred residue combinations[b]
N′–Ncap	ARG-ARG, ASP-GLY, GLY-GLN, GLU-ASP, ASN-GLU, GLN-PRO, GLY-GLY, GLY-PHE, GLY-VAL, ARG-GLU, ILE-ARG, LEU-SER, LEU-THR, LYS-GLU, MET-GLU, PHE-GLN, PRO-ASP, PRO-CYS, THR-ASN, THR-HIS, THR-LYS, TRP-ASP, VAL-ASN, VAL-GLY, VAL-THR
Ncap-N1	ALA-LEU, ASN-ARG, ASN-PRO, ASN-THR, ASP-GLY, GLN-LEU, GLU-ILE, GLU-SER, GLU-THR, HIS-GLU, ILE-PRO, LEU-ARG, LEU-LYS, LYS-HIS, LYS-LYS, LYS-THR, MET-PRO, SER-ARG, THR-ARG, THR-ASN, THR-GLU, TYR-PRO
N1–N2	HIS-LYS, ILE-HIS, ILE-ILE, LYS-ARG, MET-GLN, PHE-ASN, PHE-GLN, PHE-GLY, PRO-ILE, SER-ALA, SER-GLY, SER-ILE, SER-TYR, THR-HIS, TYR-ASN, TYR-LYS, TYR-PRO
N2–N3	ALA-GLY, ALA-PRO, ARG-GLY, ARG-TRP, ASN-ASN, ASN-ILE, ASN-SER, GLU-ILE, GLU-TYR, GLY-HIS, GLY-THR, ILE-ALA, ILE-GLY, LEU-ARG, LEU-HIS, LYS-GLU, LYS-MET, MET-ASP, MET-VAL, PRO-ARG, PRO-LEU, PRO-PHE, PRO-VAL, SER-ASN, SER-SER, THR-MET, THR-THR, TRP-ASP, TYR-GLU, TYR-LEU, VAL-GLY
N3–N4	ALA-ASP, ALA-LYS, ALA-THR, ASN-MET, CYS-GLN, CYS-ILE, GLN-ASN, GLU-LEU, GLY-VAL, HIS-ILE, ILE-ASP, ILE-LEU, ILE-LYS, ILE-MET, LEU-ASP, LEU-LYS, LEU-SER, LEU-VAL, LYS-VAL, PHE-GLU, PRO-LEU, SER-GLY, SER-LYS, THR-ASN, THR-GLY, VAL-GLU, VAL-SER

[a] The stes are mutually the nearest-neighbors.
[b] Identified by the chi-square test.

the local structure pattern. As this is only a change in preference, it does not imply that the specific physical chemical interactions exist in all members having the preferred residue combination. There are still some protein fragments which, do not rely on specific interactions involving the respective residues to stabilize the native structure while having preferred amino-acid combinations at corresponding sites.

E. Preferences for the nearest-neighbor amino acid combinations

This has been specifically discussed in the literatures [23,24]. However, few such nearest-neighbor combination preferences have been identified in a context of given local structure patterns because of method and data-set limitations. Table V lists preferred nearest neighbor combinations for the N-terminal capping sites of helices obtained from local structure pattern.

F. Local structure patterns within globular proteins

Since the correctness of local structure patterns identified by the score function correlates with Z-scores, we hypothesize that the Z-score of the native local structure could be used as a measure on the native-structure preferences of local sequence segments. Such preferences can be used as important input for further understanding of roles of local interactions in particular protein folds and in folding mechanisms, and to guide protein engineering experiments. Figure 3 shows an example of a "local native structure preference profile" for a protein. Here we observe that segments whose native local structures are outstandingly preferred by local sequences as indicated by the Z-score are well distributed in the global sequence. Some of them correspond to regular secondary structures, which might be well predicted by secondary structure prediction algorithms. However, a number of fragments cover secondary structure boundaries, suggesting that at these places, local sequence signals may play important roles in shaping the globular structures.

CC***CSSSSSSTT***TT***AAA***AAAAAAAAAAAC***CSSSSSSTT***333CCCCCCCCTT

SSSSCTT**TTAAAAAAAA**TTTCSSSSCCCTTTTCCCC***AAAAAA*AAAA**

AAAAAAACCC*SSSSSCC333*CCTTTTTC333A***AAAAAA*AAAAAAAAAA**

CCSSSSSCCCSSSCCCCCCCCSSSTTTCCCCCSSS***AAAAA*AAAAA**

ATTTTTTTTTSSSSSCCC

FIG. 3 The secondary structure composition of the lhdo protein (A stands for a-helix, S stands for strand, C stands for coil, T stands for turn, the number 3 stands for 310 helix.). Segments with local sequence-local structure matching Z-scores greater than 10 are indicated by bold letters; those with Z-score between 5 and 10 are indicated by bold italic letters.

IV. CONCLUSION

Local structure-sequence correlations in native proteins were investigated through statistical analyses. Af-

ISSN 1003-7713/DOI:10.1360/cjcp2007.20(1).71.7

ter collecting nine-residue peptide fragments from a culled data set of native protein structures and clustering them into structure patterns, we analyzed sequence preferences at individual sites and combinations of sites on the structure patterns. One novelty of this analysis is that in addition to single-site preferences normally used in conventional sequence profile analyses [25-30], we used chi-square testing to analyze inter-site correlations in amino-acid combination preferences. The statistically significant combinations are identified. For some well-studied structure patterns, such combinations cover previously summarized inter-site interaction patterns. The N-terminal capping boxes for helices are discussed in detail as examples.

We demonstrated that the single-site profiles and two-site combination preferences can be combined into a single score function, which can be used as a measure of the local structure preferences of a given sequence fragments. Including two-site combinations in this score greatly improves the significance of this measure.

How important local interactions are in determining protein structures is a fundamental issue [31,32]. By studying the chemical synthesis of polypeptide, it has been shown that the local sequence segments of about 40 amino acids lengths may be able to fully determine the structure [33]. A large number of results from secondary structure predictions indicates that good predictions can be obtained using sequence information of much shorter fragments. This paper provides a scheme to analyze local sequence-structure preferences beyond secondary structure predictions and independent-site hypotheses which limit sequence analysis. We will consider more complex combination preference and bigger database in the future. Further work along this direction can improve our understanding of how proteins fold and our ability to predict protein structures from sequences.

[1] D. Baker and A. Sali, Science. **294**, 93 (2001).
[2] K. T. Simons, R. Bonneau, I. Ruczinski and D. Baker, Proteins. Supp. **3**, 171 (1999).
[3] Y. Hou, W. Hsu, M. L. Lee and C. Bystroff, Bioinformatics **19**, 2294 (2003).
[4] I. Jacoboni, P. L. Martelli, P. Fariselli, M. Compiani and R. Casadio, Proteins. **41**, 535 (2000).
[5] E. G. Hutchinson and J. M. Thornton, Protein Sci. **3**, 2207 (1994).
[6] A. S. Yang and L. Y. Wang, Bioinformatics **19**, 1267 (2003).
[7] D. Frishman and P. Argos, Proteins **27**, 329 (1997).
[8] T. N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Bruuak, G. P. Gippert and O. Lund, Proteins **41**, 17 (2000).
[9] C. Bystroff and D. Baker, J. Mol. Biol. **281**, 565 (1998).
[10] R. L. Baldwin and G. D. Rose, Trends Biochem. Sci. **24**, 26 (1999).
[11] R. L. Baldwin and G. D. Rose, Trends Biochem. Sci. **24**, 77 (1999).
[12] K. F. Han and D. Baker, J. Mol. Biol. **251**, 176 (1995).
[13] J. Pei and N. V. Grishin, Proteins **56**, 782 (2004).
[14] O. Sander, I. Sommer and T. Lengauer, BMC Bioinformatics **7**, 14 (2006).
[15] C. Bystroff and D. Baker, J. Mol. Biol. **281**, 565 (1998).
[16] U. Hobohm, M. Scharf, R. Schneider and C. Sander, Protein Sci. **1**, 409 (1992).
[17] R. Kolodny, P. Koehl, L. Guibas and M. Levitt, J. Mol. Biol. **323**, 297 (2002).
[18] C. Bystroff, K. T. Simons, K. F. Han and D. Baker, Curr. Opin. Biotechnol. **7**, 417 (1996).
[19] M. Heinig and D. Frishman, Nucleic Acids Res. **32**, 500 (2004).
[20] N. B. Venkateswarlu and R. P. S. V. S. K. Raju, Pattern Recognition **25**, 335 (1992).
[21] N. Turner, J. Clin. Nurs. **9**, 93 (2000).
[22] R. Aurora and G. Rose, Protein Sci. **7**, 21 (1998).
[23] J. W. Wang and J. A. Feng, Protein Eng. **16**, 799 (2003).
[24] W. R. Pearson and D. J. Lipman, Proc. Natl. Acad. Sci. **85**, 2444 (1988).
[25] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lippman, Nucleic Acids Res. **25**, 3389 (1997).
[26] J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting and P. Bork, Nucleic Acids Res. **28**, 231 (2000).
[27] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. Sonnhammer, Nucleic Acids Res. **30**, 276 (2002).
[28] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann and A. Baicoch, Nucleic Acids Res. **30**, 235 (2002).
[29] J. Y. Huang and D. L. Brutlag, Nucleic Acids Res. **29**, 202 (2001).
[30] B. Goliaei and Z. Minuchehr, FEBS Lett. **537**, 121 (2003).
[31] B. Z. Lu, C. X. Wang and B. H. Wang, Chin. J. Chem. Phys. **19**, 117 (2003).
[32] W. Song, W. Z. Chen, X. T. Zhang and C. X. Wang, Chin. J. Chem. Phys. **19**, 257 (2003).
[33] D. Shortle, Protein Sci. **11**, 18 (2002).

ISSN 1003-7713/DOI:10.1360/cjcp2007.20(1).71.7