Chapter 14
Discovery of Structural Motifs Using Protein Structural Alphabets and 1D Motif-Finding Methods

Shih-Yen Ku and Yuh-Jyh Hu

Abstract Although the increasing number of available 3D proteins structures has made a wide variety of computational protein structure research possible, yet the success is still hindered by the high 3D computational complexity. Based on 3D information, several 1D protein structural alphabets have been developed, which can not only describe the global folding structure of a protein as a 1D sequence, but can also characterize local structures in proteins. Instead of applying computationally intensive 3D structure alignment tools, we introduce an approach that combines standard 1D motif detection methods with structural alphabets to discover locally conserved protein motifs. These 1D structural motifs can characterize protein groups at different levels, e.g., families, super families, and folds in SCOP, as group features.

Keywords Protein structure · Structural alphabet · Motif

14.1 Introduction

As the rapid growth of protein structural information, biologists require accurate classification to understand and rationalize the variety in proteins [1]. To ensure a more easily constructed and better comprehensible classification, it is desired that we provide only essential characteristic structural descriptions of protein functional parts. With such a classification, we can assign a novel protein to known categories, and thus predict its structures and functions. The task of extracting characteristic structural features for classification is difficult and becomes more challenging for small proteins, where the characteristic statistics are marginal owing to short protein chains, or for proteins that only share low sequence similarity.

Y.J. Hu (✉)
Department of Computer Science, Institute of Biological Engineering, National Chiao Tung University, Hsinchu, Taiwan
e-mail: yhu@cs.nctu.edu.tw

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978-1-4419-5913-3_14,

In functionally related protein families, there usually exist conserved local structural characteristics, e.g., the binding sites for metal-binding proteins. Given that the conservation in local active sites is likely to reflect similar biological functions, 3D patterns of local active sites can be used to predict the functions of previously unknown proteins [2]. These conserved structural features themselves represent significant motifs, which can be identified and described in various ways. Unlike most previous works on protein local structures, we describe a combinatorial approach to structural motif discovery. It first converts protein 3D structures into 1D structural alphabet letters, and then identifies conserved local features as 1D structural alphabet sequence motifs. There are several advantages of 1D structural alphabet over the conventional 3D co-ordinates. First, 1D representation of protein structures is more efficient in comparison and more economical in storage. Second, many commonly used 1D sequence tools can be directly applied to protein structure and sequence analysis. Third, 1D-based approaches can serve as pre-processors to filter out irrelevant proteins prior to the application of more computationally intensive 3D structure analysis tools.

14.2 Discovery of Structural Alphabet Motifs

The discovery of structural motifs can be divided into two stages. Given a structural alphabet, we can first transform a set of functionally or structurally related proteins, e.g., SCOP family, into a 1D representation. Different alphabets were derived based on different design philosophies [3–6]. Their size can vary from a dozen to nearly a hundred. They reflect different structural characteristics and have various applications. In different domains, we can adopt an appropriate structural alphabet to transform amino acid sequences or protein 3D structures into 1D structural alphabet sequences as required. In the second stage, we can apply a sequence motif detection algorithm to discover significant motifs from the 1D structural alphabet sequences. Like structural alphabets, a significant number of motif detection tools have been developed based on different objective functions, motif representations, and search strategies [7–10].

For this study, we designed a structural alphabet [6] that contains 18 letters, five of which represent the helix structure, eight for the sheet, and the rest for the coil. To discover structural motifs, we used MEME [7], which adopts an expectation maximization approach to find motifs represented as weight matrices. Unlike IUPAC-IUB codes, motifs described in weight matrices are more flexible because a weight matrix can show each alphabet letter preference in every motif position. Besides, a weight matrix can be easily transformed to IUPAC-IUB codes or regular expressions when necessary, but not vice versa. We call the motifs found by MEME from the structural alphabet sequences simple motifs. When the local properties in protein structures are too complicated, e.g., multiple binding sites or sub-domains, to capture in a simple motif, we can further combine several simple motifs into a compound motif. To avoid the computational complexity of combining matrices,

we transform simple motifs to regular expressions first, and then combine them to a compound motif. A compound motif example looks like the following.

$M_1(20,\ 50)M_2(0,\ 6)M_3$, where M_1, M_2, and M_3 are simple motifs, and the numbers in the parentheses denote the range of residue separation between motifs.

$$
\begin{aligned}
M_1 &= \text{SP[PS][SN]N[NE]EE},\\
M_2 &= \text{[WE][NE]EEACWGQS},\\
M_3 &= \text{TTTTTTTTTLK[TG][SH]WNMR[DQ]},
\end{aligned}
$$

where letters in brackets denote the possible structural alphabet letters in that particular motif position.

We show in Fig. 14.1 a general framework for structural motif discovery in which the structural alphabet and the motif finding algorithm can be replaced when necessary in different applications.

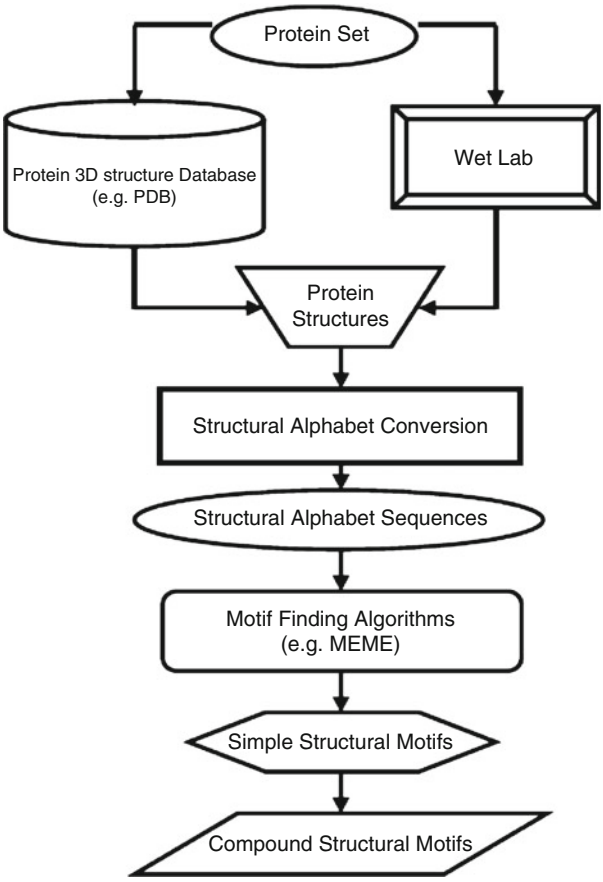

Fig. 14.1 System flow of structural motif discovery

The structural alphabet motifs can characterize the local structure features conserved in functionally related proteins. Based on the motif analysis in alphabet letter preference, alphabet letter occurrence distribution, and its significance, we may get a deeper insight of protein structures. To show the difference in alphabet conservation between structural alphabet motifs and the corresponding amino acid motifs, we showed one example motif for a SCOP family in Fig. 14.2a, which indicates that the structural alphabet motif is more conserved than the amino acid motif.

Besides alphabet conservation, we can also study the distribution of alphabet letter occurrences in each position of the structural and the amino acid motif, respectively. An example is shown in Fig. 14.2b and c. From the histograms, we can analyze the number of occurrences of each alphabet letter in a particular position of a motif. From the comparison of occurrences between structural alphabet and amino acids, we can derive the relationships between protein sequences and structures, e.g., the preference of structural alphabet for specific amino acids. Relationships of this kind about motifs can be further refined as building blocks to predict the structures of novel protein sequences.

14.3 Application Example of Structural Motifs

The C2H2 zinc finger is one of the best-studied metal-binding domains. It was first observed as a repeated zinc-binding motif with DNA-binding properties in the Xenopus transcription factor IIIA, and the term "zinc finger" is now largely used to denote any compact domain stabilized by a zinc ion [11, 12]. The domains from C2H2-like fingers consist of a β-hairpin followed by an α-helix that forms a left-handed ββα-unit, where two zinc ligands are contributed by a zinc knuckle at the end of the β-hairpin and other two ligands come from the C-terminal end of the α-helix [13, 14]. To demonstrate that our approach is capable of characterizing the structural ββα-unit, we analyzed the structural motifs discovered from 156 C2H2 zinc finger proteins in SCOP. A motif found was considered to match a (sub-)domain correctly if more than half of the residues in the (sub-)domain were included in the motif. If any simple or compound motif correctly corresponded to a (sub-)domain, we claimed that this (sub-)domain was recovered successfully (i.e., a hit). In Table 14.1, we present the compound motif found to characterize the (sub-)domains, and its coverage. The results suggested that using protein structural alphabet combined with 1D motif-finding algorithm was able to recover the structural (sub-)domains in proteins. We show some C2H2 zinc finger proteins with structural motifs marked in Fig. 14.3.

14.4 Conclusion

In this chapter, we introduced a general framework for structural motif discovery and proposed the applications of the motifs. Two major components in our framework are (1) the structural alphabet used to describe protein structures and (2) the

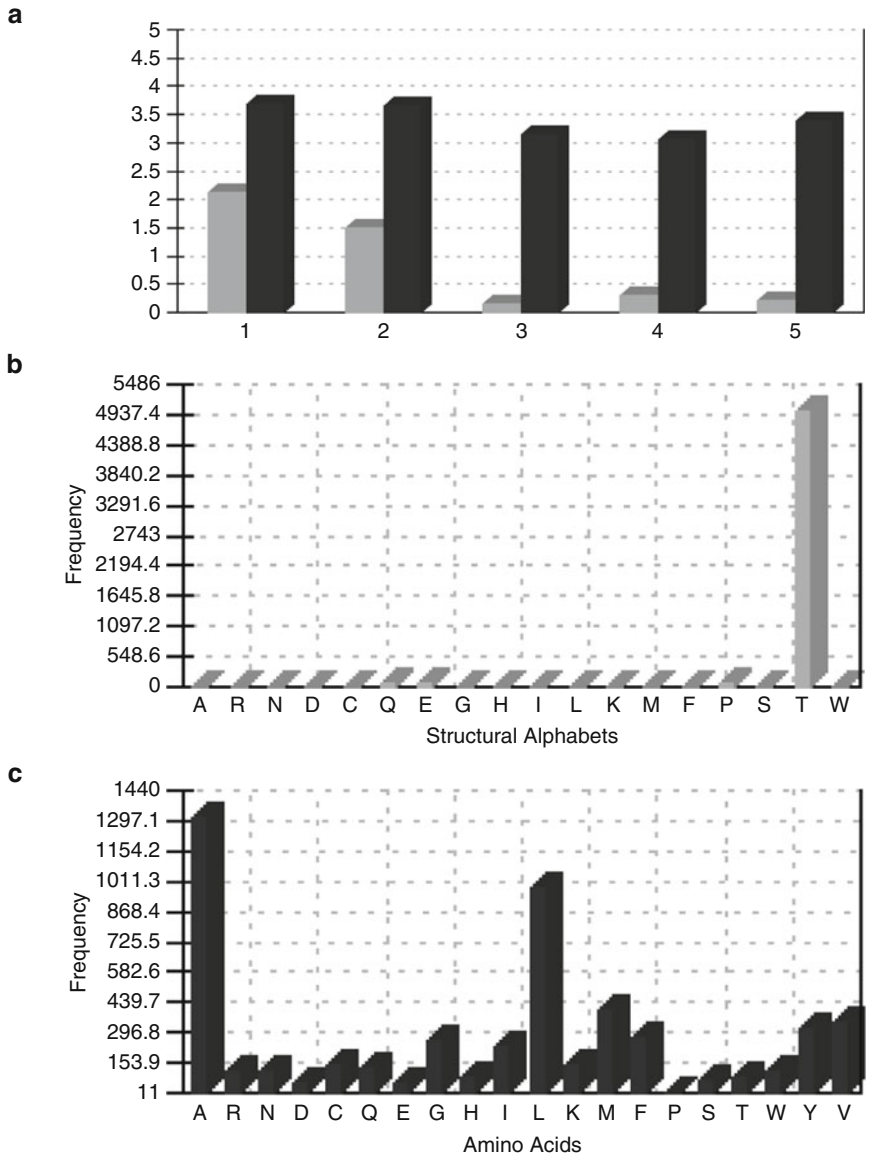

Fig. 14.2 Examples of motif analysis. (**a**) Histogram of entropy in each position of the structural alphabet motif and its corresponding amino acid motif. The *x*-axis indicates the positions, and the *y*-axis shows the entropy. Entropy of structural alphabet motif is colored in *gray*, and the amino acid motif in *black*. The lower the entropy, the more conserved the alphabet. (**b**, **c**) Histograms of alphabet letter occurrence distribution in the first position of the structural alphabet motif and its corresponding amino acid motif. The *x*-axis indicates all the alphabet letters, 18 in structural alphabet and 20 in amino acids. The *y*-axis shows the number of occurrences for a particular alphabet letter. (**b**) The distribution of structural alphabet letter occurrences in the first position. (**c**) The distribution of amino acid occurrences in the first position within the corresponding amino acid motif

Table 14.1 Summary of motifs mapping to C2H2 zinc finger ββα-unit that consists of β-hairpin and α-helix

Structural (sub-)domains	Motifs	Hit[a]	Coverage (%)[b]
β-Hairpin	[FH]CWNA[RC]QK(0-2) [GN][HE][NE]AC [AW]RQ	131	83.9
α-Helix	[GN][HE][NE]AC[AW]RQ(0-5)TTTTTT[PL] [KPL]	142	91.0
ββα-Unit	[FH]CWNA[RC]QK(0-2) [GN][HE][NE]AC [AW]RQ(0-5)TTTTTT[PL][KPL]	124	79.5
Total	–	156	100

[a]We called it a hit for a structural (sub-)domain when more than half of the (sub-)domain residues were contained in a motif. We presented the count of hits of different (sub-)domains
[b]Coverage was defined as the ratio of the count of hits to the number of zinc finger proteins, e.g., if total = 156 and Hit = 131, then coverage = 131/156 = 83.9%

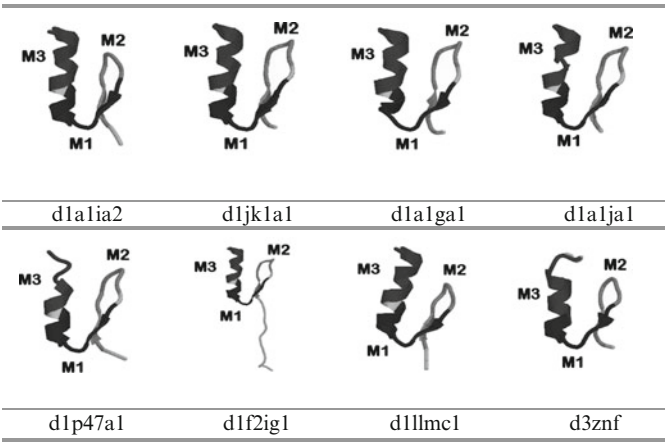

Fig. 14.3 Examples of C2H2 zinc finger protein structures. The simple motifs that map to the β-hairpin and the α-helix are highlighted in different *gray* levels, where M_1 = [GN][HE][NE]AC [AW]RQ, M_2 = [FH]CWNA[RC]QK, and M_3 = TTTTTT[PL][KPL]. The compound motif mapping to the ββα-unit is [FH]CWNA[RC]QK(0-2) [GN][HE][NE]AC[AW]RQ(0-5)TTTTTT [PL][KPL]

motif-finding algorithm used to discover significant local structure features. In our evaluation experiments, we used the structural alphabet designed in [6] and a widely used motif detection algorithm, MEME [7]. These components can be flexibly replaced with others when necessary to increase the applicability in different domains. The current results showed that using structural alphabets combined with 1D motif-finding algorithms could successfully identify biologically meaningful sub-domains in proteins.

We plan to continue the work in the following directions. First, many structural alphabets and quite a few motif detection algorithms have been developed based on

different design philosophies and application domains. We intend to incorporate other structural alphabets and motif-finding algorithm into our system. We expect to discover more kinds of motifs in a wider variety of protein structures. Second, the analysis of the distribution of alphabet occurrences and conservations within the motifs provides a different point of view from which to investigate the conserved evolutionary relationships in proteins as well as an alternative way in which to assist in protein structure prediction. We intend to design a protein function predictor using structural motifs as important features. Based on the motifs, the functions of novel proteins can be predicted by classifying them in to protein groups with known functions. Third, as many other protein structure or protein function prediction systems are available, we also plan to verify the possibility of using structural alphabet-based methods as a pre-processor. Compared with most prediction strategies typically based on 3D information, alphabet-based methods have much lower computational complexity. Thus they can help other predictors constrain the search space efficiently by filtering out irrelevant predictions in advance.

References

1. Berman HM, Battistuz T, Bhat TN, et al (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899–907.
2. Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5:355–373.
3. de Brevern AG (2005) New assessment of a structural alphabet. In Sillico Biol 5:26.
4. Camproux AC, Gautier R, Tuffery P (2004) A hidden Markov model derived structural alphabet for proteins. J Mol Biol 339:591–605.
5. Offmann B, Tyagi M, de Brevern AG (2007) Local protein structures. Curr Bioinformatics 2:165–202.
6. Ku S, Hu Y (2008) Protein structure search and local structure characterization. BMC Bioinformatics 9:349.
7. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:W369–W373.
8. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16:939–945.
9. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15:563–577.
10. van Helden J, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827–842.
11. Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol 11:39–46.
12. Iuchi S (2001) Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci 58:625–635.
13. Grishin NV (2001) Treble clef finger – a functionally diverse zinc binding structural motif. Nucleic Acids Res 29:1703–1714.
14. Wang B, Jones DN, Kaine BP, Weiss MA (1998) High resolution structure of an archaeal zinc ribbon defines a general architectural motif in eukaryotic RNA polymerases. Structure 6:555–569.