Computers in Biology and Medicine 42 (2012) 93–105

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Structural alphabet motif discovery and a structural motif database

Shih-Yen Ku [a], Yuh-Jyh Hu [a,b,*]

[a] Department of Computer Science, National Chiao Tung University, 1001 Tashuei Rd., Hsinchu, Taiwan
[b] Institute of Biomedical Engineering, National Chiao Tung University, 1001 Tashuei Rd., Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 1 June 2010
Accepted 27 October 2011

Keywords:
Protein structures
Structural alphabets
Structural motifs
Sequence motifs
Motif-finding tools

ABSTRACT

This study proposes a general framework for structural motif discovery. The framework is based on a modular design in which the system components can be modified or replaced independently to increase its applicability to various studies. It is a two-stage approach that first converts protein 3D structures into structural alphabet sequences, and then applies a sequence motif-finding tool to these sequences to detect conserved motifs. We named the structural motif database we built the SA-Motifbase, which provides the structural information conserved at different hierarchical levels in SCOP. For each motif, SA-Motifbase presents its 3D view; alphabet letter preference; alphabet letter frequency distribution; and the significance. SA-Motifbase is available at http://bioinfo.cis.nctu.edu.tw/samotifbase/.

1. Introduction

In functionally related protein groups, a number of conserved structural characteristics exist, such as binding sites of metal-binding proteins [1] and structural motifs in protein loops [2]. Because of the conservational structure of protein functional parts, the 3D patterns of local active sites can be used to predict the functions of previously unknown proteins [3]. These structurally conserved segments are important motifs, and they can be identified and described in various manners. For example, some motifs are identified based on particular structural properties, such as secondary structural content [4–6]. Nevertheless, these structural features are not always available, and therefore, to cater for this, we proposed a two-stage approach for structural motif discovery. Our approach did not require any special structural property other than the 3D coordinates.

In the first stage, our approach converts protein 3D structures into 1D structural alphabet sequences. In the second stage, it identifies conserved local segments as structural alphabet motifs. There are several advantages of structural alphabets: the first is that 1D representation of protein structures is more efficient in protein comparison and is also more economical in storage; second, many commonly used 1D sequence motif-finding tools [7–13] can be applied to protein structure and sequence analysis, and when encoded properly, for example, by using a similar amino acid alphabet, structural alphabet sequences can be treated the same as protein sequence inputs by the 1D sequence tools to find conserved motifs; and third, 1D-based approaches can serve as preprocessors to filter irrelevant proteins prior to the application of more computationally intensive 3D structure analysis tools. This study proposes a modular framework for identifying locally conserved protein structures. The components in the framework can be refined or replaced independently to improve the synergy and render our system applicable in different domains.

We identified the conserved structural motifs for each fold, superfamily, and family in SCOP [7]. We constructed the SA-Motifbase, which compiles a set of important motifs for protein structures at different hierarchical levels, namely the fold, superfamily, and family. It is easier to maintain and update than other motif databases, which store the motifs by more computationally complex comparisons of special structural features such as 3D substructures, hydrogen-bonding patterns, and residue packing [14–16]. We used several protein groups to demonstrate the applicability of the structural motifs in the SA-Motifbase.

2. Materials and methods

2.1. Discovery of structural alphabet motifs

We divided the structural motif discovery process into two stages. First, for a protein group, such as a SCOP family, we transformed the protein in its 3D structure into a structural alphabet sequence. Various structural alphabets have been developed based on different design strategies and domain knowledge [17–23]. Their size can vary from a dozen to nearly a hundred. They characterize

[*] Corresponding author at: Institute of Biomedical Engineering, National Chiao Tung University, 1001 Tashuei Rd., Hsinchu 300, Taiwan. Tel.: +886 3 5731795; fax: +886 3 5721490.
E-mail address: yhu@cs.nctu.edu.tw (Y.-J. Hu).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.10.013

different structural features and have various applications. To convert 3D structures, we adopted the alphabet used in SA-FAST because it has been proven to be effective in characterizing local protein structures [24]. Once this was accomplished, we detected the conserved motifs from the sequences. Numerous sequence motif-finding tools exist [8–14], which use different search strategies and objective functions to identify and evaluate motifs. We selected MEME [7] to discover the motifs from the proteins in each fold, superfamily, and family of SCOP because MEME is freely accessible, and it provides a convenient web-based interface. In MEME, a motif is represented as a position weight matrix. It is more expressive, and can be converted easily into a regular expression based on specified weight thresholds.

We call the motifs found by MEME *simple motifs*. We can combine multiple simple motifs into a *compound motif* to characterize more complex protein structures, such as multiple binding sites or subdomains. An example of compound motifs in the expression is shown below. To increase readability, we chose to represent the motifs in regular expressions.

$M_1(10,20)M_2(0,6)M_3$, where M_1, M_2 and M_3 are simple motifs, and the numbers within parentheses denote the range of residue separation between motifs.

$M_1 = \mathrm{SP[PN][SD]}N\mathrm{[NH]EE}$,
$M_2 = \mathrm{[WE][NE]EEACWGQS}$,
$M_3 = \mathrm{TTTTTTLK[TG][SH]WNMR[DQ]}$,

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

The system flow of the motif discovery is shown in Fig. 1. The protein 3D structures can be obtained from structure databases or biological laboratories. After converting the structures to alphabet sequences, and then detecting the conserved motifs from the sequences, the system produces simple motifs that can be further combined together to form compound motifs. Our goal of this paper is to propose a general framework and to prove its effectiveness for constructing structural motifs. Though in its current version SA-Motifbase does not allow arbitrary user inputs or replacement of system components, in Section 3 we demonstrate the modularity of our framework by applying different sequence motif-finding methods in the pipeline to detect structural motifs. To enable SA-Motifbase to accept various inputs or to substitute any component interactively requires a more advanced user interface and a more flexible interface between system components in the pipeline. The designs for both possibilities are now in progress.

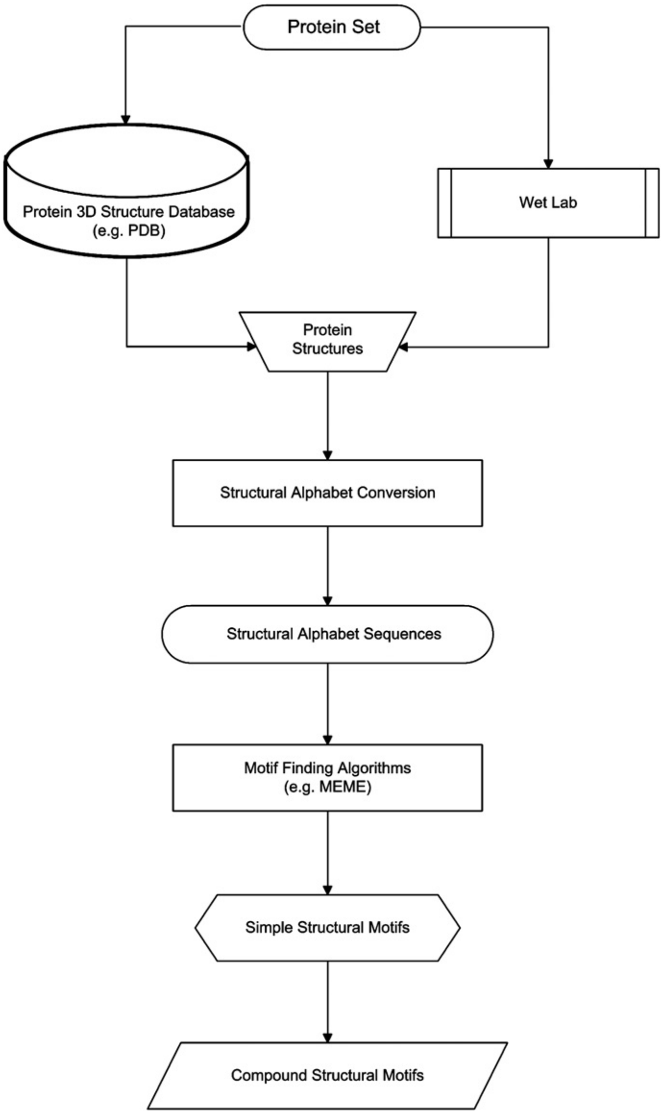

Protein Set
Protein 3D Structure Database (e.g. PDB)
Wet Lab
Protein Structures
Structural Alphabet Conversion
Structural Alphabet Sequences
Motif Finding Algorithms (e.g. MEME)
Simple Structural Motifs
Compound Structural Motifs

Fig. 1. System flow of structural motif discovery. The system flow *per se* is a modular design framework for structural motif databases. Based on this design, designers can use different protein inputs, different structural alphabets, and apply different motif-finding tools to build their own motif databases. We proposed a modular design concept to combine and organize different components required to build a structural motif database.

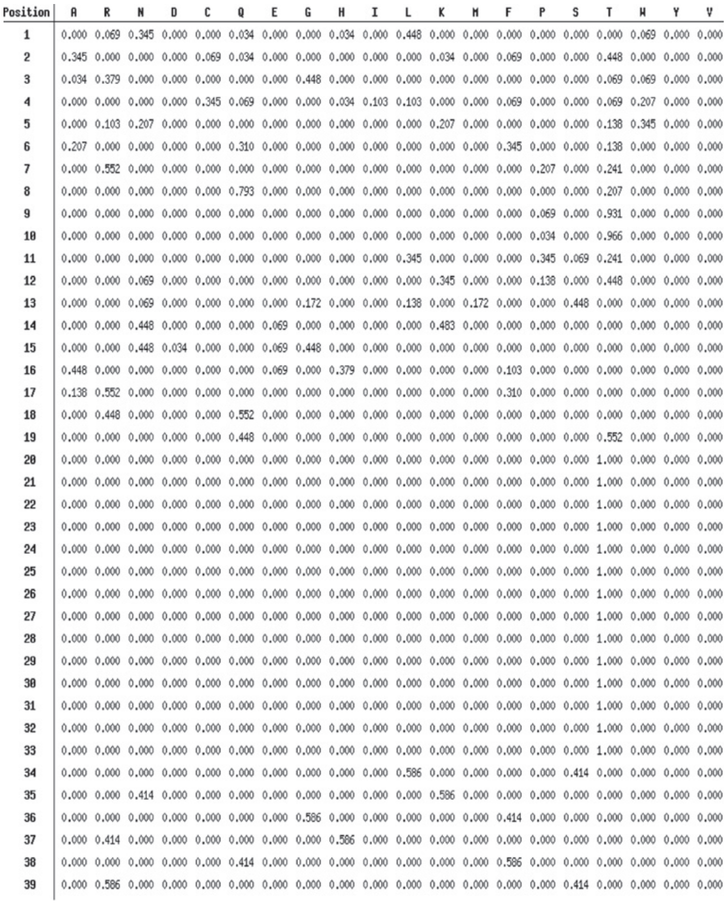
2.2. Content of SA-Motifbase

SA-Motifbase stores the structural alphabet motifs that characterize the local structural segments that are conserved in the SCOP protein hierarchy. For each motif, SA-Motifbase records its alphabet letter preference, the alphabet letter frequency distribution, and the significance. By comparing the alphabet distribution among the twenty amino acids and the structural alphabet, users of SA-Motifbase can analyze the correlation between protein sequences and structures in a particular protein group. They can also compare more easily a conserved structural motif in a protein family against others in different families in its alphabet letter preference rather than in its 3D coordinates. SA-Motifbase provides the possibility to extend the functionalities of 1D sequence analysis tools. For example, phylogenetic analysis tools [25,26] based on primary sequence similarity may be extended to analyze structural similarities when applied in structural alphabets.

In SA-Motifbase, we represent each simple motif by a regular expression and by a position weight matrix, as shown in Fig. 2. To show the difference in alphabet conservation between a structural alphabet motif and its amino acids, SA-Motifbase displays a histogram of entropy in each position within the motif. We presented a sample histogram of SA-Motifbase in Fig. 3. This example showed that the entropy of the structural alphabet was lower than that of the amino acids. It indicated that this example motif was more conserved in structure than in sequence.

For each structural alphabet motif and its amino acids, SA-Motifbase also provides a histogram to show the distribution of alphabet letter frequency in each position within the motif. We showed an example of this in Fig. 4. From this histogram, biologists can analyze the alphabet letter preference in each position within a motif. By comparing the letter preference between the structural alphabet and amino acids, we can derive a relationship between protein sequences and structures. Relationships of this type can then be further used to predict the structures of novel protein sequences.

a



Position	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.000	0.069	0.345	0.000	0.000	0.034	0.000	0.000	0.034	0.000	0.448	0.000	0.000	0.000	0.000	0.000	0.000	0.069	0.000	0.000
2	0.345	0.000	0.000	0.000	0.069	0.034	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.069	0.000	0.000	0.448	0.000	0.000	0.000
3	0.034	0.379	0.000	0.000	0.000	0.000	0.000	0.448	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.069	0.069	0.000	0.000
4	0.000	0.000	0.000	0.000	0.345	0.069	0.000	0.000	0.034	0.103	0.103	0.000	0.000	0.069	0.000	0.000	0.069	0.207	0.000	0.000
5	0.000	0.103	0.207	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.207	0.000	0.000	0.000	0.000	0.138	0.345	0.000	0.000
6	0.207	0.000	0.000	0.000	0.000	0.310	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.345	0.000	0.000	0.138	0.000	0.000	0.000
7	0.000	0.552	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.207	0.000	0.241	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.793	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.207	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.069	0.000	0.931	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.966	0.000	0.000	0.000
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.345	0.000	0.000	0.000	0.345	0.069	0.241	0.000	0.000	0.000
12	0.000	0.000	0.069	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.345	0.000	0.000	0.138	0.000	0.448	0.000	0.000	0.000
13	0.000	0.000	0.069	0.000	0.000	0.000	0.000	0.172	0.000	0.000	0.138	0.000	0.172	0.000	0.000	0.448	0.000	0.000	0.000	0.000
14	0.000	0.000	0.448	0.000	0.000	0.000	0.069	0.000	0.000	0.000	0.000	0.483	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	0.000	0.000	0.448	0.034	0.000	0.000	0.069	0.448	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	0.448	0.000	0.000	0.000	0.000	0.000	0.069	0.000	0.379	0.000	0.000	0.000	0.000	0.103	0.000	0.000	0.000	0.000	0.000	0.000
17	0.138	0.552	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.310	0.000	0.000	0.000	0.000	0.000	0.000
18	0.000	0.448	0.000	0.000	0.000	0.552	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.000	0.000	0.000	0.000	0.000	0.448	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.552	0.000	0.000	0.000
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
21	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
22	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
23	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
24	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
25	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
26	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
27	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
28	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
29	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
30	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
31	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
32	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
33	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
34	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000	0.414	0.000	0.000	0.000	0.000
35	0.000	0.000	0.414	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
36	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000	0.000	0.414	0.000	0.000	0.000	0.000	0.000	0.000
37	0.000	0.414	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
38	0.000	0.000	0.000	0.000	0.000	0.414	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000	0.000	0.000
39	0.000	0.586	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.414	0.000	0.000	0.000	0.000

b

[LN][TA][GR][CW][WKN][FQA][RTP][QT]TT[LPT][TK]S[KN][GN][AH][RF][QR][TQ]TTTTTTTTTT
TTTT[LS][KN][GF][HR][FQ][RS]

Fig. 2. A sample structural alphabet motif in SCOP a.1.1.1: (a) represented by a position weight matrix and (b) represented by a regular expression. A position weight matrix is translated into a regular expression based on specified weight thresholds.

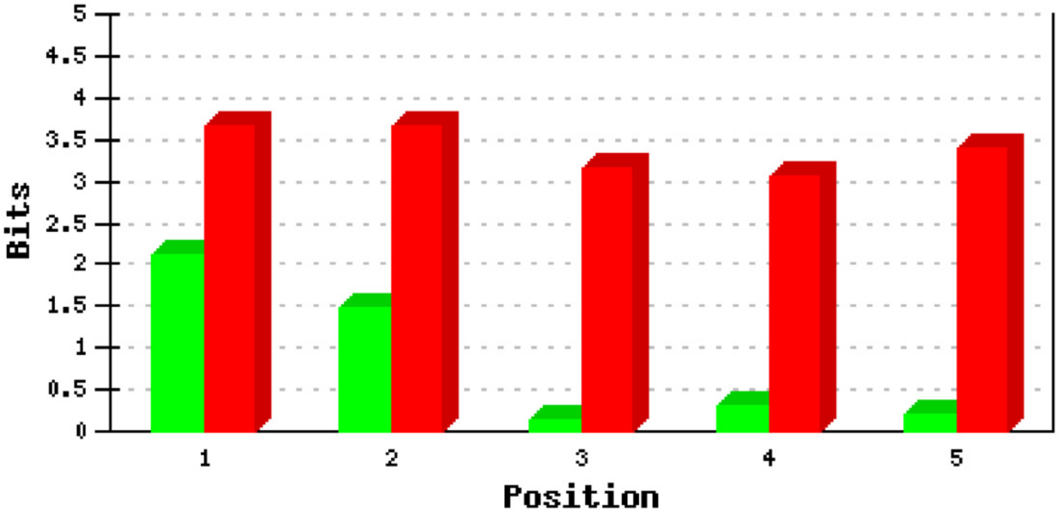


Fig. 3. A sample histogram of entropy in each position within a 5-letter motif represented in structural alphabet or amino acids. The *x*-axis indicates the positions, and the *y*-axis shows the entropy in bits. In each position, entropy in structural alphabet and amino acids is indicated as the left and the right bar, respectively. Entropy was defined as $-\sum_{i=1}^{N} p_i \lg p_i$, where N is the size of the alphabet (e.g. N=20 for amino acids), and p_i is the probability of some alphabet letter in position i. The lower the entropy, the more conserved the alphabet in that position.

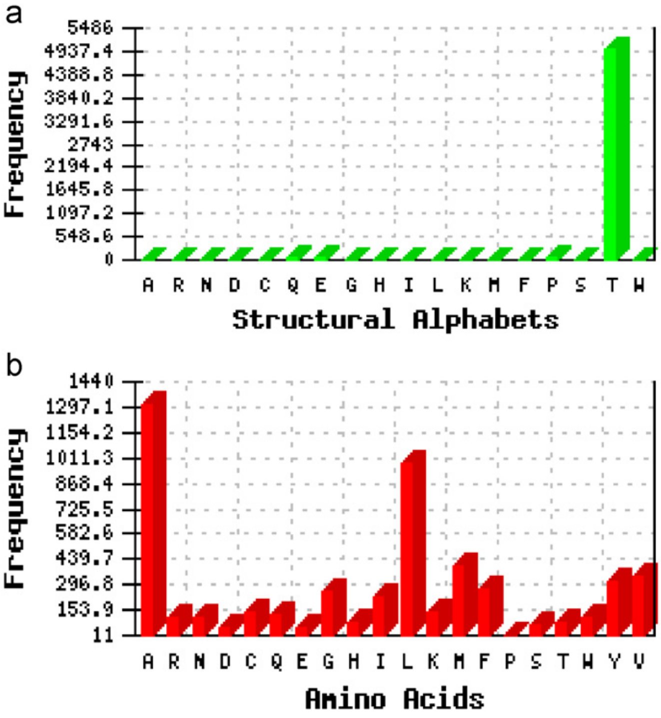


Fig. 4. Sample histograms of alphabet letter frequency distribution in the first position within an example motif represented in structural alphabet or amino acids. The *x*-axis indicates all the alphabet letters, 18 letters in structural alphabet and 20 in amino acids. The *y*-axis shows the number of occurrences for a particular alphabet letter. (a) The distribution of structural alphabet letter frequency and (b) the distribution of amino acid frequency.

Table 1
Summary of motif statistics in folds, superfamilies and families.

Meaning of statistics	**Hierarchical level**		
	Fold	**Super family**	**Family**
Number of proteins containing motifs (total proteins)	964 (1160)	1506 (1833)	2567 (3277)
Coverage (%)[a]	83.10	82.16	78.33
Total motifs[b]	8857	9899	10,040
Number of motifs/protein[c]	9	7	4
Mean size of motifs[d]	22	31	55

[a] Coverage is the number of proteins containing motifs divided by total proteins, e.g. 83.10% ≈ 964/1160.

[b] Total motifs is the total number of motifs identified at a hierarchical level, e.g. 8857 motifs were found from all the folds in SCOP.

[c] Number of motifs/protein is defined as total motifs divided by number of proteins containing motifs, e.g. 9 ≈ 8857/964.

[d] Mean size of motifs is the average motif size in the number of alphabet letters.

2.3. Motif statistics

We summarized in Table 1 the statistics of the motifs identified from the folds, superfamilies, and families in the SCOP database. The statistics include the number of proteins containing the motifs, the average number of motifs in a protein, the total number of motifs, as well as the mean size of the motifs. There are 83% and 82% of the proteins at the fold and superfamily levels that contain the structural alphabet motifs, respectively. This demonstrates that the structural segments we identified are conserved well in these protein groups. Compared with folds and superfamilies, the percentage at the family level is lower, but 78% of the proteins still show the existence of motifs. In contrast, because the size of a family is significantly smaller than that of a superfamily or a fold, family members share a greater similarity in both sequence and structure. The mean size of conserved motifs is larger at the family level than that of a superfamily or fold, as shown in Table 1.

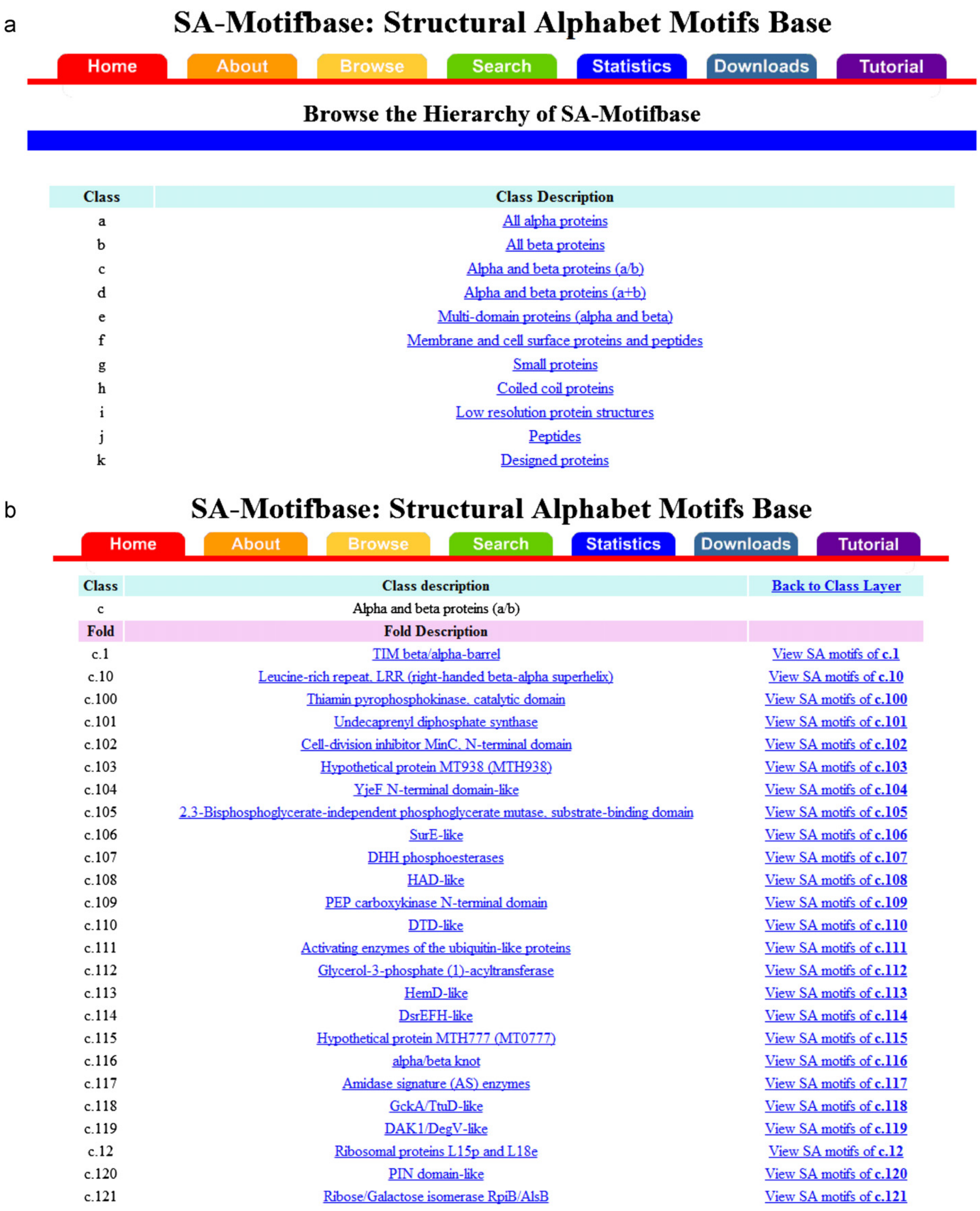
3. Results

The eight α/β motifs folded into a barrel-like structure in TIM barrel proteins were first discovered in triose-phosphate isomerase, and have since been widely analyzed [27–29]. Here, we use TIM barrel proteins as an example to demonstrate how to use the SA-Motifbase and what information the SA-Motifbase can deliver.

Fig. 5 shows a possible query on the motif information obtained from the TIM barrel fold. Users of SA-Motifbase can start tracing the protein structure hierarchy by clicking the hyperlinks to the desired level, as shown in Fig. 5(a) and (b). SA-Motifbase presents the conserved structural alphabet motifs in regular expressions, and highlights their locations in each protein, as shown in Fig. 5(c). Users can acquire the 3D view of proteins in which the structural motifs are marked for reference, as shown in Fig. 5(d). The structural motifs we identified match well to the α/β units in the barrel structure. In addition to regular expressions, users can also refer to the position weight matrix representation of the motifs. An example of these weight matrices is shown in Fig. 5(e). For further analysis of the motifs, SA-Motifbase also provides histograms (Fig. 5(e)) that visualize the motif conservation and alphabet letter preference.

In addition to TIM barrel proteins, we also presented the study of two types of metalloproteins, namely Zn^{2+}-binding proteins and Mg^{2+}-binding proteins, to further demonstrate the applicability of our system. Metalloproteins require metal cofactors in cellular biochemistry. These proteins play an important role in both intra- and extracellular catalytic activities and structural stabilization [30–32] because metal binding increases the thermal and conformational stability of small domains.

The C2H2 zinc finger is one of the most extensively 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 widely used to denote any compact domain stabilized by a zinc ion [33,34]. The domains from the C2H2-like fingers consist of a β-hairpin followed by an α-helix that together form a left-handed ββα-unit, where two zinc ligands are contributed by a zinc knuckle at the end of the β-hairpin and the other two ligands are derived from the C-terminal end of the α-helix [35,36]. To demonstrate that the proposed approach is capable of characterizing the structural ββα-unit, we analyzed the structural motifs discovered from the g.37.1 superfamily in SCOP. A motif was considered to match a subdomain correctly if over half the residues in the subdomain were included in the motif. If any simple motif or compound motif matched a subdomain, the subdomain was considered to have recovered successfully. Table 2 lists the simple or compound motifs that were found to characterize the subdomains. The results suggested that using a protein structural alphabet combined with a 1D motif-finding algorithm is able to recover the structural subdomains in proteins. A number of C2H2 zinc finger proteins, with structural motifs numbered, are shown in Fig. 6. To further demonstrate that locally conserved structures could be characterized by structural alphabet motifs, we calculated the average RMSD for the three motifs represented in Fig. 6. We first extracted the 3D structure segment corresponding to the structural motif in superfamily g.37.1, and then, for all possible structural segment pairs, we calculated their RMSD, and the cumulative distribution of the pairwise RMSD is shown in Fig. 7. Results showed that more than 94.5% of the paired RMSDs



SA-Motifbase: Structural Alphabet Motifs Base

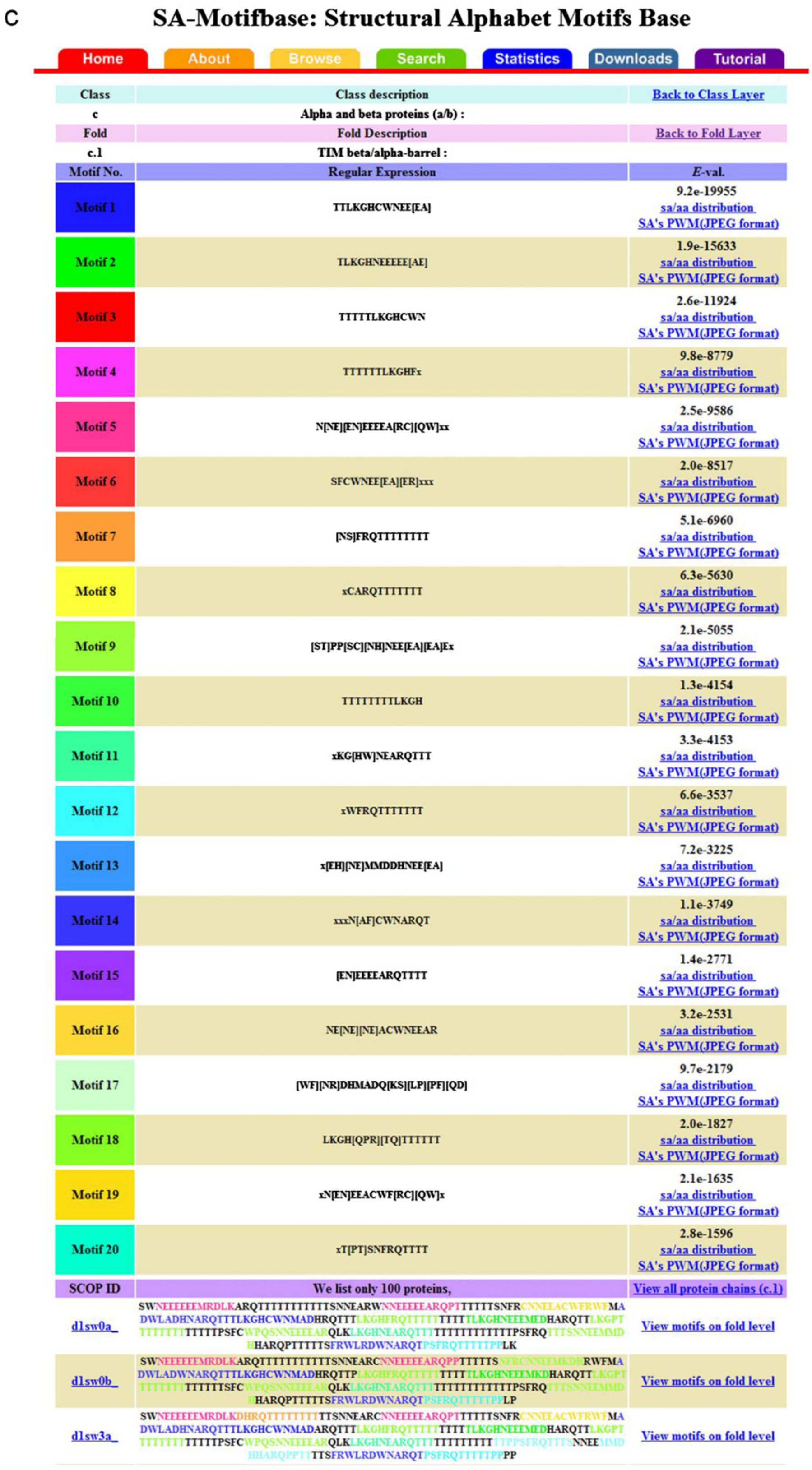
Browse the Hierarchy of SA-Motifbase

Class	Class Description
a	All alpha proteins
b	All beta proteins
c	Alpha and beta proteins (a/b)
d	Alpha and beta proteins (a+b)
e	Multi-domain proteins (alpha and beta)
f	Membrane and cell surface proteins and peptides
g	Small proteins
h	Coiled coil proteins
i	Low resolution protein structures
j	Peptides
k	Designed proteins

SA-Motifbase: Structural Alphabet Motifs Base

Class	Class description	Back to Class Layer
c	Alpha and beta proteins (a/b)	
Fold	**Fold Description**	
c.1	TIM beta/alpha-barrel	View SA motifs of **c.1**
c.10	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	View SA motifs of **c.10**
c.100	Thiamin pyrophosphokinase, catalytic domain	View SA motifs of **c.100**
c.101	Undecaprenyl diphosphate synthase	View SA motifs of **c.101**
c.102	Cell-division inhibitor MinC, N-terminal domain	View SA motifs of **c.102**
c.103	Hypothetical protein MT938 (MTH938)	View SA motifs of **c.103**
c.104	YjeF N-terminal domain-like	View SA motifs of **c.104**
c.105	2,3-Bisphosphoglycerate-independent phosphoglycerate mutase, substrate-binding domain	View SA motifs of **c.105**
c.106	SurE-like	View SA motifs of **c.106**
c.107	DHH phosphoesterases	View SA motifs of **c.107**
c.108	HAD-like	View SA motifs of **c.108**
c.109	PEP carboxykinase N-terminal domain	View SA motifs of **c.109**
c.110	DTD-like	View SA motifs of **c.110**
c.111	Activating enzymes of the ubiquitin-like proteins	View SA motifs of **c.111**
c.112	Glycerol-3-phosphate (1)-acyltransferase	View SA motifs of **c.112**
c.113	HemD-like	View SA motifs of **c.113**
c.114	DsrEFH-like	View SA motifs of **c.114**
c.115	Hypothetical protein MTH777 (MT0777)	View SA motifs of **c.115**
c.116	alpha/beta knot	View SA motifs of **c.116**
c.117	Amidase signature (AS) enzymes	View SA motifs of **c.117**
c.118	GckA/TtuD-like	View SA motifs of **c.118**
c.119	DAK1/DegV-like	View SA motifs of **c.119**
c.12	Ribosomal proteins L15p and L18e	View SA motifs of **c.12**
c.120	PIN domain-like	View SA motifs of **c.120**
c.121	Ribose/Galactose isomerase RpiB/AlsB	View SA motifs of **c.121**

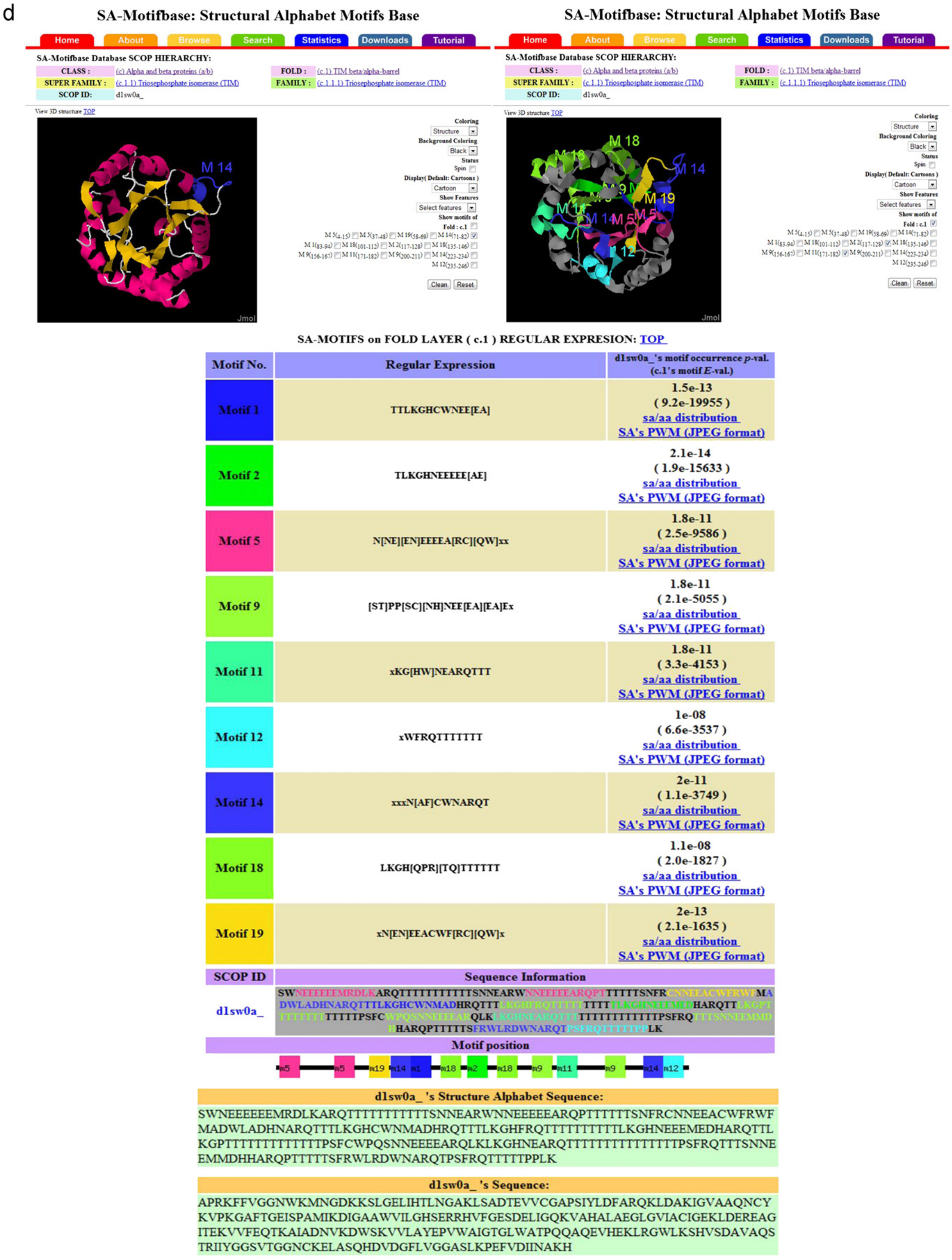
Fig. 5. Demonstration of SA-Motifbase. We used the TIM barrel fold as an example. (a) SA-Motifbase shows the protein structural hierarchy as in SCOP. Users can select the classification group of interest by clicking the hyperlink. (b) List of folds under Class c alpha and beta proteins (a/b). (c) List of conserved structural alphabet motifs in Fold c.1. (d) By clicking "View motifs on fold level," users get a 3D view of the selected protein, e.g. d1sw0a_. (e) SA-Motifbase provides the information of alphabet letter entropy and the letter frequency distribution in each position, visualized by histograms. Besides a regular expression, a position weight matrix of the motif is also provided for reference.



SA-Motifbase: Structural Alphabet Motifs Base

Class	Class description	Back to Class Layer
c	Alpha and beta proteins (a/b) :	
Fold	**Fold Description**	Back to Fold Layer
c.1	TIM beta/alpha-barrel :	
Motif No.	**Regular Expression**	***E*-val.**
Motif 1	TTLKGHCWNEE[EA]	9.2e-19955 sa/aa distribution SA's PWM(JPEG format)
Motif 2	TLKGHNEEEEE[AE]	1.9e-15633 sa/aa distribution SA's PWM(JPEG format)
Motif 3	TTTTTLKGHCWN	2.6e-11924 sa/aa distribution SA's PWM(JPEG format)
Motif 4	TTTTTTLKGHFx	9.8e-8779 sa/aa distribution SA's PWM(JPEG format)
Motif 5	N[NE][EN]EEEEA[RC][QW]xx	2.5e-9586 sa/aa distribution SA's PWM(JPEG format)
Motif 6	SFCWNEE[EA][ER]xxx	2.0e-8517 sa/aa distribution SA's PWM(JPEG format)
Motif 7	[NS]FRQTTTTTTTT	5.1e-6960 sa/aa distribution SA's PWM(JPEG format)
Motif 8	xCARQTTTTTTT	6.3e-5630 sa/aa distribution SA's PWM(JPEG format)
Motif 9	[ST]PP[SC][NH]NEE[EA][EA]Ex	2.1e-5055 sa/aa distribution SA's PWM(JPEG format)
Motif 10	TTTTTTTTLKGH	1.3e-4154 sa/aa distribution SA's PWM(JPEG format)
Motif 11	xKG[HW]NEARQTTT	3.3e-4153 sa/aa distribution SA's PWM(JPEG format)
Motif 12	xWFRQTTTTTTT	6.6e-3537 sa/aa distribution SA's PWM(JPEG format)
Motif 13	x[EH][NE]MMDDHNEE[EA]	7.2e-3225 sa/aa distribution SA's PWM(JPEG format)
Motif 14	xxxN[AF]CWNARQT	1.1e-3749 sa/aa distribution SA's PWM(JPEG format)
Motif 15	[EN]EEEEARQTTTT	1.4e-2771 sa/aa distribution SA's PWM(JPEG format)
Motif 16	NE[NE][NE]ACWNEEAR	3.2e-2531 sa/aa distribution SA's PWM(JPEG format)
Motif 17	[WF][NR]DHMADQ[KS][LP][PF][QD]	9.7e-2179 sa/aa distribution SA's PWM(JPEG format)
Motif 18	LKGH[QPR][TQ]TTTTTT	2.0e-1827 sa/aa distribution SA's PWM(JPEG format)
Motif 19	xN[EN]EEACWF[RC][QW]x	2.1e-1635 sa/aa distribution SA's PWM(JPEG format)
Motif 20	xT[PT]SNFRQTTTT	2.8e-1596 sa/aa distribution SA's PWM(JPEG format)
SCOP ID	**We list only 100 proteins,**	View all protein chains (c.1)
d1sw0a_	SWNEEEEEEMRDLKARQTTTTTTTTTTTTSNNEARWNNEEEEEARQPTTTTTTSNFRCNNEEACWFRWFMA DWLADHNARQTTTLKGHCWNMADHRQTTTLKGHFRQTTTTTTTTTTLKGHNEEEMEDHARQTTLKGPT TTTTTTTTTTTPSFCWPQSNNEEEEARQLKLKGHNEARQTTTTTTTTTTTTTTTPSFRQTTTSNNEEMMD HHARQPTTTTTSFRWLRDWNARQTPSFRQTTTTTPPLK	View motifs on fold level
d1sw0b_	SWNEEEEEEMRDLKARQTTTTTTTTTTTTSNNEARCNNEEEEEARQPPTTTTTSNFRCNNEEMKDHRWFMA DWLADWNARQTTTLKGHCWNMADHRQTTPLKGHFRQTTTTTTTTTTLKGHNEEEMKDHARQTTLKGPT TTTTTTTTTTTTSFCWPQSNNEEEEARQLKLKGHNEARQTTTTTTTTTTTTTTTPSFRQTTTSNNEEMMD HHARQPTTTTTSFRWLRDWNARQTPSFRQTTTTTPPLP	View motifs on fold level
d1sw3a_	SWNEEEEEEMRDLKDHRQTTTTTTTTTTSNNEARCNNEEEEEARQPTTTTTTSNFRCNNEEACWFRWFMA DWLADHNARQTTTLKGHCWNMADARQTTTLKGHFRQTTTTTTTTTTLKGHNEEEMEDHARQTTLKGPT TTTTTTTTTTTPSFCWPQSNNEEEEARQLKLKGHNEARQTTTTTTTTTTTTTPPSFRQTTTSNNEEMMD HHARQPPTTTTSFRWLRDWNARQTPSFRQTTTTTPPPP	View motifs on fold level

Fig. 5. (*continued*)



SA-Motifbase: Structural Alphabet Motifs Base

SA-Motifbase Database SCOP HIERARCHY:

CLASS :	(c) Alpha and beta proteins (a/b)	FOLD :	(c.1) TIM beta/alpha-barrel
SUPER FAMILY :	(c.1.1) Triosephosphate isomerase (TIM)	FAMILY :	(c.1.1.1) Triosephosphate isomerase (TIM)
SCOP ID:	d1sw0a_		

View 3D structure TOP

Coloring
Structure
Background Coloring
Black
Status
Spin □
Display(Default: Cartoons)
Cartoon
Show Features
Select features
Show motifs of
Fold : c.1 □
M 5(4-15) □ M 5(37-48) □ M 19(58-69) □ M 14(71-82) ☑
M 1(83-94) □ M 18(101-112) □ M 2(117-128) □ M 18(135-146) □
M 9(156-167) □ M 11(171-182) □ M 9(200-211) □ M 14(223-234) □
M 12(235-246) □
Clean Reset

SA-Motifbase: Structural Alphabet Motifs Base

SA-Motifbase Database SCOP HIERARCHY:

CLASS :	(c) Alpha and beta proteins (a/b)	FOLD :	(c.1) TIM beta/alpha-barrel
SUPER FAMILY :	(c.1.1) Triosephosphate isomerase (TIM)	FAMILY :	(c.1.1.1) Triosephosphate isomerase (TIM)
SCOP ID:	d1sw0a_		

View 3D structure TOP

Coloring
Structure
Background Coloring
Black
Status
Spin □
Display(Default: Cartoons)
Cartoon
Show Features
Select features
Show motifs of
Fold : c.1 ☑
M 5(4-15) □ M 5(37-48) □ M 19(58-69) □ M 14(71-82) □
M 1(83-94) □ M 18(101-112) □ M 2(117-128) ☑ M 18(135-146) □
M 9(156-167) □ M 11(171-182) ☑ M 9(200-211) □ M 14(223-234) □
M 12(235-246) □
Clean Reset

SA-MOTIFS on FOLD LAYER (c.1) REGULAR EXPRESION: TOP

Motif No.	Regular Expression	d1sw0a_'s motif occurrence *p*-val. (c.1's motif *E*-val.)
Motif 1	TTLKGHCWNEE[EA]	1.5e-13 (9.2e-19955) sa/aa distribution SA's PWM (JPEG format)
Motif 2	TLKGHNEEEEE[AE]	2.1e-14 (1.9e-15633) sa/aa distribution SA's PWM (JPEG format)
Motif 5	N[NE][EN]EEEEA[RC][QW]xx	1.8e-11 (2.5e-9586) sa/aa distribution SA's PWM (JPEG format)
Motif 9	[ST]PP[SC][NH]NEE[EA][EA]Ex	1.8e-11 (2.1e-5055) sa/aa distribution SA's PWM (JPEG format)
Motif 11	xKG[HW]NEARQTTT	1.8e-11 (3.3e-4153) sa/aa distribution SA's PWM (JPEG format)
Motif 12	xWFRQTTTTTTT	1e-08 (6.6e-3537) sa/aa distribution SA's PWM (JPEG format)
Motif 14	xxxN[AF]CWNARQT	2e-11 (1.1e-3749) sa/aa distribution SA's PWM (JPEG format)
Motif 18	LKGH[QPR][TQ]TTTTTT	1.1e-08 (2.0e-1827) sa/aa distribution SA's PWM (JPEG format)
Motif 19	xN[EN]EEACWF[RC][QW]x	2e-13 (2.1e-1635) sa/aa distribution SA's PWM (JPEG format)

SCOP ID	Sequence Information
d1sw0a_	SWNEEEEEEEMRDLKARQTTTTTTTTTTTTSNNEARWNNEEEEEARQPTTTTTTSNFRCNNEEACWFRWFMA DWLADHNARQTTTLKGHCWNMADHRQTTTLKGHFRQTTTTTTTTTTLKGHNEEEMEDHARQTTLKGPT TTTTTTTTTTTTTPSFCWPQSNNEEEEARQLKLKGHNEARQTTTTTTTTTTTTTTTTPSFRQTTTSNNEEMMD HHARQPTTTTTSFRWLRDWNARQTPSFRQTTTTTPPLK

Motif position

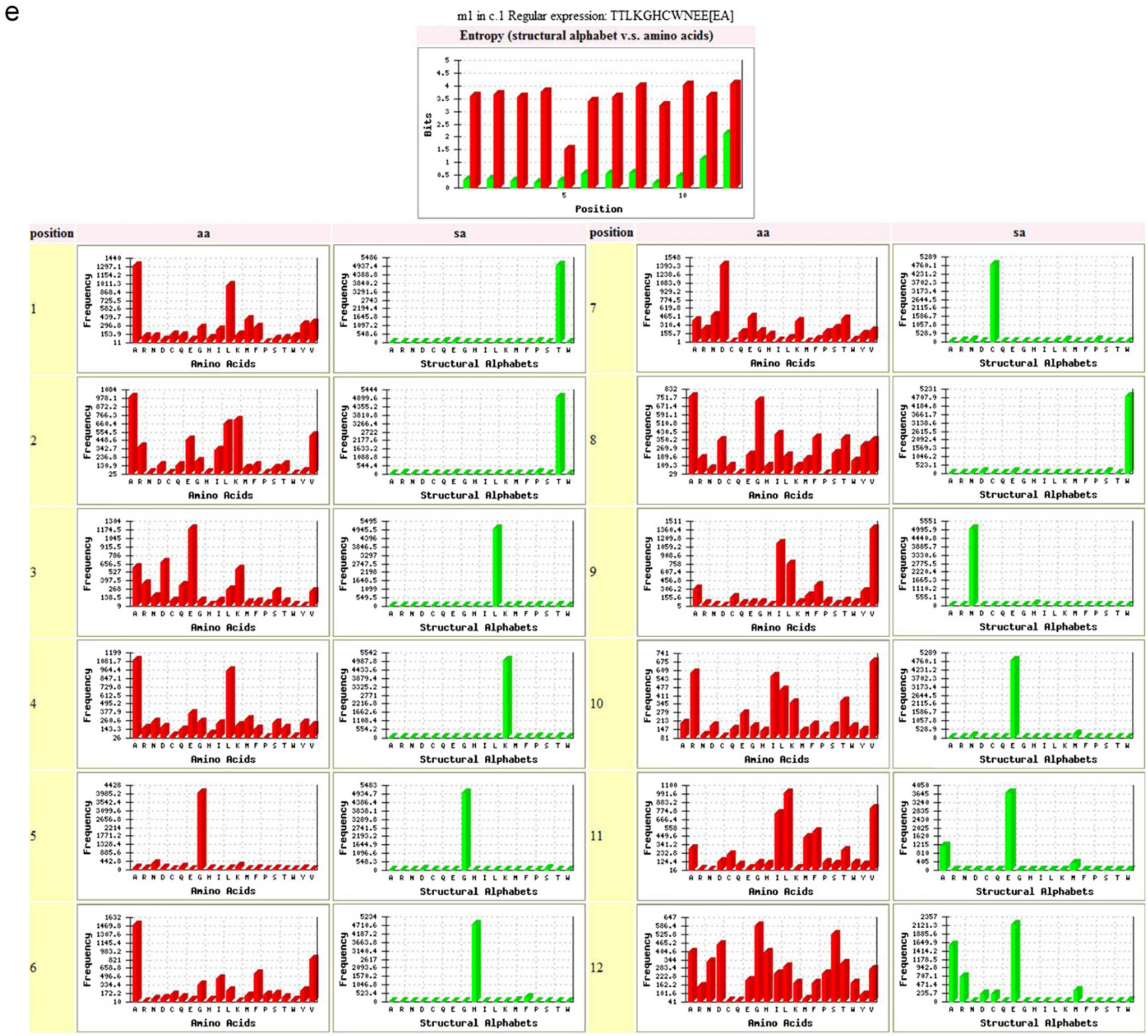
d1sw0a_ 's Structure Alphabet Sequence:

SWNEEEEEEEMRDLKARQTTTTTTTTTTTTSNNEARWNNEEEEEARQPTTTTTTSNFRCNNEEACWFRWF
MADWLADHNARQTTTLKGHCWNMADHRQTTTLKGHFRQTTTTTTTTTTLKGHNEEEMEDHARQTTL
KGPTTTTTTTTTTTTTPSFCWPQSNNEEEEARQLKLKGHNEARQTTTTTTTTTTTTTTTTPSFRQTTTSNNE
EMMDHHARQPTTTTTSFRWLRDWNARQTPSFRQTTTTTPPLK

d1sw0a_ 's Sequence:

APRKFFVGGNWKMNGDKKSLGELIHTLNGAKLSADTEVVCGAPSIYLDFARQKLDAKIGVAAQNCY
KVPKGAFTGEISPAMIKDIGAAWVILGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAG
ITEKVVFEQTKAIADNVKDWSKVVLAYEPVWAIGTGLWATPQQAQEVHEKLRGWLKSHVSDAVAQS
TRIIYGGSVTGGNCKELASQHDVDGFLVGGASLKPEFVDIINAKH

Fig. 5. (*continued*)



Position Weight Matrix of m1 in c.1 (Regularation: TTLKGHCWNEE[EA])

Position	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.005	0.001	0.000	0.002	0.000	0.007	0.007	0.000	0.000	0.000	0.002	0.001	0.000	0.001	0.006	0.000	0.966	0.003	0.000	0.000
2	0.001	0.009	0.000	0.001	0.000	0.003	0.006	0.000	0.000	0.000	0.002	0.002	0.000	0.000	0.017	0.002	0.959	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.001	0.003	0.000	0.000	0.000	0.000	0.967	0.001	0.008	0.000	0.007	0.002	0.010	0.000	0.000	0.000
4	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.976	0.000	0.003	0.011	0.001	0.006	0.000	0.000	0.000
5	0.000	0.000	0.003	0.015	0.000	0.000	0.000	0.965	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.000	0.000	0.000	0.000
6	0.002	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.922	0.000	0.003	0.000	0.012	0.051	0.003	0.000	0.001	0.002	0.000	0.000
7	0.000	0.006	0.019	0.001	0.931	0.003	0.001	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.016	0.004	0.000	0.000	0.000	0.000
8	0.000	0.000	0.015	0.019	0.001	0.000	0.020	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.024	0.000	0.921	0.000	0.000
9	0.000	0.000	0.977	0.000	0.000	0.000	0.002	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.005	0.000	0.023	0.000	0.000	0.000	0.931	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	0.221	0.000	0.000	0.001	0.003	0.000	0.713	0.000	0.000	0.000	0.000	0.000	0.061	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.305	0.132	0.000	0.043	0.043	0.000	0.415	0.000	0.000	0.000	0.000	0.001	0.058	0.000	0.000	0.000	0.000	0.003	0.000	0.000

Fig. 5. (*continued*)

Table 2
Summary of compound motifs mapping to C2H2 zinc finger ββα-unit that consists of β-hairpin and α-helix.

Structural (sub-)domain	Compound motif	SCOP g.37.1	
		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 for different (sub-)domains.

[b] Coverage was defined as the ratio of the count of hits to the total of zinc finger proteins, e.g., if total=156 and hits=131, then coverage=131/156=83.9%.

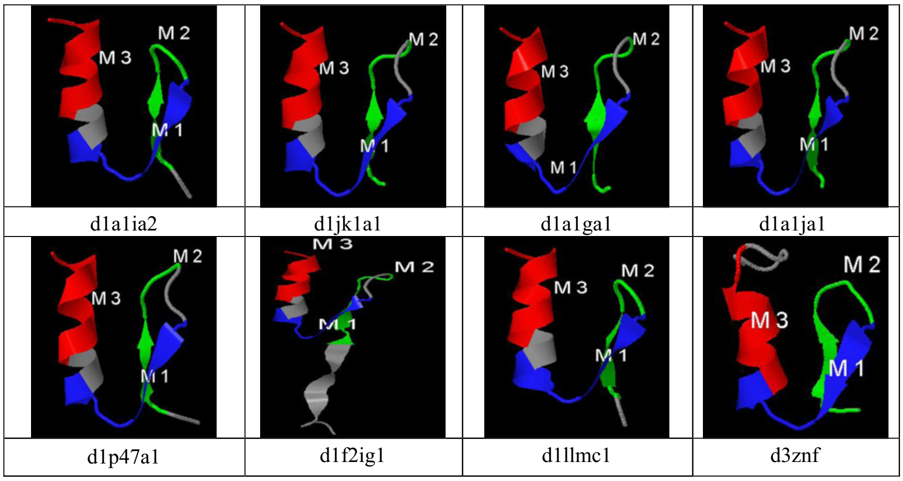


Fig. 6. Examples of C2H2 zinc finger protein structures. The simple motifs detected by MEME that map to the β-hairpin and the α-helix are numbered, where M_1=[GN][HE][NE]AC[AW]RQ, M_2=[FH]CWNA[RC]QK and M_3=TTTTTT[PL][KPL]. The compound motif mapping to the ββα-unit is [FH]CWNA[RC]QK **(0-2)** [GN][HE][NE] AC[AW]RQ **(0-5)** TTTTTT[PL][KPL].

were zero, which suggested that structural alphabet motifs reflect locally conserved structures.

A few relatively short sequence motifs have been discovered for Mg^{2+} proteins with a close sequence homology. Examples include the NADFDGD motif, found in different RNA polymerases; DNA Pol I and HIV reverse transcriptase; and the YXDD or LXDD motifs in reverse transcriptase and telomerase [28]. Nevertheless, these Mg^{2+} sequence motifs are occasionally too short to be statistically specific to Mg^{2+}-binding sites, and may escape detection easily. Unlike Zn^{2+} binding sites, Mg^{2+}-binding sites have less sequence similarity, but still offer sufficient structural similarity. Therefore, we used them to verify the capability of our approach to discover structural motifs with low sequence similarity.

Previously in [1], Dudev and Lim identified 4 first-shell structural motifs in Mg^{2+}-binding proteins. They are e(24-47)h(24)k; f(1)h(109-349)b; f(2)h(126-158)m; and k(26-29)h(1)a, where the number in parentheses indicates the number of residues separating the letters that correspond to the Mg^{2+}-bound residues. Table 3 shows the structural motifs detected to contain the Mg^{2+}-binding residues. Each of the simple motifs has an *E-value* than 3.5e−20, as defined by MEME and presented in Table 4. Unlike Dudev and Lim, who identified a structural motif based on the number of its occurrences, we applied a motif-finding tool to detect the position weight matrix motifs. We found that the position matrix motifs are more flexible and expressive than simple alphabet strings. Subsequently, by combining multiple motifs into a compound motif, we were able to detect less frequent, but highly important, structural motifs that could be missed by others, for example, m3-141-m8-19-m5 in 1wzc (see Table 3).

Thereafter, we replaced the MEME [7] with the Gibbs Sampler [8] in the system pipeline to demonstrate the feasibility of our modular design. To maintain consistency of the comparison, we tested the Gibbs Sampler on the C2H2 zinc finger proteins, and Fig. 8 shows the same protein structures that are displayed in Fig. 6, with the structural motifs numbered. The results show that both the MEME and Gibbs Sampler identified the locally conserved structures. Regardless, for the match between a local structure and a motif, the MEME's results were found to be more consistent. For example, MEME's motifs M_1 and M_2 matched the β-hairpin consistently, but the same β-hairpin was matched by the Gibbs Sampler's motifs M_1, M_3, and M_4, though with some variance. The variance addressed the difference in performance when different components were used in the proposed modular system design. Designers that adapt the modular design framework can substitute appropriate tools for the components they choose to be able to build their motif database systems in different application domains.

4. Future work and conclusion

We plan to continue this work in the following directions:

- First, numerous structural alphabets and motif detection algorithms have been developed, all based on different designs and

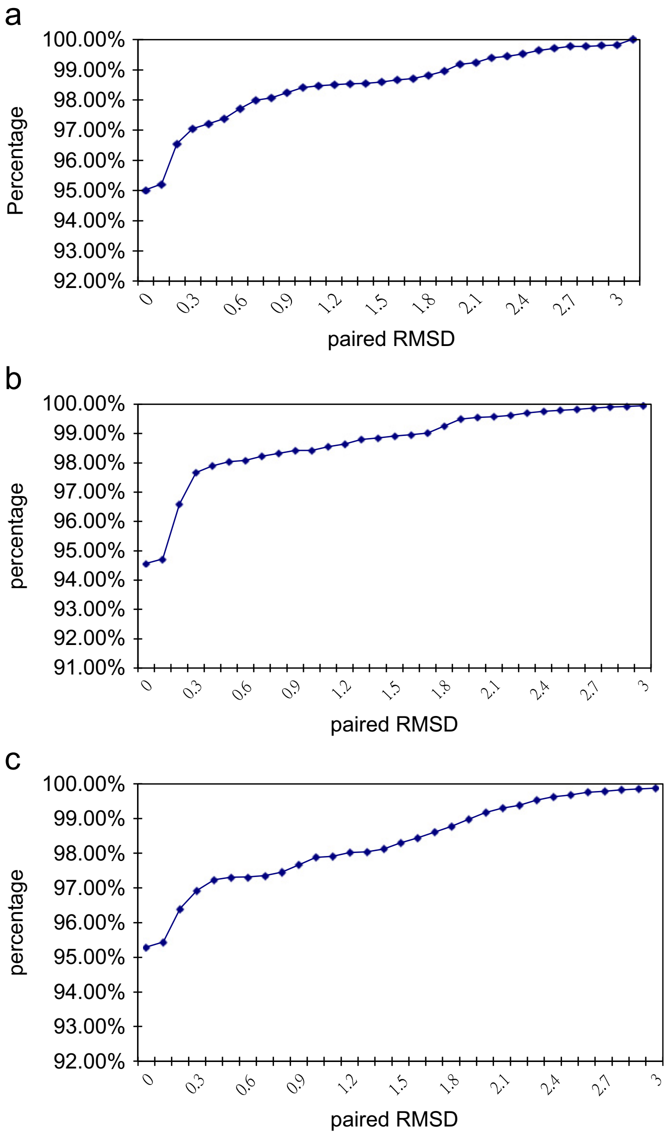


Fig. 7. Cumulative distributions of paired RMSD of motifs in C2H2 zinc finger protein structures presented in Fig. 6. (a) M_1=[GN][HE][NE]AC[AW]RQ, 95.03% of paired RMSD are zero, (b) M_2=[FH]CWNA[RC]QK, 94.57% of paired RMSD are zero, and (c) M_3=TTTTTT[PL][KPL], 95.29% of paired RMSD are zero. These suggested that the structural motifs characterized the locally conserved protein structures very well.

applied in various domains. With this work, we intend to incorporate other structural alphabets and motif-finding algorithms into our system. Therefore, we expect to discover a wider variety of motifs in the protein structures we analyze.

- Second, the alphabet letter preference and the motif conservation stored in the SA-Motifbase will allow us to investigate the evolutionary relationships within the proteins. The structural motifs can also be used as important structural features in the design of a protein function predictor. Based on the established motifs, the functions of a novel protein can be predicted by assigning it to an appropriate protein class with known functions.
- Third, numerous protein structures or function prediction systems have been designed in the past. We plan to evaluate the applicability of structural alphabet-based methods as a preprocessor because they have a substantially lower computational complexity when compared with systems relying on 3D information. Structural alphabet-based methods can constrain the search space efficiently by filtering irrelevant predictions before applying other, more computationally intensive, but also more accurate, tools.
- Lastly, to transform the proposed modular design framework for structural motif discovery into an adaptable and operational system capable of accepting various input formats and substituting different analysis tools, we plan to develop more advanced user interfaces, as well as improve the interfaces between the system components. If provided with a uniform interface, users will be able to build their own motif databases based on the same modular design framework.

This study introduced a general framework for structural motif discovery, and created a structural alphabet motif database. Two key components in our framework are as follows: (1) the structural alphabet used to describe protein structures and (2) the motif-finding algorithm used to discover significant local structural segments. The proposed method can replace these components with others to increase applicability in different domains. The experimental results showed that using structural alphabets combined with 1D motif-finding algorithms could enable successful identification of biologically meaningful subdomains in proteins. We identified the structural motifs conserved in the SCOP protein hierarchy, and constructed a structural motif database called SA-Motifbase. We demonstrated that in several protein classification groups, namely TIM barrel proteins, EGF-like proteins [24], and metalloproteins, the motifs we identified map to the known protein (sub-)domains or functional parts well. These results served to verify the relationships between protein functional parts and conserved structural segments represented as

Table 3
Summary of structural motifs containing Mg^{2+}-binding domains.

Compound Motif[a]	PDB	SCOP ID[b]	Binding residue position	Dudev and Lim's motif[c]	Functional description
m6-0-m1	1ig5	a.39	54-56-58-60	–	Calbindin d9k
m8-0-m1	1wdc	a.39.1	28-30-32-34-39	–	Scallop myosin
m8-0-m1	1yvh	a.39.1	229-231-235-240	–	cbl e3 ubiquitin protein ligase
m7-35-m13	1iq8	b.122	528-566-567-569	–	Archaeosine trna-guanine transglycosylase
m5-20-m8-38-m3	1vcl	b.42.2	177-178-218-219-265-266	–	Hemolytic lectin cel-iii
m2	1xxx	c.1.10	162-164-167	–	Dihydrodipicolinate synthase
m2-6-m1	1mdl	c.1.11	195-221-247	–	Mandelate racemase
m2-5-m1	1sjc	c.1.11	189-214-239	e-h-k	n-acylamino acid racemase
m4-4-m2-6-m1	2akz	c.1.11	244-292-317	e-h-k	Gamma enolase
m12-40-m13-1-m7	1mxg	c.1	24-80-88	–	Alpha amylase with carbohydrates
m8	1yq2	c.1.8	525-527-529	–	Beta-galactosidase
m3-137-m2	1o08	c.108.1	1008-1010-1170	f-h-b	Beta-phosphoglucomutase
m3-90-m2	1u7p	c.108.1	11-13-123	f-h-b	Magnesium-dependent phosphatase-1
m3-253-m4	1wpg	c.108.1	351-353-703	f-h-b	Sarcoplasmic/endoplasmic reticulum calcium atpase 1
m3-141-m8-19-m5	1wzc	c.108.1	8-10-169-204	–	Mannosyl-3-phosphoglycerate phosphatase
m3-105-m4	2b82	c.108.1	44-46-167	f-h-b	Class b acid phosphatase

Table 3 (*continued*)

Compound Motif[a]	**PDB**	**SCOP ID**[b]	**Binding residue position**	**Dudev and Lim's motif**[c]	**Functional description**
m3-169-m2	2c4n	c.108.1	9-11-201	f-h-b	Nagd
m8	1yl7	c.2.1	20-23-26	–	Dihydrodipicolinate reductase
m4-32-m14	1chn	c.23	13-57-59	–	Chemotaxis protein chey
m4-12-m2	1yio	c.23	12-55-57	–	Response regulatory protein
m4-32-m4	1zes	c.23	10-53-55	–	Phosphate regulon transcriptional regulatory protein phob
m4-2-m1-0-m5	1pox	c.36	447-474-476	k-h-a	Pyruvate oxidase
m4-5-m1	1umd	c.36	175-204-206	k-h-a	2-oxo acid dehydrogenase alpha subunit
m4-2-m1	1zpd	c.36	440-467-469	k-h-a	Pyruvate decarboxylase
m7-2-m1	2c3m	c.36.1	963-991-993	k-h-a	Pyruvate-ferredoxin oxidoreductase
m1-4-m4-20-m4	1dak	c.37.1.10	16-54-115	–	Dethiobiotin synthetase
m3-11-m15	1t0f	c.52	114-130-131	–	Transposon tn7 transposition protein
m1-40-m5-0-m19	2d0o	c.55.1	105-166-183	–	Thiamin pyrophosphokinase 1
m3-49-m4	1ido	c.62	142-144-209	–	Integrin
m1-109-m3	1pt6	c.62	152-154-253	–	Integrin alpha-1
m7-16-m1	1h1d	c.66	141-169-170	–	Catechol-o-methyltransferase
m8-83-m16	1jyl	c.68.1	107-216-218	–	ctp:phosphocholine cytidylyltransferase
m19-84-m13	1tw1	c.68	254-344-347	–	Beta-1-4-galactosyltransferase 1
m11-102-m9-129-m13	1ed9	c.76.1	51-155-322	–	Alkaline phosphatase
m7-94-m9-137-m13	1shq	c.76.1	37-151-310	-	Alkaline phosphatase
m7	1v71	c.79	208-212-214	–	Hypothetical protein c320.14 in chromosome iii
m2	3pmg	c.84	116-287-289-291	–	Alpha-D-glucose-1-6-bisphosphate
m1-39-m4	1tqy	c.95	307-308-355	–	Beta-ketoacyl synthase/acyl transferase
m1-32-m7	1khz	d.113.1	112-116-164	–	Adp-ribose pyrophosphatase
m1-27-m5	1ktg	d.113.1	52-56-103	–	Iadenosine tetraphosphate hydrolase
m1-65-m8	2bvc	d.128	135-219-227	–	Glutamine synthetase 1
m16-206-m9	2hgs	d.142	144-146-368	–	Glutathione synthetase
m17	1ofh	c.37	157-160-163	–	atp-dependent protease hslv
m7-1-m8	1hyo	d.177.1	733-734-753-756	–	Fumarylacetoacetate
m2-17-m1	2as8	d.3.1	56-57-59-91	–	Major mite fecal allergen der p 1
m9-38-m12	1wc1	d.58	1017-1018-1061	–	Adenylate cyclase
m14-25-m17	1iv2	d.79	8-10-42	–	2-c-methyl-d-erythritol 2-4-cyclodiphosphate synthase?
m15-67-m1	1t1s	d.81	149-151-230	–	1-deoxy-d-xylulose 5-phosphate reductoisomerase
m11-2-m13-4-m6	1ka2	d.92	269-273-299	–	m32 carboxypeptidase
m1-137-m9	1ka1	e.7.1	142-145-294	f-h-m	Halotolerance protein hal2
m1-148-m9	1nuy	e.7.1	1118-1121-1280	f-h-m	Fructose-1-6-bisphosphatase

[a] Compound motifs were composed of significant simple motifs, e.g. m24-161-m12 is a compound motif composed of simple motif m24 and m12, where 161 is the number of residues in between. The significance of a simple motif is determined by its *E-value*. The *E-value* of all simple motifs in table is less than or equal to 3.5e−20 (the smaller *E-value*, the more significant).

[b] The SCOP ID indicates where the motifs were identified from.

[c] We presented only the significant motifs determined by Dudev and Lim.

Table 4
Summary of significant motifs used in compound motifs.

Simple structural alphabet motif (regular expression)	**Motif index**	**SCOP ID**
LKGHNAR	m1	a.39
TTTSPLK	m6	a.39
KLKGHNARQ	m1	a.39.1
TTTTTTSPL	m8	a.39.1
[KT][GT]TSF	m13	b.122
[NE]EARQ	m7	b.122
N[AE][CE][WE][NE]EMRD[HF]RQSNN	m3	b.42.2
[NE]EEE[EA][ER][QM][RS][DN][FH]RQ[SP][NS]N	m5	b.42.2
[WQR][NQS][KMN][AFL][DF][HDW][NW][AF]CWNE[ME][RA][DR]	m8	b.42.2
xWFRQTTTTTTT	m12	c.1
x[EH][NE]MMDDHNEE[EA]	m13	c.1
[NS]FRQTTTTTTTT	m7	c.1
[TQ]TTTTT[LT]K[GS][FH]CWNEEARQ[PT][PT][STF][FTR]	m2	c.1.10
SFC[WN][NE]E[EA][AR][RQ][QL][SL][PK][CG][AF]RQTTTTTTLKG[SH]FCWNEEARQL[LK][GIK][HIG][TRH][QT]TTTT	m1	c.1.11
SP[PS][SN]N[NE]EE[EMR]M[MS][NDG][GN][SA][RFN][FQR][RTQ][TQ][TP]TTTTTTTT[LS][KN][GN][HA]CWN[AE][DA][CR]W[FW][RP]QPPQT	m2	c.1.11
[NQ][EK][EGA][EH][ENW]E[EA]E[AE]C[AC]R[QR][TQ]TTTTTTTTTTLKGHCWN[EN][EN][ME][ER][QE]M[SR][QF][PR][QT]T[TS]TT	m4	c.1.11
TTTxSFCWNE[EM][EA][ED]HxxxxNEE	m8	c.1.8
LKGHARQSNNEEACAR[QT]	m2	c.108.1
FC[WE][NE]EEACWGQS[NF][FR]RQx	m3	c.108.1
TTTTTTLKGH[NE]EEE[AE]xx	m4	c.108.1
[TP]TTTTTSFCWNE[EA]ARxN	m5	c.108.1
[NT]xEExxS[EN][FA]RQTTTTTT	m8	c.108.1
TTTT[LS][NK][NG][AH]CWNEA[RC]Q	m8	c.2.1
N[NE]ACWNA[RD]Q	m14	c.23
SFCWNEEEE	m2	c.23
[EN]EEEEACW[NF]	m4	c.23
LKGH[NE]EEEEEE[ME][RE][DQA][HRP][QP]	m1	c.36

Table 4 (*continued*)

Simple structural alphabet motif (regular expression)	**Motif index**	**SCOP ID**
[TS]PPSN[NE][EN]EEEEARQTT	m4	c.36
TTTTTTTS[FS][FP]C[WN]NE[EA][EA]	m5	c.36
TTTTTTLKGH[NE]EEEEEE[ME][RE][DQA]	m1	c.36.1
TTTSFRC[NC]N[EN]EEAC[QA]RQ[TP]TT	m7	c.36.1
SFCWNE[AE]E	m17	c.37
[MT][KT]G[TS][TN]TTT[TQ]TTTTTTTLKGHNEEEE[AE][CE]	m1	c.37.1.10
[AP][RP]Q[KF][RG][QTM][TK][TG]TTTTTTTTT[TP]SFCWNE[EA][EC][EA]	m4	c.37.1.10
[EN][EN]EEEEARQ[PST][NS]NxE	m15	c.52
EA[ML][QK][GD]HNEEE[EA][EC][EA][ER]	m3	c.52
EEEARQ[SN]N[NE]EEEE[EM]	m1	c.55.1
LKG[HS]R[CS][NW]NE[EM][EM]A[DR][QH]	m19	c.55.1
MD[DG]HN[EA][RE][QE]x[MN]N[DE]Dx	m5	c.55.1
SFR[LQ]KGHF[RA][WCD]NNEEEEEEA	m1	c.62
[NE]EEEEEEACAR[CQ]NFRQTTT	m3	c.62
[SN]N[NM]MRQ[TH][WS][NS][AF]RQTTTTTTT	m4	c.62
MAD[CNH][NW][EN]EMA[MD][RQ]LKG	m1	c.66
LKG[HA]C[WN]NE[ME][AM]D[QW]P[QT]	m7	c.66
SN[NC]E[NE][EN]EEEARQ[TP]TT	m13	c.68
FC[WC]N[AR]CWNEEEA[RC]Q[NT]	m19	c.68
[TP]SNN[EN]EEEEEE[MA][AR]D[HN]NEARQ[TR]	m16	c.68.1
TTTTT[TP][TP]T[LT]TT[SF][NC][NW][NE]EEEA[RC]W	m8	c.68.1
SNFC[WE]NACWNEEEEMAD[QH][RF]	m11	c.76.1
TTTT[TP]SFL[QK]G[HW][NE]EEEEARQ	m13	c.76.1
CWNEACWNEEEEMADQP[PT]S	m7	c.76.1
[WF][PR]CFRQTTTPSNNAC[AW]RQP	m9	c.76.1
TTTTTT[TL][TK][GT]PP[PT]SNNEE[AE][MEC]Q[GM]P	m7	c.79
TTT[LT][KT][GT][LM][KF][GD][HW][NE]E[EA][MAR][RCA][NQ]F	m2	c.84
LKGH[FA]RQSFCWNEEEAC	m1	c.95
TTTTTTTTTTTTTLKG[HS]	m4	c.95
MAD[HW]FRQTTTTTTTTTTLKGH	m1	d.113.1
[EP][ES][EN][EN]EEEE[AM]M[QD][GS]HNEEEEEE	m5	d.113.1
EEANAxRx[NS][NE][EN]EEEEARQTT	m7	d.113.1
LKGSF[CN]WNEEEEEEEEE[EA]EACWN	m1	d.128
LKGHNA[RC][QW][SN][EN][ENA][ER][EW]M[AM]D[WD][NH]NEE[EA]	m8	d.128
EEACWNMDDHDHN	m16	d.142
[TF]LKGHEEEEE[EA][EM]A	m9	d.142
[CE][WE][NER]EARQ[TS][PT][TP]T[PT]PSNNEE[EG]M	m7	d.177.1
[EP][LM][KN][GDE][HE]EE[AE][CE][WE]N[EC][AM]R[QC][TP]P[ST][PT][NS]	m8	d.177.1
TTTTTTTTTLKGHNEEE	m1	d.3.1
[TQP][TQ]TTTTTT[TL]LKGH[EN]ARQ	m2	d.3.1
[CS]WNEE	m12	d.58
[NE]MADH	m9	d.58
[NE]EEEEEEE	m14	d.79
[KH][GE][AH]RQTTT	m17	d.79
TTTTTTTTTLKGH	m1	d.81
[KC][GA][HR]QTTTTTT[TS]TT	m15	d.81
RQ[SN]FRQTTTTTTTTT	m11	d.92
TTTTTTTTTTTLKGH	m13	d.92
[NE]EARQTTTTTTT[TL][TK][TG]	m6	d.92
SFRWNNEEEAC[WM][KF]DQ[TP][TP]TLKGH	m1	e.7.1
[PT][PT]TTTTTTTTTLK[TG][SH]WNMR[DQ]SN	m9	e.7.1

*The *E-value* of all simple motifs in table is less than or equal to 3.5e – 20 (the smaller *E-value*, the more significant).

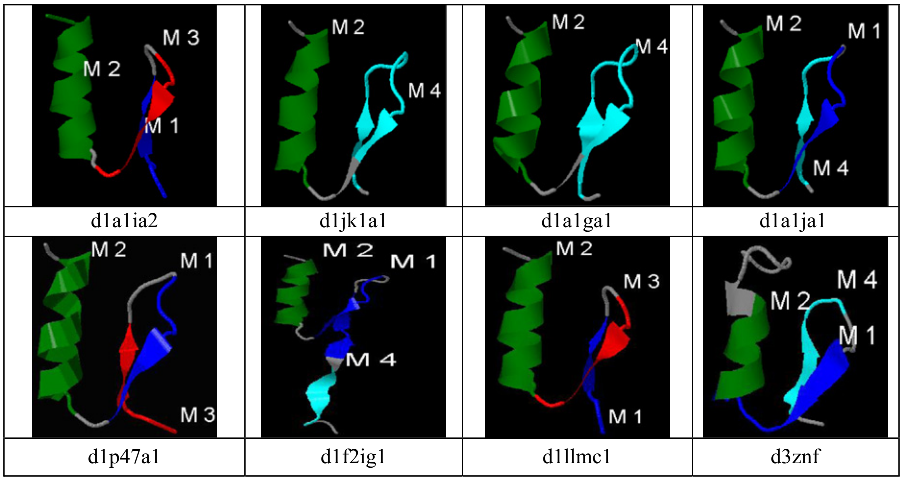


Fig. 8. Examples of C2H2 zinc finger protein structures for motif comparison. The simple motifs detected by Gibbs Sampler that map to the β-hairpin and the α-helix are numbered, where M_1=[GNDW][WF][CN][EW]NA, M_2=[AT][TR][TQ]T[TP][TL], M_3=KGH[NE]AC and M_4=[SNP][SNHW][NEW][NE][AE][CER]. The β-hairpin was matched by M_1, M_3 and M_4 inconsistently in different proteins. Compared with MEME, motifs found by Gibbs Sampler were less conserved than those identified by MEME.

structural alphabet motifs. They confirmed the capabilities of the framework for structural motif discovery and the database of structural alphabet motifs.

Availability

SA-Motifbase is available at http://bioinfo.cis.nctu.edu.tw/samotifbase/.

Conflict of interest statement

None declared.

Acknowledgments

This work was supported in part by the National Science Council, Taiwan (NSC 100-2221-E-009-146; 99-2218-E-009-002).

References

[1] M. Dudev, C. Lim, Discovering structural motifs using a structural alphabet: applications to magnesium-binding sites, BMC Bioinf. 8 (2007) 106.

[2] L. Ragad, J. Martin, G. Nuel, A.C. Camproux, Mining protein loops using a structural alphabet and statistical exceptionality, BMC Bioinf. 11 (2010) 75.

[3] R. Unger, D. Harel, S. Wherland, J.L. Sussman, A 3D building blocks approach to analyzing and predicting structure of proteins, Proteins 5 (1989) 355.

[4] S. Chakrabarti, R. Sowdhamini, Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modeling using distant relationships, FEBS Lett. 569 (2004) 31.

[5] G. Pugalenthi, P.N. Suganthan, R. Sowdhamini, S. Chakrabarti, SMotif: a server for structural motifs in proteins, Bioinformatics 23 (2007) 637.

[6] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536.

[7] T.L. Bailey, N. Williams, C. Misleh, W.W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, Nucl. Acids Res. 34 (2006) W369.

[8] C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, J. Wootton, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, Science 262 (1993) 208.

[9] J. van Helden, B. Andre, J. Collado-Vides, Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, J. Mol. Biol. 281 (1998) 827.

[10] F.P. Roth, J.D. Hughes, P.W. Estep, G.M. Church, Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation, Nat. Biotechnol. 16 (1998) 939.

[11] G. Pavesi, G. Mauri, G. Pesole, An algorithm for finding signals of unknown length in DNA sequences, Bioinformatics 17 (2001) S207.

[12] P. Pevzner, S. Sze, Combinatorial approaches to finding subtle signals in DNA sequences, in: Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology, San Diego, CA, 2000, p. 269.

[13] G.Z. Hertz, G.D. Stormo, Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, Bioinformatics 15 (1999) 563.

[14] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, G. Vriend, A database of protein structure families with common folding motifs, Protein Sci. 1 (1991) 1691.

[15] S. Chakrabarti, K. Venkatramanan, R. Sowdhamini, SMoS: a database of structural motifs of protein superfamilies, Protein Eng. 16 (2003) 791.

[16] G. Pugalenthi, P.N. Suganthan, R. Sowdhamini, S. Chakrabarti, MegaMotifBase: a database of structural motifs in protein families and superfamilies, Nucl. Acids Res. 36 (2008) D218.

[17] A.G. de Brevern, New assessment of a structural alphabet, In Sillico Biol. 5 (2005) 26.

[18] A.-C. Camproux, R. Gautier and P. Tuffery, A hidden Markov model derived structural alphabet for proteins, J. Mol. Biol., doi:10.1016/j.jmb. (2004).

[19] R. Unger, J.L. Sussman, The importance of short structural motifs in protein structure analysis, J. Comput. Aided Mol. Des. **7** (1993) 457.

[20] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg, P. Wrede, Local structural motifs of protein backbones are classified by self-organizing neural networks, Protein Eng. 9 (1996) 833.

[21] J.S. Fetrow, M.J. Palumbo, G. Berg, Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme, Proteins 27 (1997) 249.

[22] C. Bystroff, D. Baker, Prediction of local structure in proteins using a library of sequence-structure motif, J. Mol. Biol. 281 (1998) 565.

[23] B. Offmann, M. Tyagi, A.G. de Brevern, Local Protein Structures, Curr. Bioinf. 2 (2007) 165.

[24] S. Ku, Y. Hu, Protein structure search and local structure characterization, BMC Bioinf. 9 (2008) 349.

[25] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596.

[26] G. Jobb, A. von Haeseler, K. Strimmer, TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics, BMC Evol. Biol. 4 (2004) 18.

[27] N. Nagano, C.A. Orengo, J.M. Thornton, One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions, J. Mol. Biol. 321 (2002) 741.

[28] G.K. Farber, G.A. Petsko, The evolution of a/b barrel enzymes, Trends Biochem. Sci. 15 (1990) 228.

[29] C.I. Branden, The TIM barrel—the most frequently occurring folding motif in proteins, Curr. Opin. Struct. Biol. 1 (1991) 978.

[30] J.A. Cowan, Metal activation of enzymes in nucleic acid biochemistry, Chem. Rev. 98 (1998) 1067.

[31] J.A. Cowan, Biological Chemistry of Magnesium, VCH, New York, 1995.

[32] S. Bohm, D. Frishman, H.W. Mewes, Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins, Nucl. Acids Res. 25 (1997) 2464.

[33] J.H. Laity, B.M. Lee, P.E. Wright, Zinc finger proteins: new insights into structural and functional diversity, Curr. Opin. Struct. Biol. 11 (2001) 39.

[34] S. Iuchi, Three classes of C2H2 zinc finger proteins, Cell. Mol. Life Sci. 58 (2001) 625.

[35] N.V. Grishin, Treble clef finger—a functionally diverse zincbinding structural motif, Nucl. Acids Res 29 (2001) 1703.

[36] B. Wang, D.N. Jones, B.P. Kaine, M.A. Weiss, High-resolution structure of an archaeal zinc ribbon defines a general architectural motif in eukaryotic RNA polymerases, Structure 6 (1998) 555.