Constrained Multiple Structure Feature Alignment (CMSFA)*

限制性之多重結構特徵排比

蘇柏翰、周偉堯、白敦文
國立台灣海洋大學資訊工程學系
twp@mail.ntou.edu.tw

張大慈、張顥騰、周維宜
國立清華大學生命科學系
lscmdt@life.nthu.edu.tw

Abstract

With the rapid accumulation of released three-dimensional protein structure database, the importance of structural comparison parallels that of sequence alignment. It has been shown that despite primary sequence diversity, protein structures of related sequences possess a structural core of α-helices and β-sheets and vary in the loop regions. To determine the characteristic properties for each target sequence from a protein family, we have developed a fast algorithm for structure alignment based on the combination of primary sequences and three-dimensional structures. The sequence-based comparison utilizes the labeled consensus motifs to provide combinatorial features for multiple sequence alignment, and the spatial positions of the key amino acids in each of the combinational segments are assigned for the proposed constrained multiple structure feature alignment (CMSFA). The 3D co-ordinates of aligned amino acids provide data for calculating the root-mean-square deviation (RMSD) values which build the references for the detection of structurally distinct regions. In this study, RNase A P450, and ricin A protein families were employed to demonstrate the outstanding performance of the structure alignment algorithms, and the comparisons between our proposed CMSFA and several existing structural alignment tools are also described in this paper.

Keywords: sequence-structural alignment; Combinatorial features; structural features

隨著蛋白質結構資料庫的快速成長，三維蛋白質結構的比較已經逐漸的比傳統序列排比更為重要。儘管有些序列具備高度的差異性，但它們在結構上仍可以持有相似的 alpha 螺旋和 beta 平板結構，卻僅有在環狀區域上有一些變化。對於來自同一蛋白質家族中的每一個目標序列，為了能夠決定其獨特之性質，我們發展出一套快速的結構排比演算法，同時採用序列與結構之特徵做為排比之基礎。在序列分析上，先針對已加上標籤的共同序列片段，利用多重序列排比找出組合式之特徵；然後再擷取出組合式特徵上具重要化學性質的胺基酸，並設定為多重結構排比的關鍵點，再依關鍵點之位置進行快速排比。經排比後，藉由均方根差值的計算，可以用來偵測幾何結構上獨立的區域。本論文中，是以 Rnase A, P450,和 ricin A 蛋白質家族為例，可顯示出結構排比的顯著效果，另外提供與其它現存之結構排比工具之比較結果分析。

關鍵字：序列/結構排比、組合特徵、結構特徵

1 Introduction

The analysis of tertiary structure of proteins provides precious information on their biological functions. Therefore, development of an efficient and accurate bioinformatics tool for protein comparison becomes an important research topic. Currently, if the three-dimensional structure of a target protein sequence is not resolved, the homology modeling methodology is considered as one of the most reliable structure prediction methods. It is applicable, when at least one of homologous structures of the target protein is resolved, to predict 3D coordinates by aligning the target sequence and the template structure. It can be observed that the accuracy of homology model strongly depends on the precision of alignment between target and template. Generally speaking, the accuracy of conventional alignment algorithms declines sharply when the sequence identity between the target and the template proteins is lower than 45% [1]. On the other hand, if the three-dimensional structure of a target sequence is already known, then users would like to perform structure comparison in order to predict structure-function relationship from its related family sequences. Previous prediction methods focused on distinguishing possible candidates by examining the

* This work is supported by NSC-93-3112-B-007-005 and VGHUST94-G2-02-1.

presence of the appropriate primary/secondary anchor residues [2] [3]. Lately, the strategies give emphasis on pattern matching technologies including statistical, machine learning methodologies and structural information [4][5][6][7]. However, pure sequence based methods have inherent statistical limits, and the use of structural information have been shown to increase both the sensitivity in detection and accuracy in alignment. The Protein Data Bank (PDB) currently holds more than 32,400. Since the ultimate goal is to unveil the function of all proteins, it is obvious that 3D structure comparison becomes a significant task and it may reveal biologically interesting similarities which are not detectable by direct sequence alignment. Several protein structure comparison tools and many of the methodologies focus on the superposition of protein structures to alignment results. Structural comparison results lead to understanding of the evolutionary relationships and physico-chemical interactions among protein sequences. At this moment, only a few public web services are available to perform multiple protein structure alignment, such as MultiProt [8] (http://bioinfo3d.cs.tau.ac.il/ MultiProt/) which finds the common geometrical cores between the input molecules and detects high scoring partial multiple alignments for all possible number of molecules from the input; CE [9] (http://cl.sdsc.edu/ce.html) which employs the combinatorial method on aligned fragment pairs of a given length; JOY[10] (http://www-cryst.bioc.cam.ac.uk/joy/) which displays 3D structural information in a sequence alignment and helps identification of the conservation of amino acids in their specific local environments; COMPARER[11] (http://www-cryst.bioc.cam.ac.uk/COMPARER/) utilizes the DiCE structural alignment program to superimpose selected structures and provide output file in the JOY format.

In general, after multiple structures have been superimposed by three-dimensional rigid body rotation, a common measurement of structural similarity is evaluated by the RMSD between the positions of the corresponding amino acid pairs on the aligned protein structures. However, it has been argued that the RMSD measurements are ambiguous with respect to distantly related proteins [12]. One of the reasons is that portions of mismatched substructure tend to dominate the RMSD values for remotely related sequences. To avoid such a dilemma, we select the RMSD as a measuring parameter of the proposed system, and at least 30% sequence identity among protein sequences was set as the basic requirement prior to our CMSFA. This assumption is reasonable when our main goals are to perform structure alignment for a set of family protein sequences based on their sequence and tertiary structure information.

Here we present a new method that applies multiple combinatorial features for multiple structure alignments. This combinatorial features enhances both sensitivity of sequence search and quality of structural alignment. In our approach, combinatorial features of related family sequences are aligned by applying dynamic programming on labeled local consensus motifs which are searched by interval jumping approximate searching algorithms (LIJSA) [13]. These combinatorial sub-segments represent common characteristics of a protein family and positions of the corresponding key amino acids are selected for efficient and effective three-dimensional multiple structure alignment processes. Therefore, the aligned structures provide prompt identification of residues comprising substructures or surface regions that are conserved with respect to the target protein. More details of the proposed algorithms are introduced in the following sections.

2 Materials and Methods

2.1 *Problem Definitions*

The protein sequences retrieved from the C-alpha atom in the PDB files are represented as strings over the 20 amino acid set. Each residue is assigned its own three-dimensional rectangular coordinates. Let W be the set of input protein sequences in this paper. The i^{th} protein sequence in W will be denoted by W_i, and the total number will be indicated by $N=|W|$. More specifically, the W set is constructed as $W=\{W_1,W_2,\ldots,W_{N-1},W_N\}$. In this paper, a target protein is defined by $W_t \in W$ and $W_i(j)$ means the j^{th} residue in W_i. $\hat{X}(j), \hat{Y}(j)$, and $\hat{Z}(j)$ stands for the orthogonal coordinates of the j^{th} residue X, Y and Z in the unit of Angstroms. Based on the properties of hydrophobicity, hydrophilicity, and charge, this paper defines the set of amino acids with charge as follows, $CH[AA]=1, AA\in\{D,E,H,K,R\}$; otherwise $CH[AA]=0$, and AA represents one of the 20 amino acids. For example, $CH[R]=1, CH[G]=0$. As for the hydrophilic characteristic, it is specified as $HY[AA]=1, AA\in\{\mathrm{D,E,H,K,R,G,C,Y,N,Q,S,T}\}$; otherwise $HY[AA]=0$ and for instance, $HY[D]=1, HY[A]=0$. Furthermore, the paper defines the homology characteristics of the amino acids based on the aligned sequence similarities in W and indicated by $HO[\cdot]$, i.e. $HO[AA]=1$ if AA belongs to the homology set. According to the above formulated properties, the j^{th} residue in $W_i(l,k)$ can be assigned with a score, $C_p(j)$, that stands for the degree of significance of chemical properties. In the later section, the program groups the amino acids and specifies the key residues which hold the highest $C_p(j)$ from the specified set, $KR\{\cdot\} = \{W(j) \in W_i(l,k), l \le j \le k\}$. Besides, in order to evaluate degree between two aligned proteins in this research, the measurement utilizes the RMSD values. If one subgroup sequences, $\{W_1,...W_M\}$, is aligned to W_t, the RMSD value of each residue in W_t is represented as $R_k(j), 1 \le k \le M, 0 \le j \le |W_t|, k \ne t$. Finally, a general threshold function, $F_a(x) = \begin{cases} 1, x \ge a \\ 0, x < a \end{cases}$, is applied, where *x* is the variable of RMSD values or the number of identical residues in this proposed system, and *a* is the thresholding values with respect to *x*.

2.2 *System Description*

Figure 1 depicts the system configuration. The system requires importing protein sequences of a family in PDB format. There are two main phases in the proposed Constrained Multiple Structure Feature Alignment. The first phase focuses on sequence analysis which provides both clustering and combinatorial feature extraction operations.

The consensus motifs among sequences are searched prior to hierarchical clustering operations. If the sequences under analysis comprise the near neighboring proteins in addition to target protein family, the system will suggest to perform clustering operations to divide the near neighboring proteins into several subgroups for better performance in terms of combinatorial feature analysis. On the other hand, the performance of extraction of combinatorial features will be obtained with better results if the imported sequences are clustered with higher similarity in each subgroup. Once the imported sequences are clustered, the combinatorial features of each subgroup are aligned employing traditional Dynamic Programming techniques. In the next step, the key residue analysis, constrained multiple structure feature alignment (CMSFA) and related biological applications are categorized in the second phase. The key residue will be retrieved based on the characteristics of homologous, charged, and hydrophilic degree from the aligned consensus segments. Afterwards, all protein structures will be superimposed together rapidly by the geometry centers of those key residues. By means of the RMSD values between the target protein and the others, related biological applications can be performed. For example, the unique peptide motifs are acknowledged as one of the greatest interests to define sequences that antibodies may recognize with high degree of uniqueness. The above system will be described in the following section in detail.

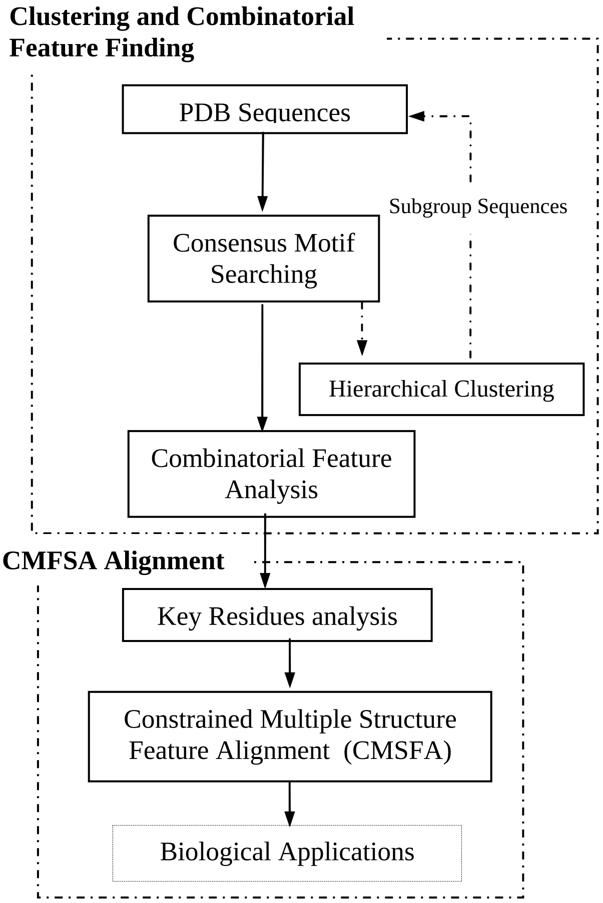
Clustering and Combinatorial Feature Finding
PDB Sequences
Subgroup Sequences
Consensus Motif Searching
Hierarchical Clustering
Combinatorial Feature Analysis
CMFSA Alignment
Key Residues analysis
Constrained Multiple Structure Feature Alignment (CMSFA)
Biological Applications

Figure 1. System Configuration

2.3 *Clustering and Combinatorial Feature Finding*

In the first phase, there are three modules which include consensus motif searching, hierarchical clustering, and combinatorial feature analysis. The first module searches consensus motifs by Ladderlike Interval Jumping Searching Algorithms (LIJSA) [13]. It is an efficient algorithm for matching variable-length and tolerant strings with linear time complexity. Users are able to determine whether clustering functions should be performed or not. If the input protein sequences are known for the homologous family that are expected to hold high structure similarity in advance, users can ignore the clustering processes and continue to execute the combinatorial feature analysis. On the other hand, when the input protein sequences comprise related neighboring proteins barring the target protein family, users are suggested to execute the clustering operations prior to combinatorial feature extraction procedures. The clustering algorithms utilize the searched consensus segments from the previous module and their respective clustering scoring matrices are calculated for grouping procedures. The agglomerative clustering algorithms are employed to cluster sequences into several subgroups, and our system takes the simple linkage, a kind of hierarchical measurement to determine which sequences should be grouped together. After the clustering operations for all of the imported protein sequences, each clustered subgroup is then individually performing consensus motif searching operations with target protein sequence followed by the combinatorial feature analysis. The combinatorial feature analysis module performs indexed multiple sequence alignment based on Dynamic Programming (DP) algorithms. The fundamental elements in DP algorithms are labeled consensus motifs instead of individual residues. The output results from this module provide combinatorial features sequentially for each subgroup family, and those features are composed of merged local consensus motifs and will enhance the important characteristics of each subgroup. Two examples of RNase A and P450 protein families are shown in Figure 2(A) and 2(B), and their combinatorial features are represented in large and uppercase amino acids.

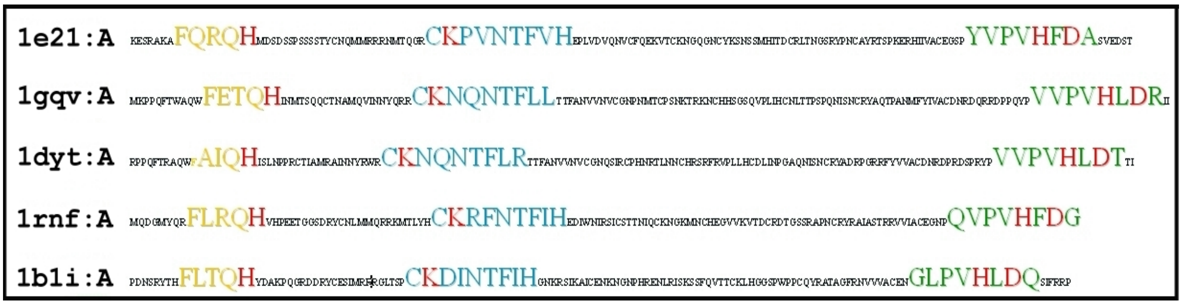

Figure 2(A) . Sequential combinatorial features of human RNase A superfamily

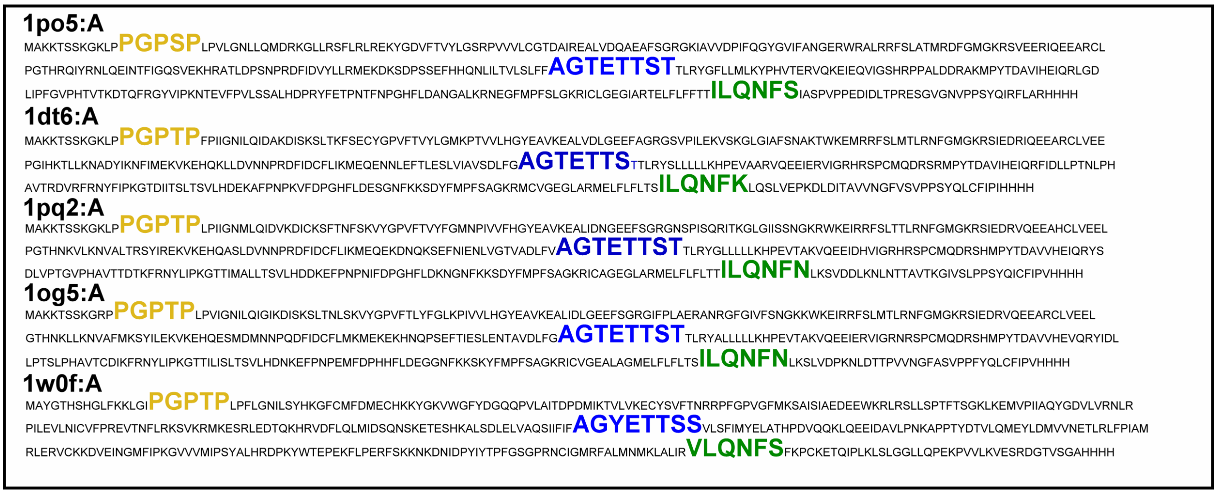

Figure2(B). Sequential combinatorial features of human P450 superfamily.

2.4 *CMSFA Alignment*

The modules in the second phase include key residue analysis, constrained 3D feature alignment, and related biological applications. According to the combinatorial features, the module of key residue analysis evaluates priority score, $c_p(j)$, of each residue for further identification. The priority score is determined by protein properties including homology ($HO[\cdot]$), charge ($CH[\cdot]$) and hydrophilicity ($HY[\cdot]$). In this procedure, identical amino acids from aligned segments are referred as candidates of key amino acids and assigned into the homologous set $HO[\cdot]$. If an amino acid in $HO[\cdot]$ is charged and possesses hydrophilicity, it is assigned with the highest score of 3. Assume the amino acid possessing charged feature only, it obtains a score of 2. The amino acid will be assigned a score of 1 when it holds hydrophilic feature only. All other amino acids without charged and hydrophilic features will not be assigned with any score by the system. Consequently, if $W_i(j)$ belongs to one consensus segment, the priority score represents its functional properties in the protein, and is formulated as $C_p(j) = HO[W_i(j)] + CH[W_i(j)] + HY[W_i(j)]$. For proteins possessing enzyme activities, the system will regard the set of residues, $KR\{\cdot\}$ possessing the highest scores in each combinatorial feature segment, as the potential key residues for further constrained multiple structure feature alignment.

Afterwards, the geometric centers of the selected key sites in each aligned consensus motif are calculated as $(\frac{\sum_{W_i(j)\in KR\{\cdot\}} \hat{X}(j)}{|KR\{\cdot\}|}, \frac{\sum_{W_i(j)\in KR\{\cdot\}} \hat{Y}(j)}{|KR\{\cdot\}|}, \frac{\sum_{W_i(j)\in KR\{\cdot\}} \hat{Z}(j)}{|KR\{\cdot\}|})$ in each subgroup sequence, and these centers are utilized to perform constrained multiple structure feature alignment. With these centers, the module will randomly choose three candidates for multiple alignments, since three spatial positions can determine a surface plane and then confirm the orientation of each structure. Based on the structure alignment, all other proteins in each subgroup family will be aligned rapidly with their fixed plane in 3D space constructed from the selected points.

3. Results

3.1 *Clustering and Sequential Combinatorial Features*

In this paper, we use different sets of structural sequences to emphasize the important features of each system module. Some of them have been enumerated in other published papers and annotated with their references. One of the major reasons we reuse those data is to compare the performance between our proposed system and others. To demonstrate the performance of clustering module and combinatorial feature for structure alignment, we use a complete list of structure all related to the reference ricin A chain, there are 31 structures[7] from PDB that possess a certain similarity to the target structure 1br6 [14]. These sequences are suggested to perform clustering operations and resulting in 6 subgroups as following. Group1: 1lp8:A, 1qi7:A, 1rl0:A; Group2: 1apa, 1gik:A, 1oql:A 1qci:A; Group3: 1abr:A, 1ce7:A, 1ggp:A, 1m2t:A, 1onk:A, 1pum:A, 1sz6:A, 1tfm:A, 2mll_A; Group4: 1ahc, 1bry:Y, 1cf5:A, 1d8v:A, 1hwn:A, 1j4g:A, 1mom, 1mrg, 1mrh, 1mrj, 1nio_A; Group5: 1dm0:A, 1r4p:A, 1r4q:L; and Group6: 1lln:A. Each of the subgroup is structurally aligned with the target structure and their RMSD values will be shown in the later section.

To describe the combinatorial features of a family sequence, we select the human ribonuclease A (RNase A) and P450 superfamily as examples. The structural information of 5 human RNase sequences and P450 sequences can be extracted from PDB. Their sequence identity and similarity percentages are listed in Table1(A) and (B) for reference. From the Table1(A), the RNase A superfamily currently contain 5 different structures with high similarity, in which RNase2 (1gqv:A) and RNase3 (1dyt:A) share 65.67% identity and 82.09% similarity. To distinguish the characteristics of each RNase, the combinatorial features are extracted rapidly to align their structures. In Figure 2(A), the sequential combinatorial feature segments are highlighted in various colors and amino acids in red colors show the key residues for 3D alignment. Interestingly, the first three amino acids labeled in red, H, K, and H, matched perfectly with the key catalytic residues in the active site. Similarly, Table1(B) describes the identity and similarity among member of human P450 protein family, and its sequential combinatorial feature segments are highlighted in various colors and shown in Figure 2(B).

Table 1(A). Identity/Similarity percentages (%) among members of RNase A superfamily.
(B) Identity/Similarity percentages (%) among members of Human P450 family.

Portein	1e21:A	1gqv:A	1dyt:A	1rnf:A	1b1i:A
1e21:A	-	33.08/56.15	32.09/56.72	45.61/75.44	36.80/65.60
1gqv:A	33.08/56.15	-	65.67/82.09	28.13/50.00	28.68/45.74
1dyt:A	32.09/56.72	65.67/82.09	-	28.70/56.48	29.41/47.06
1rnf:A	45.61/75.44	28.13/50.00	28.70/56.48	-	40.35/65.79
1b1i:A	36.80/65.60	28.68/45.74	29.41/47.06	40.35/65.79	-

(A)

Portein	1po5:A	1dt6:A	1pq2:A	1og5:A	1w0f:A
1po5:A	-	53.15/81.72	55.25/80.67	52.10/80.46	28.60/56.21
1dt6:A	53.15/81.72	-	74.37/92.02	77.52/93.70	26.84/56.26
1pq2:A	55.25/80.67	74.37/92.02	-	79.00/93.49	28.43/56.86
1og5:A	52.10/80.46	77.52/93.70	79.00/93.49	-	28.70/57.18
1w0f:A	28.60/56.21	26.84/56.26	28.43/56.86	28.70/57.18	-

(B)

3.2 Quality of Constrained 3D Multiple Structure Feature Alignment

The proposed CMSFA performs efficient and effective structure matching when the combinatorial features are available. The combinatorial features can exist due to sequences possessing similarity at a certain level. In facts, the fundamental consensus segments of combinatorial features hold tolerant characteristics in our system which guarantees the realization of combinatorial feature extraction if basic requirements of more than 30% identity are satisfied. However, if the imported sequences indeed possess diversely distributed residues, the combinatorial features may not exist and therefore CMSFA can not provide appropriate solutions. So far, under wide range of testing cases, our proposed algorithms provide superior performance in terms of accuracy and efficiency. Here, we compare the performance of CMSFA with that of well-known structural alignment systems such as DALI [15], CE [8], LGA [16], and FAST [17]. The test cases were performed on the following pairs of known protein structures [16][17]: (1df4:A, 1qce:A), (1hx8:A, 1hg5:A), (1oyc:_, 2tmd:A), (1af6:A, 1a0t:_), and (2sim:_, 1nsb:A). [Four character PDB codes followed by a colon and the chain identifier identify the protein polypeptide chains whereas proteins with unassigned polypeptide chains are symbolized by an underscore (_).] The comparison results are shown in Table 2. All the RMSD values for our CMSFA are less than or equal to those of the existing programs and the number of matched residues successfully identified are greater than or equal to those of the best conditions of other algorithms. In multiple alignment circumstances, we took human ribonuclease A (RNaseA) superfamily, P450, and ricin A chain as examples to compare with the public COMPARER system. In Figure 3 (A), the original 3D structures of five RNaseA protein sequences are revealed whereas the aligned results of 3D structures calculated by CMSFA and COMPARER are shown in Figure 3 (B) and 3 (C) respectively. In Figure 4 (A), the original 3D structures of five P450 protein sequences are revealed whereas the aligned results of 3D structures calculated by CMSFA are shown in Figure 4(B). Similarly, the system performs clustering operation prior to structure alignment, and here shown the original and aligned structures of the first group of the ricin A chain family in Figure 5(A) and (B). To display the precision of aligned results, the number of alignment residues and corresponding distance measurements are calculated and shown. Here we take RNase A superfamily as an example. In Table 3 and Table 4, the average RMSD values, matched residues, and standard deviations for each pair of sequences of RNase A superfamily are displayed. Each column represents different target sequences and each row denotes the aligned results with respect to the other members of the RNase A superfamily. In the last, ricin A chain related sequences was used as another example. As mentioned above, 31 ricin related sequences were clustered into 6 subgroups, and the target sequence was aligned by CMSFA with respect to each subgroup. The compared results with each subgroup are shown in Table 5 which contains the information including: M/N (the numbers of matched residues out of the total residues of target sequence): RMSD (average RMSD values), S.D. (Standard deviations of RMSD), and Similarity (Sequence identity and similarity percentages). In this example, there are three protein sequences clustered in the first group, and performed by CMSFA with a ricin A target sequnce. The number of matched residues ranges from 243 to 246, if selected the 1br6:A sequences as the target sequence. The average RMSD values, standard deviation, and pairwise sequence similarity range from 2.24 to 2.49, 1.08 to 1.11, and 59.29 to 61.33 respectively. The aligned structure of the group1 is shown in Figure 5.

Table 2. Comprasion of structure alignments for 5 pairs of proteins. Protein1 is fixed as the target structure, and the protein2 is allowed to rotate. Each aligned protein pair is represented by N/RMSD, where N is the matched numbers of equivalent residues and RMSD represents average values of RMSD for matched residues. The last column provides the sequence similarity derived from FASTA.

Portein1	Protein2	DALI	FAST	CE	LGA	**CMSFA**	Similarity
1df4:A	1qce:A	57/1.5	57/1.2	57/1.6	57/0.9	57/1.0	52.83%
1hx8:A	1hg5:A	258/1.1	255/1.1	249/0.8	256/1.0	263/1.0	89.23%
1oyc:_	2tmd:A	323/2.6	284/2.3	354/3.0	324/2.1	354/2.9	52.14%
1af6:A	1a0t:_	367/2.5	323/1.8	355/1.9	344/2.3	378/2.5	55.03%
2sim:_	1nsb_A	289/3.2	236/3.0	275/3.0	269/2.6	289/3.0	54.55%

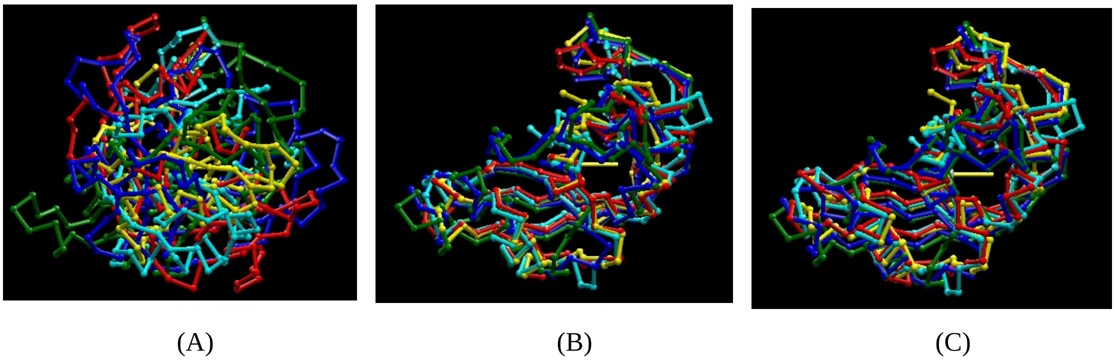

Figure 3. Five human ribonuclease A (RNaseA) superfamily proteins are depicted by CMSFA system. (A) The original RNaseA protein structures are displayed in different colors. (B) and (C) show the proteins aligned by the CMSFA and COMPARER system respectively. The five proteins (1e21:A, 1gqv:A, 1dyt:A, 1rnf,:A ,.and 1b1i:A) are individually displayed in red, green, blue, yellow, and cyan.

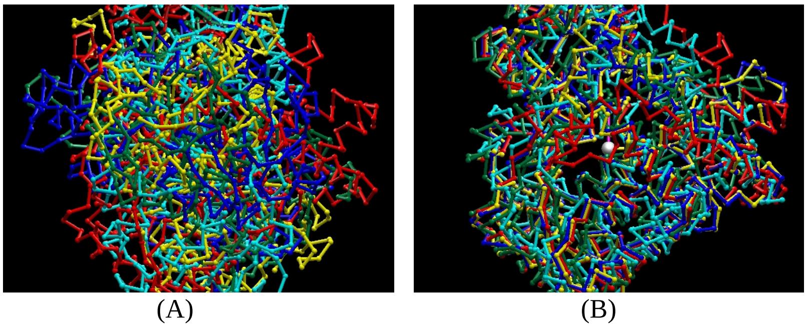

Figure 4. Five human P450 superfamily proteins are depicted by CMSFA system. (A) The original P450 protein structures are displayed in different colors. (B) The five proteins (1po5:A, 1dt6:A, 1pq2:A, 1og5,:A , and 1w0f:A) are aligned by the CMSFA system and displayed in red, green, blue, yellow, and cyan respectively.

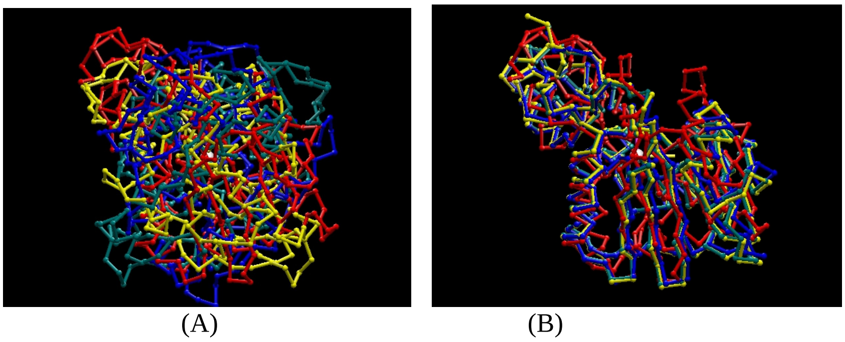

Figure 5. The combination of target sequence (1br6:A) and the first grouped sequence of ricin A related proteins (1lp8:A, 1qi7:A, and 1rl0:A) are depicted by CMSFA system. (A) The original Group 1 ricin A chain protein structures are displayed in different colors. (B) The five proteins of Group 1 are aligned by the CMSFA system and displayed in red, blue, green, and yellow respectively.

Table 3(A). Data are represented in K(M/N) format, where K is average RMSDs between two aligned sequences (unit in Å), M represents number of align ed residues, and N denotes the number of residues in target protein (Rnase A).

		Target proteins				
	RNase A	1e21:A	1gqv:A	1dyt:A	1rnf:A	1b1i:A
Compared proteins	1e21:A	-	1.96/(115/135)	2.07/(117/133)	1.58/(116/120	1.78/(113/123)
	1gqv:A	1.87/(113/128)	-	1.54/(133/133)	1.83/(117/120)	2.11/(113/123)
	1dyt:A	1.97/(115/128)	1.44/(134/135)	-	2.02/(120/120)	2.14/(116/123)
	1rnf:A	1.61/(117/128)	1.98/(121/135)	2.05/(122/133)	-	1.76/(117/123)
	1b1i:A	1.69/(113/128)	2.32/(122/135)	2.22/(123/133)	1.58/(116/120)	-

Table 3(B). Data are represented in K(M/N) format, where K is average RMSDs between two aligned sequences (unit in Å), M represents number of align ed residues, and N denotes the number of residues in target protein (P450).

		Target proteins				
	P450	1po5:A	1dt6:A	1pq2:A	1og5:A	1w0f:A
Compared proteins	1po5:A	-	2.84 /(387/473)	1.73/(418/476)	1.64/(411/475)	2.61/(411/485)
	1dt6:A	2.79/(381/476)	-	2.74/(427/476)	2.73/(429/475)	2.96/(388/485)
	1pq2:A	1.77/(419/476)	2.76/(427/473)	-	1.05/(461/475)	2.41/(432/485)
	1og5:A	1.64/(409/476)	2.74/(429/473)	1.07/(463/476)	-	2.45/(428/485)
	1w0f:A	2.64/(411/476)	3.01/(403/473)	2.38/(430/476)	2.51/(433/475)	-

Table 4(A). Standard deviations values of RMSD between two Rnase A proteins (unit in Å).

		Target proteins				
	RNase A	1e21:A	1gqv:A	1dyt:A	1rnf:A	1b1i:A
Compared proteins	1e21:A	-	1.15	1.11	1.58	1.12
	1gqv:A	1.09	-	0.86	1.07	1.20
	1dyt:A	1.02	0.72	-	1.06	1.18
	1rnf:A	1.06	1.19	1.08	-	1.14
	1b1i:A	1.03	1.33	1.19	1.96	-

Table 4(B). Standard deviations values of RMSD between two P450 proteins (unit in Å).

		Target proteins				
	P450	1po5:A	1dt6:A	1pq2:A	1og5:A	1w0f:A
Compared proteins	1po5:A	-	1.06	0.96	1.03	1.00
	1dt6:A	1.03	-	0.98	1.03	1.03
	1pq2:A	1.00	1.00	-	0.62	0.90
	1og5:A	1.03	1.03	0.65	-	0.90
	1w0f:A	1.03	1.06	0.88	0.95	-

Table 5. CMSFA analysis of ricin A related family.

	Group1	Group2	Group3	Group4	Group5	Group6
M/N	[243,246]/267	[196,258]/267	[174,211]/267	[218,250]/267	[238,239]/267	139/267
RMSD(Å)	[2.24,2.49]	[1.50,3.13]	[3.07,3.20]	[2.47,3.03]	[2.45,2.60]	3.06
S.D.(Å)	[1.08,1.11]	[0.88,1.09]	[1.07,1.29]	[0.93,1.07]	[1.12,1.22]	1.1
Smilarity	[59.29,61.33]	[63.98,67.87]	[65.04,67.84]	[66.11,70.19]	[52.94,55.20]	56.48

4. Conclusion

Combinatorial feature analysis of protein family provides important characteristics from sequence alignment. Key residues in combinatorial feature segments can be selected by their chemical properties and provide significant information for performing constrained multiple structure feature alignment. Although the quality of our alignment method could be limited by the degree of sequence similarity, the system involves hierarchical clustering algorithms to enhance their similarity relationships. For the ricin A protein (1br6), related proteins are suggested to cluster into six groups to be aligned with target sequence seperately. Based on the clustering analysis, we can successfully perform the structure alignment as other programs. For the case of human RnaseA and P450 protein families, our approaches also correctly explores key residues information. From these results, our proposed system is shown to be able to yield a fine alignment with their combinatorial features.

5. References

[1] Jaroszewski, L., Li, W. and Godzik, A., In search for more accurate alignments in the twilight zone, *Protein Sci., 11:1702-1713, 2002.*

[2] Collins, E.J., D.N. Garboczi, and D.C. Wiley, Three-dimensional structure of a peptide extending from one end of a class I MHC binding site. *Nature*, 371(6498): p. 626-9, 1994

[3] Ruppert, J., J. Sidney, E. Celis, R.T. Kubo, H.M. Grey, and A. Sette, Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell*, 74(5): p. 929-37, 1993

[4] Chang, H.T., T.C. Fan, M.D.T. Chang, T.W. Pai, B.H. Su, and P.C. Wu, Unique peptide prediction of RNase family sequences based on reinforced merging algorithms. *Proceedings of 3rd Asia-Pacific Bioinformatics Conference*, 289-298, ISBN 1-86094-477-9, 2005

[5] Donnes, P. and A. Elofsson, Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3(1): p. 25, 2002

[6] Gulukota, K., J. Sidney, A. Sette, and C. DeLisi, Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol*, 267(5): p. 1258-67, 1997

[7] Zhou, C.L.E., A.T. Zemla, D. Roe, M. Young, M. Lam, J.S. Schoeniger, and R. Balhorn, Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin. *Bioinformatics*, 21(14): p. 3089-96, 2005

[8] Diamond R., On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Sci*., 1:p. 1279–87, 1992

[9] Guda C., S. Lu, E.D. Scheeff, P.E. Bourne, and I.N. Shindyalov, CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res*, 32(Web Server issue):W100-3, 2004

[10] Mizuguchi, K., C.M. Deane, T.L. Blundell, M.S. Johnson and, J.P. Overington, JOY: protein sequence-structure representation and analysis. *Bioinformatics*, 14:p. 617-23, 1998

[11] Sali,A. and T.L. Blundell, Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, 212:p. 403-428, 1990

[12] Carugo, O. and S. Pongor, A normalized root mean square distance for comparing protein three dimensional structures. *Protein Sci*, 10: p.1470-73, 2001

[13] Pai, T.W., M.D.T. Chang, J.H. Chu, and H.L. Tai, Ladderlike Stepping and Interval Jumping Searching Algorithms for DNA Sequences. *APBC*, p.93-98, 2004

[14] Yan, X., T. Hollis, M. Svinth, P. Day, A.F. Monzingo, G.W. Milne, and J.D. Robertus, Structure-based identification of a ricin inhibitor. *J Mol Biol*, 266, 1043, 1997

[15] Holm,L. and C. Sander, Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 223(1):p. 123-38, 1993

[16] Zemla, A., LGA - a Method for Finding 3D Similarities in Protein Structures. *Nucleic Acids Research*, 31(13):p.3370-4, 2003

[17] Zhu, J. and Z. Weng, FAST: a novel protein structure alignment algorithm. *Proteins.*, 58(3):p. 618-27, 2005