The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

A new algorithm for sequential and non-sequential protein multiple structure alignment

Lin Wang[1,2,*] Wen-Juan Zhang[3]

[1]Institute of Applied Mathematics, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100080, China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Basic Courses, Tianjin Foreign Studies University, Tianjin 300204, China

Abstract A fundamental issue in studies of protein structure and function is how to identify the conserved common core in multiple protein structures. The existing algorithms work well for proteins family such as those in HOMSTRAD but are not satisfactory for general multiple structure alignment problems, especially in some challenging cases, such as the occurrence of circular permutations. In this paper, an efficient approach called SANA-mult for multiple structure alignment is presented. Specifically, the alignment problem is first mapped on to a mixed integer programming problem via introducing a structural template, and then the algorithm decomposes the problem into two subproblems, i.e. solving the pairwise alignment and updating the template chain. We show that the proposed method can obtain sequential and non-sequential solutions for multiple structure alignment in an accurate manner, which is competitive or superior to the existing methods. The effectiveness of the new algorithm SANA-mult is tested using various protein structure sets and benchmark examples.

Keywords Protein structure; multiple structure alignment; circular permutation; sequential alignment; non-sequential alignment.

1 Introduction

Multiple structure alignment can aid in protein structure classification [1], understanding evolutionary conservation and divergence [2] and their correlation with sequences [3]. A fundamental issue in structural biology is how to identify the conserved structural common core via protein multiple structure alignment.

For most multiple structure alignment algorithms, the alignment results preserve the sequence order [4, 5, 6]. Such comparisons may miss important relationships because sequence order-dependent algorithms may disguise complex evolutionary events such as circular permutations. In this paper to overcome this problem, we present a new algorithm called SANA-mult to align multiple protein structures so as to identify the conserved structural common core in a more accurate and reliable manner. Specifically, we

*Email: linwang@amss.ac.cn

formulate the multiple structure alignment as a mixed integer programming problem. The proposed algorithm not only ensures the local convergence but also is able to handle both sequential and non-sequential alignments. The computational experiments show that our method works well for protein family in HOMSTRAD [7] and even protein family with circular permutations, in contrast to the existing methods.

2 Methods

In this paper, we propose a new approach to compare multiple protein structures with respect to both sequence order-independent and sequence order-dependent based on a pairwise structure alignment algorithm SANA [Lin Wang, Ling-Yun Wu, Xiang-Sun Zhang and Luonan Chen, "SANA: an algorithm for sequential and non-sequential protein structure alignment", submitted]. Next, we first give a brief description on SANA, and then focus on the multiple structure alignment that exerts all-against-one pairwise alignment by importing a template chain.

2.1 A pairwise structure alignment approach

SANA considers two AFPs' geometric compatibility by comparing aligned residues in two corresponding sequence neighborhood pairs, where AFP means Aligned Fragment Pair. In each sequence neighborhood pair, the aligned residues are detected by means of dynamic programming. The two AFPs are considered as geometrically compatible if their corresponding sequence neighborhood pairs have several identical aligned residues. We chain two AFPs if they are geometrically compatible, sequential and non-overlapped. Because geometric compatibility is related to residue indices, two AFPs that are far in sequence order cannot be joined and a maximal connected component finding algorithm is adopted to obtain several connected components according to the number of nodes in descendent order. The results from these connected components are the alignments in core regions and as the initial alignments to be refined.

For the several initial alignments, we refine each of them through solving the following programming which does not consider sequence-order constraint [8]. We assign the correspondence between aligned residues in each initial alignment to the variable s and iteratively solve the programming until convergence to obtain the refined alignment and the corresponding transformation. Then for the several refined alignments we choose the alignment having the least objective function value as the final non-sequential alignment.

$$min \quad \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} s_{ij}(|A+RX_i-Y_j|^2-\lambda^2) \tag{2.1}$$

$$s.t. \quad \sum_{i=1}^{n_x} s_{ij} \le 1 \quad for \quad j=1,...,n_y \tag{2.2}$$

$$\sum_{j=1}^{n_y} s_{ij} \le 1 \quad for \quad i=1,...,n_x \tag{2.3}$$

$$s_{ij} \in \{0,1\} \tag{2.4}$$

where X_i, Y_j are the coordinates of proteins X and Y, n_x and n_y are their protein chain lengths, and λ is a parameter. s is the assignment variables, A and R are respectively translation variables and rotation variables.

To get sequential alignment we have the following steps. First we substitute each transformation to the objective function and implement dynamic programming to minimize the objective function to obtain a sequential alignment. Then for the resulted se-

quential alignments we also select the alignment having the least objective function value as the final sequential alignment.

2.2 Multiple structure alignment

In this subsection, we first model the multiple structure alignment as a mixed integer programming problem, then show a detailed iterative algorithm for solving the problem.

2.2.1 Optimization model representation

By importing a structural template Y and generalizing the model (2.1)-(2.4) of pairwise alignment, we can formulate the multiple structure alignment problem as the following mixed integer programming (MIP).

$$min \quad \sum_{k=1}^{N}\sum_{i=1}^{n_x^k}\sum_{j=1}^{n_y} s_{ij}^k(|A^k + R^k X_i^k - Y_j|^2 - \lambda^2) \tag{2.5}$$

$$s.t. \quad \sum_{i=1}^{n_x^k} s_{ij}^k \le 1 \quad for \quad k = 1,...,N, \quad j = 1,...,n_y \tag{2.6}$$

$$\sum_{j=1}^{n_y} s_{ij}^k \le 1 \quad for \quad k = 1,...,N, \quad i = 1,...,n_x^k \tag{2.7}$$

$$s_{ij}^k \in \{0,1\} \tag{2.8}$$

where N is the number of protein chains, X_i^k is the coordinate of protein X^k, n_x^k is its protein chain length, n_y is set as the longest protein chain length, λ is a positive parameter which has the same scale as RMSD, with assignment variables $s = (s^1,...,s^N)$, transformation variables $A = (A^1,...,A^N)$ and $R = (R^1,...,R^N)$, and continuous variables Y_j. Notice that the objective function is a simple summation of pairwise alignments between proteins chains and the template chain.

From the form of MIP, it is clear that the optimal A^i, R^i, s^i and A^j, R^j, s^j, for $i \neq j$, are independent of each other when Y is given. So the variable A, R, s can be solved as pairwise structure alignment respectively with given template chain Y. In this paper, we adopt such a decomposition scheme by using the pairwise structure alignment algorithm SANA to solve the pairwise alignment problem, and the proposed multiple structure alignment method is called SANA-mult.

2.2.2 Updating template chain

We always choose the longest chain as the initial consensus structure. Then based on the KKT condition of the MIP for Y, we can analytically derive the updated rule for Y:

$$Y_j = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n_x^k} s_{ij}^k[A^k + R^k X_i^k]}{\sum_{k=1}^{N}\sum_{i=1}^{n_x^k} s_{ij}^k} \quad for \quad j = 1,\cdots,n_y \tag{2.9}$$

That is the template chain is updated as the average of the transformed coordinates.

2.2.3 Finding sequential and non-sequential alignment

From the above analysis, clearly our method mainly includes two phases. In the first phase, for a given template chain Y, we perform all-against-one pairwise alignments using SANA, i.e. the alignments between each protein chain and the template chain, which give the matching s and the transformation A, R. As described in SANA, we can obtain both

sequential and non-sequential pairwise alignments. In the second phase, for the given (s,A,R) obtained in the first phase, we update the template chain Y according to eqn.(2.9) and return the latest Y to the first phase. The iterative process between the two phases continues until convergence. We terminate the iteration if $\sum_{k=1}^{N}|D^{k(m)}-D^{k(m-1)}|\leq\varepsilon$, where $D^{k(m)}$ is the optimum value of the objective function (2.1) with respect to the k-th chain at the m-th iteration.

2.2.4 Convergence analysis

The decomposition of the algorithm actually ensures the local convergence. We next prove the convergence of the proposed algorithm. Let s^m, A^m, R^m be the solution of the first phase at the m-th iteration with a template chain Y^{m-1}. Then we have $\sum_{k=1}^{N}\sum_{i=1}^{n_x^k}\sum_{j=1}^{n_y}s_{ij}^{k(m)}(|A_k^{(m)}+R_k^{(m)}X_i^k-Y_j^{(m-1)}|-\lambda^2)\leq\sum_{k=1}^{N}\sum_{i=1}^{n_x^k}\sum_{j=1}^{n_y}s_{ij}^{k(m-1)}(|A_k^{(m-1)}+R_k^{(m-1)}X_i^k-Y_j^{(m-1)}|-\lambda^2)$. Substituting S^m, A^m, R^m to eqn.(2.9), and let the solution be $Y^{(m)}$. Then because of the KKT condition of the MIP for Y, it holds for $\sum_{k=1}^{N}\sum_{i=1}^{n_x^k}\sum_{j=1}^{n_y}s_{ij}^{k(m)}(|A_k^{(m)}+R_k^{(m)}X_i^k-Y_j^{(m)}|-\lambda^2)\leq\sum_{k=1}^{N}\sum_{i=1}^{n_x^k}\sum_{j=1}^{n_y}s_{ij}^{k(m)}(|A_k^{(m)}+R_k^{(m)}X_i^k-Y_j^{(m-1)}|-\lambda^2)$ which shows that the value of the objective function $D(S^{(m)},A^{(m)},R^{(m)},Y^{(m)})$ always decreases with the iteration of the computation. Since the solution space of s is a finite set, the convergence condition will be satisfied to terminate the computation.

3 Results and Discussion

The algorithm designed in this article is implemented in C++. Section 3.1 reports the validation of our algorithm (SANA-mult) in sequential case on several protein family in HOMSTRAD. Section 3.2 shows that our non-sequential alignments are effective for alignments of proteins that have circular permutations.

3.1 Comparing with existing algorithms on several sets of HOMSTRAD

We benchmark the performance of our algorithm (SANA-mult) in sequential case against three popular algorithms MultiProt [9], POSA [5] and Matt [6] using protein family in HOMSTRAD as shown in Table 1. Notice that MultiProt has both sequential and non-sequential alignment options, like SANA-mult; we compare against the option of the sequence order in this subsection. Matt has both unbent and bent alignment options; we compare against the option of unbent alignment. POSA shows no flexibility for aligning these protein family. We use two commonly used indices to measure the quality of a multiple structure alignment: the number of residue positions that contribute to the conserved structural core [10] (where structural core is defined as a set of residues that can be simultaneously superimposed with small structural variation), as well as average pairwise RMSD of the conserved core. Clearly, it is a multi-objective optimization problem: the goal is to minimize the RMSD of the conserved core while maximizing the number of the residues placed in the conserved core.

The alignments of the several protein family are shown in Table 1. Compared with POSA, clearly our algorithm has a larger core size but with a lower RMSD value on the second family. Notice that we use the default parameter λ=6.0 in SANA-mult in this

section. Table 1 also shows that SANA-mult is superior or competitive to other multiple structure alignment algorithms on protein family of HOMSTRAD.

Table 1: Alignment results of protein family in HOMSTRAD

Protein family	pdb-num	*aveLen*	SANA-mult $\overline{m}/\overline{RMSD}$	MultiProt $\overline{m}/\overline{RMSD}$	POSA $\overline{m}/\overline{RMSD}$	Matt $\overline{m}/\overline{RMSD}$
cyt3	6	110	84/1.64	71/1.19	92/2.24	91/1.95
ghf7	4	399	347/1.51	313/0.94	336/1.65	354/1.81
tim	10	249	235/1.19	220/0.91	241/1.39	243/1.37
cytc	11	111	94/1.18	90/0.95	92/1.16	97/1.56
ricin	7	254	232/1.34	208/0.99	233/1.47	236/1.48
asprs	3	470	375/2.13	309/1.67	380/2.58	388/2.42

aveLen: the average length of protein chains in a family;
$\overline{m}$: the number of residue positions in the common core;
$\overline{RMSD}$: the average pairwise RMSD of the conserved core

3.2 Detecting sequence order-independent structural similarity

Circular permutation is a phenomenon of fold changing occurred in evolution of a protein structure that results in the N and C terminus transferring to a different position. Protein structures that are related to circular permutations could exhibit sequence order-independent structural similarity, where polypeptide fragments are inconsistent with the linear order of the protein sequence. In this section, we show the ability of SANA-mult in non-sequential case in detecting sequence order-independent structural similarity. In order to obtain similar RMSD values with other algorithms we set parameter λ=3.0 in SANA-mult in this subsection.

Protein set 1 includes 1glh_, 1byh_, 1cpn_ and 1ajkA, which are taken from SCOP family "Glycosyl hydrolases family 16". The former two proteins as a group includes 1glh_ and 1byh_, 1cpn_ and 1ajkA are as two other groups respectively, where the three groups of proteins are related to circular permutations. MASS [11] is designed for detecting structural similarity without considering the sequence order. In contrast to non-sequential algorithms, POSA is designed for the preserved sequence order. We compare SANA-mult in non-sequential case, MultiProt in non-sequential case, MASS and POSA on the protein set 1. The alignment results of these proteins by different algorithms are shown in Table 2, which illustrates that the non-sequential algorithms align the multiple chains with circular permutations well than the sequential algorithms. Figure 1 is the superposition of the four proteins by SANA-mult.

Another circular permutation family (set 2) that are aligned include five proteins [12]. They are Glycine betaine-binding proteins 2b4lA, 1r9lA and 1sw1A, cyteine regulon transcriptional activator cysb 1al3_ and molybdate-binding protein 1amf_. We compare SANA-mult in non-sequential case, MultiProt in non-sequential case, MASS and POSA on the above protein set. Note that we compare against POSA with the unbent alignment result. The results by different algorithms are summarized in Table 2. Compared with MASS, SANA-mult finds more aligned residue positions in core but with similar RMSD values. Compared with MultiProt, SANA-mult has a weaker RMSD but with a larger core size. Figure 2 shows the alignment of the multiple structures by SANA-mult in non-sequential case. Figure 2(c) shows the match among a part of residues in the common

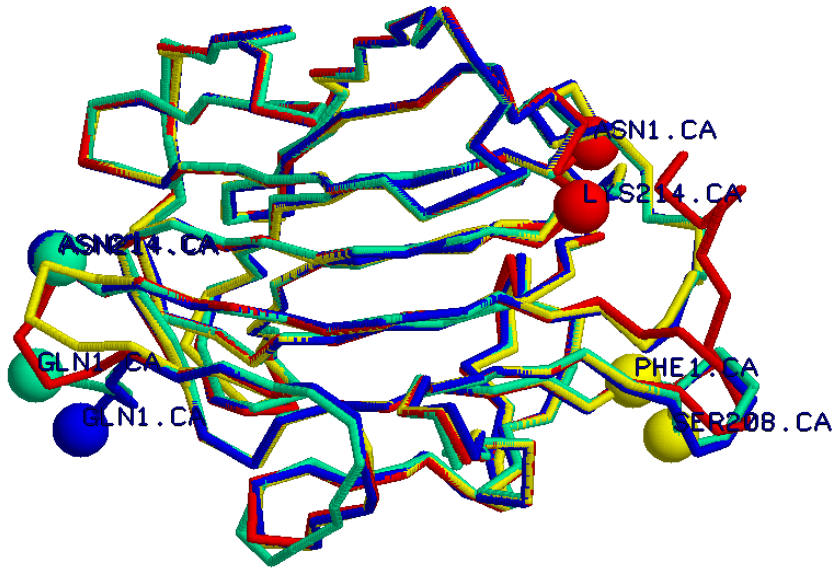


Figure 1: The superposition of four proteins 1glh_ (blue line), 1byh_ (green line), 1cpn_(yellow line), and 1ajkA(red line) by SANA-mult in non-sequential case. The two terminal residues of the four structures are indicated by labeling their residue names respectively. Notice that the terminal residues of the blue chain (1glh_) and green chain (1byh_) are aligned with the middle of the yellow chain (1cpn_), and the terminal residues of the yellow chain (1cpn_) are aligned with the middle of the red chain (1ajkA)

core, and the residues in functional sites are highlighted in orange. Clearly it illustrates that SANA-mult aligns the functional sites well and the sequence order of residues in functional sites is not necessarily conserved.

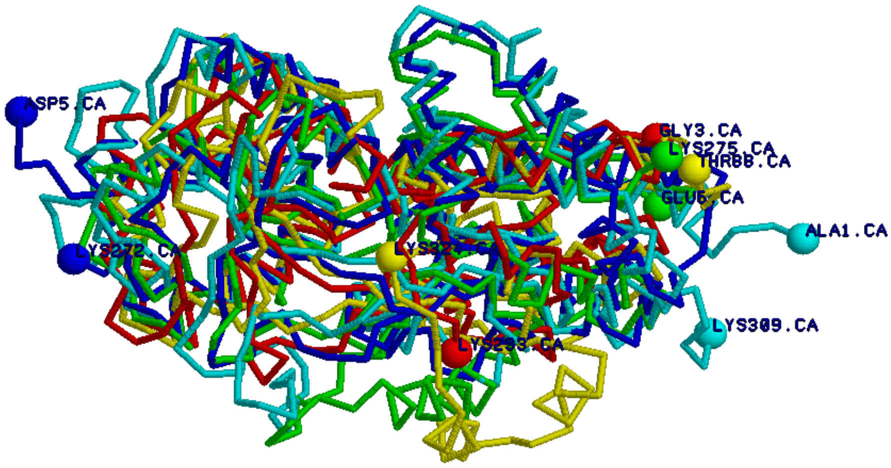
Table 2: Comparison of multiple protein structures with circular permutations

Protein sets	pdb-num	*aveLen*	SANA-mult $\overline{m}/\overline{RMSD}$	MultiProt $\overline{m}/\overline{RMSD}$	MASS $\overline{m}/\overline{RMSD}$	POSA $\overline{m}/\overline{RMSD}$
set 1	4	213	195/0.57	195/0.51	194/0.51	121/1.21
set 2	5	263	107/2.04	81/1.72	35/2.1	62/3.00

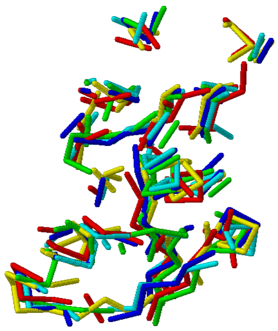
aveLen: the average length of protein chains in a family;
$\overline{m}$: the number of residue positions in the common core;
$\overline{RMSD}$: the average pairwise RMSD

4 Conclusion

In this paper, we developed a new method (SANA-mult) for protein multiple structure alignment, which can handle two different alignment manners, i.e. sequential and non-sequential alignments. Numerical results show that SANA-mult is superior to other algorithms in the quality of alignments on the protein family in both data of HOMSTRAD



(a)



2b41A	V176	E181	I182	···	Q203	A206	M209	···	A220	W225	L226
1r91A	S15	E20	T21	···	E43	Y46	V48	···	F60	W65	T66
1sw1A	P16	Q18	Y19	···	E40	T45	L46	···	L59	Y63	T64
1al3_	T105	Q103	A104	···	Q128	P131	T132	···	F145	T270	T269
1amf_	A11	L13	T14	···	S36	S39	T41	···	L54	Y196	E195

N228	H230	Y233	···	D22	S25	M28	···	T71	W72	T73	···	E97
P67	H69	M72	···	N137	W140	E143	···	Y187	W188	T189	···	S234
G65	T66	I71	···	P144	E145	F146	···	P189	Y190	T191	···	Y214
S268	E150	A151	···	T225	F201	R204	···	I243	A244	S245	···	H165
V194	D59	K61	···	E150	V123	G126	···	V169	Y170	G171	···	N83

(b) (c)

Figure 2: The comparison among multiple proteins 2b4lA, 1r9lA, 1sw1A, 1al3_ and 1amf_ by SANA-mult in non-sequential case. (a) shows the full alignment of multiple proteins differentiated by colors. They are marked by blue, cyan, green, olive and red respectively, and the two terminal residues of these proteins are indicated by labeling their residue names respectively. (b) is the core alignment of multiple proteins. (c) is the match among a part of residues in core alignment. The functional sites are marked by orange.

and others with circular permutations. There are no limitations of our method on protein domains which do not keep the sequence order but have spatial similarity. In addition, the analysis of the numerical results shows that SANA-mult is also able to find similar functional sites in underlying primary sequences.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (NSFC) under Key Research Grant No.10631070, Research Grant No.60503004, and JSPS and NSFC under JSPS-NSFC collaboration project. The authors thank Prof. Xiang-Sun Zhang and Prof. Luonan Chen for helpful discussions and suggestions.

References

[1] Liu X., Zhao Y.P., Zheng W.M.: CLEMAPS: Multiple alignment of protein structures based on conformational letters. Proteins, **71** (2007) 728–736

[2] Xie L., Bourne P.E.: Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci USA, **14** (2008) 5441–5446

[3] Edgar R., Batzoglous S.: Multiple sequence alignment. Curr Opin Struct Bio, **16** (2006) 368–373

[4] Zhou T.S., Chen L., Tang Y., Zhang X.S.: Aligning multiple protein structures by deterministic annealing. Journal of Bioinformatics and Computational Biology, **3(4)** (2005) 837–860

[5] Ye Y., Godzik A.: Multiple flexible structure alignment using partial order graph. Bioinformatics, **21** (2005) 2362–2369

[6] Menke M., Berger B., Cowen L.: Matt: Local flexibility aids protein multiple structure alignment. PLoS Computational Biology, **4(1)** (2008) e10

[7] Mizuguchi K., Deane C., Blundell T.L., Overington J.P.: HOMSTRAD: A database of protein structure alignments for homologous families. Protein Science, **11** (1998) 2469–2471

[8] Chen L., Wu L.Y., Wang Y., Zhang S., Zhang X.S.: Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Structural Biology, **6** (2006) 18

[9] Shatsky M., Nussinov R., Wolfson H.: A method for simultaneous alignment of multiple protein structures. Proteins, **56** (2004) 143–156

[10] Eidhammer I., Jonassen I., Taylor W.R.: Structure comparison and structure patterns. Journal of Computational Biology, **7** (2000), 685–716

[11] Dror O., Benyamini H., Nussinov R., and Wolfson H.: Multiple structural alignment by secondary structures: algorithm and applications. Protein Science, **12** (2003) 2492–2507

[12] Lo W.C., Lyu P.C.: CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships. Genome Biology, **9** (2008) R11