Advances in Computing 2012, 2(2): 23-28
DOI: 10.5923/j.ac.20120202.04

Comparison of Structure Based Sequence Alignment Programs for Protein Domain Superfamilies with Multiple Members

A. Gandhimathi, Anu G. Nair, R. Sowdhamini*

National Centre for Biological Sciences (TIFR), UAS-GKVK Campus, Bellary Road, Bangalore, 560065, India

Abstract Structure comparison is used to reveal the similarity between protein structures. Every method has its own strength and weakness and the assessment parameters need to be appropriate to the original question on performance of the method. Here, we have assessed three multiple structure-based sequence alignment programs and compared their results. The results suggest that superfamily members which have low sequence identity (<40%) can be aligned using flexible structure alignment methods followed by methods which consider multiple structural features like COMPARER. This kind of structural analysis protocol appears to produce more relevant results, due to consideration of large number of structural features, rather than pure geometric features.

Keywords Structure Alignment, Outliers, Domain swapping, protein evolution, distant relationships

1. Introduction

Protein sequence alignments are important in understanding the structural, evolutionary and functional relationship between proteins[1]. It will be more challenging to perform alignments in distantly related proteins owing to high sequence divergence. Computation of structure-based alignment is a delicate task, but can give rise to more reliable alignments at distant relationships, when compared to pure sequence alignment[2].

Any method for protein structure alignment needs to balance coverage versus accuracy. Some methods align the core of a protein at very high accuracy (i.e., very low RMSD) and very low coverage (i.e., omitting loop regions), while some methods prefer to increase the coverage (i.e., include the loop regions in the alignment) by compromising on the accuracy (i.e., increasing the RMSD) [3-7]. Structure alignment programs are preferred since they have high accuracy, high coverage and fast execution to cope with the increasing number of structures. Certainly, maximizing the biological relevance of a result is going to be the most desirable outcome in a majority of the cases.

In general, there are a number of structure alignment programs available. Some structure alignment programs are developed for aligning a pair of 3D structures called as pairwise structure alignment programs (PStA) and some programs are used for multiple structure alignment (MStA). Examples for PStA methods are LSQMAN[8], FATCAT[9], MINRMS[10], and Dali[11]. MStA Programs are MASS[12], Matt[13], MultiProt[14], MUSTANG[15], POSA[16], SALIGN[17] and 3DCOMP[18]. COMPARER[19] is another structure-based sequence alignment program which considers various structural features to recognize the structural core and variable regions to guide the presence of gaps and to obtain reliable alignments. But the program needs initial equivalencies from any of the alignment program or through graphical inspection. Any sequence or structure alignment can be refined through COMPARER to obtain final alignment.

In a previous study, we investigated alignment accuracy of several frequently used structure based sequence alignment methods[20].Three thousand and fifty-two alignments of 218 pairs of protein domain structural entries, with <40% sequence identity, belonging to different structural classes, of diverse domain sizes and length-rigid/variable domains were performed using 12 programs. These programs were compared and assessed by three structural parameters such as root mean square deviation, secondary-structural content and equivalences. The biological and functional relevance of such alignments were examined for some examples. From this study, we concluded that FATCAT, MATT, DALI, MINRMS and LSQMAN programs perform equally in most of the cases. In many cases, LSQMAN fails to improve the percentage of secondary-structure equivalences. This study helped us to select the suitable tool MINRMS for aligning two member superfamilies, where only two domains exist in the superfamily with <40% identity[21].

* Corresponding author:
mini@ncbs.res.in (R. Sowdhamini)
Published online at http://journal.sapub.org/ac

Table 1. Properties of Multiple structure alignment methods

Method	Alignment	Flexibility	Multi chain usage	Output information
Mutiprot	Core	No	All	Structure
POSA	Core	Yes	First	Structure
MASS	Core	No	All	Structure
MUSTANG	Full	No	Rejected	Sequence and Structure
MATT	Full	Yes	Split	Sequence and Structure
COMPARER	Full, require initial equivalencies	No	All	Sequence

This paper describes the continuation of such analyses and the intention of this study was to select and work on a suitable program for aligning superfamilies containing multiple members (3-239) and varying size (a.a 21-1419).This kind of study helps to understand the strength and limitations of particular programs.

2. Analysis of Multiple Structure Alignment Tools

In case of multiple structure alignment, a program that can produce high-quality alignment, flexibly handle multi-chain structures, process a large number of input structures and provide a consistent output format (like exactly reference to PDB residue numbers) is desirable. For a broader applicability, proper treatment of local conformational variability is probably of utmost importance. We have originally started with the comparison of six different methods for the alignment of multiple members. Our main interest was to understand the structural and functional relationship of superfamilies where the members have <40 identity. Hence, we would prefer structure-based sequence alignment and consider the whole domain for alignment. Based on our criteria, only two programs are suitable for multiple structure alignment (Table.1). We have assessed the two programs to choose the best program to align aforementioned task.

3. Features of Structure Alignment Programs

It is always a challenging task to compare all the existing programs since they have limitations at various levels. Hence it is better to compare the programs which are suitable for particular type of data. In our analysis, we found MATT, MUSTANG and COMPARER programs are suitable (Table 1) by highlighting the limitations of other programs for aligning domains at superfamily level. MUSTANG aligns residues on the basis of similarity in patterns of both residue–residue contacts and local structural topology. MATT allows local flexibility between fragments: small translations and rotations are temporarily allowed to bring sets of aligned fragments closer, even if they are physically impossible under rigid body transformations.

Current methods for structure comparison and alignment usually focus on optimizing geometrical similarities between two or more structures. The alignment of superfamily members are based on the conservation of structural features such as secondary structures, hydrogen bonding and solvent accessibility. Hence, apart from geometrical features, parameters that reflect secondary, tertiary, possibly quaternary features and evolutional information can help in finding the most relevant alignment between structures. Thus, COMPARER is one such method that takes account of all the above mentioned parameters.

4. Assessment Parameters

In the earlier study[20], we had employed three methods such as RMSD, POSSE (Percentage of Secondary Structure Equivalences) and number of fitted points for assessing the alignment quality and accuracy. These parameters are well explained by our earlier work. We have used one more parameter AGSS (Assessment of gaps in secondary structure) which again depends on the structural information representing the number of gaps introduced within secondary structures during the alignment. If there are more gaps within secondary structure blocks, then the quality of the alignment is questionable.

AGSS describes the number of gaps which are introduced within secondary structures in the course of alignment.

$$AGSS = -[(G_1 + G_2 + \ldots\ldots + G_n)/n]$$

where,

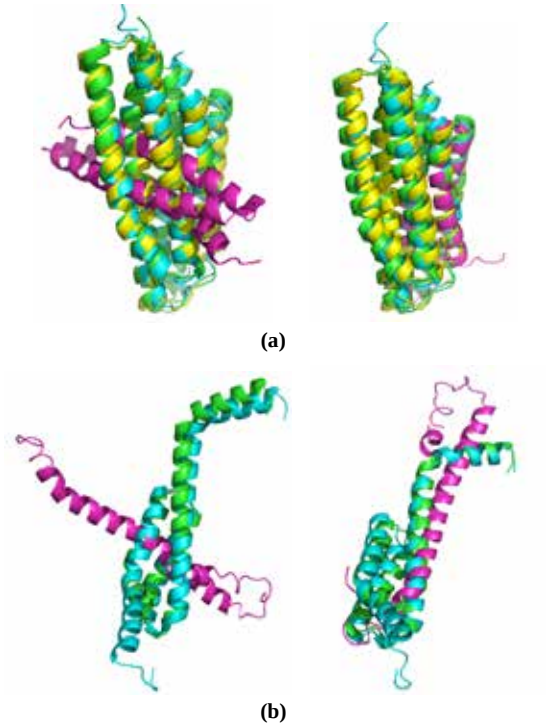
n = number of regular secondary structures in the alignment,

G_k = (no. of gaps in K)/(length of K, where K is the k^{th} secondary structure.

5. Structure Based Sequence Alignment

Initially, superfamily members are aligned by MATT, MUSTANG and the alignment is annotated by JOY program[22]. Since COMPARER needs initial equivalences and guided tree for the alignment, we primarily compared MATT and MUSTANG for initial alignment. There are total of 731 superfamilies are aligned using both the methods (data not shown). MATT can handle all kinds of superfamily like total number of domains (2-239) and size of the domain. But, MUSTANG is not suitable for handling multi-chain proteins. Around 86 superfamilies have high RMSD between super-

imposed structures and failed to align large superfamilies due to some technical issues. For performing refined alignment using COMPARER, initial equivalences are taken from both MUSTANG and MATT. After the COMPARER alignment, MNYFIT[23] is used to obtain superimposed structures. There are 25 superfamilies, belonging to different classes, were used for assessing the accuracy of structure comparison and alignment by MATT, MUSTANG and COMPARER. The alignments were checked using the assessment parameters as mentioned above.



(a)

(b)

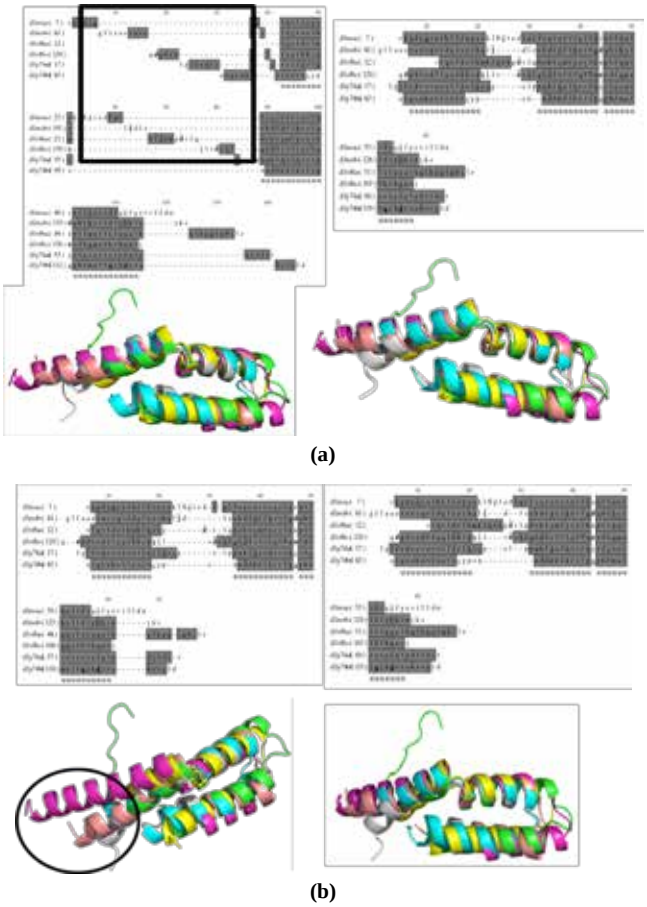
Figure 1. MUSTANG and MATT comparison. Protein domains that belong to a superfamily are shown in the best fit form after superposition.(a) Superfamily Phou-like (SCOP code: 109755) (b) Superfamily Heat shock protein 70kD (HSP70), C-terminal subdomain (SCOP code: 100934). Result of superposition obtained by MUSTANG is shown to the left and that obtained by MATT is shown to the right. In both these examples, results obtained by MATT are better

MUSTANG failed in the alignment of two out of 25 superfamilies (shown in Figure 1) as it was not able to provide optimal superposition and find the equivalent elements. All programs fared well in all the assessment parameters, except the number of gaps included within secondary structural elements (Table 2). MATT introduces lot of gaps in the alignment compared to MUSTANG. However, COMPARER was preferred for refined final alignment (Figure 2).

MATT is able to handle all possible cases which did not give any technical failures, but it adds more gaps to the alignment and the derived alignment fares low according to our fourth assessment parameter, AGSS. COMPARER was able to handle all possible cases and the derived alignment fared well for all the assessment parameters. Therefore, an alignment protocol with an initial alignment using a flexible program like MATT and realignment by COMPARER works better in all superfamilies considered.

5.1. Predicting Structurally Deviant Members

The suggested protocol provides good alignment accuracy with low RMSD. It still permits us to find structurally deviant members of the superfamily which are called as outliers. Glutathione synthetase ATP-binding domain-like superfamily (SCOP code: 56059) contains 22 domains with low sequence identity (<40%). Alignment protocol using MATT followed by COMPARER yielded satisfactory structure-based sequence alignment of these superfamilies, leaving only two members (d1eucb2,d2nu7b2) with high RMSD (>5.5Å). These two members belong to Succinyl-CoA synthetase, beta-chain, N-terminal domain family (Figure 3). The next member (d1kbla3) with high RMSD belongs to pyruvate phosphate dikinase family. It is already reported that two families (succinyl-CoA synthetase and pyruvate phosphate dikinase) from the glutathione synthetase ATP-binding domain-like superfamily are slightly more different than the other families. These two families display the same manner of binding the nucleotide inside the active site[24]. In these cases, structural alignment of these outlier members was still possible.



(a)

(b)

Figure 2. shows the structure alignment of L27 domain (SCOP ID 101288). Fig 2(a) shows MATT derived alignment and the structural superposition (left) and COMPARER refined alignment and the structural superposition (right). More number of gaps are introduced in the MATT alignment shown in highlighted box. Fig 2(b) shows the same as Fig. 2(a) but for MUSTANG-derived alignment. Alignment is improved, as shown in highlighted circle, with less number of gaps

Table 2. Comparison of MATT, MUSTANG and COMPARER using Four Assessment Parameters

S. NO	Super Family	MATT				MUSTANG				COMPARER			
		RMSD (Å)	SST (%)	Matches	AGSS	RMSD	SST	Matches	AGSS	RMSD	SST	Matches	AGSS
1	101288(a)	1.54	67.6	36	-2.31	2.49	72.1	43	-0.32	1.71	67.6	37	0.0
2	109885(a)	1.91	69.7	91	-0.37	2.75	73.8	99	-0.06	2.15	67.6	97	-0.18
3	89028(a)	1.42	71.7	114	-0.69	2.27	76.3	112	-0.36	2.37	74.4	114	-0.15
4	63570(a)	2.23	49.1	47	-1.57	3.97	47.9	48	-1.08	2.92	45.8	59	0.0
5	101386(a)	2.14	32.7	35	-2.01	3.42	44.7	28	-1.55	2.54	32.1	44	-0.24
6	47446(a)	0.70	67.8	55	-0.79	2.78	62.5	64	-0.19	0.82	71.9	57	0.0
7	109755(a)	1.25	67	91	-0.36	-	-	-	-	1.12	69.2	85	0.0
8	100934(a)	4.67	81.1	51	-1.04	-	-	-	-	3.04	76.8	56	0.0
9	50911(b)	1.53	35.2	87	-0.85	1.68	37.6	94	-0.31	1.75	38.4	93	-0.218
10	50118(b)	2.87	51.0	44	-2.24	3.36	42.4	53	-0.42	3.91	41.9	61	-0.19
11	141571(b)	0.82	16.8	106	-0.15	4.70	15.3	115	-0.12	0.87	16.5	107	-0.03
12	51294(b)	1.47	39.7	119	-0.70	1.44	40.3	107	-0.32	1.57	41.2	116	-0.27
13	51206(b)	1.75	30.7	117	-1.18	1.36	29.0	111	-0.51	1.37	28.6	109	-0.08
14	63380(b)	1.87	32.9	133	-1.65	1.80	38	137	-0.52	1.89	38.9	135	-0.512
15	90209(a+b)	1.12	7.4	27	-0.17	1.47	7.1	27	-0.40	1.13	7.4	26	0.0
16	54665(a+b)	1.14	51.5	97	-0.05	1.04	51.5	92	-0.06	1.17	52.6	93	0.0
17	103190(a+b)	1.77	19.7	45	-1.74	2.65	28.8	51	-0.74	2.98	24.6	55	-0.35
18	160631(a+b)	1.43	50.4	118	-0.19	1.50	51.2	111	-0.07	1.54	50.4	126	0.0
19	100879(a+b)	1.59	17.6	67	-1.19	1.72	21.9	69	-0.24	1.95	18.1	75	-0.40
20	54292(a+b)	2.36	39.5	157	-0.98	2.10	16.1	95	-0.82	2.10	15.2	95	-0.30
21	102114(a/b)	2.74	53.6	158	-0.49	4.02	51.7	183	-0.44	4.93	47.2	215	0.0
22	51730(a/b)	1.85	56.3	168	-0.41	4.77	58.1	217	-0.19	4.75	53.3	226	-0.02
23	102405(a/b)	1.59	43.5	85	-1.29	2.55	43.2	105	-0.23	2.08	42.9	94	-0.04
24	51366(a/b)	2.34	28.6	212	-2.76	2.44	32.2	219	-0.96	2.52	35.5	218	-0.244
25	68923(a/b)	1.31	54.1	158	-0.30	1.50	53.8	155	-0.20	2.50	53.2	163	-0.11
26	57581(small)	1.03	34.5	51	-0.25	1.83	33.3	52	-0.11	0.98	32.2	50	0.0
27	56645(Multi)	2.09	39.5	157	-0.99	2.55	41.5	140	-0.53	2.65	43.6	162	-0.14

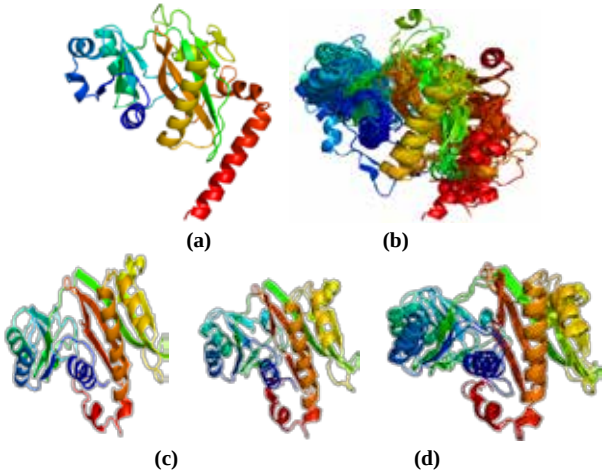
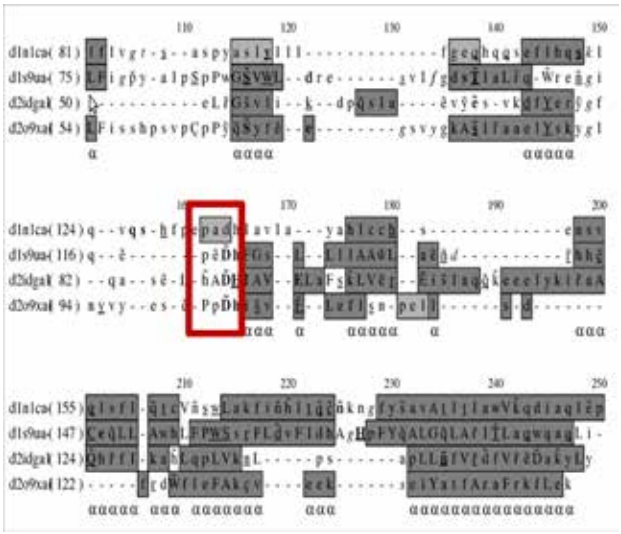


Figure 3. Identifying structurally deviant members in Glutathione synthetase ATP-binding domain-like superfamily**.** (a) & (b) shows the representative structure of this superfamily and alignment of 20 protein domains. (c) &(d) shows the two outliers belonging to Succinyl-CoA synthetase family

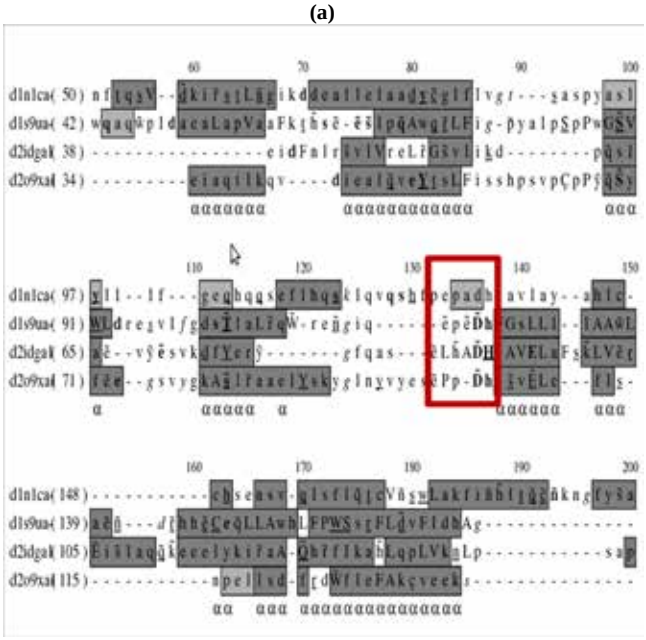
5.2. Domain Swapping –TorD like Superfamily

Domain swapping is an important and interesting phenomenon which refers to two or more proteins exchanging equivalent parts of their structures to form intertwined oligomers, inclusive of dimers[25]. Sequence-based alignment tools may be inadequately sensitive to detect such evolutionarily distantly related members. Moreover, conventional structure comparison algorithms are weak to detect global similarities between proteins related by domain swapping due to considerable difference in the structure of swapped and non swapped forms[5,26].

TorD like superfamily in PASS2 dataset has four members, of which three are having closed monomer and one member (d1n1c_) is in open domain swapped conformation[27]. The 3D structures show a long loop region separating the N- and C-terminal domains of the proteins. This loop retains the highly conserved 'E(Q)PxDH' motif[28] as equivalent when aligning a ''closed'' monomer and its domain-swapped ''open'' homologue. Many structure alignment programs tend to yield an alignment of monomers with only one domain of the swapped dimer aligned. MATT is one such program which gives alignment of monomers with one domain, but COMPARER refined alignment retains both the swapped as well as non-swapped domains as equivalent. Such biologically meaningful alignment proves to be helpful to understand the relationship between monomers and swapped dimers. MUSTANG takes care of domain swapping problem in the alignment. Apart from this, MUSTANG and COMPARER refined alignment retains the linker motif in the alignment as equivalent regions both in the monomer and domain-swapped form despite their structural differences (Figure 4).



(a)



(b)

Figure 4. (a) shows the alignment of TorD superfamily by MUSTANG. MUSTANG is able to align the linker conserved motif (E(Q)PxDH) of swapped and non-swapped domains. Figure 4(b) shows COMPARER-refined alignment where initial equivalences are from MATT. MATT fails to align the conserved motif although MATT followed by COMPARER gives rise to an alignment where the conserved motif is equivalent

6. Conclusions

Structure-based sequence alignment methods are crucial for understanding distantly related proteins. We have mainly compared structure comparison tools which give rise to consistent and reliable alignment of multiple members that in turn can enable better understanding of the structural, evolutionary, functional relationships between distantly related proteins at the superfamily level. The aim of this study is not to rank or benchmark the different methods, but instead to recognise the differences in the results and the challenges that remain.

We found that an alignment protocol with MATT followed by COMPARER works well for most of the superfamilies. Although MUSTANG performed equally well in many cases and can handle alignment of swapped domain with monomers, it failed to align two superfamilies in our test dataset of 25 superfamilies that belong to alpha-class. MATT could handle all 25 superfamilies, but it introduces lot of gaps in the alignment. Simple RMSD measures are insufficient to recognize good quality alignments. Structure comparison programs like MATT face challenges in swapped domain examples, unlike MUSTANG. However, in all superfamilies studied, COMPARER is efficient in improving the alignment and in recognising the highly deviant members, namely the outliers, of the superfamily.

ACKNOWLEDGEMENTS

A.G is supported by Senior Research Fellowship from the Council of Scientific and Industrial Research (CSIR) Government of India.

REFERENCES

[1] Sierk ML, Kleywegt GJ, “Deja vu all over again: finding and analyzing protein structure similarities”, Structure, vol.12, no.12, pp.2103-2111, 2004.

[2] Carugo O, “Recent progress in measuring structural similarity between proteins”, Current Protein & Peptide Science, vol.8, no.3, pp.219–241, 2007.

[3] Godzik A, “The structural alignment between two proteins: is there a unique answer?” Protein Science, vol.5, no.7, pp.1325 –1338, 1996.

[4] Christoph B, Christine S. S, Peter L, “Accuracy analysis of multiple structure alignments”, Protein Science, Vol.18, PP.2027-2035, 2009.

[5] Liu W, Srivastava A, Zhang J, “A Mathematical Framework for Protein Structure Comparison.” PLoS Computational Biology, vol. 7, no.2, e1001075, 2011.

[6] Aysam G, Knapp E, “GIS: a comprehensive source for protein structure similarities”, Nucleic Acids Research, Vol. 38, pp. W46–W52, 2010.

[7] Joseph A.P, Srinivasan N, Alexandre G, “Improvement of protein structure comparison using a structural alphabet” Biochimie vol.93,pp. 1434-1445, 2011.

[8] Kleywegt G, “Use of non-crystallographic symmetry in protein structure refinement”, Acta Crystallographica D, Biological Crystallography, vol.52, no.4, pp.842–57, 1996.

[9] Ye Y and Godzik A, “Flexible structure alignment by chaining aligned fragment pairs allowing twists”, Bioinformatics, vol.19, no.2, pp.ii246-ii255, 2003.

[10] Jewett AI, Huang CC, Ferrin TE, “MINRMS: an efficient algorithm for determining protein structure similarity using

root-mean-squared-distance", Bioinformatics, vol.19, no.5, pp.625-634, 2003.

[11] Holm L, Sander C, "Protein structure comparison by alignment of distance matrices", Journal of Molecular Biology, vol.233, no.1, pp.123-138, 1993.

[12] Dror O, Benyamini H, Nussinov R, Wolfson H, "MASS: multiple structural alignment by secondary structures", Bioinformatics vol.19, no.1, pp.i95–i104, 2003.

[13] Menke M, Berger B, Cowen L, "Matt: local flexibility aids protein multiple structure alignment", PLoS Computational Biology, vol.4, no.1, pp.88-99, 2008.

[14] Shatsky M, Nussinov R, Wolfson HJ, "A method for simultaneous alignment of multiple protein structures", Proteins, vol.56, no.1, pp.143–156, 2004.

[15] Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM, "MUSTANG: a multiple structural alignment algorithm", Proteins, vol.64, no.3, pp.559–574, 2006.

[16] Ye Y, Godzik A, "Multiple flexible structure alignment using partial order graphs", Bioinformatics, vol.21, no.10, pp.2362–2369, 2005.

[17] Madhusudhan M,S. et al. Alignment of multiple protein structures based on sequence and structure features. Protein Engineering Design &. Selection, vol. 22, pp. 569–574, 2009.

[18] Sheng W, Jian P,Jinbo X, "Alignment of distantly related protein structures: algorithm,bound and implications to homology modeling", vol. 27, no.18,pp. 2537–2545, 2011.

[19] Sali A, Blundell TL, "Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming", Journal of Molecular Biology, vol.212, no.2, pp.403-428, 1990.

[20] S. Kalaimathy, R. Sowdhamini and K. Kanagarajadurai, "Critical Assessment of Structure-based Sequence Alignment Methods at Distant Relationships", Briefings in Bioinformatics, vol.12, no.2, pp.163-175, 2010.

[21] Gandhimathi A, Nair A and Sowdhamini R, "PASS2.4: An update of database of structure-based sequence alignments of structural domain superfamilies", Nucleic Acids Research, vol.40, no.D1, pp.D531-D534, 2012.

[22] Mizuguchi K, Deane C. M, Blundell T.L, Johnson M. S and Overington J. P, "JOY: protein sequence-structure representation and analysis", Bioinformatics, vol.14, no.7, pp.617–623, 1998.

[23] Sutcliffe M. J, Haneef I, Carney D and Blundell T. L, "Knowledge based modeling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures", Protein Engineering, vol.1, no.5, pp.377–384, 1987.

[24] Dinescu A, Cundari TR, Bhansali VS, Luo JL and Anderson ME, "Function of conserved residues of human glutathione synthetase: implications for the ATP-grasp enzymes", The Journal of Biological Chemistry, vol.279, no.21, pp.22412-22421, 2004.

[25] Bennett M. J, Schlunegger MP and Eisenberg D, "3D domain swapping: a mechanism for oligomer assembly", Protein Science, vol.4, no.12, pp.2455–2468, 1995.

[26] Chia-Han C,et.al, "Detection and Alignment of 3D Domain Swapping Proteins Using Angle-Distance Image-Based Secondary Structural Matching Techniques", PLoS ONE, vol.5, no.10, e13361,2010.

[27] Samuel T, Chantal Nivol,Catherine B, Marianne I,Isabelle M, Vincent M and Jean-Pierre S, "A Novel Protein Fold and Extreme Domain Swapping in the Dimeric TorD Chaperone from Shewanella massilia", Structure, vol.11, no.2, pp.165–174, 2003.

[28] Olivier G, Vincent M, Chantal I, "Multiple roles of TorD-like chaperones in the biogenesis of molybdoenzymes", FEMS Microbiology Letters, vol.297, no.1, pp.1–9, 2009.