STRATEGIES OF NON-SEQUENTIAL PROTEIN STRUCTURE ALIGNMENTS

AYSAM GUERLER
guerler@chemie.fu-berlin.de

ERNST-WALTER KNAPP
knapp@chemie.fu-berlin.de

Freie Universität Berlin, Institute of Chemistry and Biochemistry, Fabeckstrasse 36a, 14195 Berlin, Germany

Due to the large number of available protein structure alignment algorithms, a lot of effort has been made to define robust measures to evaluate their performances and the quality of generated alignments. Most quality measures involve the number of aligned residues and the RMSD. In this work, we analyze how these two properties are influenced by different residue assignment strategies as employed in common non-sequential structure alignment algorithms. Therefore, we implemented different residue assignment strategies into our non-sequential structure alignment algorithm GANGSTA+. We compared the resulting numbers of aligned residues and RMSDs for each residue assignment strategy and different alignment algorithms on a benchmark set of circular-permuted protein pairs. Unfortunately, differences in the residue assignment strategies are often ignored when comparing the performances of different algorithms. However, our results clearly show that this may strongly bias the observations. Bringing residue assignment strategies in line can explain observed performance differences between entirely different alignment algorithms. Our results suggest that performance comparison of non-sequential protein structure alignment algorithms should be based on the same residue assignment strategy.

Keywords: non-sequential structure alignment methods; protein structures; measurement.

1. Introduction

Protein structure alignment approaches are of great importance for the analysis of protein function, structure and evolution. A large number of structure alignment programs have been developed in the recent years to find a solution to this NP-hard problem (see for instance [1]). A representative set of sequential alignment programs is DaliLite [2], K2 [3], CE [4] and TM-align [5]. In addition to these sequential structure alignment methods, also non-sequential structure alignment methods have been developed i.e. TOPOFIT [6], MASS [7], GANGSTA [8], GANGSTA+ [9] (see Table 1) and others [10, 11] to mention just a few. Due to the large number of available programs and the variety of underlying concepts and strategies, selecting the “best” method for a certain application is difficult. Generally, a comparison of protein structure alignment algorithms is carried out on the basis of reasonably difficult and representative protein pair benchmark sets [12]. The resulting protein structure alignments obtained for a selected set of protein pairs can be evaluated with a wide variety of protein structure alignment scoring functions i.e. Z-Score [2], TM-score [5] and SAS [13]. These scoring functions have in common to consider two major properties of a protein structure alignment, which are the number of aligned residue pairs between the protein pair and the corresponding root mean square deviation (RMSD) of the C_α atoms of the two structures. Many scoring functions consider also other characteristic quantities such as total number of unaligned sequence segments (gaps) or the amino acid identities of aligned residues to achieve an improved and more reliable measure of the alignment quality. Similar to sequence alignment approaches [14, 15] a p-value (see also Z-Score [2]) can be evaluated, defining the probability that an unrelated protein structure pair obtains by chance a specific score. Scoring functions can be constructed to be insensitive versus the size of proteins, as for instance the TM-score [5]. The biological relevance of a scoring function can be analyzed by probing its ability to recognize protein families or alternative relations for a given set of proteins [16]. For these tests high-quality protein structure databases like e.g. CATH [17, 18] and SCOP [19] are used as a reference. Kolodny et al. [13] studied the problem to find scoring functions that are suitable to compare and measure the quality of protein structure alignments obtained with different algorithms and proposed a set of four scoring functions i.e. SI, MI, SAS and GSAS. To judge the performance of different alignment tools the benchmark of protein structure pairs and the choice of the scoring function are critical.

An issue, which is particularly important for the evaluation of non-sequential structure alignment methods, is the underlying strategy of residue assignment. Although often disregarded, this issue can have a significant influence on the results of the alignment algorithms as we will show in the present study. To illustrate these influences, firstly we selected a minimal set of five individual residue assignment options. They comprise the following features: (α) to align residues located everywhere in the structure or only within secondary structure elements (SSE) (i.e. α-helix and β-strand); (β) to align two SSEs also in reverse orientation (i.e. N-terminus of the SSE from one protein is aligned on the C-terminus of the corresponding SSE of the second protein); (γ) to make only unique or also shared (fuzzy) residue assignments. In the latter case the residue of one protein can be assigned to more than one residue in the other protein; (δ) to assign only residues belonging to the same SSE type or to ignore the SSE type in these assignments; (ε) to assign residues without gaps or to assign also isolated residue pairs. These residue assignment options [(α)-(ε)] are combined to a set of five residue assignment strategies [(1) – (5)] as defined in Table 2.

To analyze the influences of these different alignment strategies [(1) – (5)], we implemented them in GANGSTA+ [9]. We like to point out that this study can also be done with several other protein structure alignment tools mentioned above. We use GANGSTA+ in this application since with our own method we have the procedures better under control. As a benchmark set five circular permuted protein structure pairs were considered. To obtain reference values for the residue assignment strategies [(2) – (5)], we applied the non-sequential protein structure alignment approaches TOPOFIT [6], MASS[7], GANGSTA+ [9] and GANGSTA [8] (see Table 1) on the same five protein pairs. We like to emphasize that this work does not intend to compare the overall performances of these non-sequential structure alignment methods, but to illustrate the effect of their specific residue assignment strategies on the number of aligned residues and the RMSD.

Firstly, the results reveal an essential effect of the implemented residue assignment strategies on the number of aligned residues and the RMSD. Secondly, these results align well to the reference values, observed by applying the corresponding protein structure alignment method for each residue assignment strategy. Concluding, this shows that any performance comparison of protein structure alignment methods has to consider the variations in the underlying residue assignment strategies. Furthermore, considering these variations can explain observed performance differences between entirely different non-sequential structure alignment algorithms.

2. Methods

2.1. *Benchmark of circular permuted protein structure pairs*

The benchmark dataset for the comparison of non-sequential protein structure alignment algorithms consists of five circular permuted protein pairs. These are the protein pairs with PDB [20] id (a) 1RIN [21] / 2CNA [22], (b) 1GLH [23] / 1CPN [24], (c) 1EXG [25] / 1TUL [26], (d) 1RHG [27] / 1BCF [28], (e) 1IHW [29] / 1SSO [30] (see Table 1). The same dataset is used to illustrate the effect of different residue assignment strategies with GANGSTA+ (see Table 2 and Figure 1).

2.2. *Non-sequential protein structure alignment tools*

Four non-sequential protein structure alignment approaches were considered in this study (listed in Table 1). GANGSTA [8] operates hierarchically on two stages. First, a genetic algorithm is employed to optimize the SSE assignment between two protein structures by maximizing a contact map overlap based on the specific GANGSTA objective function (GOF). Second, the SSE assignment is transferred to the residue level. GANGSTA ignores segments with loop and coil structure and operates in sequence direction only, i.e. SSEs in a protein pair are not aligned in reverse sequential order. GANGSTA assigns residues only, if they belong to the same SSE type and allows no gaps in the same SSE. GANGSTA+ [9] is the successor of GANGSTA and employs an efficient combinatorial approach for the SSE assignment to maximize the GOF. Thus, both methods share the same optimization target on the SSE level. GANGSTA+ transfers the SSE assignment to the residue level by a heuristic point matching potential function, which minimizes the distances between aligned residue pairs. In addition, GANGSTA+ refines and extends the residue assignment from SSEs to loop and coil segments and can also detect similarities between SSEs assigned in reverse sequential order. With default setting GANGSTA and GANGSTA+ assign sequential patches of residue pairs belonging to the same SSE types only. Both allow database searches. TOPOFIT [6] is based on geometric hashing. Its online service allows database searches. TOPOFIT aligns residues including loops and coils in both sequence directions. Single residue alignment is allowed. It does not distinguish between SSE types, when maximizing the geometrical similarities of a protein structure pair. MASS [7] is an algorithm designed for multiple protein structure alignment. Residue assignments occur similar to TOPOFIT, i.e. both include loops and coils and are able to detect structure similarities in opposite sequence orientation. In addition, both algorithms allow gaps in the residues aligned to the same SSE pair. In contrast to TOPOFIT, MASS distinguishes the SSE types of the aligned residues, ensuring that they are aligned type consistently.

3. Results

3.1. *Application of different non-sequential protein structure alignment tools*

We applied TOPOFIT [6], MASS [7], GANGSTA [8] and GANGSTA+ [9] on the benchmark of five circular permuted protein pairs. TOPOFIT and MASS yielded an average SAS of 2.0 Å [SAS definition [13]: SAS = (100*RMSD) / $N_{aligned}$ with $N_{aligned}$ number of aligned residues and RMSD in Å] (see [6, 7] for details on computing time). With default setting GANGSTA+ yielded a slightly larger average SAS = 2.5 Å. The CPU time required by GANGSTA+ is less than 1s [1.6 GHz AMD/OPTERON] per protein structure pair. For GANGSTA the structure alignment results took about 10s per protein pair on average [1.6 GHz AMD/OPTERON]. With its intrinsic GOF and scoring function for the residue assignment GANGSTA employed a different protein structure alignment strategy designed to detect more distant structural similarities and is not extending the residue assignment into the loop segments. In the present benchmark the degree of structure similarity was high, but GANGSTA performed less well yielding an average SAS of 5.5 Å only. Table 1 depicts the resulting RMSD and $N_{aligned}$ of the considered methods for the benchmark of five protein structure pairs.

Table 1. Protein structure alignment results on a benchmark set of five protein structure pairs.

	protein 1	**protein 2**	**TOPOFIT**	**MASS**	**GANGSTA+**	**GANGSTA**
a)	1RIN:A(180)[a]	2CNA:_(237)[a]	152/1.09[b]	164/1.2[b]	147/1.36[b]	56/0.78[b]
b)	1GLH:_ (214)	1CPN:_ (208)	206/0.49	206/0.49	205/0.48	97/0.26
c)	1EXG:_ (110)	1TUL:_ (102)	52/1.79	60/1.9	43/2.02	41/3.5
d)	1RHG:A(145)	1BCF:A (158)	109/1.40	106/1.7	108/1.38	100/2.5
e)	1IHW:A (52)	1SSO:_ (62)	35/1.47	39/1.7	31/1.70	16/2.4

[a] PDB id: domain id according to SCOP (number of residues)

[b] number of aligned residues / RMSD in Å

3.2. *Application of different residue assignment strategies with GANGSTA+*

We generated protein structure alignments for the benchmark of the five circular permuted protein structure pairs with GANGSTA+ using five different residue assignment strategies, listed in Table 2. Hereby, we intended to capture the major differences in the residue assignment strategies of the non-sequential structure alignment methods employed in the section above. The resulting number of aligned residues and the RMSD values for the five protein pairs are displayed in Figure 1. Strategy (1) comprises the least constrained residue assignment, thus yielding the smallest (optimal) average SAS = 1.7 Å. Relatively small SAS = 2.0 Å were also obtained with strategy (2) and (3) where residue pairs were assigned uniquely as done by TOPOFIT and MASS. Additional constraints requiring that residues belonging to the same pair of SSEs must be aligned without gaps correspond to the default setting of GANGSTA+ [9]. The SAS values (average SAS = 2.5 Å) for this strategy (4) are shown as orange colored symbols in Figure 1. The alignment can be further restricted such that coils are ignored and SSEs cannot be aligned in reverse orientation [strategy (5)], which corresponds to the default setting of GANGSTA and yields the average SAS = 4.0 Å. Strategies (2) and (3) differ only in the type consistency of the aligned SSEs, which is fulfilled for (3) but not for (2). In the present application this algorithmic difference did not effect the results, while in other studies significant effects were observed varying this option [9]. Only, the results for the protein structure pair (d) 1RHG / 1BCF remained invariant for different residue assignment strategies. This contrasts with the results for the other four protein structure pairs, which exhibit large variations in the number of aligned residues and the corresponding RMSD (see Figure 1).

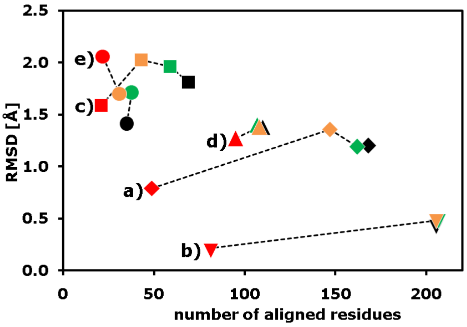

Fig. 1. Results of protein structure alignments with GANGSTA+ using different residue assignment strategies see Table 2 [black (1), green (2,3), orange (4), red (5)] on the benchmark set of five protein structure pairs listed in Table 1 (dashed lines to guide the eye). The protein pairs are (a) 1RIN [21] / 2CNA [22] (diamonds), (b) 1GLH [23] / 1CPN [24] (inverted triangles), (c) 1EXG [25] / 1TUL [26] (squares), (d) 1RHG [27] / 1BCF [28] (triangles), (e) 1IHW [29] / 1SSO [30] (circles).

Table 2. Different residue assignment strategies used for non-sequential protein structure alignment. A detailed description of the assignment options (α)-(ε) is given in the introduction.

residue assignment options /strategies	**1**	**2**	**3**	**4**	**5**
(α) includes loops and coils in the alignment	x	x	x	x	
(β) detects sequence reverse alignments	x	x	x	x	
(γ) assigns residue pairs uniquely		x	x	x	x
(δ) assigns residues of same SSEs type			x	x	x
(ε) assigns residues gapless on SSE				x	x

4. Conclusion

This study illustrates that a balanced evaluation of protein structure alignments generated with commonly used methods can be difficult due to algorithmic differences regarding the assignment strategy of residues. Bringing these differences in line is essential for a fair comparison of protein structure alignment tools. We applied the originally developed alignment program GANGSTA [8] on a set of five circular permuted protein structure pairs. GANGSTA needs about 10s CPU time [AMD/OPTERON at 1.6 GHz] per protein pair, but yielded the largest average SAS value compared to the three other considered non-sequential structure alignment algorithms. There are mainly two constraints used by GANGSTA, which contribute to the larger SAS values. GANGSTA assigns only residues belonging to SSEs and strictly ignores residues in loops and coils. Another reason for the reduced performance of the genetic algorithm in the original GANGSTA method is its tendency to not fully explore the search space. Therefore, the genetic algorithm in GANGSTA was replaced by a more reliable combinatorial approach in GANGSTA+ [9].

The residue assignment strategies used by the other alignment tools are generally less constrained than in GANGSTA [8]. The least constrained strategy (1) of fuzzy residue assignment (a single residue of one protein is assigned to more than one residue in the other protein) should be analyzed in more detail in future work. This strategy might improve detection of conserved residues in protein structure alignment. However, a comparison of the alignment quality of protein structures obtained from different alignment tools, based on the number of aligned residues and RMSD can be particularly misleading in this case. GANGSTA+, TOPOFIT and MASS performed similar on the considered benchmark (average SAS between 2.0 Å and 2.5 Å), yielding consistently very good protein structure alignments, although each of the alignment tools is based on entirely different optimization algorithms. These results illustrate the efficiency of currently available protein structure alignment approaches in solving non-sequential structure alignment problems. In a second case study, we generated protein structure alignments of the same benchmark set using GANGSTA+ with varying residue assignment strategies. These strategies have a significant effect on the number of aligned residues and the corresponding RMSD (see Figure 1). One has to keep in mind that with its default settings GANGSTA+ aims to generate alignments with consistent and complete SSE assignments and ignores loops and coils in the initial stage of optimizing protein structure alignments (see details in [9]). However, we implemented different residue assignment strategies as employed by other non-sequential structure alignment methods (see Table 2). In this way GANGSTA+ seems to reproduce the results of TOPOFIT and MASS for the considered benchmark faithfully. Allowing shared residue assignments, GANGSTA+ yields a further decrease of the resulting SAS values yielding an average SAS of 1.7 Å.

This study underlines the strength of current non-sequential protein structure alignment tools. These are capable to detect sophisticated similarities with SSEs in reverse orientation, circular permuted protein pairs and can consider shared residue assignments. However, these capabilities cause difficulties, when carrying out performance comparisons between different methods. Comparing the number of aligned residues or the RMSD can be very misleading, even if the same protein pair is considered. Since, these two properties are very essential to almost every scoring function a comparison of different methods can be carried only if the underlying residue assignment strategies are taken into account.

Acknowledgments

This work was supported by the International Research Training Group (IRTG) on "Genomics and Systems Biology of Molecular Networks" (GRK1360, Deutsche Forschungsgemeinschaft (DFG)).

References

[1] Wikipedia, [http://en.wikipedia.org/wiki/Structural_alignment_software].

[2] Holm, L., Park, J., DaliLite workbench for protein structure comparison, *Bioinformatics*, 6:566-7, 2000.

[3] Szustakowski, J., Weng, Z., Protein structure alignment using a genetic algorithm, *Proteins*, 38(4):428-440, 2000.

[4] Shindyalov, I. N., Bourne, P. E., Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *PEDS*, 11:739-747, 1998.

[5] Zhang, Y., Skolnick, J., TM-align: A protein structure alignment algorithm based on TM-score, *Nucleic Acids Research*, 33:2302-2309, 2005.

[6] Ilyin, V., Abyzov, A., Leslin, C., Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point, *Protein Science*, 13:1865-1874, 2004.

[7] Dror, O., Benyamini, H., Nussinov, R., Wolfson, H. J., MASS: multiple structural alignment by secondary structures, *Bioinformatics*, 19:95-104, 2003.

[8] Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., Knapp, E. W., Connectivity independent protein-structure alignment, *BMC Bioinformatics*, 7(510), 2006.

[9] Guerler, A., Knapp, E. W., Novel protein folds and their non-sequential structural analogs, *Protein Science*, 17:1374-1382, 2008.

[10] Bystroff, Y., Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins, *Bioinformatics*, 7:1010–1019, 2005.

[11] Teichert, F., Bastolla, U., Porto, M., SABERTOOTH: protein structural alignment based on a vectorial structure representation, *BMC Bioinformatics*, 8(425), 2007.

[12] Fischer, D., Elofsson, A., Rice, D., Eisenberg, D., Assessing the performance of fold recognition methods by means of a comprehensive benchmark, *Pac. Symp. Biocomput.,* 1:300-318, 1996.

[13] Kolodny, R., Koehl, P., Levitt, M., Comprehensive Evaluation of Protein Structure Alignment Methods, *Journal of Molecular Biology*, 346:1173-1188, 2005.

[14] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25(17):3389-3402, 1997.

[15] Henikoff, S., Henikoff, J. G., Amino acid substitution matrices from protein blocks, *PNAS*, 89:10915-10919, 1992.

[16] Day, R., Beck, D. A. C., Armen, R. S., Daggett, V., A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary, *Protein Science*, 12:2150-2160, 2003.

[17] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., Thornton, J. M., CATH- A Hierarchic Classification of Protein Domain Structures, *Structure*, 5(8):1093-1108, 1997.

[18] Pearl, F. M. G., Lee, D., Bray, J. E., Sillitoe, I., Todd, A. E., Harrison, A. P., Thornton, J. M., Orengo, C. A., Assigning genomic sequences to CATH, *Nucleic Acids Research*, 28(1):277-282, 2000.

[19] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C., SCOP, *J. Mol. Biol.*, 247:536-540, 1995.

[20] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., The Protein Data Bank, *Nucleic Acids Research*, 28:235-242, 2000.

[21] Rini, J. M., Hardman, K. D., Einspahr, H., Suddath, F. L., Carver, J. P., X-ray crystal structure of a pea lectin-trimannoside complex at 2.6 A resolution, *J.Biol.Chem.*, 268:10126-10132, 1993.

[22] Reeke, G. N., Becker, J. W., Edelman, G. M., The covalent and three-dimensional structure of concanavalin A. IV. Atomic coordinates, hydrogen bonding, and quaternary structure, *J.Biol.Chem.*, 250:1525-1547, 1975.

[23] Keitel, T., Meldgaard, M., Heinemann, U., Cation binding to a Bacillus (1,3-1,4)-beta-glucanase. Geometry, affinity and effect on protein stability, *Eur. J. Biochem.*, 222:203-214, 1994.

[24] Hahn, M., Piotukh, K., Borriss, R., Heinemann, U., Native-like in vivo folding of a circularly permuted jellyroll protein shown by crystal structure analysis, *Proc.Natl.Acad.Sci.USA*, 91:10417-10421, 1994.

[25] Xu, G. Y., Ong, E., Gilkes, N. R., Kilburn, D. G., Muhandiram, D. R., Harris-Brandts, M., Carver, J. P., Kay, L. E., Harvey, T. S., Solution structure of a cellulose-binding domain from Cellulomonas fimi by nuclear magnetic resonance spectroscopy, *Biochemistry*, 34:6993-7009, 1995.

[26] Holden, H. M., Wesenberg, G., Raynes, D. A., Hartshorne, D. J., Guerriero, V., Rayment, I., Molecular structure of a proteolytic fragment of TLP20, *Acta Crystallogr.*, 52:1153-1160, 1996.

[27] Hill, C. P., Osslund, T. D., Eisenberg, D., The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors, *Proc. Natl. Acad. Sci. USA*, 90:5167-5171, 1993.

[28] Frolow, F., Kalb, A. J., Yariv, J., Structure of a unique twofold symmetric haem-binding site, *Nat.Struct.Biol.*, 1:453-460, 1994.

[29] Lodi, P. J., Ernst, J. A., Kuszewski, J., Hickman, A. B., Engelman, A., Craigie, R., Clore, G. M., Gronenborn, A. M., Solution structure of the DNA binding domain of HIV-1 integrase, *Biochemistry* 34:9826-9833, 1995.

[30] Baumann, H., Knapp, S., Lundback, T., Ladenstein, R., Hard, T., Solution structure and DNA-binding properties of a thermostable protein from the archaeon Sulfolobus solfataricus, *Nat.Struct.Biol.*, 1:808-819, 1994.