Accuracy analysis of multiple structure alignments

Christoph Berbalk,¹ Christine S. Schwaiger,² and Peter Lackner^{1*}

¹Department of Molecular Biology, University of Salzburg, Salzburg, Austria

²Department of Biophysics and Biochemistry, Center for Biomembrane Research, Stockholm University, Stockholm, Sweden

Received 15 April 2009; Accepted 8 July 2009

DOI: 10.1002/pro.213

Published online 20 July 2009 proteinscience.org

Abstract: Protein structure alignment methods are essential for many different challenges in protein science, such as the determination of relations between proteins in the fold space or the analysis and prediction of their biological function. A number of different pairwise and multiple structure alignment (MStA) programs have been developed and provided to the community. Prior knowledge of the expected alignment accuracy is desirable for the user of such tools. To retrieve an estimate of the performance of current structure alignment methods, we compiled a test suite taken from literature and the SISYPHUS database consisting of proteins that are difficult to align. Subsequently, different MStA programs were evaluated regarding alignment correctness and general limitations. The analysis shows that there are large differences in the success between the methods in terms of applicability and correctness. The latter ranges from 44 to 75% correct core positions. Taking only the best method result per test case this number increases to 84%. We conclude that the methods available are applicable to difficult cases, but also that there is still room for improvements in both, practicability and alignment correctness. An approach that combines the currently available methods supported by a proper score would be useful. Until then, a user should not rely on just a single program.

Keywords: multiple structure alignment; alignment accuracy; reference structural alignments; structural bioinformatics

Introduction

Determining structural relationships between proteins is fundamental in biological and biochemical science to classify proteins, to analyze and predict protein function, or to support the prediction of experimentally yet undetermined protein structures.

Several computational methods have been developed for the comparison of protein structures. A few of them have been especially optimized to quickly search protein structure databases, such as PDB,¹ SCOP,² or CATH,³ for similarities to a given query 3D structure.^{4–6} The majority of the structure comparison methods have been developed to calculate the alignment of a pair of 3D structures (pairwise structure

alignment, PStA) to determine the structurally and hence potential evolutionary or functionally equivalent residues or regions. Examples for PStA methods are the frequently used methods CE,⁷ Dali,⁸ VAST,⁹ or LGA,¹⁰ or some recent developments like GANG-STA+,¹¹ RAPIDO,¹² or TopMatch.¹³ A smaller number of methods have been implemented for multiple structure alignment (MStA), for example, CBA,¹⁴ MASS,¹⁵ Matt,¹⁶ MultiProt,¹⁷ MUSTANG,¹⁸ or POSA.¹⁹

Computation of structural alignments is a far more complex problem than the computation of sequence alignments. Sequence methods arrange strings such that as much as possible similar symbols are aligned while minimizing the number of insertions and deletions. For the construction of structural alignments, the goal is to find equivalent residues in terms of the location in 3D space, whereas the type of the aligned amino acids plays a minor role. The problem further increases with the appearance of permutations in the sequential order of structural motifs and/or local or

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Peter Lackner, Department of Molecular Biology, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria. E-mail: peter.lackner@sbg.ac.at

global conformational variations, such as loops or domain movements in related proteins.²⁰

There are two basic questions that are tried to be answered by structure comparison methods. First, is there significant similarity between two structures? This is the main issue in database searches.^{21,22} And second, given a set of protein structures, where are the equivalent regions or residues located in the proteins? This second point is strongly related to alignment accuracy.

From structural alignment, one aims to draw conclusions about functional sites or similar biologically important properties. Incorrect alignments may lead to wrong conclusions. Furthermore, structure alignments are being used to characterize a protein family by a sequence profile.^{23,24} Inaccurate alignments will decrease sensitivity and selectivity when such profiles are being used to query sequence databases for remote homologous family members. Knowledge about the alignment quality to be expected from a method is desirable for a user.

In a previous study, we investigated alignment consistency and alignment correctness of several frequently used PStA methods.²⁵ The major difficulty was the lack of a reliable automatically generated golden standard of correctness. To overcome this problem, a set of curated reference alignments has been compiled. Almost all PStA methods work well when the structural similarity is high, which is the case in many protein families. Consequently, we concentrated on difficult cases taken from literature or from the SISYPHUS database of multiple structural alignments for proteins with nontrivial relationships.²⁰ It turned out that there are significant differences in the alignment correctness between different methods.

Here, we extended this investigation to the performance of MStA methods, concentrating on alignment accuracy. We included the methods POSA, MultiProt, MASS, MUSTANG, and Matt into the study. When possible their pairwise accuracy has been tested and compared to the pairwise methods used in our previous study.²⁵ For the evaluation of MStA accuracy, a new test set was derived from the SISYPHUS database. We report a numerical ranking of the methods and discuss limitations and failures of the MStA methods.

Brief overview of the investigated MStA methods

The methods differ in terms of the representation of the 3D structure, similarity measures applied, and the algorithms used to build the MStA. Alignment construction is either done simultaneously (MASS, MultiProt) for all input structures or it follows a guiding tree calculated from pairwise structure similarity (POSA, MUSTANG) or a similar iterative strategy (Matt).

MASS decomposes structures into secondary structure elements, treats them as vectors in 3D space,

and applies geometric hashing to calculate sequence order independent alignments. The SSE alignment defines an initial superimposition. Then, MASS extends this to global alignments by searching C-alpha atom positions close in space. Reducing the structural complexity to secondary structure elements and applying geometric hashing makes MASS a comparably fast method.

The other methods operate on atom coordinates, where a residue is represented by its C-alpha atom. Generally, the methods try to detect similarity in fragments first. Subsequently, different strategies are applied to build the multiple alignment. In all cases, rigid body superimposition is applied, at least, to define fragment similarity. MultiProt identifies fragments similar in all input structures simultaneously. This again provides a set of different initial superimposition solutions from which different global alignments are built. The highest scoring one is reported. POSA and MUSTANG build the MStA from a set of pairwise alignments. POSA, in contrast to MUSTANG, allows for twists between consecutively aligned fragments. Similar to POSA, Matt also allows for rotations and translations when connecting fragment pairs to a pairwise alignment while keeping the allowed transformations small to account for local structural variation only. Then, iteratively, pairwise alignments are merged to multiple alignments.

From the user point of view, the methods differ in their principle ability to account for flexibility and permutations when constructing the MStA. Furthermore, a method either provides a core alignment, that is, the regions that are structurally conserved in all members of a family or a full alignment, where in addition structurally equivalent regions in subsets of family members are calculated. For further details on features and limitations, refer to the Results section and see Table I. Other MStA methods^{14,26} were considered for this study, but frequent program failures did not allow for an adequate evaluation.

Results

The MStA methods included in the study were first classified by the way the input structures are processed, the type of the final alignment, and the format the alignment is presented to the user. For the evaluation of the alignment accuracy, three test sets representing cases of proteins that are difficult to align were prepared. The RIPC and the SISY-pairwise set²⁵ were used to evaluate the alignment accuracy if the methods were supplied with a protein pair. These two sets have been introduced in a previous work on the comparison of PStA methods.²⁵ The SISY-multiple set contains protein ensembles and was used to assess the multiple alignment accuracy. A score for alignment column correctness and a score for aligned residue pair correctness were applied for the statistical analysis.

Table I. *Properties of the Tested Methods*

Method	Alignment ^a	Sequence order ^b	Flexibility ^c	Chain usage ^d	Output ^e
POSA	Core	Yes	Yes	First	Sequence
MultiProt	Core	Yes	No	All	PDB resnbr.
MASS	Core	No	No	All	PDB resnbr.
MUSTANG	Full	Yes	No	MC rejected	Sequence
Matt	Full	Yes	Yes	Split	Sequence

^a Type of alignment, either core or full. In a core alignment only those positions that are equivalent in all input structures are reported. A full alignment additionally reports equivalent residues in subsets of the input structures.

^b Preservation of sequence order. “No” means that permutations are allowed.

^c Indicates if a method is able to handle conformational variations.

^d Chain usage describes the processing of input files. “First” means that only the first chain in the PDB file is considered as an input molecule. “All” means that all chains in the file define the input molecule. “MC rejected” means that input files that contain more than one chain are rejected. And “split” means that each chain is considered as a separate independent molecule.

^e The output of the alignment is either given only by an amino acid sequence usually in fasta format or by fully qualified PDB residue numbers (note that MultiProt omits insertion codes).

Features of the MStA methods

The MStA methods differ considerably in their abilities and technical limitations. In Table I, a classification of the MStA methods is shown. The considered properties are as follows: sequence order preservation in the resulting alignment, handling of flexible regions, and whether a method reports a core or a full alignment. A core alignment represents those regions that are equivalent in all input structures, whereas a full alignment additionally considers equivalences in subsets of two or more proteins. Other useful features are the output of PDB residue numbers for exact mapping of the alignment to the structures or the consideration of multiple chains in an input structure.

Test set features

The test sets are listed in the electronic Supporting Information and can be viewed on our web pages (<http://biwww.che.sbg.ac.at/RSA>). The RIPC set is composed of pairs of SCOP domains exposing repetitions, indels (insertions and deletions of considerable size), permutations, and conformational variability.²⁵ Each pair is annotated with R, I, P, C or a combination thereof. The SISY-pairwise and multiple set were investigated for cases exhibiting permutations, indels, or conformational variability. The structures in the SISY-sets are complete PDB chains. The SISY-pairwise cases are named by the corresponding chains. SISY-multiple set cases are named by the respective SISYPHUS alignment identifier (format is AL at the beginning followed by eight digits).

All structures have been examined for containing insertion codes or alternate location indicators or if they are composed of multiple chains. A certain method implementation might inadequately handle such a peculiarity as shown later.

Evaluation of alignment accuracy

The preparation of the test as described in the Methods section results in 23 pairs for the RIPC set, 130 pairs for the SISY-pairwise set, and 106 ensembles for the SISY-

multiple set (we refer to these as the complete sets). During the evaluation, we encountered program failures in several cases. We decided to exclude protein pairs or ensembles if one of the methods fails, as our goal was to find out how good a method’s concept works rather than how solid the implementation of the concept is. This results in 22, 98, and 61 cases for RIPC-, SISY-pairwise, and SISY-multiple, respectively. We call these the consolidated sets. All results presented later refer to the consolidated sets. Detailed results for both the types of data sets are provided as electronic Supporting Information.

MStA alignment accuracy is measured with the scores Q_C and Q_P (see Methods). They are differently sensitive to alignment errors. Q_C determines if all structures are correctly aligned in a certain column. If only one structure is shifted, the complete column is considered to be wrong. The maximum value $Q_C = 100\%$ means that the methods core alignment completely agrees with the reference core alignment. The Q_P score, in contrast, measures all pairwise correctly aligned positions with respect to the reference core alignment. Single failures are less penalized.

When applied to pairwise alignments, the scores Q_C and Q_P deliver the same result. We then use only the symbol Q_P .

Pairwise alignment accuracy

For pairwise alignment accuracy evaluation, we use updated versions of the RIPC and SISY test set (here called SISY-pairwise set) as proposed in Ref. 25. The most accurate PStA program is DALI and most accurate MStA programs are Matt and MUSTNG, all having comparable median Q_P scores. The difference in median Q_P scores between the highest and lowest accuracy method is 29.45% (28.19% between mean Q_P scores) on the SISY-pairwise set and 20% (10.59% between mean Q_P scores) on the RIPC-pairwise set. Figures 1 and 2 show box-and-whisker plots for Q_P for the two sets. In the electronic Supporting Information, detailed results for all protein pairs are provided.

Figure 1. Comparison to reference alignments in the RIPC set. Box-and-whisker plot of the distribution of the percentage of agreement Q_p obtained for the six PStA and four MStA methods.

Interestingly, for almost every pair in the test set, there is at least one method that provides a good solution, where at least half of the reference is correctly aligned. An exception in the SISY-pairwise set is the protease inhibitor pair 3ssiA-4sgbI, which contains a circular permutation and shows some conformational flexibility. In the RIPC set, the exception is the pair d1qq5a-d3chya. SCOP classifies the domains as Flavodoxin-like folds. Similar to the aforementioned example, a circular permutation is involved. The structural core superimposes well, domain d1qq5a however has some additional secondary structure elements. The

detailed structural relations between these domains have been previously discussed.²⁷ The two cases resemble a combination of problems, namely permutation plus conformation flexibility, which are beyond the capabilities of the tested PStA methods. For further examples and details, we refer to our previous analysis.²⁵

Multiple alignment accuracy

To investigate the accuracy of MStAs, we assembled the SISY-multiple set (see Methods for further details). SISYPHUS defines the three categories homologous,

Figure 2. Comparison to reference alignments in the SISY-pairwise set. Box-and-whisker plot of the distribution of the percentage of agreement Q_p obtained for the six PStA and four MStA methods.

Figure 3. Comparison to reference alignments in the SISY-multiple set. Box-and-whisker plot of the distribution of the percentage of agreement obtained for the five MStA methods. On the left side, the percentage of correctly aligned columns Q_c , and on the right side, the percentage of correctly aligned residue pairs Q_p is shown.

fold, and fragment and assigns each multiple alignment case to one of them. In the first two categories, the similarity is defined between complete chains or domains. The fragment category contains members that are related by rather small fragments or sets of fragments composed of ~20–50 amino acids representing functional equivalent regions, such as a nucleophile elbow motif. However, the reference alignments are always core alignments. This means that only those regions are included, where all structures have equivalent residues in common. The distribution of Q_c and Q_p scores for the five MStA methods is shown in Figure 3.

We observe differences in the average accuracy of the methods up to 30%. The median values for the core alignment accuracy score Q_c from high to low are 90.14% (mean 75.83%) for Matt, 85.19% (mean 71.33%) for POSA, 84.38% (mean 71.74%) for MUSTANG, 70.37% (mean 58.20%) for MultiProt, and 38.89% (mean 43.65%) for MASS. The median values for the pairwise score Q_p from high to low are 94.23% (mean 85.44%) for Matt, 92.86% (mean 85.91%) for MUSTANG, 88.39% (mean 77.19%) for POSA, 76.19% (mean 64.41%) for MultiProt, and 53.00% (mean 51.39%) for MASS.

Matt, MUSTANG, and POSA are the leading methods in the comparison. The distribution of the accuracy scores of the five MStA methods was compared using a Wilcoxon paired rank test.

Regarding Q_c , the differences between Matt, POSA, and MUSTANG are of low significance ($P > 0.02$). The differences between all other methods are considerably significant with $P < 10e^{-5}$. When apply-

ing the test to Q_p , only the difference between Matt and MUSTANG is not significant ($P = 0.5$). This indicates large differences in the method alignments.

Now, if we only consider the best method within a certain family, the median Q_c score increases to 91.67% (mean 83.41%) and the median Q_p score to 96.23% (mean 90.62%). These numbers reflect the accuracy one can reach if the optimal method is selected for each alignment problem. In the electronic Supporting Information, detailed results for all MStA methods are provided.

Common achievements and problems of the MStA methods

We were interested in parameters that influence the alignment accuracy. Therefore, the correlations between Q_c and the length of the reference alignment (EQR), the average size of the input structures (L), and the ratio EQR/L were analyzed. For EQR versus Q_c and L versus Q_c , there is no correlation. Weak but significant correlation (Spearman's rank correlation coefficient $R = 0.66$) is detected for Q_c versus EQR/L . Methods work less accurate if the equivalent region is small compared to the size of the input structures, which is especially the case in the fragment category. However, although the methods investigated are not optimized for detection of fragment similarity, some of them are also able to find such relations occasionally, for example, POSA and Matt.

To investigate the influence of the secondary structure composition, each family was assigned to either class A (mainly alpha), B (mainly beta), M (mixed alpha and beta), or F (few secondary structure

Figure 4. (A) Reference alignment of SISYPHUS family AL10063410. Left: Sketch of the different points of circular permutation (light gray stars) relative to the first protein. Right: Superimposition according to the reference alignment (only the equivalent regions are shown). There are small but numerous distortions in the β -strands. (B) Reference alignment of SISYPHUS family AL10050875. Left: Sketch of the strand swapping appearing in this family. Right: Superimposition according to the reference alignment (only the equivalent regions are shown). The arrow indicates the region where the strand swapping appears in the 3D structures. In both the families, structural relations are too complex to be properly handled by automated alignment methods. Molecular graphics were prepared with UCSF-Chimera.²⁸

elements). For details about the assignment procedure see Methods. Then, we averaged for each class the maximum reached accuracy. Class A has the highest average of about 90% followed by B and M with 84 and 82%, respectively. There is no class F member in the consolidated data set. In sum, structures containing mostly alpha helix in the core regions are more accurately aligned than others.

Collective failure was observed in the case of the homologous family AL10063410 [Fig. 4(A)], described as LuxS/MPP/ThrRS/AlaRS common domain in SISYPHUS. First, this family exhibits a number of circular permutations. Second, there are small distortions in

the helical regions, and third, the domains are embedded in relatively large chains. A similar situation is found in AL00051904, which comprise FAD/NAD(P)-binding domains. Of the MStA tested methods, only MASS does not preserve sequence order and thus has in principle the ability to correctly align such cases, but the combination of difficulties impedes this. Regarding permutation, one should keep in mind that sequence order preserving methods should be able to report at least partially correct alignments. Either the N-terminal or the C-terminal permuted part could be used by these methods in the alignment. In the worst case, the point of permutation is exactly in the

middle of the reference. Then, a sequence preserving method could do at most a 50% correct alignment. In fact, sequence order preserving methods occasionally deliver rather good alignments for families, which include members containing permutations (e.g., Matt or MUSTANG for the pseudouridine synthase family AL00100877).

The most problematic case without a permutation is observed in the calycin family AL10050875 [Fig. 4(B)], a β -barrel structure comprising strand swapping, a structural rearrangement where two adjacent strands exchange their order in the related proteins. In the region of the strand swapping the architecture changes from antiparallel to a parallel sheet and the strands additionally exchange their positions in 3D space. The most successful method for this family is POSA with $Q_C = 31\%$.

Highest alignment accuracy is found for families where the structural core superimposes nicely, the number of members is low, and the extent of indels or additional domains is low, for example, AL00101215, AL00051392, or AL00102031.

To summarize these data: there is no single parameter that determines the alignment accuracy; it is the combination of difficulties which let MStA methods fail to compute accurate alignments. However, in many cases, there is at least one method delivering a good result.

Program deficiencies

So far, only cases where all MStA methods provide an alignment have been discussed. However, in the complete SISY-multiple set, there are 45 cases where one or more methods fail. One problem is the different behavior on multichain input. MUSTANG rejects such input. POSA always takes the first chain. Matt splits multichain input into single chains and treats each chain as an independent structure, and therefore the result files cannot be evaluated. In many cases, it was not clear why a certain method fails. This indicates implementation errors in the respective method. Neglecting the multichain cases, POSA failed in 15 cases followed by MASS with 10, MUSTANG with 9, and MultiProt with 4 cases. Matt is the only program in the test that did not show any technical failures.

One problem in this study was the diverse representation of the structural alignment in the methods output. The PDB format defines residue number, insertion code, and chain identifier, which together unambiguously refer to a residue. Unfortunately, this is not being used consistently in the output of the methods. POSA only provides an internal residue number relative to the first residue in the PDB file. MASS provides a complete PDB residue reference for the alignment. MultiProt also provides residue numbers but omits insertion codes. MUSTANG and Matt only provide the corresponding sequence alignment. Matt by default also includes amino acids in the out-

put alignment, which do not have coordinates in the PDB files. For a user, it can be difficult to assign a concrete PDB residue from the alignment output. For our purpose, the methods' results were transcribed into an XML format. The files can be obtained from our web pages (<http://biwww.che.sbg.ac.at/RSA>).

Availability of test sets and evaluation software

All data set lists are provided in the Supporting Information materials. The complete accuracy evaluation environment and corresponding documentation can be obtained from our web pages (<http://biwww.che.sbg.ac.at/RSA>). It includes: (i) all structures in PDB format, (ii) an example Python script for the execution of PStA/MStA methods, and (iii) the Python script for the accuracy scores calculation. Furthermore, we provide an extension for the UCSF Chimera molecular modeling system to visualize reference and method alignments.

Discussion

Pairwise and MStA programs are valuable tools in protein science. There are many different methods based on different measures, techniques, and algorithms available for the community. In general, all methods deliver good and consistent results when the input structures have high structural similarity. However, if the structures to be aligned show large insertions/deletions, permutations, or conformational variability, different methods vary considerably in terms of alignment quality. Most critical are cases with combinations thereof, for example, permutation plus conformational variation. The average correctness of the best MStA method in the test equals 76% correct core alignment. Within the consolidated MStA test set, there are only two cases where all methods provide a totally wrong alignment. These particular protein families contain members showing sequentially permuted structural elements. Note that the test set is not representative for any kind of protein structure classification. Therefore, for an average structural family in SCOP or CATH, the alignment accuracy is expected to be higher than what we observe here.

The analysis has shown that there is room for improvements for structure comparison methods concerning usability and applicability. From the usability point of view, a more flexible handling of multichain structures, processing an unlimited number of input structures, and providing a consistent output format which exactly refers to PDB residue numbers is desirable. For broader applicability, proper treatment of local conformational variability is probably of most importance. The Matt Methods seem to be superior to other methods in this respect. In almost all test cases, there is at least one method that provides a good alignment. Combining different alignment approaches into a single program supported by an automated scoring could improve the alignment quality. Until

such a method is implemented, it seems important for a user to apply different tools and to manually compare their results.

Methods

Preparation of the test data sets

The pairwise data sets RIPC- and SISY-pairwise are taken from our previous study on PStAs methods.²⁵ The RIPC set has been updated for SCOP 1.73 and includes 23 pairs of domains. The SISY-pairwise data set is now based on SISYPHUS version 1.3. From within a SISYPHUS family, we checked all possible pairs for exposing a circular permutation or large conformational variability or large indels. If none of these features has been observed, the pair with the lowest sequence similarity was selected. This gives 130 protein pairs. The SISY-multiple dataset is derived from SISYPHUS version 1.3. For each family, we applied these rules:

1. Exclude members that are no longer in PDB.
2. Exclude members where a certain chain is used several times in the family. This happens, for example, if duplications occurred in a certain family member.
3. If there are more than two members left, add the family to dataset. This gives 106 families, composed of 3–119 members.

We call these the complete sets. During the evaluation of the PStA and MStA methods program failures occurred. We decided to remove these cases from the complete sets and call the remaining the consolidated sets. The consolidated sets are composed of 22 pairs for RIPC, 98 pairs for SISY-pairwise, and 61 for SISY-multiple.

For the RIPC set, the input structures to a certain MStA method were SCOP domains taken from the ASTRAL database.²⁹ For the SISY-sets, the respective complete PDB chains were used. Note that SISYPHUS occasionally defines multichain constructs. Then, at least one family member consists of two or more chains. In this case, all corresponding chains were collected into the input PDB file.

Assignment of secondary structure classes

The goal for the assignment was to determine the predominant type of secondary structure in the aligned regions as defined by SISYHPUS rather than to classify the complete input structures. Therefore, we could not use predefined class assignment by SCOP or CATH. Moreover, the input for a certain MStA program was not necessarily a domain, which additionally impedes to adopt SCOP classes.

Consequently, the program STRIDE³⁰ was used to assign secondary structural elements in the input structures. A rather simple scheme was applied to assign the aligned regions to A (mainly alpha), B

(mainly beta), M (mixed alpha and beta), and F (few secondary structure elements).

We define a as the percentage of alpha and b as the percentage of beta, both with respect to length of the aligned regions in SISYHPUS. Then, class F is assigned if $(a + b) < 40\%$. Otherwise, class A is assigned if $a/(a + b) \geq 75\%$ or B if $b/(a + b) \geq 75\%$. The remaining cases are considered as class M.

Calculation and parsing of PStA and MStA method results

When available, we used the standalone version of the alignment methods, otherwise the corresponding web service. In any case the default parameter settings of the different methods are used. Unfortunately, the output of some methods does not provide the PDB residue numbers but only sequence information. In that case the sequence information was mapped back to the PDB residue number by substring matching or by a semiglobal Dynamic Programming alignment. Then, the alignments with full PDB residue number information have been stored in a XML format as described in our web pages (<http://biwww.che.sbg.ac.at/RSA>).

Measures of alignment correctness

Two alignment quality scores as proposed by different authors^{31,32} were applied. The first score, Q_C , (previously called CS) counts the percentage of columns identical between the method and the reference alignment, C_{columns} . The MStA reference alignments are core alignments of length L_R . Score Q_C thus reflects how much of the alignment core is reproduced by a method.

$$Q_C = \frac{C_{\text{columns}}}{L_R} * 100 \quad (1)$$

Q_C is a rigorous measure. If only one structure in the input set is shifted in the alignment, Q_C is zero. The second score, Q_P , (previously called developer score³² or SPS³¹) counts for correctly aligned residue pairs.

$$Q_P = \frac{C_{\text{pairwise}}}{R_{\text{pairwise}}} * 100 \quad (2)$$

Here, C_{pairwise} is the number of correctly aligned residue pairs and R_{pairwise} is the number of aligned residue pairs in the reference alignment. For core reference alignments $R_{\text{pairwise}} = L_R * n_S * (n_S - 1)/2$, where n_S is the number of proteins in the input set. In case of comparing PStAs, Q_C and Q_P deliver the same result. We use the symbol Q_P when evaluating pairwise alignments. The SISY-pairwise set contains entries with alternative reference alignments for a certain protein pair. In that case, we compare to all alternatives and report the highest Q_P value.²⁵

Acknowledgment

The authors thank F.S. Domingues and C.X. Weichenberger for helpful discussion.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
2. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
3. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res* 37:D310–D314.
4. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst D* 60:2256–2268.
5. Carugo O, Pongor S (2002) Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J Mol Biol* 315:887–898.
6. Rogen P, Fain B (2003) Automatic classification of protein structure by using Gauss integrals. *Proc Natl Acad Sci USA* 100:119–124.
7. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11:739–747.
8. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233:123–138.
9. Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. *Curr Opin Struct Biol* 6:377–385.
10. Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31:3370–3374.
11. Kolbeck B, May P, Schmidt-Goenner T, Steinke T, Knapp E-W (2006) Connectivity independent protein-structure alignment: a hierarchical approach. *BMC Bioinformatics* 7:510.
12. Mosca R, Brannetti B, Schneider TR (2008) Alignment of protein structures in the presence of domain motions. *BMC Bioinformatics* 9:352.
13. Sippl MJ, Wiederstein M (2008) A note on difficult structure alignment problems. *Bioinformatics* 24:426–427.
14. Ebert J, Brutlag D (2006) Development and validation of a consistency based multiple structure alignment algorithm. *Bioinformatics* 22:1080–1087.
15. Dror O, Benyamini H, Nussinov R, Wolfson H (2003) MASS: multiple structural alignment by secondary structures. *Bioinformatics* 19 (Suppl 1):i95–i104.
16. Menke M, Berger B, Cowen L (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol* 4:e10.
17. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. *Proteins* 56:143–156.
18. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins* 64:559–574.
19. Ye Y, Godzik A (2005) Multiple flexible structure alignment using partial order graphs. *Bioinformatics* 21:2362–2369.
20. Andreeva A, Prlic A, Hubbard TJP, Murzin AG (2007) SISYPHUS—structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res* 35:D253–D259.
21. Sierk ML, Pearson WR (2004) Sensitivity and selectivity in protein structure comparison. *Protein Sci* 13:773–785.
22. Novotny M, Madsen D, Kleywegt GJ (2004) Evaluation of protein fold comparison servers. *Proteins* 54:260–270.
23. Bernardes JS, Dávila AMR, Costa VS, Zaverucha G (2007) Improving model construction of profile HMMs for remote homology detection through structural alignment. *BMC Bioinformatics* 8:435.
24. O'sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* 340:385–395.
25. Mayr G, Domingues FS, Lackner P (2007) Comparative analysis of protein structure alignments. *BMC Struct Biol* 7:50.
26. Lupyán D, Leo-Macias A, Ortiz AR (2005) A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* 21:3255–3263.
27. Ridder IS, Dijkstra BW (1999) Identification of the Mg²⁺-binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY. *Biochem J* 339:223–226.
28. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612.
29. Chandonia J-M, Hon G, Walker NS, Conte LL, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res* 32:D189–D192.
30. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23:566–579.
31. Thompson JD, Plewniak F, Poch O (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res* 27:2682–2690.
32. Sauder JM, Arthur JW, Dunbrack RL (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40:6–22.