Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Assessing 3D scores for protein structure fragment mining

This article was published in the following Dove Press journal:
Open Access Bioinformatics
10 July 2010
Number of times this article has been viewed

Frédéric Guyon[1]
Pierre Tufféry[1,2]

[1]MTi, INSERM UMR-S973, Université Paris Diderot-Paris 7, Paris, France; [2]RPBS, Université Paris Diderot-Paris 7, Paris, France

Abstract: Quantifying the 3D similarity between two proteins is a difficult task that has motivated the assessment of several 3D scores. New developments in protein modeling and analysis have led to the emergence of new interest towards mining structures at the local level. We assess in the context of fragment mining several dissimilarity scores. We revisit the concept of mirror conformation previously introduced at the level of complete structures and extend it to the more local level. We also consider an explicit criterion measuring the fragment boundary discrepancies. Whereas classical criteria such as the root mean square deviation (RMSd) fail to identify similar shapes in a consistent way, we show that local mirror and boundary mismatch filtering greatly supplements classical scores to select significant matches. The geometrical conditions defined by such criteria can be considered as signatures of fragment similarity. Furthermore, it is possible to tune the degree of similarity depending on the size of the mirrors accepted. This results in a more intuitive perception of the concept of similarity, and opens new perspectives for the rapid mining of large collections of structures.

Keywords: protein fragments, similarity, distance, mining

Introduction

Similarity is a central concept for the analysis of protein structure and function, underlying protein structure classification,[1–5] protein structure prediction,[6–7] modeling performance,[8–11] and protein function annotation.[12,13] Depending on the goal, similarity can be considered at the level of complete structures, considering classical structural alignment[14–18] or un-sequential alignment,[19] at the level of fragments,[20,21] or for the search for motifs involving amino acids not consecutive in the sequence, using atomic positions[22–27] or molecular shape.[28,29]

In a general manner, the focus of similarity search has progressively moved during the recent years from the complete protein domain level to the more local level, both for functional annotation and structure prediction. At the level of fragments, the scope of this study, there is evidence that recurrent conformations occur at the local level in protein structures.[21,30,31] Indeed, fragment seeds are used by some approaches to align complete structures[14,18] and fragment assembly has become a major paradigm in structure prediction.[32] Presently, local structure is the level at which modeling encounters limits in order to get the most accuracy related to significant functional arrangements.[33] As a consequence, the efficient search for similar fragments in large structure collections has become a concern. It is unclear, however, how the approaches developed for complete structures are relevant for shorter sizes, since the objectives in the search differ. For instance, whereas local deviations are expected at the level of complete

Correspondence: Pierre Tufféry
RPBS, Université Paris Diderot-Paris 7, INSERM UMR-S973 and F75205, Paris, France
Tel +33 157 278 374
Fax +33 157 278 372
Email pierre.tuffery@univ-paris-diderot.fr

structures, these cannot, in general, be so much tolerated for fragments. It is often important that the fragments collected as significantly similar not only exhibit a globally similar shape, but also that atomic details of the backbone, in particular with respect to the preservation of hydrogen bond patterns be quantified. As well, one would usually prefer that similar fragments result in similar side chain orientations. In a general manner, it should be possible to avoid too large local deviations, and boundary conditions compatible with loop closure should also be taken into account.

Considering measures of protein structure similarity, two predominant categories can be distinguished. The first relies on the structural alignment of the structures from which pairs of corresponding residues are identified. One widely used criterion in this context is the alpha-carbon root mean square deviation (cRMSd) criterion. This long used criterion has well known flaws, long discussed in the literature,[34] among which, cRMSd dependence on the alignment length, possibly large deviations between aligned positions despite low global values, and above all, a poor classification performance for medium range cRMSd values. To address these limitations, several studies have proposed different normalizations, several assuming the proteins as globular, to make the cRMSd independent of the fragment size.[35–38] For instance, Maiorov and Crippen[35] have proposed to consider the radius of gyration of the proteins, whereas Carugo and Pongor[36] have proposed a normalization of the cRMSd taking into account the number of amino acids. Another measure introduced in the late 90s in the context of the Comparative Assessment of Structure Prediction (CASP) is the GDT-TS.[8] An underlying assumption of this score is that the structures to compare should be close. Given the superimposition of two protein models, it combines the fraction of the residues aligned at different distance thresholds to produce a score between 0 and 1, where values of 1 would correspond to perfect modeling. Nevertheless, this score still faces the flaw that it achieves poor discriminative performances for high GDT-TS values,[39] ie, for very close models, and to overcome this difficulty Sadreyev and coworkers[10] have very recently proposed to add to the GDT-TS score a penalty term to repulse non-equivalent residues.

The second category of scores does not rely on a prior superposition. It is based on the comparison of the inter-atom distance matrices of each protein.[40] Among these, the DALI score[41] is probably the most well known. It is based on the comparison of the contact matrices of each protein. This score is used to perform a structural alignment and then to derive a Z-score with normalization depending on the observed distribution or protein size and inter-residue distance. It has been applied successfully to the large scale detection of protein similarity in the well known DALI server.[42] Although these scores are easy to compute, one flaw is that they cannot distinguish between mirror images because a structure and its mirror images have the same distance matrices, therefore such a score can be zero for very dissimilar fragments.

Last, but not least, concerns about structural similarity are not only qualitative, but also quantitative. Several groups have tackled the difficult question of the significance of the 3D alignment of structures. In the mid 90s, Maiorov and Crippen[43] proposed that two proteins are significantly similar when their cRMSd is smaller than that obtained from the mirror conformation. As a consequence, a significant level of cRMSd is the value below which the mirrored cRMSd cannot be less than the cRMSd. In the late 90s, Levitt and Gerstein[44] introduced a score based on inter-protein distance matrix to assess the statistical significance of the 3D alignments, and carried out an extreme value distribution analysis as is performed to study sequence alignment score significance. This score has been further adapted by Zhang and Skolnick into the TM-Score.[11] Very recently, Wrabl and Grishin[45] have the statistics associated with different scores and proposed an approach involving the cRMSd, the radius of gyration and the thinnest molecular dimension.

Here, we study the properties of various scores applied to the mining of collections of structures to search for small fragments. We consider dissimilarity scores between proteins fragments from 8 to 20 residues based on atom coordinates of alpha-carbons. Two families of scores are studied: scores based on a superimposition of the fragments, named deviation scores, and scores comparing inter-atom distance matrices, named distortion scores. We revisit and extend the concept of mirrors first presented by Maiorov and Crippen[43] to assess the significance of RMSd on globular proteins to the level of fragments. We use local mirrors as geometric criteria to assess the similarity of fragments. We also consider an explicit criterion measuring the fragment boundary discrepancies. We show with large benchmarking, and we illustrate in some test cases, that both criteria can be combined to greatly improve the accuracy and the efficiency of the 3D-scores for short protein fragment mining.

Material and methods

Datasets

In order to assess the similarity criteria we have used a reduced benchmark of 976 protein domains subset SCOP 1.37 at 40% identity (PDB40), as selected by Lindahl and Elofsson.[46] This benchmark consists in domains with few errors, well

diversified, and covering the PDB although not too large. We have selected 12 domains consisting in three samples from the first four classes of SCOP (all alpha, all beta, alpha + beta, alpha/beta), and we have performed a search of all the fragments of these 12 test domains against the 976 domains of the benchmark, forbidding self comparisons. We have considered fragment length between 8 and 20 residues, randomly selecting one length for each protein pair. This resulted in more than 300 million fragment comparisons from which have been computed the four dissimilarity scores, the presence of global and local mirrors of length 5, 7, 9, 11, and varying values of boundary conditions penalty. Such calculations are tractable since we do not perform the effective superimposition of the fragments but only compute score values. Typical search times are on the order of only 12 hours for 300,000,000 comparisons. To generate the illustrations the size of the data was, however, out of the memory capabilities of the programs, and we have sampled subsets of 3,000,000. Tests show that the resulting curves are independent on such sampling.

In order to illustrate the behavior of some scores, we have also performed searches against the complete ASTRAL compendium[4] filtered at 40% sequence identity.

Deviation scores

Superimposition calculation

Given two sets of aligned coordinates, the superimposition consists in computing the optimal rigid body translation and rotation which moves one set of atoms onto the second one. cRMSd is the most classical one which minimizes the sum of squared euclidean distances between aligned pairs of atoms. The two sets of atom coordinates are represented by two $3 \times N$ matrices X^1 and X^2, with X^1_{ij} (resp. X^2_{ij}) denoting the i-th coordinate of the j-th atom. The coordinate deviation is, using the Frobenius norm of matrices. The first step of superimposition consists in centering the two structures. After this translation, we have: $\Sigma_j X^1_{ij} = \Sigma_j X^2_{ij} = 0$ for all i, $1 \leq i \leq 3$. Next step is to find the rotation matrix which minimizes:

$$\min_R \| RX^1 - X^2 \|^2 \tag{1}$$

(Mathematically, a rotation matrix is characterized by $R^TR = Id$ [R is a unitary matrix] and $\det(R) = 1$.) And then the cRMSd is given by:

$$cRMSd\left(X^1, X^2\right) = \sqrt{\frac{1}{N} \| RX^1 - X^2 \|^2}$$

Many algorithms have been proposed in the past to solve this problem.[47–53] The solution of (1) can be formulated as follows:

$$cRMSd^2\left(X^1, X^2\right) = \frac{1}{N}\left(\| X^1 \|^2 + \| X^2 \|^2 - 2\left(s\sigma_1 + \sigma_2 + \sigma_3\right)\right) \tag{2}$$

where σ_1 are the three singular values of the 3×3 matrix $X^1 (X^2)^T$ with $0 \leq \sigma_1 \leq \sigma_2 \leq \sigma_3$ and s is the sign of the determinant of $X^1 (X^2)^T$. The latter solution implies the diagonalization of the matrix in order to compute the three singular values. The now classical approach represents rotations with quaternions.[52] Quaternion method requires the computation of the maximal eigenvalue of a 4×4 matrix which is more efficiently obtained, for instance, using a power iteration method. However, the formulation (2) provides useful information for our analysis (see the section related to mirror detection).

cRMSd normalization

Following Maiorov and Crippen,[35] we have considered a normalization which is not based on statistical considerations but on a geometrical measure of the fragments, the radius of gyration, which is given by:

$$\rho(X) = \sqrt{\frac{1}{N} \| X \|^2}$$

with X centered. From (2) we have:

$$cRMSd^2\,(X^1, X^2) \leq \rho^2(X^1) + \rho^2(X^2)$$

Hence, the normalized cRMSd (nRMSd):

$$nRMSd\left(X^1, X^2\right) = \frac{cRMSd\left(X^1, X^2\right)}{\sqrt{\rho^2\left(X^1\right) + \rho^2\left(X^2\right)}} \tag{3}$$

This dissimilarity score ranges from 0 to 1. Note that Maiorov and Crippen proposed a slightly different normalization:

$$sRMSd\left(X^1, X^2\right) = \frac{cRMSd\left(X^1, X^2\right)}{\sqrt{2\rho^2\left(X^1\right) + 2\rho^2\left(X^2\right) - cRMSd\left(X^1, X^2\right)}}$$

According to Maiorov and Crippen,[35] this score is independent of scaling and is minimal for proteins with equal radii of gyration.

Distortion scores

Distance Matrix Distortion

We first consider the Distance Matrix Distortion (DMD) score introduced by Levitt.[40] In essence, it is conceptually close to the cRMSd, but applied to the difference between two matrices of internal distances. As the cRMSd, it varies in theory from 0 to infinity.

$$DMD\left(X^1, X^2\right) = \sqrt{\frac{2}{N\left(N-1\right)} \sum_{i,j>i} \left(D^1_{ij} - D^2_{ij}\right)^2}$$

with N = number of $\alpha-$carbon atoms, D^1_{ij} (resp. D^2_{ij}) distance between atoms i and j of the structure X^1 (resp. X^2).

Mean Distance Matrix Distortion

In order to compare with the nRMSd, we also assess a normalized version of the DMD (mDMD), to vary between 0 and 1:

$$mDMD\left(X^1, X^2\right) = \frac{2}{N\left(N-1\right)} \sum_{i,j>i} \frac{|\, D^1_{ij} - D^2_{ij} \,|}{|\, D^1_{ij} + D^2_{ij} \,|}$$

Detection of mirror symmetry

The coefficient s in (2) carries a very interesting piece of information: whenever it is negative, the mirrored structure of X^1 is better superimposed on X^2. If S is a symmetry matrix (hence det(S) = −1) about any axis, then we have:

$$cRMSd\,(SX^1, X^2) \leq cRMSd\,(X^1, X^2)$$

This fact has been exploited by Maiorov and Crippen[44] to derive a significant level of cRMSd between two globular proteins. For any symmetry transformation given by a matrix S (S^T = Id and det(S) = −1), we have:

$$cRMSd^2\,(SX^1, X^2) = cRMSd^2\,(X^1, X^2) + 4s\sigma_1\,(X^1\,(X^2)^T)$$

and for any scores based on distance matrices, we have:

$$d(SX^1, X^2) = d(X^1, X^2)$$

A fragment represented by a centered coordinate matrix X^2 is mirror similar to the fragment given by X^1 whenever:

$$\det(X^1(X^2)^T) < 0$$

We say that a structure X has a mirror if, relative to a query structure, X is better superimposed onto the query after mirroring.

Detection of local mirror symmetry

We say that a fragment X^2 contains a local mirror (relative to X^1) or is locally mirror-similar to X^1 whenever the fragment sub-matrix X^2_{loc} (resp. X^1_{loc}) composed of a continuous subset of rows of X^2 (resp. X^1) whenever:

$$\det(X^1_{loc}(X^2_{loc})^T) < 0$$

We have considered odd mirror lengths varying from 5 to 11 residues, called in the following l-mirrors (l = mirror length). Mirrors of length 3 do not exist, as 3 points are planar and imply a null determinant.

Boundary conditions

Optimal superimposition averages the deviation between two fragments, and boundary discrepancies cause only a moderate increase to the cRMSd and other deviation scores. Fragments with a significantly low cRMSd or distortion can present boundary mismatches with the query. The same kind of discrepancy can also occur for DMD scores, the deviation between the relevant distances being average over a large amount of distances. To tackle this difficulty, we also consider a second geometrical condition to measure distance mismatches between the two end points of the queried and searched fragments.

$$s_{ij} = \frac{|\, D^1_{ij} - D^2_{ij} \,|}{|\, D^1_{ij} + D^2_{ij} \,|}$$

$$d_{bound} = s_{1N} + s_{1;N-1} + s_{2;N} + s_{2;N-1} + s_{3;N} + s_{3;N-2}$$

The boundary conditions are satisfied whenever the boundary distance is below a given threshold.

ROC analysis

ROC curves[54,55] represents the False Positive Rate (FPR) versus the sensitivity (or True Positive Rate) relatively to the geometrical conditions. The FPR is the proportion of positive results (dissimilarity below a given cutoff) among geometrically dissimilar fragments (not passing the mirror and boundary conditions). It is an estimation of the probability of accepting a dissimilar fragment:

$$FPR = Pr\,(d \leq d_0 |\, geom.cond.\,fulfilled)$$

FPR is also equal to one minus the specificity.

The sensitivity (or TPR) is the proportion of selected fragments among those having a correct geometry:

$$TPR = Pr\,(d \leq d_0 |\, geom.cond.not\ fulfilled)$$

Overall efficiency of a dissimilarity score is measured by the area under the ROC curve (AUC). AUC is the probability that a positive fragment fulfilling the geometrical conditions has a better score than a negative one,[54] and we use this criteria as a sensitivity measure of the 3D distances to the geometry of the fragments. A straight line (FPR = TPR) and an AUC = 0.5 means that the score is completely independent of the geometrical conditions, and the fragment mining process is no more effective than a purely random selection.

Precision-recall curves[55] can also be used when there is a large skew between positive and negative fragment proportions. This is not the case for the mirror conditions. For instance, globally mirrored and non-mirrored fragments are

equally distributed. However, when considering boundary conditions, positive and negative fragment proportions are highly unbalanced with a small number of positive ones. Precision-recall curves represents the precision versus the recall. Precision is given by:

$$Prec = Pr\,(geom.cond.fulfilled\,|d \leq d_0)$$

and the recall is equal to the sensitivity.

Geometrical significance

Maiorov and Crippen[43] have considered the presence of global mirror as a mean to assess the significance of cRMSd between globular proteins. Figure 1 shows that mirror cannot be present below a fixed cRMSd value. They also estimated this cRMSd threshold for globular proteins scaled to have their radius of gyration set to one, and obtained $(2/3)^{1/2}$. We generalize this definition of significance by also considering local mirrors and boundary conditions, and call it geometrical significance. The geometrical significance depends on the mirror length and on the boundary threshold.

In the next section, we study the geometrical significance of the 3D dissimilarity scores by evaluating the sensitivity and specificity of scores to geometrical conditions.

Results

Local mirror occurrences

We first discuss the general behavior of the scores and the occurrence of mirror conformations. Figure 1 illustrates the results obtained by systematic superimposition of fragments over all possible fragments of 976 protein domains, for the four scores. Not surprisingly, one notes a positive correlation between the scores. For instance, the correlation factor between the cRMSd and DMD (Figure 1a) is equal to 0.88. Normalized scores (Figure 1b) are less correlated with a correlation factor of 0.70, and this results in larger clouds of points. The correlations are slightly increased when considering presence of mirrors (correlation factors of 0.9 and 0.75 in Figures 1a and 1b, respectively) and much more increased when taking boundary conditions into account (B.C. < 0.25) (correlation factors of 0.95 and 0.9 in Figures 1c and 1d, respectively). Looking at the localization of the mirrors, one observes that the matches corresponding to global mirror conformations of the query (blue regions) are associated with large deviation score (cRMSd, nRMSd) values and that for low values, no mirror exists: there is a cutoff value below which fragments cannot be mirror-similar. This is fairly comprehensible since a mirror implies differences in the geometry that result in cRMSd penalty. Looking at the local mirrors, we observe that the occurrences of local mirrors (here 7-mirrors) encompass the area of the observed global mirror, but also outflanks this area towards lower cRMSd values. For the deviation scores, these results suggest that considering mirror effects is discriminate for medium range values, and that considering more local mirror effects will result in focusing the search for similar fragments towards low score values. Note, however, that the plot can be misleading since outside the low score regions, there is in fact an equal distribution of mirrored and non-mirrored fragments.

The DMD and mDMD dissimilarity scores do not exhibit such a clear cutoff for global mirror. Comparing the deviation and distortion scores, we first observe that two similar fragments can have small DMD scores and simultaneously large cRMSd values. This implies that distortion scores are, compared to deviation scores, less efficient for fragment mining at intermediate level of similarity. Secondly, distortion scores are efficiently supplemented by local mirrors: matches with no local mirror have significantly lower DMD scores. For all four scores we observe (Table 1) that filtering the search, by forbidding more and more local mirrors, effects results by significantly increasing similarity and the mean score is significantly lower for pairs of fragments with no local mirrors at all. However, for distortion scores, only for the very small values of the scores do we observe no mirror. Indeed, distortion scores theoretically cannot distinguish between global mirror and non-mirror superimposition. Although it is in theory possible that the absence of mirror conformation comes from the fact that we could not perform an all-proteins-against-all experiment, this fraction is well populated. This suggests that in proteins, mirror effects do not occur for very similar fragments.

Figure 2 reports the results of a ROC analysis. The AUCs are reported in Table 2. The ROC curves shows that DMD and mDMD are almost completely insensitive to global mirror, as are the cRMSd and nRMSd for larger values of the scores. The AUC increases when forbidding smaller mirrors, up to values of over 0.9, which shows that all scores are sensitive and correlated to local mirror elimination. A lower mirror size implies a more stringent geometrical condition. This shows that the absence of local mirror can be considered as a signature of similarity between fragments.

Fragment boundary conditions

From our assessment (not shown), values of the criterion resulting in satisfactorily overlapping boundaries are below 0.5. Figure 1 (middle) illustrates the distribution of fragments with various boundary conditions. It mostly shows

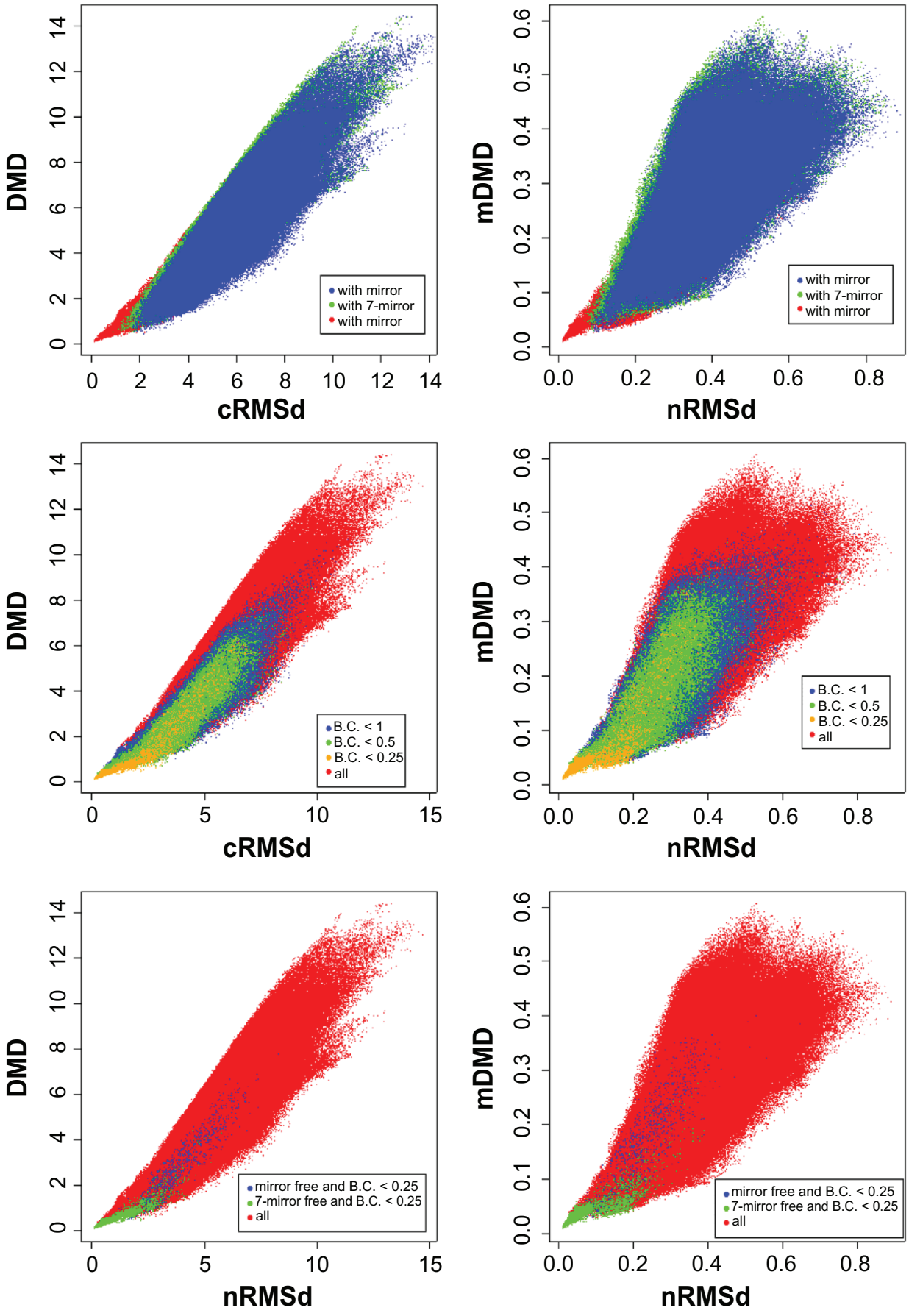

Figure 1 Mirror effects and boundary conditions for deviation and distortion scores.

Table 1 Scores mean and standard deviation for different l-mirror conditions (global – complete fragment length, 11, 9, 7, 5)

		All	Global	11	9	7	5
cRMSd	mean	5.23	4.94	4.18	3.79	3.51	2.39
	s.d.	1.528	1.551	1.270	1.363	1.388	1.548
nRMSd	mean	0.367	0.346	0.326	0.296	0.277	0.192
	s.d.	0.10	0.10	0.10	0.10	0.10	0.11
DMD	mean	4.47	4.33	3.58	3.26	2.77	1.81
	s.d.	1.79	1.82	1.59	1.65	1.59	1.54
mDMD	mean	0.2715	0.2640	0.2328	0.2116	0.1817	0.1176
	s.d.	0.084	0.0890	0.0892	0.0937	0.0924	0.0813

that boundary conditions alone are not specific enough to select structurally similar fragments. Selecting fragments with low boundary deviations is not a sufficient condition for similarity. Indeed, and not surprisingly, it is possible to reach, for fragment less than 20 residues, cRMSD deviations by over 5Å with boundary score less than 0.5. We also observe that some fragments with low cRMSd values can show boundary discrepancies. However, we observe a different behavior for deviation and distortion scores. Whereas it is possible to have boundary discrepancies even for low deviation score values, this is not observed for distortion scores. This is not surprising, since the boundary distances corresponding to fragment limits enter directly into the score formulation. Finally, we observe that fragments which fulfil more stringent boundary conditions have on average significantly lower scores. This is clear from Figure 3 that shows the precision-recall analysis associated with the boundary conditions. Overall, all scores show the same tendencies,

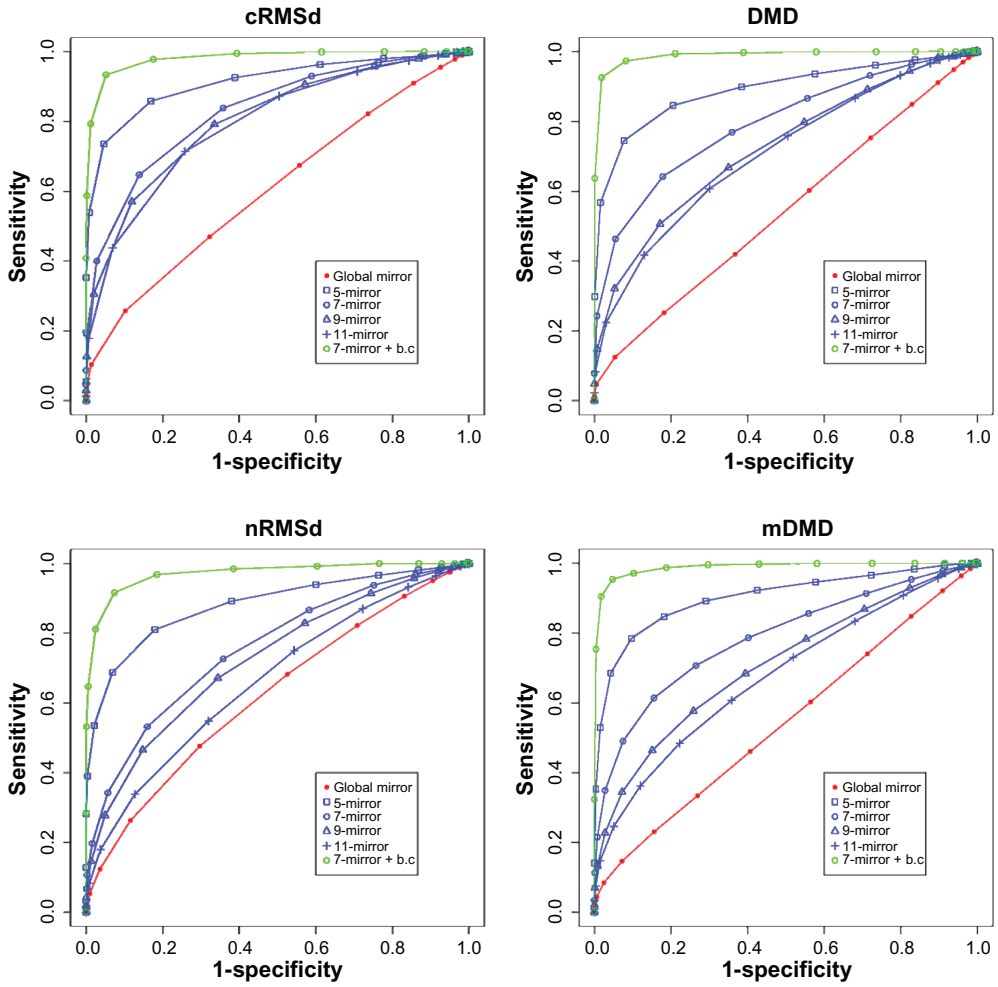

Figure 2 ROC analysis of mirror conditions.

Table 2 AUC for different l-mirrors (global – complete fragment length, 11, 9, 7, 5) and one combination: 7-mirror and boundary conditions less than 0.4

AUC	Global	11	9	7	5	7;0.4
cRMSd	0.6069	0.7940	0.8068	0.8323	0.9139	0.9814
nRMSd	0.6225	0.6638	0.7239	0.7556	0.8816	0.9714
DMD	0.5429	0.7080	0.7238	0.7941	0.8920	0.9890
mDMD	0.5434	0.6764	0.7098	0.7873	0.9035	0.9884

and the distortion scores (DMD and mDMD) show the best performance in detecting boundary mismatches. One notes however, the poorer performance of the cRMSd compared to the nRMSd. The cRMSd appears to be more tolerant to boundary mismatches and therefore less sensitive to boundary conditions. It seems that the normalization including the radius of gyration has a marked effect on the boundaries, although the relationship remains to further explore.

Combining mirror occurrence and fragment boundary conditions

Figure 1 (bottom) illustrates the combination of both mirror and boundary effects. It shows that combining mirror and boundary conditions results in a very tight structural similarity. Indeed, dramatic effect is observed by simply combining the removal of global mirrors and boundary conditions. For all scores, a very focused similarity area can be delimited considering local mirrors. As shown in Figure 2, close to perfect behavior is obtained combining 7-mirror and boundary conditions. All dissimilarity scores are highly specific and sensitive to combined geometrical conditions, the best performance being obtained for the mDMD score.

To assess in real case the feasibility to tune the search for similar fragments we have performed searches against the complete Astral compendium at 40% sequence identity. Figure 4 illustrates, for three different types of fragments – a beta-hairpin (top), a random coil (second line), and a helix-turn-strand motif (third line) – the effect of accepting or not local mirrors. All the matches displayed have cRMSd values less than 2.5Å. It is visible from left to right that filtering of the matches by increasingly not accepting more local mirrors results in matches closer to the query. Although only the alpha carbons were considered, the orientations of all the backbone atoms are also closer to that of the query, suggesting that both compatible main chain and side chain directions are retained. Finally on the last line, we illustrate the effect of filtering on the boundary conditions for fragment having no 7-mirror. Remarkably, some matches for larger cRMSd values but having no 7-mirror exist as well (bottom right). However their boundary scores are rather large.

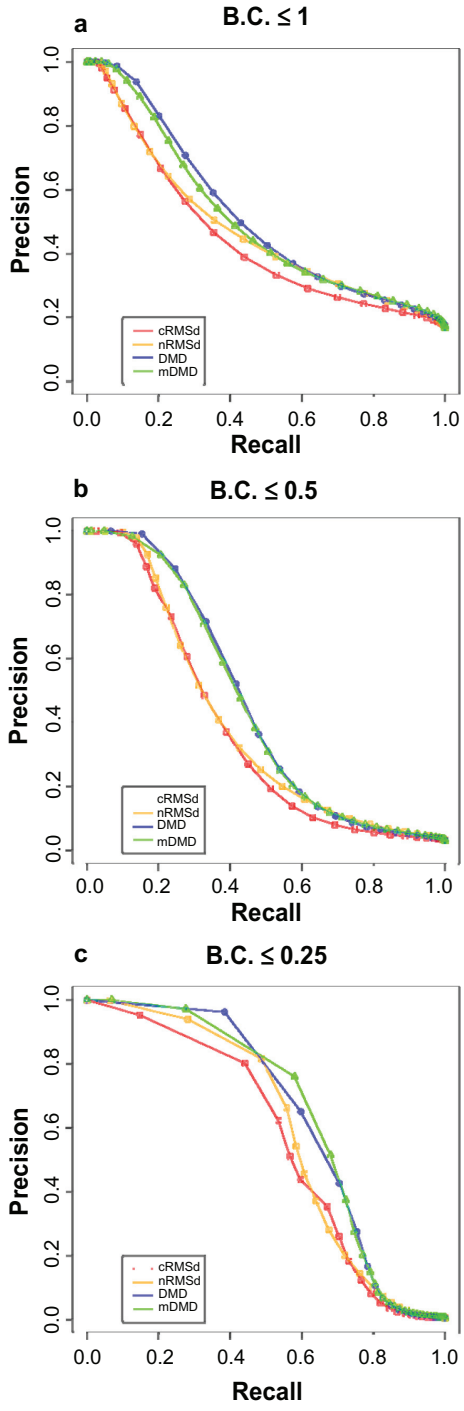

Figure 3 Precision-recall analysis of boundary conditions (B.C.).

This highlights the importance of considering both absence of mirror and boundary condition satisfaction.

Discussion

The results of the present assessment show that distortion scores (DMD, mDMD) can perform as efficiently as deviation scores (cRMSd, nRMSd) in the context of fragment mining. In such context, the filtering of the matches on the basis

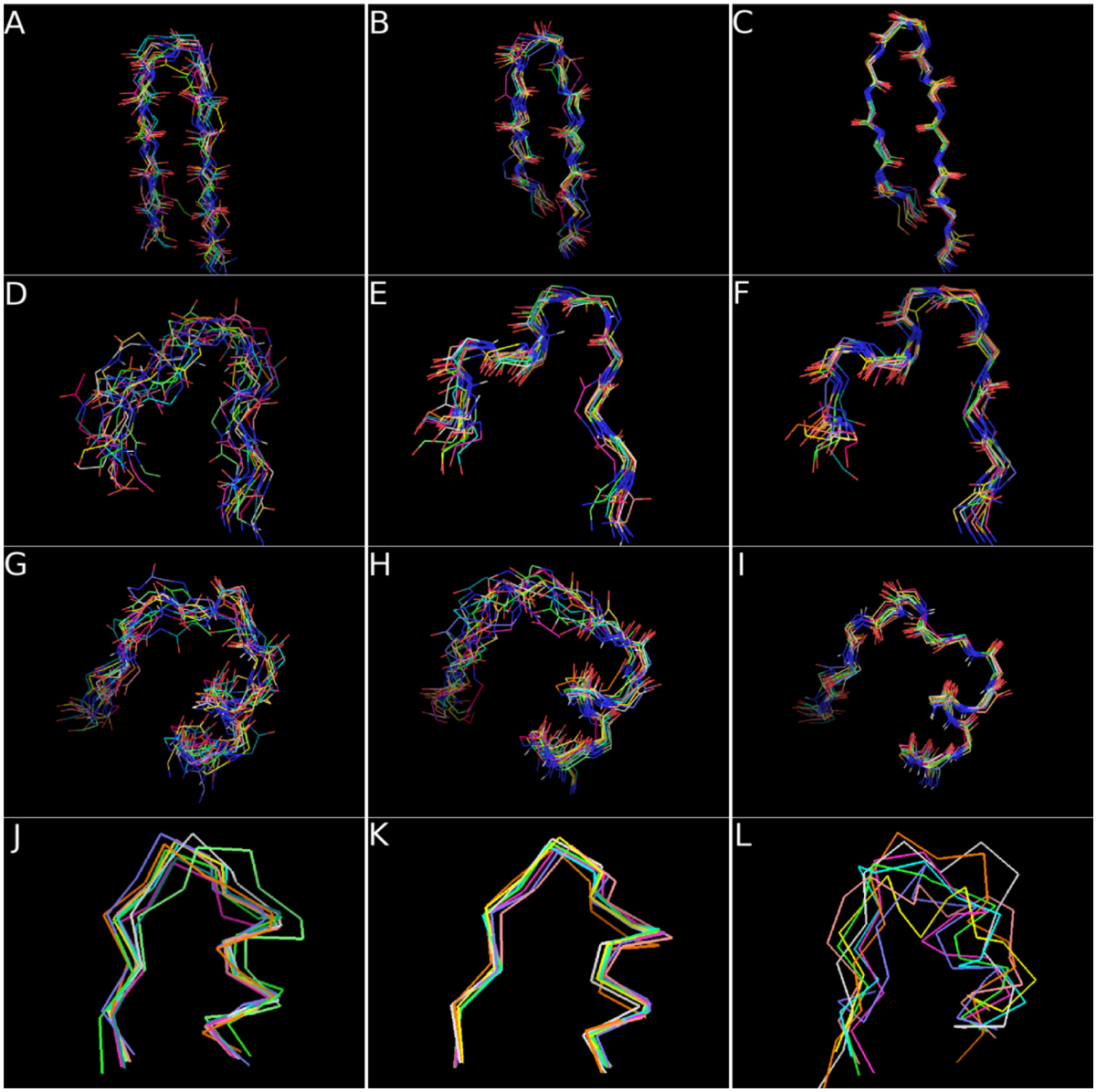

Figure 4 Considering local mirrors. For three different topologies of loops – beta-hairpin (top) random coil (middle) and alpha-turn-strand (HTB) (bottom – we illustrate the results obtained accepting mirror over complete fragment (left), 7 residue mirror only (middle) or no 7 residue mirror (right). For HTB, no fragment was found accepting mirror over complete fragment, we illustrate by an 11 residue mirror. All the fragments have RMSd values less than 2.5Å. On the last line (left, middle), we also report illustrate the impact of the boundary conditions. On the right the fragments that have no 7-mirror, but have larger RMS deviations (2.5–3Å).

of local mirror elimination is an efficient way to focus on matches that correspond more intuitively to the concept of similarity: both local shapes and fragment conformation including complete backbone gain in consistency compared to the query. A more pronounced effect is observed by combining local mirror and boundary conditions. Among the four scores, the cRMSd, probably the most widely used, appears not to be the most efficient in regards of our results. In particular, it behaves less efficiently for boundary conditions.

Interestingly, it is possible to identify score values below which no or a given fraction of false positives exist. Some of these are reported in Table 3 for different fragment lengths. The 0% values correspond to the values below which no mirror is observed, and thus can be related to the absolute value proposed by Maiorov and Crippen[43] to assess significance. We observe that the cutoff value can be very low. For instance, for mDMD no clear cutoff is observed, the values being very small. We prefer to define significant the values with a probability less than α to observe fragments with incorrect geometrical conditions. Table 3 also reports values for α value of 5%. For global mirror, we note that the significance cutoff tends, for

Table 3 Significance thresholds depending on fragment lengths and significance level

	Length	10	12	14	16	18	20
Global mirror							
cRMSd	0%	1.69	1.93	2.50	2.65	2.93	3.06
	5%	3.02	3.51	3.85	4.16	4.47	4.78
nRMSd	0%	0.09	0.09	0.10	0.13	0.15	0.15
	5%	0.22	0.24	0.25	0.26	0.27	0.28
DMD	0%	0.51	0.58	0.67	1.05	1.21	1.26
	5%	1.80	2.15	2.39	2.62	2.90	3.17
mDMD	0%	0.03	0.03	0.03	0.06	0.07	0.07
	5%	0.12	0.15	0.16	0.16	0.17	0.18
7-mirror and BC							
cRMSd	0%	0.11	0.13	0.20	0.47	0.61	0.65
	5%	2.22	2.80	3.24	3.65	4.01	4.35
nRMSd	0%	0.01	0.01	0.01	0.03	0.03	0.03
	5%	0.17	0.20	0.21	0.23	0.24	0.25
DMD	0%	0.08	0.11	0.13	0.25	0.27	0.29
	5%	1.31	1.69	2.04	2.38	2.71	3.02
mDMD	0%	0.01	0.01	0.01	0.02	0.02	0.02
	5%	0.09	0.11	0.13	0.14	0.16	0.17

Notes: Top: the significance is based on a no-global mirror condition. Bottom: the significance is based on a no- 7-mirror condition and boundary score less than 0.5.

nRMSd and mDMD, to be independent on fragment length compared to the un-normalized scores (cRMSd, DMD). This tendency is also observed for a combination of 7-mirror and boundary condition. This remains however to be further assessed. Overall, it remains that it is possible to tune the degree of similarity by controlling the size of the local mirror and the stringency of the boundary conditions.

Conclusion

We have assessed various scores in the context of fragment search. Particularly, we have explored the impact of considering local mirrors and boundary conditions as a filter of the matches. Our results clearly show that such filters are relevant to remove matches involving unrelated shapes. The geometrical conditions defined by such criteria can be considered as signatures of fragment similarity and greatly supplement classical scores such as the cRMSd or DMD. Combined with them, the distortion scores behave as well as the deviation scores. Interestingly, we also observe better independence to fragment length for the normalized scores. It follows that such criteria could be candidate for the design of a mining strategy based on them alone. Indeed, it becomes possible to tune the level of similarity desired in a very intuitive manner, which is more difficult to achieve with classical criteria such as the cRMSd. This clearly opens new perspectives for the rapid mining of large collections of structures.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res*. 1994;22:3600–3609.
2. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247:536–540.
3. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH – a hierarchic classification of protein domain structures. *Structure*. 1997;5:1093–1108.
4. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. 2000;28:254–256.
5. Sam V, Tai CH, Garnier J, Gibrat JF, Lee B, Munson PJ. Towards an automatic classification of protein structural domains based on structural similarity. *BMC Bioinformatics*. 2008;9:74.
6. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins*. 1995;23:356–369.
7. Sanchez R, Sali A. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol*. 1997;7:206–214.
8. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins*. 1999;Suppl 3:22–29.
9. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*. 2000;16:776–785.
10. Sadreyev RI, Shi S, Baker D, Grishin NV. Structure similarity measure with penalty for close non-equivalent residues. *Bioinformatics*. 2009;25:1259–1263.

11. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins*. 2004;57:702–710.
12. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. *Q Rev Biophys*. 2003;36:307–340.
13. Shah PK, Aloy P, Bork P, Russell RB. Structural similarity to bridge sequence space: finding new families on the bridges. *Protein Sci*. 2005;14:1305–1314.
14. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*. 1998;11:739–747.
15. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*. 2002;11:2606–2621.
16. Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31:3370–3374.
17. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33:2302–2309.
18. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics*. 2008;9:531.
19. Mizuguchi K, Go N. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng*. 1995;8:353–362.
20. Akutsu T, Onizuka K, Ishikawa M. Rapid protein fragment search using hash functions based on the Fourier transform. *Comput Appl Biosci*. 1997;13:357–364.
21. Samson AO, Levitt M. Protein segment finder: an online search engine for segment motifs in the PDB. *Nucleic Acids Res*. 2009;37: D224–D228.
22. Fischer D, Bachar O, Nussinov R, Wolfson H. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J Biomol Struct Dyn*. 1992;9:769–789.
23. Escalier V, Pothier J, Soldano H, Viari A. Pairwise and multiple identification of three-dimensional common substructures in proteins. *J Comput Biol*. 1998;5:41–56.
24. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. *J Mol Biol*. 2003;326:1307–1316.
25. Brakoulias A, Jackson RM. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins*. 2004;56:250–260.
26. Nebel JC, Herzyk P, Gilbert DR. Automatic generation of 3D motifs for classification of protein binding sites. *BMC Bioinformatics*. 2007;8:321.
27. Jakuschev S, Hoffman D. A novel algorithm for macromolecular epitope matching. *Algorithms*. 2009;2:498–517.
28. Sommer I, Muller O, Domingues FS, Sander O, Weickert J, Lengauer T. Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*. 2007;23:3139–3146.
29. Sael L, Li B, La D, et al. Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins*. 2008;72:1259–1273.
30. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*. 1998;281:565–577.
31. Camproux AC, Gautier R, Tufféry P. A hidden markov model derived structural alphabet for proteins. *J Mol Biol*. 2004;339:591–605.
32. Bujnicki JM. Protein-structure prediction by recombination of fragments. *Chembiochem*. 2006;7:19–27.
33. Weinhold N, Sander O, Domingues FS, Lengauer T, Sommer I. Local function conservation in sequence and structure space. *PLoS Comput Biol*. 2008;4:e1000105.
34. Mizuguchi K, Go N. Seeking significance in three-dimensional protein structure comparisons. *Curr Opin Struct Biol*. 1995;5:377–382.
35. Maiorov VN, Crippen GM. Size-independent comparison of protein three-dimensional structures. *Proteins*. 1995;22:273–283.
36. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci*. 2001;10: 1470–1473.
37. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. *Biopolymers*. 2001;59:305–309.
38. Prasad JC, Comeau SR, Vajda S, Camacho CJ. Consensus alignment for reliable framework prediction in homology modeling. *Bioinformatics*. 2003;19:1682–1691.
39. Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*. 2003;53 Suppl 6:436–456.
40. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*. 1976;104:59–107.
41. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol*. 1993;233:123–138.
42. Holm L, Sander C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci*. 1995;20:478–480.
43. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol*. 1994;235:625–634.
44. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A*. 1998;95:5913–5920.
45. Wrabl JO, Grishin NV. Statistics of random protein superpositions: *P*-values for pairwise structure alignment. *J Comput Biol*. 2009;15:317–355.
46. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol*. 2000;295:613–625.
47. McLachlan AD. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr A*. 1972;A28:656–657.
48. Kabsch W. A solution for the best rotation to relate two sets of vecctors. *Acta Crystallogr A*. 1976;32:922–923.
49. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A*. 1978;34:827–828.
50. Sippl MJ. On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. *J Mol Biol*. 1982;156:359–388.
51. Lesk AM, Levitt M, Chothia C. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng*. 1986;1:77–78.
52. Zuker M, Somorjai RL. The alignment of protein structures in three dimensions. *Bull Math Biol*. 1989;51:55–78.
53. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. *Fold Des*. 1996;1:123–132.
54. Fawcett T. An introduction to ROC analysis *Pattern Recognition Letters*. 2006;27:861–874.
55. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Cohen W, Moore A, editors. Proc. ICML'06: Proceedings of the 23rd international conference on Machine learning; 2006 June 25–29; Pittsburgh, PA. New York: ACM; 2006. p. 233–240.