Title

Comparative analysis of methods for evaluation of protein models against native structures

Journal

Bioinformatics, 35(6)

ISSN

1367-4803

Authors

Olechnovič, Kliment

Monastyrskyy, Bohdan

Kryshtafovych, Andriy

et al.

Publication Date

2019-03-15

DOI

10.1093/bioinformatics/bty760

Peer reviewed

Bioinformatics, 35(6), 2019, 937–944
doi: 10.1093/bioinformatics/bty760
Advance Access Publication Date: 29 August 2018
Original Paper

OXFORD

Structural bioinformatics

Comparative analysis of methods for evaluation of protein models against native structures

Kliment Olechnovič[1], Bohdan Monastyrskyy[2], Andriy Kryshtafovych[2] and Česlovas Venclovas[1,*]

[1]Institute of Biotechnology Life Sciences Center Vilnius University, Saulėtekio 7, Vilnius, LT 10257, Lithuania and [2]Genome Center UC Davis, 451 Health Sciences Drive, Davis, CA 95616, USA

*To whom correspondence should be addressed
Associate Editor: Alfonso Valencia

Received on March 30, 2018; revised on August 4, 2018; editorial decision on August 25, 2018; accepted on August 28, 2018

Abstract

Motivation: Measuring discrepancies between protein models and native structures is at the heart of development of protein structure prediction methods and comparison of their performance. A number of different evaluation methods have been developed; however, their comprehensive and unbiased comparison has not been performed.

Results: We carried out a comparative analysis of several popular model assessment methods (RMSD, TM-score, GDT, QCS, CAD-score, LDDT, SphereGrinder and RPF) to reveal their relative strengths and weaknesses. The analysis, performed on a large and diverse model set derived in the course of three latest community-wide CASP experiments (CASP10–12), had two major directions. First, we looked at general differences between the scores by analyzing distribution, correspondence and correlation of their values as well as differences in selecting best models. Second, we examined the score differences taking into account various structural properties of models (stereochemistry, hydrogen bonds, packing of domains and chain fragments, missing residues, protein length and secondary structure). Our results provide a solid basis for an informed selection of the most appropriate score or combination of scores depending on the task at hand.

Contact: ceslovas.venclovas@bti.vu.lt

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Measuring similarity between different conformations of the same protein is a common though far from trivial task in computational structural biology. This task is particularly important in the protein structure prediction field, because both development of structure prediction methods and their benchmarking depend on comparison of modeled and native (reference) protein structures. For more than two decades, advances in protein structure prediction have been monitored by the community-wide CASP experiments (Moult *et al.*, 2018). The progress in protein structure prediction has also stimulated the development of methods for model accuracy evaluation. However, as the problem of protein structure comparison is multiparametric in nature (Kufareva and Abagyan, 2012), it is impossible to arrive at a universally acceptable single measure that can tell the whole story about the modeled structure. That's why a well-rounded assessment of model-target similarity should include various conceptually different measures. Knowing distinctive properties of the measures may help select their combinations that would be most suitable for a given task such as developing and benchmarking structure prediction methods, or identifying models with desired characteristics for specific biomedical applications.

To date, multiple reference-based model evaluation scores have been proposed. Depending on the design and implementation details, the measures can be categorized into several binary classes: superposition-based or superposition-free; based on rigid-body (global) similarity or local similarity of constitutive regions; considering all atoms or only selected subsets of atoms (e.g. Cα atoms). This study is dedicated to comprehensive comparison of scores

commonly used by method developers for benchmarking their modeling approaches and CASP assessors for evaluating submitted models. To this end, we compared Root-Mean-Square Deviation (RMSD) (Kabsch, 1976), Template Modeling (TM) score (Zhang and Skolnick, 2004), Global Distance Test (GDT) score (Zemla *et al.*, 1999, 2001), Contact Area Difference (CAD) score (Olechnovič *et al.*, 2013, 2014), Local Distance Difference Test (LDDT) (Mariani *et al.*, 2013), SphereGrinder (Kryshtafovych *et al.*, 2014; Lukasiak *et al.*, 2015), Recall, Precision and F-measure (RPF) score (Huang *et al.*, 2012, 2014) and Quality Control Score (QCS) (Cong *et al.*, 2011).

To avoid subjective judgment, we did not treat any score as a 'gold standard', i.e. no measure was considered superior to others. Instead, our analyses aimed at comparing specific characteristics of the scores that could be considered as desirable in general. Probably one of the most desired features of a measure is its ability to give advantage to models with higher fraction of accurately modeled residues, without explicitly penalizing for inaccurate regions. This feature is also important for methods development as it encourages the construction of complete models. Another desirable property is the potential of a measure to address flexibility of specific regions or relative orientation of structural domains. Whereas in CASP this issue is partly circumvented by splitting multidomain proteins into rigid evaluation units upon manual inspection, such a recipe is not acceptable for automatic model evaluation systems such as CAMEO (Haas *et al.*, 2018). As protein structure prediction continues to progress, an important feature of a good score is the ability to promote realistic stereo chemical features of the structural model. Examples of other looked-for features are independence on protein size and secondary structure content.

2 Materials and methods

Considered scores are described in Table 1 and in more detail in Supplementary Data. To simplify comparison of the scores some of them were rescaled or transformed to fit within the same (0, 1) range. For GDT-TS/HA and SphereGrinder, fractions were used instead of percentages. RMSD was transformed to a (0, 1) range score (tRMSD) similarly as proposed earlier (Levitt and Gerstein, 1998) using the following equation: $tRMSD = 1/(1+(RMSD/10)^2)$. With this transformation, identical structures ($RMSD = 0$) result in $tRMSD = 1$, and those with large RMSD get tRMSD close to zero. Importantly, the relative rank of models is the same according to both scores.

The scores were compared on merged CASP10–12 data. The entire dataset includes 349 domains derived from both single- and multidomain targets (127 490 models) as well as 73 intact multidomain targets (27 879 models). In some analyses, we used a subset of models submitted on 'predictable' targets, or a subset of 'good' models only. Predictable targets were defined as those having at least 15 'good' models scoring above the GDT-TS threshold of 40 (268 targets, 93 271 models in total, 54 182 of them 'good'). The list of considered targets and the values of all the scores for corresponding models are available in the data archive: https://kliment.bitbucket.io/refscores/data.zip. The structures of targets and models are available at the Prediction Center (http://predictioncenter.org/download_area/).

Stereochemical features were evaluated using MolProbity (Chen *et al.*, 2010). Hydrogen bonds were identified with HBplus (McDonald and Thornton, 1994). Performance of scores on multidomain structures was compared using methodology proposed by Grishin and colleagues (Kinch *et al.*, 2011). The approach is based on comparison of raw scores for the full structure with the weighted sum of scores for individual domains. Similarities of scores were visualized using the multi-dimensional scaling (MDS) method (Mardia, 1978), which takes a dissimilarity matrix as an input and outputs a set of points such that the distances between them quantify the extent of the dissimilarity.

3 Results

3.1 How similar are the scores among themselves?

To answer this question, we checked empirical distributions of the scores, correspondence of their values, correlation between score-specific model rankings and agreement in selecting the best model(s). These analyses were carried out on the single domains dataset to avoid distortion of the results by conceptually different behavior of various methods on multidomain targets (analyzed further).

3.1.1 Empirical distribution of scores

Histograms of values for every analyzed score show that in general score distributions differ (Fig. 1).

RMSD/tRMSD and QCS have clear bimodal distributions, whereas distributions of GDT-TS, TM-score and in part LDDT only hint at bimodal character. CAD-AA exhibits roughly a bell-shaped distribution in a relatively narrow range of values. RPF spreads the

Table 1. Brief description of the analyzed scores

Score	Range[a]	What is measured	Superposition
RMSD (global)	0, ∞	Mean distance between corresponding atoms (Cα or all atoms)	Yes; global
TM-score (global)	0, 1	Mean distance between corresponding Cα atoms scaled by a length-dependent distance parameter	Yes; global
GDT-TS/HA (global)	0, 100	Mean percentage of Cα atoms that fit under 4 distance thresholds: GDT-TS: 1, 2, 4, 8 Å; GDT-HA: 0.5, 1, 2, 4 Å	Yes; global (four independent ones)
QCS (global)	0, 1	Agreement between the length and relative orientation of secondary structure elements, and Cα distances.	No
SphereGrinder (local)	0, 100	Mean percentage of residues whose neighborhoods (spheres) fit under 2 and 4 Å RMSD threshold	Yes; local (atoms within 6 Å from Cα)
CAD-score (local)	0, 1	Similarity of interatomic contact areas (all atoms or their subsets)	No
LDDT (local)	0, 1	Mean fraction of preserved all-atom distances using 4 tolerance thresholds (0.5, 1, 2, 4 Å) within the 15 Å inclusion radius	No
RPF (local)	0, 1	Normalized F-measure derived from distances between N and C atoms within the 9 Å inclusion radius	No

[a]For all scores, except RMSD, higher values correspond to more similar structures; RMSD values do not have a fixed upper limit.

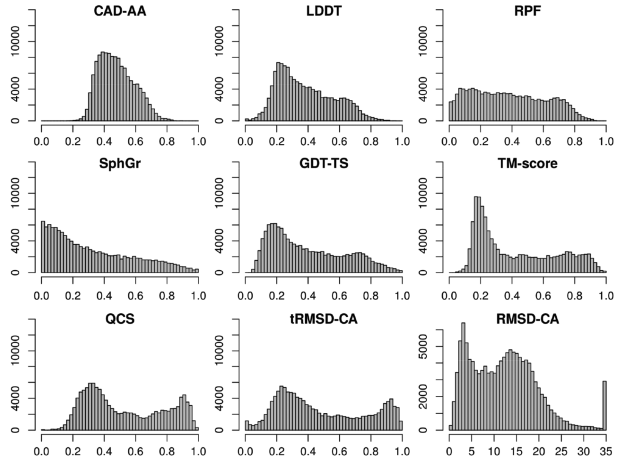

Fig. 1. Empirical distribution of score values on the single-domain dataset. Horizontal axis indicates score values, vertical axis—frequency of the value occurrence. For RMSD/tRMSD, GDT and CAD-score only representative versions are shown [Cα RMSD/tRMSD, GDT-TS and all atom CAD-score (CAD-AA)]

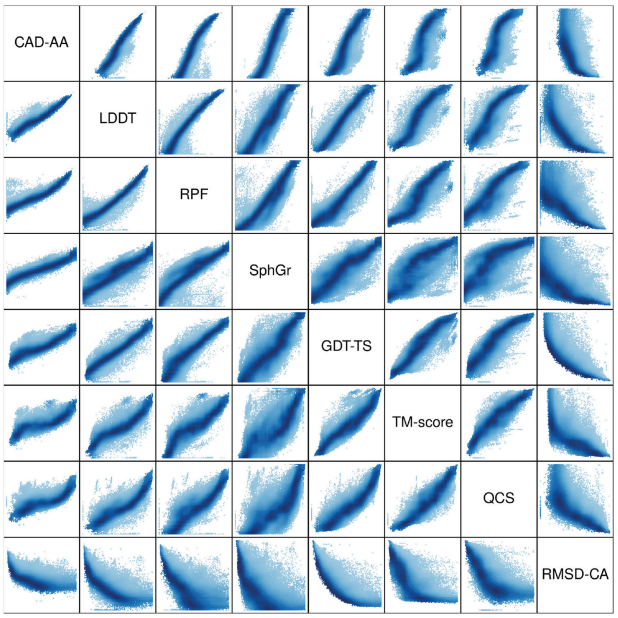

Fig. 2. Correspondence of scores. Scatter plots for 'A' and 'B' score pairs. Horizontal direction represents values of score 'A', vertical direction represents score 'B'. Increasing color intensity represents the increasing local density of values

models almost evenly along the wide range of values, and SphereGrinder monotonically assigns better values to fewer models. Among common trends is the dominance of low scores, reflecting the nature of the CASP dataset. Additional variants of the analyzed scores show similar distributions (Supplementary Fig. S1). Distribution of all-atom RMSD/tRMSD is nearly identical to that of Cα RMSD/tRMSD. The distribution of a more stringent version of GDT (GDT-HA) is shifted towards lower values. Similarly, CAD-score variants CAD-AS (all atom-side chain) and CAD-SS (side chain-side chain) display shifts towards smaller values and sharper peaks, indicating a more stringent evaluation of model accuracy. The observed differences in the distribution of different scores preclude direct comparison of their raw values. However, a common technique of converting raw values to Z-scores (separately for every reference structure) leads to similarly distributed values that can be directly compared or combined (Supplementary Fig. S2).

3.1.2 Correspondence between values of different scores

To study what values can be expected for score 'B' when score 'A' is within a given range, we produced scatter plots for every pair of scores. Results are shown in Figure 2 (in more detail in Supplementary Figs S3 and S4) with darker color representing the higher local density of points (values).

From the figure, it is immediately apparent that the correspondences of scores show high heterogeneity. Some correspondences are relatively well-defined (values are scattered narrowly) whereas others are not (scattered widely). Some of the best-defined correspondences are among the local scores CAD-AA, LDDT and RPF. Conversely, RMSD shows poorly defined correspondence with all the scores. SphereGrinder (local score) is an interesting case. It shows a poorly defined correspondence with global scores (GDT-TS, TM-score and QCS) and even with two of the local scores (LDDT and RPF). However, the correspondence between SphereGrinder and CAD-AA is one of the least ambiguous. Another observation is that most correspondences are asymmetric and in many cases non-linear. As an example, let us consider the correspondence between TM-score and local scores CAD-AA, LDDT and RPF. In all these cases the correspondences are sigmoidal in shape and asymmetric. Thus, if we use TM-score as a primary scoring measure (score 'A'), the corresponding values of the three scores (score 'B') are relatively well defined. In contrast, when original scoring is done using CAD-AA, LDDT or RPF, a fairly sharp transition of TM-score values corresponding to ∼0.45 of CAD-AA and ∼0.4 in case of LDDT and RPF makes it difficult to predict which values of TM-score can be expected in this range of local scores.

3.1.3 Correlation of scores

Since the correspondence of values for majority of score pairs is non-linear (Fig. 2), we used Spearman's rank correlation analysis, suitable for both linear and non-linear correlations (Altman and Krzywinski, 2015). We calculated correlation coefficients by considering models of individual reference structures (targets) separately and then averaging the coefficients using Fisher's Z-transformation (Fisher, 1915). Correlation analysis was applied to all models in the 'predictable' single-domain dataset.

Results of this analysis (Fig. 3, Supplementary Fig. S5) show that the scores are highly correlated, except for RMSD. Also, local scores and global scores tend to show higher correlation within their own group. For example, the correlations between LDDT and RPF scores (local) and between GDT-TS and TM-score (global) for 'predictable' targets are correspondingly 0.94 and 0.97, while their cross-correlations are lower (0.88–0.90) (Fig. 3A). If RMSD is not considered, RPF shows the most consistent correlation with the other scores, ranging from 0.90 to 0.94. Visual summary of the results is provided in Fig. 3B, where proximity of the points quantifies correlation between the corresponding scores.

3.1.4 Selecting a better model out of two

Although the scores (with the exception of RMSD) are highly correlated, they might disagree on which of the two models is more accurate. Since difference in accuracy scores is usually meaningless for models with low score values, where 'similarity' of structures is typically attributed to random fit of secondary structure elements, we carried out our analysis on 'good' models for 'predictable' targets.

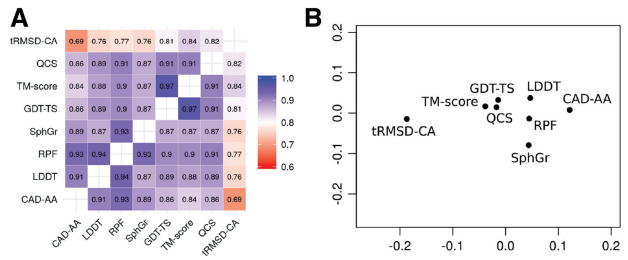

Fig. 3. Correlation of scores. (**A**) Spearman's rank correlation coefficients computed by averaging per target values. Coloring ranges from blue (high correlation) to red (low correlation). (**B**) Clustering of scores according to their correlations using multi-dimensional scaling (MDS) (Color version of this figure is available at *Bioinformatics* online.)

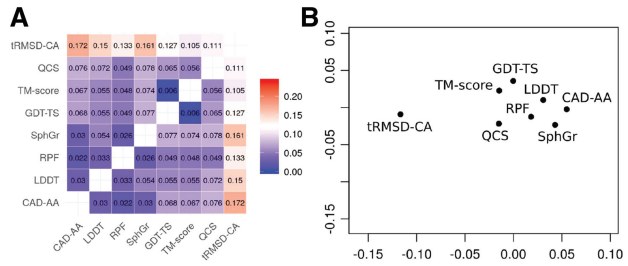

Fig. 4. Score differences in selecting a better model out of two. (**A**) Fractions of model pairs, where the disagreement between scores exceeds the tolerance threshold. Differences are colored from blue (smallest) to red (largest). (**B**) Clustering of scores based on the analysis of model pairs with conflicting ranking (Color version of this figure is available at *Bioinformatics* online.)

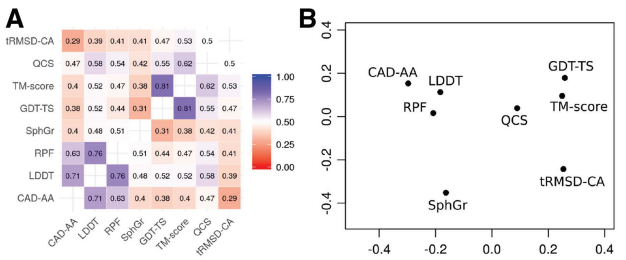

Fig. 5. Agreement between the scores in selecting the best model out of many. (**A**) Average fraction of the same selections. (**B**) Clustering of scores according to their agreement in selecting the best model

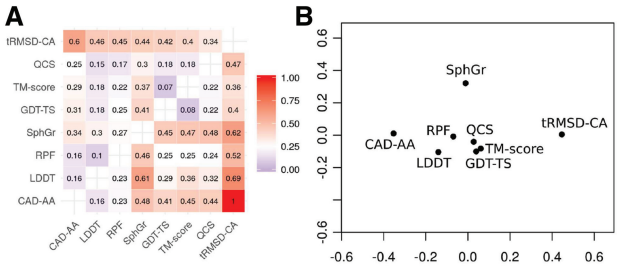

Fig. 6. Score differences in selecting the best model out of many. (**A**) Average losses of Z-score computed by score B (vertical axis) when the best model is selected using score A (horizontal axis). Z-score losses are colored from blue (smallest) to red (largest). Note the overall asymmetry of Z-score loss for score pairs. (**B**) Clustering of scores according to average Z-score losses (Color version of this figure is available at *Bioinformatics* online.)

To reduce noise resulting from minor differences between models and intrinsic uncertainties of scores, we introduced a tolerance threshold for defining models of comparable accuracy. The threshold is defined as the 25% quantile of all the differences calculated for a score and roughly corresponds to 2% of the score range for all scores except for SphereGrinder, where it is about 3%. For example, a model with a score of 0.65 is considered of similar accuracy to models scoring in the range of 0.63–0.67 (0.62–0.68 in the case of SphereGrinder).

Similarly to the correlation analysis, RMSD displays the largest disagreement with the other scores (Fig. 4). Local scores, in particular CAD-AA, LDDT and RPF, show very close agreement between themselves. The largest disagreement between them is only 3% (Fig. 4A). Among local scores SphereGrinder tends to show slightly higher disagreement, especially with LDDT (up to 5%). The overall best agreement is shown by global scores GDT-TS and TM-score. They disagree on only 0.6% of model pairs. QCS shows not particularly strong, but fairly balanced agreement with both local and global scores. The graphical illustration of score agreement is provided in Figure 4B. All-atom RMSD, GDT-HA and CAD-score variants (CAD-AS, CAD-SS) behave similarly to their representative variants (Supplementary Fig. S6).

3.1.5 Selecting the best model(s) out of many

Most common task for any score is to select the best model out of a set. Therefore, we next tested how scores differ/agree in performing this task. We used all models from the dataset of 'predictable' single domains, so that for each target reasonably accurate models are available. For every target, we identified the highest scoring models according to each of the scores and then compared these top-scoring models for all pairs of scores. It is not unusual for more than one model to have identical values according to a given score. Therefore, we considered that scores agree in the selection of the best model if they have at least one common model with the best value by both scores. The best agreement was shown by CAD-AA, LDDT and RPF as one group and GDT-TS and TM-score as another group (Fig. 5).

Differences in the choice of the best model do not tell whether these models are of comparable accuracy or not. Therefore, we next asked how quantitatively different are the selections made by different scores. To this end, we calculated Z-scores using one of the raw scores (let us say score 'B') and selected the best model. We then selected the best model for the same target using another score (score 'A') and recorded the loss of Z-score assigned using score 'B' (loss is zero if it is the same best model). We performed this for all targets using all pairwise combination of scores and calculated average Z-score losses (Fig. 6). As an example let us consider the CAD-AA/GDT-TS pair. If models are selected using the CAD-AA (score 'A'), then the GDT-TS (score 'B') Z-score loss is 0.32. If models are selected with GDT-TS (score 'A') the loss of CAD-AA Z-score is 0.4. As in the case of model pairs (Section 3.1.4), selection of the best model out of many using RMSD (score 'A') yields the largest losses no matter which score 'B' is used to calculate Z-score values. Consistent with relatively good agreement in selecting the best model, LDDT, CAD-AA and RPF show small losses of Z-score in relation to each other (from 0.1 to 0.2). However, in this test, LDDT and RPF show comparable agreement with global scores, whereas CAD-AA does not. SphereGrinder, also of local nature, is significantly different from all three. Selection of models using SphereGrinder results in larger Z-score losses against LDDT, CAD-AA or RPF (0.4–0.6) than against global scores (0.3–0.4). GDT-TS and TM-score show negligible difference in selecting best models. Selection by QCS is similar to that by GDT-TS and TM-score.

Comparison of scores based only on the single best model may be fairly stringent. Therefore, we repeated the analysis for the top-10 scoring models. In this case different scores tend to agree better and, accordingly, Z-score losses are in general smaller. At the same time, the relationship between the scores remains essentially the same (Supplementary Figs S7 and S8).

3.2 How strongly do the scores favor models with realistic stereochemical features?

It has been observed that sometimes models with high assessment scores may have systematically distorted stereochemical parameters such as bond lengths and angles, and may be physically unrealistic in general (Sadreyev *et al.*, 2009). To compare how the scores promote physical realism of models, we selected model pairs for which different scores disagree on a better model (as in Section 3.1.4 above) and employed an independent method, MolProbity, to 'judge' these conflicting rankings. MolProbity does not evaluate the accuracy of a model; instead it evaluates how well stereochemical features of the model conform to those derived from high quality experimental structures. Thus, we asked which score more often ranks models within the conflicting pair in the same way as MolProbity does. For this analysis, we only considered 'good' models of 'predictable' targets as it makes no sense to evaluate stereochemistry of grossly incorrect structures. We performed the analysis by considering only conflicting pairs with differences larger than a defined threshold (25% quantile of all the difference values) for both the considered scores and MolProbity values. Results of this analysis are shown in Figure 7.

The simplest way to analyze the results is to look at the values of score 'A' along the horizontal axis. Values over 0.5 (shades of blue) show better agreement with MolProbity than score 'B' and *vice versa*. Two extremes are clearly identifiable: CAD-AA and RMSD. CAD-AA agrees with the MolProbity score better than any other score. The CAD-AA ranking of models in conflicting pairs is favored by MolProbity over that of other scores in more than 80% of cases (except for the CAD-AA/LDDT conflicting pairs, where the CAD-AA/MolProbity agreement is 62%). At the other extreme, the RMSD's selection of better models gets the lowest support from MolProbity. In between these extremes, LDDT and SphereGrinder receive relatively strong support, being second and third after CAD-AA. If we look at the individual components of the MolProbity score (clashes, side chain rotamer outliers and backbone Ramachandran), the overall picture does not change significantly (Supplementary Fig. S9): CAD-AA remains supported better than other scores by any individual MolProbity component.

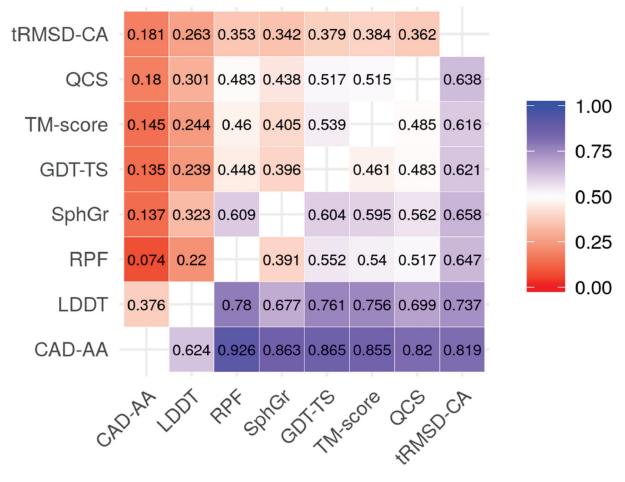

Fig. 7. Conflicting rankings of model pairs 'judged' using MolProbity. Fractions of conflicting model pairs for which MolProbity supports score 'A' (vertical axis) over score 'B' (horizontal axis). The MolProbity support is colored from blue (largest) to red (smallest) (Color version of this figure is available at *Bioinformatics* online.)

Other variants of CAD-score (CAD-AS and CAD-SS) agree with MolProbity in slightly lesser degree than CAD-AA, but still better than all the other scores (Supplementary Fig. S9). High accuracy version of GDT (GDT-HA) agrees with MolProbity slightly better than GDT-TS, whereas the agreement of all-atom RMSD is about the same as Cα RMSD.

3.3 Do the scores promote accurate reproduction of the hydrogen bonds?

An extensive network of hydrogen bonds is a common feature of folded globular proteins. Typically, in proteins there is only a small fraction of buried unsatisfied hydrogen bond donors and acceptors. Therefore, we decided to test to which degree the scores support accurate reproduction of hydrogen bonds of the target in corresponding models. To this end, we chose the same model pairs with conflicting ranking by different scores as in the MolProbity test (above). Only in this case we asked which of the two models has missed fewer hydrogen bonds present in the target structure. In addition to all hydrogen bonds we separately considered non-local hydrogen bonds (minimal sequence separation of six residues). Similarly to the MolProbity test, we analyzed only conflicting pairs with significant differences in both the score values and the number of missed hydrogen bonds. The results of this test are provided in Figure 8. CAD-AA agrees best with more accurate hydrogen bonding network, whereas RMSD agrees worst. This is true both for all and for only non-local hydrogen bonds. The results also show that in general the local scores support the accurate reproduction of hydrogen bonding network more strongly than the global scores. Additional variants of CAD-score, GDT and RMSD behave similarly to their representative ones (Supplementary Fig. S10).

3.4 How robust are the scores in the case of multidomain proteins?

Many proteins are composed of multiple domains. Their relative orientation often may not be biologically relevant. However, some scores, in particular those based on structure superposition, are known to be quite sensitive to the differences in domain orientation. We compared suitability of the scores to evaluate models of multidomain proteins automatically, without splitting them into individual domains. We took all the models of multidomain targets and performed the analysis using the so-called Grishin plots (see Section 2). If the difference between the full structure score and the weighted

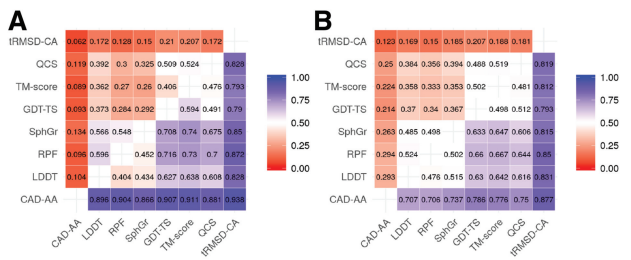

Fig. 8. Conflicting rankings of model pairs 'judged' by the number of reproduced hydrogen bonds. The scores are compared using differences larger than a defined threshold (**A**) for all hydrogen bonds and (**B**) for only non-local hydrogen bonds

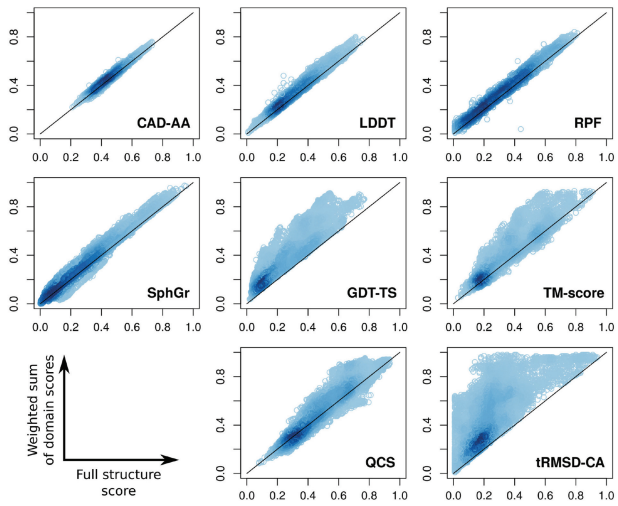

Fig. 9. Grishin plots reflecting sensitivity of different scores to the relative orientation of protein domains. Horizontal axis indicates the score values for the full structure, vertical axis indicates weighted sum of scores for individual domains

sum of domain scores is small, then the measure is insensitive to the domain orientation and evaluation of models can be performed without splitting the target structure into domains.

Figure 9 shows Grishin plots for all models of multidomain targets pooled together. One can see that different scores indeed show very different sensitivity to domain orientation. For local scores the deviations from the diagonal line are relatively small indicating that a given model gets similar value independently of whether it is evaluated as a full structure or as separate domains. Global scores show relatively large deviations from the diagonal, indicating that they are not suitable for evaluation of models corresponding to multidomain structures without parsing them into rigid domains. Results for all the score variants are shown in Supplementary Figure S11.

3.5 How do the scores deal with local structural deviations

Protein models commonly have structural fragments such as small subdomains, termini, domain linkers or long loops deviating from the corresponding ones in the target. These locally deviating fragments can be classified into two major groups: (i) fragments that have fairly accurate local structure, but are packed (oriented) differently compared to the native structure and (ii) fragments that have incorrect local structure. These two types of local deviations may be considered as special cases of domain orientation and domain structure modeling problems, respectively. Figure 10 and Supplementary Table S1 illustrate such cases on the example of target T0663 from CASP10.

The C-terminal region in two models (Fig. 10A and B) is correctly predicted as α-helical, but its orientation is different and incorrect in both. Local scores evaluate both models as of similar accuracy (Supplementary Table S1), whereas global scores, e.g. GDT-TS, deem the model in orange (Fig. 10A) to be significantly more accurate than the model in yellow (Fig. 10B). This is due to the fact that the C-terminal fragment of orange model is oriented, although incorrectly, somewhat less differently compared to the native structure. The C-terminal fragment of the third model (Fig. 10C, grey) has even local structure incorrect, and this is recognized by both local and global scores, although RMSD exaggerates the error to the

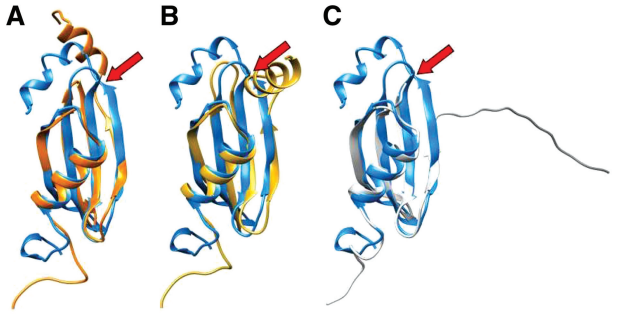

Fig. 10. Different types of local structural deviations. CASP10 target T0663-D1 (blue) superimposed with three models: (**A**) TS301_5 (orange), (**B**) TS301_3 (yellow) and (**C**) TS476_4 (grey). Red arrow indicates the C-terminal region featuring different types of deviation in the models (Color version of this figure is available at *Bioinformatics* online.)

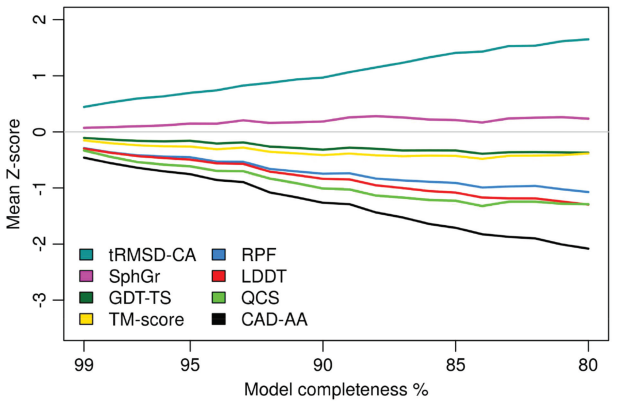

Fig. 11. Response of the scores to model completeness. Horizontal axis indicates the model completeness as the percentage of residues modeled. Vertical axis indicates mean Z-score for models of the same degree of completeness. Z-scores are averaged using left-sided sliding window of 15%, e.g. values at 80% are averages of 95–80% range

point that it may be difficult to distinguish the local error from the entirely wrong model. If we consider the truncated structure (Fig. 10, red arrow), the two models (Fig. 10A and B) are of very similar accuracy by all the scores, including RMSD (Supplementary Table S1).

3.6 Do the scores favor complete models?

An effective reference-based score should favor the construction of complete structural model. In other words, exclusion of polypeptide chain fragments or residues from the model generally should not produce a better score. To test how scores differ in this regard, we binned 'good' models for single-domain targets according to the completeness and looked at the behavior of each score. Since we combined together models for different targets, we used a sliding window of completeness to smooth the trend lines (Fig. 11). The RMSD score shows clear improvement as the models get less complete. This is not surprising since the smaller set of superimposed residues can be expected to have a smaller mean deviation. SphereGrinder is also somewhat different from the remaining scores as it stays about the same in the case of incomplete models. The remaining scores all display a downward trend with CAD-AA penalizing incomplete models most. Using smaller sliding windows does not change the picture significantly (Supplementary Fig. S12). Based on this analysis it is apparent that if a mixture of complete

and incomplete models are evaluated against the reference structure, RMSD would be a poor choice as a score.

3.7 How do the scores depend on the protein length and secondary structure?

Dependence of a given score on the nature of the protein is not important when the models are scored against the same target structure. However, if one aims to make a broader generalization using a large variety of target structures it is useful to know how a given score depends on protein properties such as length and/or secondary structure. It makes sense to investigate how the scores depend on protein length and secondary structure using only targets for which most models are reasonably accurate. To perform these tests we took single-domain targets, for which models have mean GDT-TS $\geq$50% (after removing model outliers with Z-score <-2). To investigate the length-dependence, we sorted targets by their length and calculated mean Z-score values for every target using every score. To get smoother trends, we averaged values for each target and its neighbors using a centered sliding window. Most scores do not show any obvious dependency on the protein size (Supplementary Fig. S13). A clear exception is TM-score, which generally produces higher values as proteins get larger. Two other scores, QCS and SphereGrinder, also differ from the remaining ones. QCS tends to produce lower scores for short proteins. SphereGrinder shows larger fluctuation than the rest and tends to produce lower values for proteins of up to about 200 residues and higher values for proteins over 250 residues long.

To investigate dependency of scores on secondary structure type, we took the same dataset and ordered target proteins according to their secondary structure content from mostly β-structural to mostly α-helical. As in the size-dependency test, we calculated mean values for each score using a sliding window (Supplementary Fig. S14). All scores behave similarly in the case of mostly β-proteins and proteins with mixed structure (α/β and $\alpha+\beta$), but are in general lower for the former compared to the latter. Also, with the increase of α-helical content, the scores show increasing divergence.

4 Discussion

Our comparative study of selected reference-based model evaluation scores aimed at answering two major questions: (i) how similar/different are these scores in general and (ii) how they compare with each other when specific structural properties are taken into account. The score similarity analyses revealed that RMSD differs most. This should not be surprising as RMSD values are heavily influenced by large errors whereas other scores focus on the accurate regions, or are local by nature. Similarity between the remaining scores is considerably higher, in particular within two groups of scores. One group consists of global scores GDT-TS and TM-score, whereas another group is formed by local scores LDDT, RPF and CAD-AA. GDT-TS and TM-score are extremely highly correlated and show negligible differences in selecting the best model out of two or out of many. LDDT, RPF and CAD-AA are also highly correlated and nearly always agree on a better model within a given model pair. Even if the best models selected by LDDT, RPF and CAD-AA from a set of multiple models are not always the same, their accuracy is comparable. Where these three scores differ, it is in their agreement with global scores (e.g. with GDT-TS). The best models selected using LDDT and RPF typically are also among the best according to GDT-TS, TM-score or QCS. In the case of CAD-AA, the selected best models differ by a larger margin. Among local scores, SphereGrinder is the most distinct, especially in selecting the best model(s) from a model set. QCS, which considers both global and local structural features, does not show clear affinity with either global or local scores. Thus, taking into account correlation and distinct modes of model selection, all the analyzed scores can be roughly divided into five groups according to the mutual similarity: (i) RMSD, (ii) GDT-TS and TM-score, (iii) CAD-AA, LDDT and RPF, (iv) SphereGrinder and (v) QCS.

Score differences by themselves tell nothing about the efficacy of scores. Therefore, our second aim was to investigate how the scores compare with each other when specific structural properties, namely, stereochemical parameters, hydrogen bonds, differences in domain orientation, locally deviating chain fragments, model completeness, protein size and secondary structure type are taken into account.

Among highly preferable features of a score is the ability to promote physical realism or 'protein likeness' of models, including good stereochemistry and accurately reproduced hydrogen bonding network. The emphasis on realistic physico-chemical features is especially important in the development and comparison of methods for high accuracy modeling and/or structure refinement. When disagreements between the scores were judged in the light of stereochemical criteria, CAD-AA model rankings were supported by MolProbity far better than those of any other score. Taking into consideration all atoms may be one of the reasons, since both LDDT and SphereGrinder (all-atom scores) showed better agreement with MolProbity than RPF, which uses only N and C atoms. Nonetheless, the use of all atoms cannot entirely account for this phenomenon. Other CAD-score variants (CAD-AS and CAD-SS) use subsets of residue atoms, yet they also agree with MolProbity better than any other score. When disagreements in pairwise model rankings were considered with hydrogen bonds in mind, the CAD-AA rankings again agreed best with more complete hydrogen bond network. This was true regardless of whether all or only non-local hydrogen bonds were taken into account. These findings suggest that interatomic contact areas (CAD-score) perhaps are better suited than distances (other scores) to promote physical realism of models.

Proteins having multiple domains are very common in nature. Thus, it is important to identify scores that are not overly sensitive to relative domain orientation and therefore would be suitable to evaluate models of multidomain proteins automatically. A related problem is how to properly assess locally deviating structural regions or subdomains. When we tested the sensitivity of scores to the relative domain orientation, we found a clear distinction between local and global scores. Local scores (CAD-AA, LDDT, RPF and SphereGrinder) are largely insensitive to domain orientation, whereas global ones (RMSD, GDT-TS, TM-score and QCS) are. We showed that local scores are also better suited to deal with local structure deviations. Notably, local scores can evaluate the accuracy of the deviating region independently of its relative orientation. In contrast, global scores focus mostly on the relative orientation of the deviating region regardless of the accuracy of its local structure. This distinction between local and global scores has important implications not only for benchmarking protein structure prediction methods, but also for estimation of model accuracy. If superposition-based global scores are not always reliable in the assessment of multidomain proteins, they also cannot be effective points of reference in estimating global accuracy of multidomain proteins. Assessment of per-residue accuracy (confidence) estimation faces similar issues. Let us assume that the chain fragment has accurate local conformation but wrong relative orientation. Should it be considered accurate or wrong? It depends on whether the exact orientation is biologically relevant or not. Regardless of the answer

it is apparent that the accuracy of local structure should be taken into account, and this can only be done by involving local scores.

In the protein structure prediction field, a commonly held view is that the construction of complete structural models should be encouraged. In other words, exclusion of a chain fragment from the model should be penalized. We compared the behavior of scores depending on the model completeness and found that RMSD is the weakest score in this respect as incomplete models have a larger chance of getting better scores. The remaining scores, except SphereGrinder, which is mostly neutral to model completeness, all penalize incomplete models, albeit to different extent.

When scoring models for the same target protein, the length and the secondary structure composition of a protein affect all models equally. However, these properties of proteins may introduce bias in evaluation of models on multi-target datasets. According to our results, of all the scores, TM-score shows the largest dependency on protein size. TM-score values for small proteins are lower whereas for large ones are higher. Similar but much less pronounced trend is displayed by QCS. If secondary structure is concerned, a common feature of all the scores is slightly higher values for mixed (α/β and $\alpha+\beta$) structures compared with mostly β-structures. In addition, scores increasingly diverge with the increasing α-helical content.

Collectively, the analyses performed in this study provide a comprehensive picture for every considered score and help to better understand the relationship between scores. Since there may be very different tasks involving reference-based model assessment, it is difficult to suggest specific scores or their combinations that would be best fitting in every case. On the other hand, some general observations can be made. Whenever two or more scores are utilized in model evaluation it is inappropriate to directly compare or combine their raw values, because different scores show distinct distributions, have different ranges and in most cases display asymmetric correspondence of values (Figs 1 and 2). However, transforming raw values into Z-scores largely takes care of these issues (Supplementary Fig. S2). Another observation relates to the selection of best model(s). Our results show that the superposition-free local scores (LDDT, RPF and CAD-score) make selections that are more consistent compared to the selections made by global scores such as GDT-TS or TM-score. This can be seen in the Z-score loss analysis (Fig. 6, Supplementary Fig. S8). It shows that models selected by these local scores on average are closer to the top models identified by the global scores compared to the opposite scenario, i.e. selecting models with the global scores and looking how close to the top they are according to the local scores. Local scores are also better in dealing with multidomain structures and local deviations as well as in promoting realistic stereochemistry and hydrogen bonding. Consistent with our results, a recent study has found that the use of local scores (CAD-score or LDDT) as training targets may benefit the estimation of model accuracy (Uziela *et al.*, 2018). All these findings suggest that superposition-free measures possess features that make them very useful for description and comparison of protein structures in general and inherently more appropriate for analysis of local structural details of models in particular. On the other hand, global scores are well suited for identifying overall correct topology and might be more informative for assessing differences in domain orientation or deviations of subdomains, loops and tails.

In conclusion, we hope that this comparative study will be useful not only for benchmarking and comparing protein structure prediction methods but also for many other endeavors in computational structural biology.

Funding

This work was supported by the Research Council of Lithuania [S-MIP-17-60 to K.O. and Č.V.]; and the US National Institute of General Medical Sciences (NIGMS/NIH) [GM100482 to B.M. and A.K.].

Conflict of Interest: none declared.

References

Altman,N. and Krzywinski,M. (2015) Association, correlation and causation. *Nat. Methods*, **12**, 899–900.

Chen,V.B. *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 12–21.

Cong,Q. *et al.* (2011) An automatic method for CASP9 free modeling structure prediction assessment. *Bioinformatics*, **27**, 3371–3378.

Fisher,R.A. (1915) Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, **10**, 507–521.

Haas,J. *et al.* (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, **86**, 387–398.

Huang,Y.J. *et al.* (2014) Assessment of template-based protein structure predictions in CASP10. *Proteins*, **82** Suppl 2, 43–56.

Huang,Y.J. *et al.* (2012) RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Res.*, **40**, W542–W546.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **32**, 922–923.

Kinch,L.N. *et al.* (2011) CASP9 target classification. *Proteins*, **79**, 21–36.

Kryshtafovych,A. *et al.* (2014) CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins*, **82** Suppl 2, 7–13.

Kufareva,I. and Abagyan,R. (2012) Methods of protein structure comparison. *Methods Mol. Biol.*, **857**, 231–257.

Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA*, **95**, 5913–5920.

Lukasiak,P. *et al.* (2015) SphereGrinder—reference structure-based tool for quality assessment of protein structural models. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. pp. 665–668.

Mardia,K.V. (1978) Some properties of clasical multi-dimensional scaling. *Commun. Stat.*, **7**, 1233–1241.

Mariani,V. *et al.* (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**, 2722–2728.

McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.

Moult,J. *et al.* (2018) Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins*, **86** Suppl 1, 7–15.

Olechnovič,K. *et al.* (2013) CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins*, **81**, 149–162.

Olechnovič,K. *et al.* (2014) The CAD-score web server: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes. *Nucleic Acids Res.*, **42**, W259–W263.

Sadreyev,R.I. *et al.* (2009) Structure similarity measure with penalty for close non-equivalent residues. *Bioinformatics*, **25**, 1259–1263.

Uziela,K. *et al.* (2018) Improved protein model quality assessments by changing the target function. *Proteins*, **86**, 654–663.

Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, Suppl 3, 22–29.

Zemla,A. *et al.* (2001) Processing and evaluation of predictions in CASP4. *Proteins*, Suppl 5, 13–21.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.