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

Supplementary data

Kliment Olechnovič¹, Bohdan Monastyrs’kyy², Andriy Kryshtafovych² and Česlovas Venclovas¹

¹ Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio 7, Vilnius LT-10257, Lithuania
² Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, USA

Corresponding author:
Česlovas Venclovas
E-mail: ceslovas.venclovas@bti.vu.lt
Supplementary Methods

Extended description of the analyzed scores

1. Description of global scores (RMSD, GDT-TS/GDT-HA, TM-score and QCS)

1.1 RMSD

Root-Mean-Square-Deviation (RMSD) is one of the most commonly used structure comparison scores. It is defined as follows:

\[ RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d_i^2}, \]

where

- \( N \) - number of equivalent atoms in the two structures;
- \( d_i \) - distance between equivalent atoms after optimal superposition of the structures.

Lower RMSD scores indicate smaller average deviations between corresponding atoms and hence better overall models. High accuracy models typically have RMSD<2Å when superimposed onto the native structure, while models for difficult targets are usually 10Å or further away.

1.2 GDT-TS and GDT-HA

Global Distance Test Total Score (GDT-TS) (Zemla, et al., 1999; Zemla, et al., 2001) is a rigid-body superposition-based score defined as:

\[ GDT\text{-}TS = \frac{1}{4} \frac{N_{d_1} + N_{d_2} + N_{d_3} + N_{d_4}}{N}, \]

where

- \( N \) - number of residues in the reference structure (target);
- \( N_{d_n} \) - number of corresponding Cα-atom pairs in the two structures that could be superimposed within distance cutoffs of \( d_1=1\text{Å} \), \( d_2=2\text{Å} \), \( d_3=4\text{Å} \), and \( d_4=8\text{Å} \), each using independently optimized superposition.

In contrast to RMSD, where all residues (including the superposition outliers) contribute to the score, GDT-TS accounts for only well-modeled residues (those modeled to 8Å or better) thus identifying models with a larger portion of accurately modeled structure, regardless to the scale of errors in poorly modeled regions.

GDT-TS values can be computed using the LGA webserver:
http://proteinmodel.org/AS2TS/LGA/lga.html

GDT-HA score (HA stands for High Accuracy) is calculated according to the GDT_TS formula (above), except that the distance cutoffs are halved: \( d_1=0.5\text{Å} \), \( d_2=1\text{Å} \), \( d_3=2\text{Å} \), \( d_4=4\text{Å} \). This score is slightly better suited for the evaluation of high homology targets, which are usually very close to native conformations of proteins.
1.3 TM-score

TM-score (Zhang and Skolnick, 2004) is a variation of the Levitt–Gerstein (LG) score (Levitt and Gerstein, 1998). TM-score considers all the distances between corresponding Cα-atoms after optimal superposition and uses protein size-dependent scaling:

$$\text{TM-score} = \text{Max} \left[ \frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left( \frac{d_i}{d_0} \right)^2} \right]$$

$L_N$ - length of the reference structure;
$L_T$ - number of corresponding residue pairs in the two structures;
$d_i$ - distance between $i$-the pair of equivalent residues in the superimposed structures;
$d_0$ - scale to normalize the match difference, where $d_0 = 1.24 \sqrt{L_N - 15} - 1.8$;
‘Max’ denotes the maximum value after optimal spatial superposition.

Differently from the GDT-TS, the TM-score is calculated from a single superposition and accounts for all corresponding residue pairs (not limited to the residue pairs fitting under a set of cutoffs).

TM-score web server and standalone software: [https://zhanglab.ccmb.med.umich.edu/TM-score/](https://zhanglab.ccmb.med.umich.edu/TM-score/)

1.4 QCS

Quality Control Score (QCS) (Cong, et al., 2011) examines the topological similarity of two structures and is the weighted sum of the following six components:
(1) Correct prediction of Secondary Structure Elements (SSE) measured by the length of SSEs;
(2) Relative position of pairs of SSEs measured by the distances between representative points on the SSEs;
(3) Relative angle between SSEs pairs;
(4) Distances between the Cα atoms in the key contacts between SSEs;
(5) The handedness of SSE triplets;
(6) Cα contact score.

The score was developed to mimic human expert scores based on visual inspection of ab initio models.

QCS standalone software: [http://prodata.swmed.edu/QCS/](http://prodata.swmed.edu/QCS/)

2. Description of local scores (CAD-score, LDDT, RPF and SphereGrinder)

2.1 CAD-score (CAD-AA, CAD-AS, CAD-SS)

CAD-score (Olechnovič, et al., 2013; Olechnovič and Venclovas, 2014) measures similarity between protein model and the native structure by quantifying differences in their residue-residue contact areas (CAD). The contact areas are calculated on the per-atom basis using the
Voronoi tessellation of protein structures. CAD-score can be calculated by considering all residue atoms and/or their standard subsets (main chain, side chain). In this study, CAD-AA (CAD-score based on all atom-all atom contacts), CAD-AS (CAD-score based on all atom-side chain contacts) and CAD-SS (CAD-score based on side chain-side chain contacts) are considered.

\[
\text{CAD-score} = 1 - \frac{\sum_{(i,j) \in G} \min \left( |T_{(i,j)} - M_{(i,j)}|, T_{(i,j)} \right)}{\sum_{(i,j) \in G} T_{(i,j)}}
\]

\( G \) - set of all the contacting residue pairs \((i,j, i \neq j)\) in the reference (target) structure
\( T_{(i,j)} \) - contact area of a residue pair in the reference (target) structure
\( M_{(i,j)} \) - contact area of a residue pair in the structure to be compared (model)

CAD-score web server and standalone software: [http://bioinformatics.ibt.lt/cad-score/](http://bioinformatics.ibt.lt/cad-score/)

### 2.2 LDDT

LDDT is based on the comparison of all-atom distance maps of model and target structures (Mariani, et al., 2013). The LDDT scoring function is the percentage of preserved distances between all pairs of atoms of the target structure that are closer in space than a predefined cutoff (15 Å by default). A distance is considered preserved in the model if it is within a certain tolerance threshold from the corresponding distance in the target. The final score is the average of the percentages of the preserved distances under four distance tolerance cutoffs: 0.5, 1, 2 and 4 Å.

\[
LDDT = \frac{1}{4} \frac{N_{0.5} + N_1 + N_2 + N_4}{L}
\]

where

L - number of atom pairs that do not belong to the same residue and are no further apart than 15 Å in the target structure;

\( N_d \) - number of atom pairs in the model, whose inter-atom distances deviate by no more than \( d = 0.5, 1, 2 \) and 4 Å from the corresponding distances in the target (from among the set of L atom pairs).


### 2.3 RPF

RPF (Recall, Precision and F-measure) score estimates similarity of two structures based on the similarity of their inter-atom distance patterns. The RPF algorithm was originally developed for assessing accuracy of NMR structures (Huang, et al., 2005) and later adopted for evaluation of template-based models (Huang, et al., 2014). RPF scores are computed using the following procedure. Each of the compared structures is represented as a graph with all N and C atoms as vertices. Two vertices are connected by an edge if the distance between them is smaller than the predefined parameter (9 Å by default). The agreement between the two structures is then established by comparing the differences between the two graphs. Edges that are present in both target and model are true positives (TP), those present only in
target are false negatives (FN), and those present only in model are false positives (FP). From these, the Recall, Precision, and F2-measure are computed:

\[
\text{precision} = \frac{TP}{TP + FP}, \\
\text{recall} = \frac{TP}{TP + FN}, \\
F2 = 5 \times \text{precision} \times \text{recall} / (4 \times \text{precision} + \text{recall}).
\]

The final RPF score between the model and the target is a normalized F2 score calculated according to the formula:

\[
RPF (Mdl, Targ) = \frac{F2 (Mdl, Targ) - F2 (Rand, Targ)}{1 - F2 (Rand, Targ)},
\]

where

\[F2 (Rand, Targ)\] is calculated by comparing the distance networks of a random coil and the target structure.

### 2.4 SphereGrinder

SphereGrinder (SphGr) is an all-atom fitness measure for estimation of global model-target similarity based on the local similarity of their substructures (Kryshtafovych, et al., 2014; Lukasiak, et al., 2015). For every residue, the RMSD score is calculated on sets of all atoms inside the spheres of a selected radius centered on the corresponding Cα atoms in model and target. For a selected sphere radius, the SphGr-score reports percentage of the residue-centered substructures for which the adjusted RMSD score (normalized by the ratio of predicted atoms) is below the selected cutoff. The cumulative SphGr score is calculated according to the formula:

\[
\text{SphGr} = \frac{1}{2} \cdot \frac{N_2 + N_4}{N},
\]

where

\[N\] - number of residues in the reference structure (target);
\[N_d\] - number of 6Å-radius spheres, which can be superimposed under the \(d\) Å RMSD cutoff.

## Supplementary Table and Figures

### Table S1. Score comparison for selected models of CASP10 target T0663-D1

<table>
<thead>
<tr>
<th>Model</th>
<th>CAD-AA</th>
<th>LDDT</th>
<th>RPF</th>
<th>SphGr</th>
<th>GDT-TS</th>
<th>TMscore</th>
<th>QCS</th>
<th>RMS-CA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full domain (53-138)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0663TS301_3</td>
<td>0.63</td>
<td>0.61</td>
<td>0.65</td>
<td>55.81</td>
<td>63.95</td>
<td>0.64</td>
<td>89.56</td>
<td>7.659</td>
</tr>
<tr>
<td>T0663TS301_5</td>
<td>0.61</td>
<td>0.61</td>
<td>0.66</td>
<td>51.16</td>
<td>68.61</td>
<td>0.67</td>
<td>89.79</td>
<td>5.736</td>
</tr>
<tr>
<td>T0663TS476_4</td>
<td>0.52</td>
<td>0.51</td>
<td>0.55</td>
<td>41.86</td>
<td>60.76</td>
<td>0.60</td>
<td>83.06</td>
<td>13.936</td>
</tr>
<tr>
<td><strong>Truncated domain (53-124)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0663TS301_3</td>
<td>0.67</td>
<td>0.68</td>
<td>0.74</td>
<td>83.33</td>
<td>76.39</td>
<td>0.73</td>
<td>91.65</td>
<td>3.68</td>
</tr>
<tr>
<td>T0663TS301_5</td>
<td>0.66</td>
<td>0.68</td>
<td>0.73</td>
<td>84.03</td>
<td>77.78</td>
<td>0.75</td>
<td>91.63</td>
<td>3.42</td>
</tr>
<tr>
<td>T0663TS476_4</td>
<td>0.58</td>
<td>0.60</td>
<td>0.65</td>
<td>69.44</td>
<td>72.57</td>
<td>0.69</td>
<td>86.34</td>
<td>3.68</td>
</tr>
</tbody>
</table>
Figure S1. Empirical distribution of values for all considered scores and their variants on a single-domain data set. Horizontal axis indicates score values, vertical axis – frequency of the value occurrence.
**Figure S2.** Empirical cumulative distribution function (ECDF) plots for different scores on a single-domain set. (A) ECDF plots for raw score values. (B) ECDF plots for Z-scores, calculated separately for every reference structure.
Figure S3. Correspondence of values for every pair of score ‘A’ and score ‘B’ presented as scatter plots. Horizontal axis in each plot indicates values of score ‘A’, vertical axis indicates values of score ‘B’. Increasing color intensity represents a higher local density of values.
**Figure S4.** Correspondences of score values for GDT-TS/HA and local scores (including variants of CAD-score) presented as scatter plots for ‘A’ and ‘B’ score pairs. Horizontal axis in each plot indicates values of score ‘A’, vertical axis indicates values of score ‘B’. Increasing color intensity represents a higher local density of values.
Figure S5. Correlation of scores. (A) Spearman’s rank correlation coefficients for the considered scores, including their variants. Correlation coefficients were 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).
Figure S6. Score differences in selecting a better model out of two for all the considered scores, including their variants. (A) Fractions of model pairs, where the disagreement between scores exceeds the tolerance threshold. Differences are colored from blue (smallest) to red (largest). (B) Graphical representation of score differences using MDS.
Figure S7. Agreement between the scores in selecting either the single best model or the 10 best models out of many. (A) Average fraction of common single best models selected by score pairs. (B) Graphical representation of corresponding score differences using MDS. (C) Average fraction of common models within 10 best models selected by score pairs (fraction value of 1 corresponds to 10 common models). (D) Graphical representation of corresponding score differences using MDS.
Figure S8. Score differences in selecting either the best model or 10 best models out of many for all the considered scores, including their variants. (A) Average losses of Z-score computed by score ‘B’ (vertical axis) when the single 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) Graphical representation of corresponding score differences using MDS. (C) Average losses of Z-score computed by score ‘B’ (vertical axis) when the 10 best models are 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. (D) Graphical representation of corresponding score differences using MDS.
Figure S9. 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 tolerance threshold was applied to both the considered scores and MolProbity values. MolProbity’s support is colored from blue (largest) to red (smallest). Results are shown for (A) for MolProbity score, (B) for interatomic clashes, (C) for rotamer outliers and (D) for Ramachandran outliers.
**Figure S10.** Conflicting rankings of model pairs “judged” by the number of reproduced hydrogen bonds. The scores are compared using differences larger than the defined threshold (A) for all hydrogen bonds and (B) for only non-local hydrogen bonds.
Figure S11. Grishin plots reflecting sensitivity of different scores and their variants 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.
Figure S12. 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 5%, 10% and 15% (for example, if the sliding window is 5%, the points at 80% are averages of the 85-80% range).
Figure S13. Dependence of the scores on protein length. Horizontal axis indicates the target length. Vertical axis indicates mean values of models for targets of a given length (averaged by including 5, 10 and 20 neighbors on both sides). Left panels show score curves of actual values, right panels show score curves overlaid onto each other by centering them on the horizontal axis.
**Figure S14.** Dependence of the scores on the secondary structure type. Horizontal axis indicates targets arranged by the increasing $\alpha$-helical content from largely $\beta$-structural (blue) to mixed (green) to mostly $\alpha$-helical (red). Vertical axis shows averaged score values. Per-target values of every considered score were averaged using two different sliding windows (5 and 10 neighbors on each side). Left panels show score curves of actual values, right panels show score curves overlaid onto each other by centering them on the horizontal axis.
References:


