PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

VoroMQA: Assessment of protein structure quality using interatomic contact areas

Kliment Olechnovič[1,2] and Česlovas Venclovas[1]*

[1] Institute of Biotechnology, Vilnius University, Saulėtekio 7, LT-10257 Vilnius, Lithuania

[2] Faculty of Mathematics and Informatics, Vilnius University, Naugarduko 24, LT-03225 Vilnius, Lithuania

ABSTRACT

In the absence of experimentally determined protein structure many biological questions can be addressed using computational structural models. However, the utility of protein structural models depends on their quality. Therefore, the estimation of the quality of predicted structures is an important problem. One of the approaches to this problem is the use of knowledge-based statistical potentials. Such methods typically rely on the statistics of distances and angles of residue-residue or atom-atom interactions collected from experimentally determined structures. Here, we present VoroMQA (Voronoi tessellation-based Model Quality Assessment), a new method for the estimation of protein structure quality. Our method combines the idea of statistical potentials with the use of interatomic contact areas instead of distances. Contact areas, derived using Voronoi tessellation of protein structure, are used to describe and seamlessly integrate both explicit interactions between protein atoms and implicit interactions of protein atoms with solvent. VoroMQA produces scores at atomic, residue, and global levels, all in the fixed range from 0 to 1. The method was tested on the CASP data and compared to several other single-model quality assessment methods. VoroMQA showed strong performance in the recognition of the native structure and in the structural model selection tests, thus demonstrating the efficacy of interatomic contact areas in estimating protein structure quality. The software implementation of VoroMQA is freely available as a standalone application and as a web server at http://bioinformatics.lt/software/voromqa.

Proteins 2017; 85:1131–1145.

Key words: protein structure prediction; protein structure modeling; estimation of model quality; additively weighted Voronoi diagram; Voronoi tessellation of balls; knowledge-based statistical potential.

INTRODUCTION

The ability to predict protein three-dimensional (3D) structure from sequence is one of the most important and challenging problems in computational biology. Protein structure prediction methods tackling this problem are being developed continuously and in many cases they can produce models that are close to the native structure. The performance of such methods is systematically assessed during community-wide CASP experiments[1,2] that not only reveal successes, but also point out the bottlenecks in the field of protein structure prediction. One of the most prominent bottlenecks is the model quality assessment (QA). Current structure prediction methods typically produce multiple models for a given protein, and then QA methods are used to identify the best model and to estimate how realistic the model is. However, according to the results of recent CASP experiments,[3,4] model quality assessment remains a difficult task and there is a clear need for better QA methods.

There are several classes of QA methods. Among them, single-model methods, that is, methods that can evaluate a single structural model without the need to analyze a diverse ensemble of other models, are particularly well-suited for the practical use outside of CASP-like settings. Some of the most successful single-model QA methods, for example, ProQ2[5] and QMEAN,[6] are meta-methods that combine several sources of information about an input structure. Such meta-methods often employ machine learning techniques to produce a single generalized quality score out of several lower-level scores such as the estimates of free energy and agreement scores

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

*Correspondence to: Česlovas Venclovas, Institute of Biotechnology, Vilnius University, Saulėtekio 7, LT-10257 Vilnius, Lithuania. E-mail: ceslovas.venclovas@bti.vu.lt

Received 18 November 2016; Revised 13 January 2017; Accepted 21 February 2017
Published online 6 March 2017 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25278

that tell how well some of the observed structural features, such as secondary structure and residue solvent-accessibility, correspond to the sequence-based predictions. A viable approach for creating a better QA method is designing better techniques to combine available scores, another approach is to design better independent scores that perform well on their own or become useful components of meta-methods.

Prominent examples of independent QA methods are knowledge-based statistical potentials. Over the last twenty years or so a number of different statistical potentials have been developed. Most of them rely on statistics of pairwise interaction distances,[7–12] some also utilize information about interaction angles.[13–15] However, distance-based metrics may not necessarily be best-suited for the description and analysis of physical properties of protein structure.

Recently, we have developed CAD-score, a new reference score, which uses differences in interatomic contact areas instead of distance differences to measure the similarity between a model and the native protein structure.[16] It has turned out that CAD-score is highly correlated with traditional distance-based reference scores such as GDT_TS,[17] at the same time displaying stronger preference for physical realism in protein structure. Moreover, CAD-score provides a possibility to treat single-domain, multidomain and multi-subunit structures in exactly the same way. The simplicity and robustness of CAD-score suggested that contact areas may also be effective in developing knowledge-based potentials. In fact, the first attempt to employ contact areas as a foundation for knowledge-based potentials was made over a decade ago by McConkey et al.[18] Contact areas in both CAD-score and the McConkey method are derived from variants of the Voronoi tessellation. Voronoi and related tessellation methods proved to be an effective means in the analysis of various structural features,[19–23] including the identification of physical contacts that could be utilized in deriving distance-based statistical potentials.[24–26] However, to the best of our knowledge, the study by McConkey et al. so far has been the only QA method based on tessellation-derived contact areas. Their method achieved respectable results in discriminating native protein structures from decoys; however, perhaps mainly due to the lack of publicly available software implementations, the prospects of applying contact areas for the assessment of protein structural models remained largely unexplored. Using the same concept of employing contact areas as a starting point, we have developed a new all-atom statistical potential-based method for protein structure quality assessment. The new method, VoroMQA ("Voronoi tessellation-based Model Quality Assessment"), considers protein structure as a set of balls corresponding to heavy atoms and characterize interactions through interatomic contact areas derived from the Voronoi tessellation of atomic balls.[27] Here, we present description of the method and compare its performance with both statistical potentials and composite model quality assessment scores.

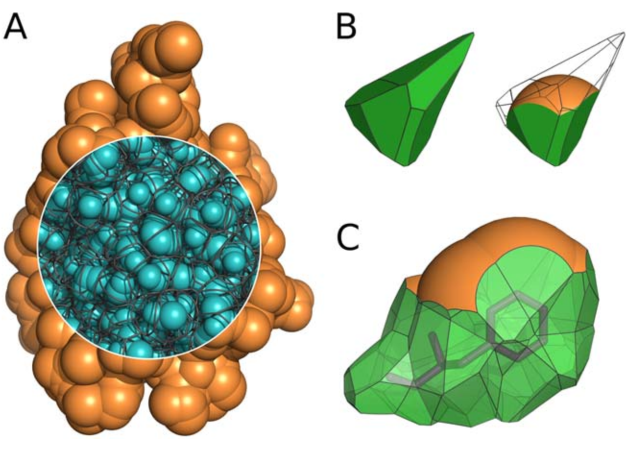

Figure 1

(A) Edges of the Voronoi cells constrained inside the solvent accessible surface of a protein structure. (B) Cutting a Voronoi cell with a sphere corresponding to the rolling probe surface results in constrained Voronoi faces and SAS patches. (C) An integral surface of a phenylalanine residue constructed by combining atomic contact surfaces. [Color figure can be viewed at wileyonlinelibrary.com]

MATERIALS AND METHODS

Construction of contacts

Given a protein structure, it can be represented as a set of atomic balls, each ball having a van der Waals radius depending on the atom type. A ball can be assigned a region of space that contains all the points that are closer (or equally close) to that ball than to any other. Such a region is called a Voronoi cell and the partitioning of space into Voronoi cells is called a Voronoi tessellation. Two adjacent Voronoi cells share a set of points that form a surface called a Voronoi face. A Voronoi face can be viewed as a geometric representation of a contact between two atoms. However, if a pair of contacting atoms is near the surface of a protein structure, the corresponding Voronoi face may extend far away from the atoms. Here, this problem is solved by constraining the Voronoi cells of atomic balls inside the boundaries defined by the solvent accessible surface (SAS) of the same balls, as illustrated in Figure 1(A,B). The resulting constrained Voronoi faces and SAS patches can be combined into integral surfaces of larger components of protein structure, for example, amino acids [Fig. 1(C)]. Construction of interatomic contact surfaces is implemented as part of the Voronota software.[27] The construction procedure uses triangulated representations of Voronoi faces and spherical surfaces. Contact areas are calculated as the areas of the corresponding triangulations.

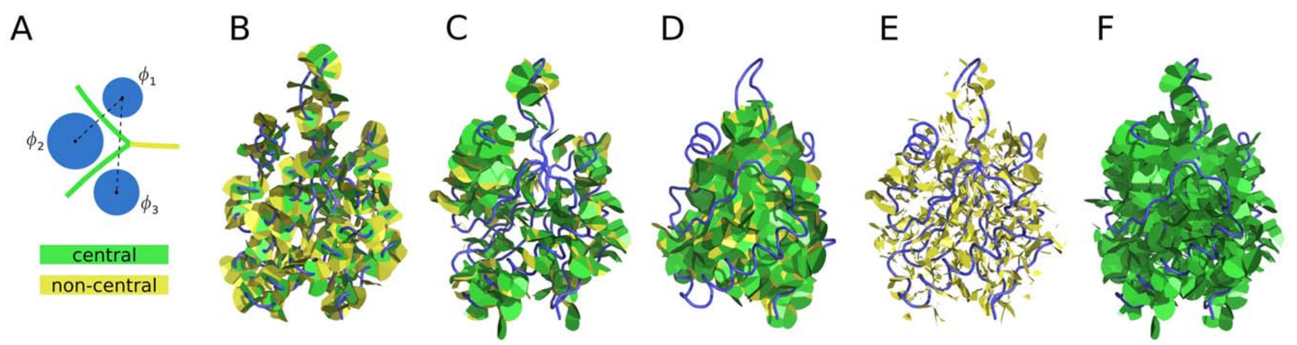

Figure 2

(A) 2D illustration of central and noncentral contacts: the contact between balls ϕ_1 and ϕ_3 is noncentral, the other contacts are central. (B) Central (green) and noncentral (yellow) contacts for sequence separation 1. (C) Central and noncentral contacts for sequence separation from 2 to 6. (D) Central and noncentral contacts for sequence separation >6. (E) Only noncentral contacts for sequence separation >1. (F) Only central contacts for sequence separation >1. The PDB ID of the protein structure used in this figure is 1T3Y. [Color figure can be viewed at wileyonlinelibrary.com]

In this study the Voronoi tessellation-based analysis is also used to describe the centrality of contacts. Given a pair of contacting atoms, the contact between them is called *central* if the line segment connecting the centers of the atoms intersects the corresponding constrained Voronoi face. Otherwise, the contact is called *noncentral*. The definition of central and noncentral contacts is illustrated in Figure 2(A). Another categorization of contacts used in this work is based on the sequence separation between the residues of the contacting atoms. It is illustrated in Figure 2(B–F) in combination with the centrality-based categorization.

Definition of the quality scoring method

Interatomic and solvent contact areas may be used to evaluate quality of protein structural models by employing the idea of a knowledge-based statistical potential as was first shown by McConkey et al.[18] Our method is aimed to employ the same principle using more elaborate contact descriptions and to be able to produce both local (atom-level) and global (structure-level) scores in a fixed range of values from 0 to 1.

In order to formulate our method, the first step is to define a set of possible contact types. Let $A=\{a_0, a_1, \ldots, a_n\}$ be a set of atom types and $C=\{c_0, c_1, \ldots, c_m\}$ be a set of contact categories. A contact type is described by a tuple $(a_i, a_j, c_k) \in A\times A\times C$, which is equivalent to (a_j, a_i, c_k) because contacts are undirected. The atom type a_0 represents solvent and the contact category c_0 represents solvent-accessible areas, therefore a_0 and c_0 always come together and the set of all possible contact types can be narrowed down to $T=([A\setminus a_0]\times[A\setminus a_0]\times[C\setminus c_0])\cup([A\setminus a_0]\times\{a_0\}\times\{c_0\})$.

A contact type can be assigned a pseudo-energy value $E(a_i, a_j, c_k)$ calculated from the corresponding expected and observed probabilities:

$$E(a_i, a_j, c_k)=\log\frac{P_{\exp}(a_i, a_j, c_k)}{P_{\text{obs}}(a_i, a_j, c_k)} \tag{1}$$

The probability values can be estimated empirically using the contact area values calculated for a learning set of high-quality experimentally determined protein structures. Let $S(a_i, a_j, c_k)$ be a sum of all the areas of the contacts of type (a_i, a_j, c_k) observed in the learning set. Also, let us define that if $(a_i, a_j, c_k) \in T$, then $S(a_i, a_j, c_k)=0$. Let S_{sol} and S_{int} be sums of solvent and interatomic contact areas, respectively:

$$S_{\text{sol}}=\sum_{1\le i\le n} S(a_i, a_0, c_0) \tag{2}$$

$$S_{\text{int}}=\sum_{1\le i\le n}\sum_{1\le j\le i}\sum_{1\le k\le m} S(a_i, a_j, c_k) \tag{3}$$

Then the observed probability of the contact type (a_i, a_j, c_k) is defined as the following ratio of areas:

$$P_{\text{obs}}(a_i, a_j, c_k)=\frac{S(a_i, a_j, c_k)}{S_{\text{int}}+S_{\text{sol}}} \tag{4}$$

The corresponding expected probability should represent how often the contacts of the same type would occur in a set of randomly folded structures of the same sequences as in the learning set. It is estimated using the observed probabilities of the isolated components of the contact type (a_i, a_j, c_k):

$$P_{\exp}(a_i, a_j, c_k)=\begin{cases} P_{\text{obs}}(a_i)\cdot P_{\text{obs}}(c_0) & \text{if } j=0 \\ P_{\text{obs}}(a_i)\cdot P_{\text{obs}}(a_j)\cdot P_{\text{obs}}(c_k) & \text{if } j\ge 1, i=j \\ P_{\text{obs}}(a_i)\cdot P_{\text{obs}}(a_j)\cdot 2\cdot P_{\text{obs}}(c_k) & \text{if } j\ge 1, i\ne j \end{cases} \tag{5}$$

$$P_{\text{obs}}(a_i)=\frac{\sum_{0\le j\le n}\sum_{0\le k\le m} S(a_i, a_j, c_k)}{2S_{\text{int}}+S_{\text{sol}}} \tag{6}$$

$$\mathrm{P_{obs}}(c_k)=\frac{\sum_{0\le i\le n}\sum_{0\le j\le i}\mathrm{S}(a_i,a_j,c_k)}{\mathrm{S_{int}}+\mathrm{S_{sol}}} \tag{7}$$

Having the derivation of pseudo-energy values defined using Eqs. (1–7), let us describe how the derived values are used for scoring protein structures. In order to assign a quality score to a single atom ϕ, a set of related contacts Ω_ϕ is selected. Atom-related contacts are defined as not only the immediate contacts of the considered atom, but also all the contacts of the neighboring atoms. A normalized pseudo-energy value $E_\mathrm{n}(\Omega_\phi)$ is computed using the information known about each contact $\omega\in\Omega_\phi$, namely the contact area (area_ω) and the contact type ($\mathrm{type}_\omega\in T$):

$$E_\mathrm{n}(\Omega_\phi)=\frac{\sum_{\omega\in\Omega_\phi}\mathrm{E}(\mathrm{type}_\omega)\cdot\mathrm{area}_\omega}{\sum_{\omega\in\Omega_\phi}\mathrm{area}_\omega} \tag{8}$$

An atom quality score $\mathrm{Q_a}(\Omega_\phi)\in[0,1]$ is defined using the Gauss error function:

$$\mathrm{Q_a}(\Omega_\phi)=\frac{1}{2}\left(1+\mathrm{erf}\left(\frac{\mathrm{E_n}(\Omega_\phi)-\mu_{\mathrm{type}_\phi}}{\sigma_{\mathrm{type}_\phi}\sqrt{2}}\right)\right) \tag{9}$$

The values of μ (mean) and σ (standard deviation) are estimated for each atom type from the normalized pseudo-energy values calculated for the atoms in the learning set of protein structures.

Given a set Φ of all atoms in a protein structure, a global structure quality score $Q_\mathrm{g}(\Phi)$ is defined as a weighted arithmetic mean of the atoms quality scores:

$$Q_\mathrm{g}(\Phi)=\frac{\sum_{\phi\in\Phi}Q_\mathrm{a}(\Omega_\phi)\cdot\mathrm{weight}_\phi}{\sum_{\phi\in\Phi}\mathrm{weight}_\phi} \tag{10}$$

The weights here indicate how deep each atom is buried inside a structure: solvent-accessible atoms have weight 1, their direct contacting neighbors have weight 2, the neighbors of the direct neighbors have weight 3, and so on.

The quality score of a residue is defined as an average of quality scores of its atoms. A sliding window with four residues on both sides is used to smooth residue scores along the sequence. Let us denote an unsmoothed residue score at position n as $Q_\mathrm{r}(n)$, then the corresponding smoothed value $W_\mathrm{r}(n)$ is computed as a normalized weighted sum of the scores of the neighboring residues:

$$W_\mathrm{r}(n)=\frac{\sum_{-5<m<5}Q_\mathrm{r}(n+m)\cdot(5-|m|)}{\sum_{-5<m<5}(5-|m|)} \tag{11}$$

Implementation of the quality scoring method

Implementation of the method requires a protein structure dataset (learning dataset) for collecting data on interatomic contacts, the set of atom types and the set of contact categories. Protein structures for the learning set were obtained from the Protein Data Bank[28] (www.rcsb.org). Only protein structures solved by X-ray at better than 2.5 Å resolution were considered. The set was limited to monomeric or oligomeric (up to 12 subunits) proteins with each chain longer than 99 residues. Proteins solved in complex with nucleic acids, membrane proteins, proteins with modified polymeric residues were excluded. From the remaining structures only representatives at 50% sequence identity were retained. For each of the resulting PDB entries (totaling 12825 as of 2015.06.11), the structure of the first biological assembly was used for deriving contact areas. Only nonbonded contacts between atoms of different residues were considered.

In the case of multi-chain biological assemblies the set of derived contacts is redundant. To remove this redundancy, the contact areas are multiplied by the ratio $\frac{N_u}{N}$, where N is the total number of chains and N_u is the number of unique protein chains.

Twenty standard amino acids have 167 different heavy atom names, however, seven atom pairs are interchangeable because of the molecular symmetry: Arg NH1 and NH2, Asp OD1 and OD2, Glu OE1 and OE2, Phe CD1 and CD2, Phe CE1 and CE2, Tyr CD1 and CD2, Tyr CE1 and CE2. Therefore, the final set contains 160 distinct atom types, plus one special type representing solvent.

As for the set of contact categories, a hybrid scheme is used: solvent contacts are treated separately; each nonsolvent contact is categorized as either near or far depending on the sequence separation between the residues of the contacting atoms; each nonsolvent contact is categorized as either central or noncentral as illustrated in Figure 2(A). This results in 5 distinct categories: "solvent", "near and central", "near and noncentral", "far and central", "far and noncentral". During the method learning stage, when the empirical probabilities are computed, a contact is considered far if the corresponding sequence separation is >6: this is done to separate the contacts that may be largely induced by the close sequence proximity of the contacting residues from the contacts that are more likely to occur because they are favorable. During the method application stage, when calculating normalized pseudo-energies of atoms using Eq. (8), only far or solvent contacts are considered, but the sequence separation threshold for contacts considered as far is lowered so that only contacts between the atoms of residues adjacent in sequence are categorized as near. This allows to take into account the vast majority of contacts while excluding the ones that are likely to appear in a structural model regardless of its correctness.

When estimating the probabilities of the contact categories using Eq. (7), we tried two datasets for input: the learning set of high quality structures and a set of lower

Table I

Observed Probabilities of the Contact Categories Estimated for The Learning Set Of High Quality Structures (P_{obs}^{high}) and the Set of Lower Quality Structures Comprised of CASP Models (P_{obs}^{low})

Category	P_{obs}^{high}	P_{obs}^{low}
Near and central	0.159	0.147
Near and noncentral	0.168	0.165
Far and central	0.225	0.166
Far and noncentral	0.056	0.052
Solvent	0.392	0.470

quality structures that was comprised of the models of the monomeric targets from CASP8,[29] CASP9,[30] and CASP10.[1] Table I contains the two resulting sets of probability values, the most prominent difference between them being the solvent contact probabilities, meaning that the lower quality structures are not as well packed as the high quality ones. We reasoned that random protein-like structures should also be packed worse than the native protein structures, therefore for Eq. (5) we employed the probabilities of the contact categories that were estimated from the set of lower quality structures.

The last required information is the mean and standard deviation values used in Eq. (9). These values were calculated for each atom type after applying Eq. (8) to every atom in the learning set of protein structures.

The VoroMQA software is available both as a standalone application and as a web-server at http://bioinformatics.lt/software/voromqa. Our standalone software does not require any third-party programs or libraries to work. However, in some cases it may be beneficial to employ an external tool to rebuild the side chains in input protein structures before evaluation as this may reduce chances of overly penalizing structural models that have good backbone but poor side-chain packing.

Expected scores for native protein structures

The implemented method was used to compute the global quality scores of the protein structures in the original learning set to estimate what scores to expect from realistic structural models. Each structure was evaluated twice: first time, using the default method configuration, and second time, after rerunning the learning stage with the structure of interest removed from the learning set. The mean difference between the first and the second global scores was <0.00028, for 99% of the structures the difference was <0.0006, the maximum observed difference was 0.0038. This allows us to conclude that the performance of the method is largely insensitive to the presence or absence of any single structure in the learning set.

The summary of global quality scores calculated by the second procedure is presented in Figure 3. Plot (A) shows the empirical distribution of global scores leading to the following observations: 1) it is unlikely for a realistic protein structure to have a global score lower than 0.3 or >0.7, and 2) a global quality score is not heavily dependent on the prevailing type of secondary structure. Plot (B) shows that, on average, smaller protein structures receive slightly lower global quality scores than larger structures, and the variance is greater for smaller structures. Another aspect of the method is that the scoring time scales linearly with the structure size, as illustrated in Figure 3(C).

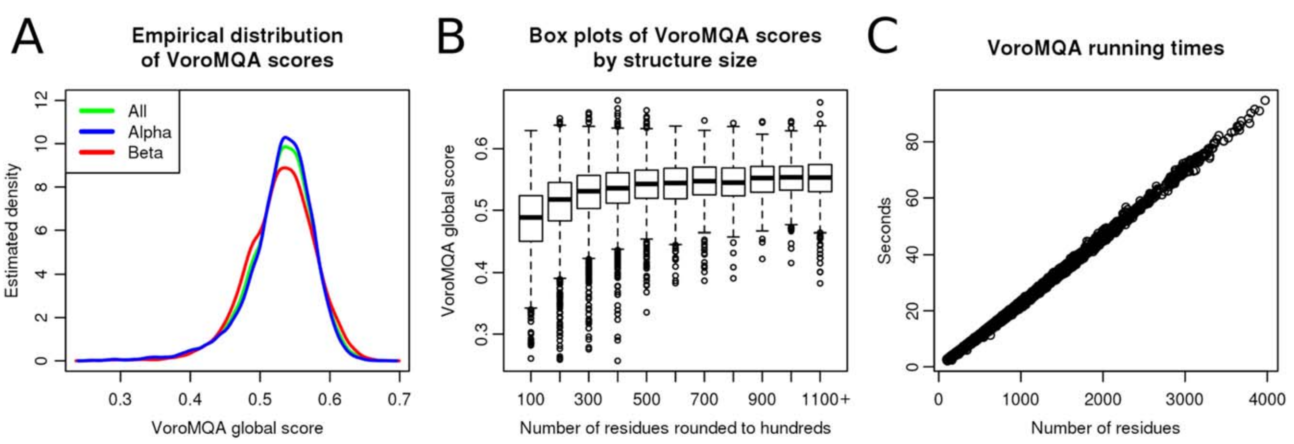

Figure 3

Recap of the global quality scores calculations performed for the protein structures in the learning set. (A) Estimated empirical density functions of the scores for all the structures (green), the structures with prevailing alpha helices (blue) and the structures with prevailing beta sheets (red). (B) Box plots of global scores for different thresholds of structure sizes (the rightmost box plot also covers all the structures from the learning set that have >1100 residues). (C) Software running times plotted against the corresponding structure sizes (the test was performed using CPU Intel® Xeon® E5-2670 v3 @ 2.30 GHz). [Color figure can be viewed at wileyonlinelibrary.com]

A note on the older version of the method

The initial simplified variant of the VoroMQA method was tested in the QA category of the CASP11 experiment.[4] Compared to the current version, the older one did not utilize contact categories, did not distinguish between different atom types when converting from pseudo-energy values to atomic quality scores and did not assign weights to the atomic scores when calculating global quality scores. Also, only single-chain protein structures from the PISCES[31] database were used in the learning stage of the older version. To assess the effect of the differences between the older and the newer versions of VoroMQA, we recorded the results achieved by both versions in the tests described later in this article. When describing the test results, the older version of VoroMQA is denoted as "VoroMQA-old" and the current version is denoted as "VoroMQA-new" or simply "VoroMQA".

Datasets used to assess the performance of the method

In order to analyze the ability of VoroMQA to select a native structure from a set of its models of varying quality, we downloaded the target and model structures from the last four CASP experiments (CASP8-11). We did not consider targets that correspond to individual subunits of obligatory protein complexes representing biologically unrealistic oligomeric state or those solved with poor resolution. Consequently, we used 140 CASP targets that conform to the following criteria: a target must correspond to a PDB entry that has at least one single-chain biological assembly and the experimental method used to determine the structure must be X-ray crystallography with resolution better than 2.5 Å. For every target, all the available complete models were downloaded, excessive regions in models were trimmed to exactly match the target structure. Scores for target and model structures were then computed using the old and the new variations of VoroMQA as well as DOOP,[12] GOAP,[15] and dDFIRE.[13]

In order to analyze the ability of VoroMQA to evaluate protein structural models, we used the CASP11 Quality Assessment (QA) data. We considered all the 88 targets that were used in the official assessment of CASP11 QA results,[4] but, after manual inspection, excluded four of them (T0775, T0787, T0799 and T0813) because their structures are single chains pulled out from obligatory oligomers. For the remaining 84 targets, the following data was downloaded from the CASP11 website: the structures of server models (best150 and sel20 sets[3]); the corresponding reference-based quality scores (GDT_TS,[17,32] LDDT,[33] SphGr (SphereGrinder)[34] and CAD_AA[16]); the available global scores calculated by the single-model quality assessment methods (MULTICOM-CLUSTER,[35] MULTICOM-NOVEL,[36] ProQ2,[5] ProQ2-refine,[37] Wang_SVM,[38] Wang_deep_{1,2,3}[38] and the old version of VoroMQA). The latest version of our method was applied to produce two more scores for each model: VoroMQA-new (calculated for the unmodified input structure) and VoroMQA-new-sr (calculated after rebuilding the side-chains of the input structure using SCWRL4[39]). Additionally, we computed the following statistical potential-based scores: DOOP, GOAP, and dDFIRE. We compiled the retrieved and the computed scores together and removed duplicate model entries and entries with at least one score absent. The final combined tables of scores characterize 11627 models from the best150 sets and 1583 models from the sel20 sets (the tables are available for download from the VoroMQA web page).

The best150 and sel20 sets, composed at the time by CASP11 organizers, differ in both their size and nature. For each target, the best150 set contains the best150 models selected using a consensus-based QA algorithm, while the sel20 set contains 20 diverse models selected based on a clustering of all the available models for the target. We reasoned that it may also be interesting to perform tests on sets that are small (like sel20) but contain better models (like best150). Therefore, in addition to the best150 and sel20 sets for each target, we used sets of models produced by the three well-performing prediction servers: BAKER-ROSETTASERVER,[40] Zhang-Server,[41] and QUARK.[42] We dubbed these sets "BZQ15" because only up to 15 models of the three servers are available for each CASP11 target. Each BZQ15 set simulates the real-life scenario, when a researcher needs to choose the best model from a few generated by several well-known servers.

Reference-based scores used to assess the model selection capabilities

For assessing the VoroMQA performance when selecting the best model from a set of models of the same target, we chose to employ the same four reference-based scores (GDT_TS, LDDT, SphGr and CAD_AA) used in the official CASP11 QA assessment. However, each of the four scores focuses on somewhat different structural properties and they often disagree in deciding which model is closer to the native structure. Moreover, each score has a degree of uncertainty so that the score difference for close models may not always be significant.[43] To take care of these issues, we additionally introduced a simple tournament-based methodology described below.

Let us take two models a and b of the same target. Let us say that a "wins" against b if all the four reference-based scores are higher for a than for b. If there is a disagreement, for example, if $\text{GDT\_TS}_a > \text{GDT\_TS}_b$ but $\text{LDDT}_a < \text{LDDT}_b$, then the outcome of the duel between a and b is a draw. Using the defined rules, all the possible duels are executed for the models in the input data set. For each model the numbers of wins, draws and losses are recorded. The results of the performed "tournament" are used as a basis for our ensuing analysis.

A straightforward way to utilize the tournament results is to assess how well a QA method is able to select the best model out of two. Let us consider a set of models M, let N be the total number of nondraw duels among the elements of M and N_p be the number of nondraw duels that the QA method correctly predicts the winner for. Then the QA method performance can be quantified using the agreement percentage score:

$$\text{Agreement-score}(M) = N_p / N \cdot 100\% \qquad (12)$$

The next step is to assess the ability of a QA method to select the best model out of many. In our tournament-based framework we define the true best model of a target as the model with the highest number of won duels. If two models have the same number of wins, the one with more draws, that is, less losses, is considered better. We combine the numbers of wins and the numbers of draws into a single score called Wins-score. Let us consider a target t that has N_t models and its best model b has w_b wins with d_b draws. Given a model m of t that has w_m wins with d_m draws, let us define Wins-score score for m. To ensure that even a single win has more weight than $N_t - 1$ draws, numbers of wins are multiplied by N_t in the following formula:

$$\text{Wins-score}(m) = \frac{w_m \cdot N_t + d_m}{w_b \cdot N_t + d_b} \qquad (13)$$

Wins-score(m) can range from 0 (when m has no wins and draws) to 1 (when m is the best model). The score can be interpreted as a measure of success achieved by model m compared to the remaining $(N_t - 1)$ models of t. We use the Win-scores of the models selected by a QA method to quantify the ability of the method to select the best possible models.

Summarizing Agreement-scores (or Wins-scores of selections) for multiple different targets can be done by calculating their mean value. However, when comparing the performances of two different QA methods, a simple comparison of the corresponding mean values is not sufficient as it lacks the information about the significance of the difference. We use the Wilcoxon signed-rank test[44] to assess whether two sets of per-target scores come from two populations with different means. We chose this particular test and not the paired Student's t-test because we cannot assume that a population of Agreement-scores (or a population of selection Wins-scores) is distributed normally. We first run the two-sided Wilcoxon test: if the computed p-values is sufficiently small, that is, the two population means differ significantly, then the one-sided version of the test is used to check if the first population mean is likely larger.

RESULTS

Overview of testing procedures

We tested the performance of VoroMQA in several ways which are outlined in this section and presented in detail in the subsequent sections.

Firstly, we focused on global VoroMQA scores and assessed if the method is able to distinguish a native structure from its decoys. We used data from several CASP experiments to form sets of decoys: such sets are comprised of models of various quality generated by a variety of structure prediction methods. We compared the performance of VoroMQA with the performance of three other methods that are based solely on analyzing geometric features and applying knowledge-based statistical potentials, namely DOOP,[12] GOAP,[15] and dDFIRE.[13]

Next, we analyzed how VoroMQA global scores computed for models relate to the observed differences between models and the native structures (targets). To this end, we used CASP11 structural models and the corresponding reference-based quality scores. As the official assessment of CASP11 QA results[4] was done using primarily GDT_TS,[17,32] LDDT,[33] SphGr (Sphere-Grinder)[34] and CAD_AA (CAD-score$_{AA}$),[16] the same four scores were also employed in our study.

During another test we analyzed the ability of VoroMQA to select the best model out of several or many. For this test we applied the four reference-based scores and the newly introduced tournament-based methodology (see "Materials and methods"), which allows multiple reference-based scores to be considered simultaneously. In addition to DOOP, GOAP and dDFIRE, we compared VoroMQA with single-model quality assessment methods that participated in CASP11, namely MULTICOM-CLUSTER,[35] MULTICOM-NOVEL,[36] ProQ2,[5] ProQ2-refine,[37] Wang_SVM,[38] and Wang_deep_{1,2,3}.[38] Unlike VoroMQA, these methods employ additional data such as secondary structure and solvent accessibility predictions, in effect incorporating evolutionary information derived from homologous sequences. Therefore, matching or surpassing their performance may be considered a serious challenge for VoroMQA.

The last testing procedure was dedicated to the local scoring. We used data from the CAMEO project (www.cameo3d.org)[45] to investigate some properties of VoroMQA local scores in relation with reference-based local scores.

Selecting native structures from sets of decoys

We tested VoroMQA alongside DOOP, GOAP and dDFIRE according to the ability to distinguish a native structure amidst a variety of its models (decoys) using the data corresponding to the 140 monomeric targets

Table II

Results of the Target Selection Ability Analysis Performed for the Set of 140 Monomeric Targets from CASP8, CASP9, CASP10, and CASP11. The "Missed Targets" Column Values Show How Many Times Each QA Method Failed to Distinguish a Target Structure Among its Models. The "Mean *z*-scores" Column Values Show the Average z-scores of the QA Scores of the Target Structures

Method	Missed targets	Mean z-score
VoroMQA-new	8	3.19
DOOP	8	3.00
GOAP	16	2.87
VoroMQA-old	27	2.67
dDFIRE	46	2.09

from CASP8-11 experiments (see "Materials and methods" for details).

The performance of each structure evaluation method was assessed by counting how many times a native (target) structure was missed. Also, differences between the target scores and the corresponding model mean scores were computed and converted to z-scores. According to the number of missed native structures VoroMQA performed on par with DOOP and surpassed the others. The summary of the results is presented in Table II, the per-target results are shown in Supporting Information Table S1.

Relationship between VoroMQA global scores and model quality

As we have shown above (Figure 3), VoroMQA global scores do not significantly depend on either prevalent secondary structure content or protein size. Thus in principle, it should be possible to decide if a computational model is close to the native structure solely on the basis of the VoroMQA global score. Figure 3(A) shows that a vast majority of high quality experimentally determined structures have VoroMQA scores >0.4. Also, almost none of the native structures have VoroMQA scores <0.3. Following these observations, we computed empirical distribution densities of the four reference-based quality scores (GDT_TS, LDDT, SphGr and CAD_AA) of the CASP11 models that have VoroMQA-new scores in the intervals $(0, 0.3)$, $[0.3, 0.4]$ and $(0.4, 1)$. The results, shown in Figure 4, allow us to formulate the following simple rule for interpreting a VoroMQA-new

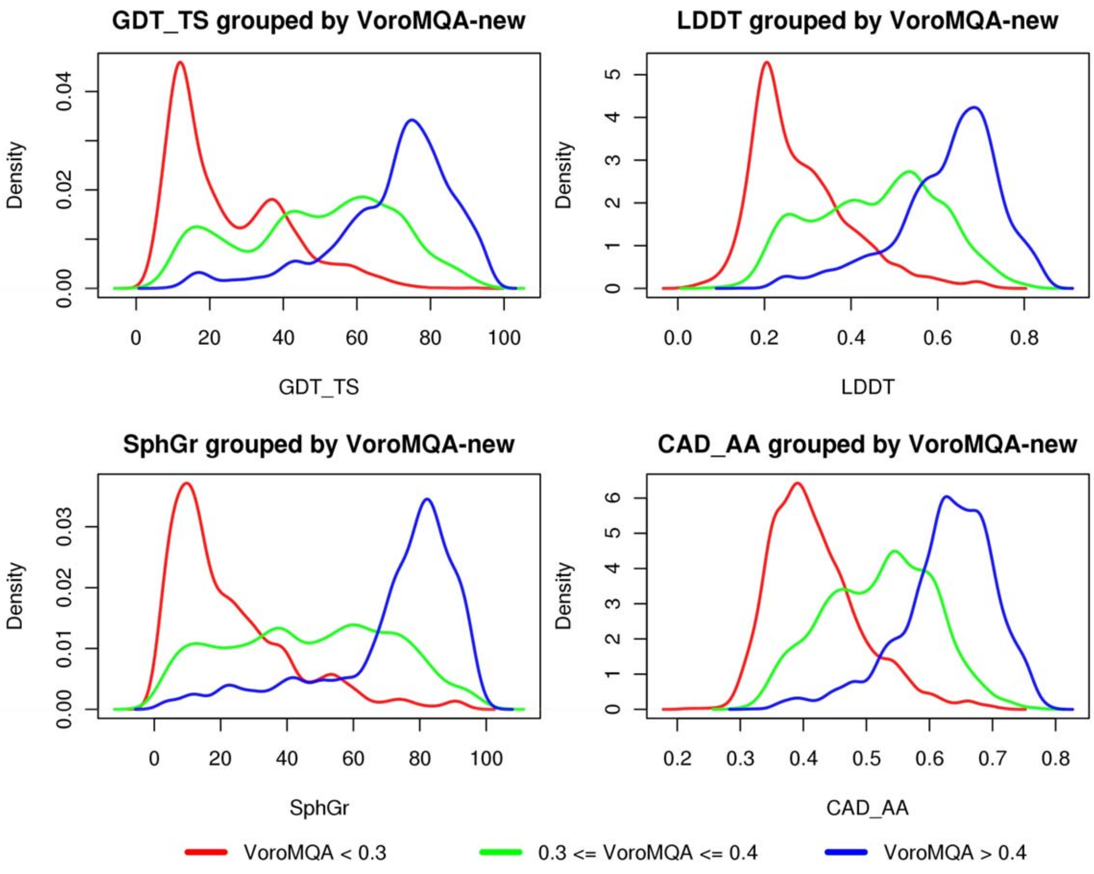

Figure 4

Empirical distribution densities of GDT_TS, LDDT, SphGr and CAD_AA scores of the CASP11 models that have VoroMQA-new scores in intervals $(0, 0.3)$, $[0.3, 0.4]$ and $(0.4, 1)$: the corresponding lines are colored in red, green and blue, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

Table III

Agreement Percentage Scores Calculated for Best150 sets of Models from CASP11: Second Column Contains Scores for all the Targets Combined, Third Column Contains Mean Per-Target Scores. The Table is Sorted by the Third Column

Method	Total %	Mean %
VoroMQA-new-sr	82.50	82.70
VoroMQA-new	81.80	82.16
GOAP	80.11	80.57
VoroMQA-old	79.70	80.48
MULTICOM-NOVEL	79.66	80.25
ProQ2-refine	78.69	79.40
MULTICOM-CLUSTER	78.76	79.21
ProQ2	78.13	78.86
dDFIRE	77.73	78.43
DOOP	76.09	76.57
Wang_SVM	74.42	75.24
Wang_deep_2	72.12	72.83
Wang_deep_3	71.65	72.30
Wang_deep_1	71.57	72.19

value v of a protein structural model: if $v < 0.3$, then the model is likely bad; if $v > 0.4$, then the model is likely good; if $v \in [0.3, 0.4]$, then the model quality cannot be reliably classified as bad or good using VoroMQA alone. This rule is most useful when just a single model is available. Results for CASP11 models also showed that VoroMQA-new and VoroMQA-new-sr global scores are highly correlated (Pearson correlation coefficient is about 0.98). Therefore, the same rule can be applied for both scores.

Results of the per target analysis of CASP11 data

In model selection tests, we first analyzed how different QA scores perform on best150 sets using the tournament-based methodology. For each available QA method, we calculated agreement percentage scores for all the targets combined and for every target separately. When considering all the possible 799703 pairs of models, only for 425877 (53%) of them all four reference-based scores (GDT_TS, LDDT, SphGr and CAD_AA) agree which model out of the two is better. The middle column in Table III shows how often different QA scores agree with the unanimous judgment of all four reference-based scores, the last column shows the average per-target agreement percentages, that is, mean Agreement-scores. Table IV shows the p-values calculated by applying the Wilcoxon signed-rank test to compare VoroMQA-new-sr with the other methods according to per-target Agreement-scores. Considering the significance level threshold of 0.05, VoroMQA-new-sr significantly outperformed all the others, except for VoroMQA-new (results of the analogous test for VoroMQA-new are presented in Supporting Information Table S2).

Next, we asked each available QA method to select a single model from the best150 set of every target. The mean Wins-score, GDT_TS, LDDT, SphGr and CAD_AA values of the selected models are presented in Table V along with the corresponding mean z-scores. Wins-scores and z-scores were calculated considering only the models from best150 sets. While Table V shows that VoroMQA-new-sr and VoroMQA-new performed relatively well, the achieved advantage over other methods is mostly not significant. Supporting Information Table S3 shows the p-values calculated by applying the Wilcoxon signed-rank test to compare the Wins-scores of the models selected by the VoroMQA-new-sr method with the corresponding results achieved by the other methods (results of the analogous test for VoroMQA-new are presented in Supporting Information Table S4). The p-values >0.05 indicate that all the scores from VoroMQA, ProQ2 and MULTICOM families, as well as GOAP and DOOP scores, demonstrate very similar model-selection abilities when analyzing models from best150 sets using our tournament-based methodology.

Additionally, we performed the analysis based on Agreement-scores and Wins-scores on BZQ15 sets (Supporting Information Tables S5–S10). The overall trends are similar to those of best150 sets but there are some differences. Most notably, VoroMQA-new-sr performed significantly better than VoroMQA-new indicating that side-chain rebuilding was particularly beneficial for scoring models from BZQ15 sets. This is consistent with the fact that BAKER-ROSETTASERVER differs considerably from Zhang-Server and QUARK in the side-chain positioning quality.[46] Therefore, rebuilding side-chains before scoring apparently helps to level significant differences in side-chain packing leading to improved results. Also, in the test based on Agreement-score, VoroMQA-new-sr did not significantly outperform MULTICOM-NOVEL.

Table IV

Results of the Wilcoxon Signed-rank Test Applied to Compare the Agreement Percentage Scores Achieved by the VoroMQA-new-sr Method for Best150 Sets of Models from CASP11 with the Corresponding Agreement-Scores Achieved by the Other Methods. The Table is Sorted by the Middle Column. All the *p*-values are Rounded Up to the Five Decimal Places. Blue Color is Used to Indicate Methods that Performed Significantly Worse than VoroMQA-new-sr. [Color table can be viewed at wileyonlinelibrary.com]

Method	*p*-value (two -sided)	*p*-value (one-sided)
VoroMQA-new	0.20292	0.10146
VoroMQA-old	0.01182	0.00591
MULTICOM-NOVEL	0.00558	0.00279
ProQ2-refine	0.00041	0.00020
GOAP	0.00024	0.00012
ProQ2	0.00016	0.00008
MULTICOM-CLUSTER	0.00004	0.00002
Wang_SVM	0.00000	0.00000
Wang_deep_3	0.00000	0.00000
Wang_deep_2	0.00000	0.00000
Wang_deep_1	0.00000	0.00000
DOOP	0.00000	0.00000
dDFIRE	0.00000	0.00000

Table V

Mean Per-target Scores of the Models Selected by Various QA Methods from Best150 sets of Models from CASP11. Each Numeric Cell Contains Two Slash-separated Values: the Mean Reference-based Score of the Selected Models and the Corresponding Mean *z*-scores. The Top Five Rows Show the Results Obtained Using Reference-based Scores, That Is, Results That are Close to Ideal. The Table Is Sorted by The Wins-Score Values. The best scores obtained by QA methods are indicated using gray background

Method	Wins-score	GDT_TS	LDDT	SphGr	CAD_AA
Wins-score	1/2.66	55/2.11	0.544/2.19	58.3/2.26	0.586/2.23
LDDT	0.939/2.42	54.3/1.93	0.549/2.33	57.8/2.14	0.584/2.13
CAD_AA	0.902/2.29	53.5/1.72	0.536/1.99	57.3/2.08	0.594/2.46
SphGr	0.861/2.1	54.1/1.85	0.522/1.7	59.7/2.55	0.58/2.04
GDT_TS	0.857/2.1	56.2/2.4	0.522/1.7	56.7/1.97	0.575/1.85
VoroMQA-new-sr	0.735/1.63	50/1.17	0.497/1.28	53/1.54	0.566/1.67
MULTICOM-CLUSTER	0.717/1.58	48.9/0.94	0.495/1.28	51/1.27	0.563/1.57
VoroMQA-new	0.713/1.56	50/1.09	0.496/1.26	51.8/1.32	0.56/1.48
VoroMQA-old	0.704/1.52	48.9/1.03	0.492/1.25	51.1/1.34	0.559/1.51
ProQ2-refine	0.703/1.51	49.3/1	0.495/1.26	52.1/1.39	0.562/1.56
MULTICOM-NOVEL	0.695/1.49	49.4/1.07	0.49/1.19	51.8/1.38	0.564/1.64
GOAP	0.694/1.49	49.8/0.89	0.501/1.27	52.2/1.22	0.568/1.66
ProQ2	0.691/1.46	49.9/0.98	0.496/1.21	52.2/1.34	0.561/1.51
DOOP	0.681/1.46	48.8/0.81	0.499/1.27	50.7/1.05	0.564/1.55
Wang_SVM	0.616/1.2	47.5/0.73	0.474/0.83	49.4/1.02	0.546/1.11
Wang_deep_2	0.568/0.99	47.2/0.65	0.471/0.72	49.8/0.96	0.545/1.05
Wang_deep_1	0.546/0.91	46.3/0.49	0.464/0.6	48.8/0.87	0.542/0.95
dDFIRE	0.542/0.97	46.1/0.33	0.471/0.75	48.4/0.78	0.553/1.28
Wang_deep_3	0.519/0.8	46.6/0.46	0.463/0.53	48.7/0.78	0.542/0.93

Finally, we analyzed sel20 sets, and the detailed results are presented in Supporting Information Tables S11–S16. Both VoroMQA-new and VoroMQA-new-sr performed significantly worse than ProQ2-refine and MULTICOM-NOVEL in Agreement-score-based testing and significantly worse than ProQ2-refine and ProQ2 in Wins-score-based testing. Overall, for sel20 sets, every method that is based solely on analyzing geometric features and applying statistical potentials (GOAP, DOOP, dDFIRE and all the VoroMQA variations) achieved worse results than the best-performing composite methods incorporating evolutionary information in the form of predicted features such as secondary structure or solvent accessibility: this was definitely not the case for best150 and BZQ15 sets. We thus asked if additional information can be decidedly beneficial when evaluating sel20 models. Considering that every sel20 set was formed to contain models as different from each other as possible, there may be cases when incorrect models can be identified simply by being significantly different from a reasonably reliable model produced by some homology-based structure prediction server. To test this surmise we defined a simple QA method, dubbed "HHpred-agreement", that evaluates models by comparing them with a model produced by the HHpred server[47] (HHpredA in CASP11) using TM-score.[48] Higher TM-scores were considered to represent better models. Supporting Information Tables S17–S19 show how HHpred-agreement performed in selecting models from sel20, best150 and BZQ15 sets: HHpred-agreement performed very similarly as ProQ2-refine and ProQ2 for sel20, but much worse than all the other tested QA scores for best150 and BZQ15. We also defined a meta-score, named "VoroMQA-new-and-HHpred-agreement", which is simply an unweighted geometric mean of VoroMQA-new and HHpred-agreement scores. An analogous meta-score was also defined for VoroMQA-new-sr. As shown in Supporting Information Tables S17–S19, the two meta-scores achieved top spots in the Wins-score-based ranking of QA methods for sel20 sets and performed relatively well (although not as well as the original VoroMQA-new-sr) for best150 and BZQ15 sets. To sum up, for sets of models similar to sel20, using just VoroMQA may not be as effective as using it in conjunction with additional information derived using sequence homologs. Results of our analysis also raise concerns about whether sel20 sets in CASP11 represent real-life model selection challenges, because the relatively good performance of the HHpred-agreement score suggests that it may be more efficient just to get a single model from HHpred (or some other well-performing homology-based server) instead of selecting a model from a small but very diverse (sel20-like) set.

In our analysis we concentrated on the ability of considered QA methods to identify best models and so far neglected the correlation analysis, that is, the calculation of coefficients of correlation between QA scores and reference-based model evaluation scores. Correlation analysis alone is a poor indicator of the method's performance, however, it may provide useful insights and is traditionally used for CASP data.[4] For consistency, we also performed correlation analysis for best150, sel20, and BZQ15 sets (Supporting Information Tables S20–S22). Both VoroMQA-new and VoroMQA-new-sr showed top results for best150 sets, but not for sel20. Also,

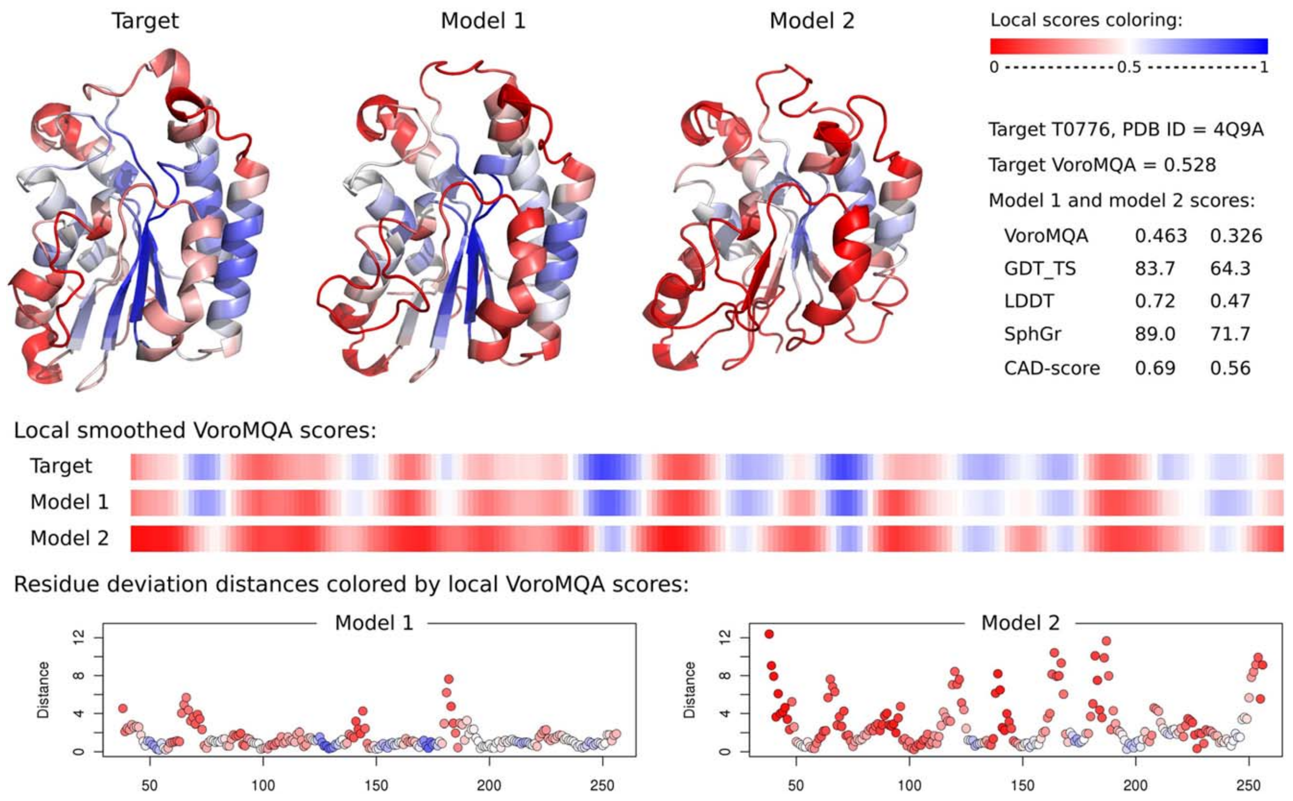

Figure 5

Local VoroMQA scores calculated for T0776 target structure and two its models using VoroMQA web-server. The cartoon structural representations are colored by smoothed per-residue VoroMQA scores. The corresponding one-dimensional color-coded profiles are shown in the middle part of the figure. Residue distance deviations (in angstroms) colored by smoothed per-residue VoroMQA scores are plotted in the bottom part of the figure.

VoroMQA-new-sr showed top results for BZQ15 sets. Overall, the results of the correlation-based analysis are consistent with those of tournament-based tests. In addition, correlation analysis showed a positive trait of Wins-score: the four reference-based scores (GDT_TS, LDDT, SphGr and CAD_AA) correlate better with Wins-score than with each other.

Local scoring

VoroMQA global scores are directly derived from the atom-level VoroMQA scores, so while testing global VoroMQA scores we also indirectly tested VoroMQA local scoring capabilities, at least the cumulative effect of atomic VoroMQA scores. Another possible way of testing local scoring is investigating how local VoroMQA scores conform to some reference-based local scores. However, due to the nature of our method, this approach is not easily applicable as illustrated with the local scoring example in Figure 5. The figure shows residue-level VoroMQA scores of a native (target) structure and its two models, the first model is better than the second one according to all four reference-based scores. The VoroMQA global scores correctly rank both models by their deviation from the native one. The color-coded VoroMQA local scoring profiles support the judgment of the global scores, because for most of the residue positions the local VoroMQA scores get lower as the global model quality gets lower. However, low absolute values of local VoroMQA scores do not necessarily correspond to low reference-based local scores. For a simple example, let us consider just the native structure. Its local VoroMQA scores are not homogeneous despite all the residue positions being correct. This is so, because different residues are not in equally favorable contact environments. Similarly, when considering a modeled structure different from the native one, low VoroMQA score for a single residue does not necessarily mean that the structural position of the residue is incorrect. This is illustrated in the bottom part of Figure 5 with the plots of residue distance deviations obtained from LGA[32] structural alignments and colored by the corresponding VoroMQA scores: some of the well-aligned residues have low VoroMQA scores. Another observation from the same plots is that the positions of the residues with higher VoroMQA scores tend to be well-predicted. To check if this is true in general, we used data from the CAMEO project (www.cameo3d.org).[45]

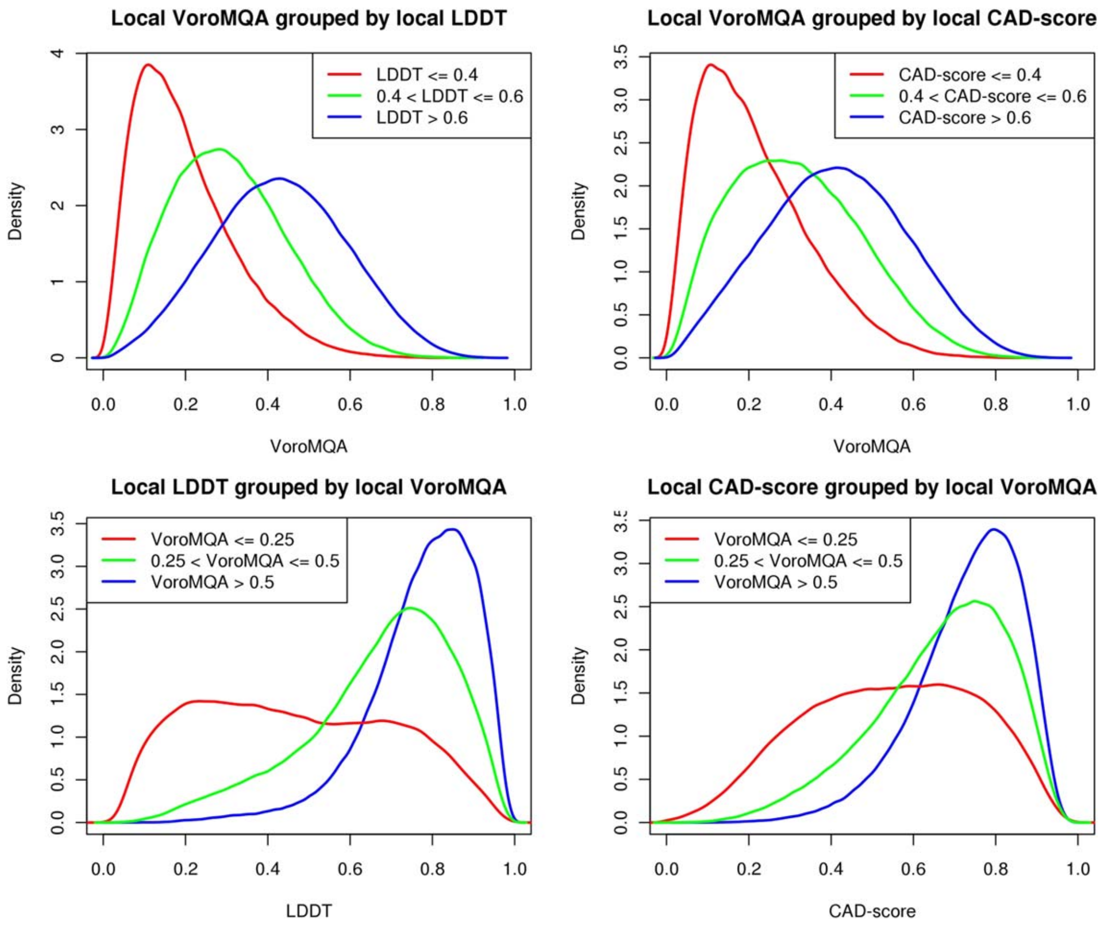

Figure 6

Empirical distribution densities of scores obtained from the CAMEO “1-year” dataset. Top row: VoroMQA local scores grouped by the corresponding reference-based local scores, that is, LDDT and CAD-score. Bottom row: Reference-based local scores grouped by the corresponding VoroMQA local scores. [Color figure can be viewed at wileyonlinelibrary.com]

We had the latest version of VoroMQA entered to the CAMEO model quality estimation category under name “VoroMQA_v2” since August 2015, thus we were able to download “1-year” (weeks from 2015.10.17 to 2016.10.08) dataset and analyze it to investigate the relations between VoroMQA local scores and the corresponding LDDT and CAD-score local scores. We looked at the empirical distribution of VoroMQA local scores that correspond to the three classes of reference-based local scores: low, average and high. Similarly, we looked at the distributions of reference-based local scores that correspond to low, average and high local VoroMQA scores. The results, presented in Figure 6, prompt us to make two important observations: if the local VoroMQA score for a residue in the model is >0.5, the residue is likely well-predicted; if the local VoroMQA score is low (0.25 or less), the accuracy of the residue position is uncertain, because it may mean either incorrectly predicted position or correctly predicted position in unfavorable environment. The latter point is illustrated in Figure 5 showing that even a native structure can have regions with relatively low local VoroMQA scores.

Overall, VoroMQA local scores are most useful when analyzed along with the manual inspection of the protein structure. For example, let us inspect models 1 and 2 in Figure 5. In model 1 most of the low-scoring residues correspond to solvent-accessible regions while many of the high-scoring ones are buried in the core of the structure, in model 2 the low-scoring regions cover larger parts of the structure and the core is scored much lower than in model 1. These observations allow us to conclude that model 1 is better than model 2 even without considering global scores.

DISCUSSION

VoroMQA is an all-atom knowledge-based protein structure scoring method. It is important to emphasize that the scoring function of VoroMQA was not optimized or trained in any way to better correspond to any of the reference-based protein structure accuracy measures such as RMSD, GDT_TS or CAD-score. Only an unsupervised learning procedure was applied taking experimentally determined structures of protein biological units (assemblies) as the source of structural information. Also, VoroMQA does not use any additional predictive features, for example, predicted secondary structure or solvent accessibility that are typically derived using multiple sequence homologs. In other words, only protein 3D structure is needed for its assessment with VoroMQA. Accordingly, VoroMQA falls into the category of statistical energy potentials. However, in contrast to most statistical potentials that are distance-based, our method uses interatomic contact areas. The choice of contact areas offers several advantages. Contact areas not only define physical interactions but also implicitly take into account their strength. Moreover, contact areas make it possible to treat interactions within the protein structure and interactions with solvent in the same way. Interactions of protein atoms with solvent are considered as just another type of contacts. In addition, the use of contact areas allows efficient normalization of pseudo-energy values, so that they can be converted into quality estimates ranging from 0 to 1. This means that the VoroMQA scores are largely independent of the type or the size of an input protein structure.

We tested the performance of VoroMQA by the ability to identify the native structure among the decoys (computational models) in a test typical for statistical potentials. In addition, we explored how well VoroMQA is able to select models by their similarity to the native structure according to different scenarios. Whereas the task of selecting native structure is unambiguous, the evaluation of model selection by their similarity to the native structure is not. There are at least two reasons why evaluation of methods for model selection is not trivial, especially in cases when differences between models are small. One of the reasons is the uncertainty of any reference score.[43] Another reason is that it is quite common for different reference scores to disagree about the exact model ranking. To test the ability of VoroMQA and other methods to select models closest to the native structure we chose the same four reference scores used by the official assessment of model accuracy estimation methods in CASP11,[4] namely, a rigid-body measure (GDT_TS) and three local-structure-based scores (LDDT, CAD-score, and SphereGrinder). However, instead of analyzing the results of these four scores separately,[4] we devised a simple procedure that enabled us to combine all four scores and in so doing to avoid the two problems mentioned above. The main idea of this procedure is that one model is considered to be better (closer to the native structure) than the other one only if all four reference scores agree to that unanimously. Based on this idea we introduced two scores, Agreement-score and Wins-score, and used them throughout the study for performance comparison of different methods. We believe that by including multiple reference scores simultaneously this procedure provides a robust way for comparing model quality estimation methods. We also believe that this evaluation scheme might be useful for comparing other type of prediction methods as well.

In our tests VoroMQA consistently outperformed DOOP, GOAP and dDFIRE that are similarly based on all-atom statistical potentials, but use distances rather than contact areas. The outcome of these tests is rather unexpected, taking into account that both GOAP and dDFIRE feature orientation-dependent potentials whereas DOOP potentials include the dependence on the backbone torsion angles. In contrast, VoroMQA does not include any terms associated with either conformation preferences of the main chain or orientation-dependence of side chains. This may suggest that contact areas are perhaps more suitable compared to distances in identifying native structure and scoring near-native conformations.

We also tested VoroMQA alongside with model quality assessment methods that in addition to the actual structure utilize various predictions derived using evolutionary information and rely heavily on using machine learning to predict reference-based model quality scores. As VoroMQA does not use any additional information, comparison with such composite methods puts VoroMQA at disadvantage. Despite this, the tests showed that VoroMQA often outperformed these composite methods, especially in the one-out-of-two model selection scenario. VoroMQA achieved top results when tested on the roughly prefiltered sets of CASP11 models, that is, the sets comprised of models produced by the top three prediction servers (BZQ15 sets defined in this article) or the sets comprised of models selected using a simple consensus-based algorithm (best150 sets provided by the CASP11 organizers). It has been previously observed that the side-chain remodeling may lead to improved model selection.[49] Indeed, the rebuilding of side chains for best150 and BZQ15 model sets has further improved VoroMQA results suggesting that significant differences in side-chain packing may conceal the main chain similarities.

The only case where VoroMQA (with or without side-chain remodeling) was more prone to make mistakes compared to the composite QA scores was when faced with the CASP11 sel20 sets that were composed to contain not better models, but models as different from each other as possible. Such sets, however, hardly represent any real-life model selection scenario. Moreover, we found that the relatively poor performance of VoroMQA in this type of setting could be rescued by simple combination of the VoroMQA score and the evolutionary information in the form of HHpred template-based models. This observation suggests that VoroMQA can be easily incorporated into composite scoring functions.

VoroMQA global scores are directly derived from atom-level scores, so the relatively good results achieved by our method in model selection tests are direct implications of the VoroMQA local scoring capabilities. However, it should be emphasized that local VoroMQA scores of a structural model indicate how energetically favorable or unfavorable the local region is, and not how much it deviates from the corresponding region in the native structure. A native protein structure has a combination of both energetically favorable (for example, hydrophobic core) and unfavorable regions (for example, active sites, protein-protein binding sites or solvent-exposed loops). Therefore, even a very accurate structural model will have regions with low VoroMQA scores that will closely reproduce the pattern observed for the native structure (see Figure 5). In general our tests indicate that high local VoroMQA scores usually correspond to accurate structural regions. In contrast, low local VoroMQA scores do not necessarily imply that the corresponding region is unrealistic. It may just be one of the regions in a less favorable environment. In other words, VoroMQA local scores could be used to classify the structure into the accurate regions and those with the uncertain accuracy. In practice, the VoroMQA local scoring perhaps would be most useful in qualitative analysis performed in conjunction with the manual inspection of the protein structure.

CONCLUSIONS

VoroMQA represents an all-atom statistical scoring function for the estimation of protein structure accuracy. VoroMQA shows robust performance both in recognition of the native structure among decoys and in selecting best models. The use of interatomic contact areas instead of distances might be one of the reasons for relatively good results. Thus, VoroMQA might be a valuable addition to the available set of model quality assessment methods, not only because of strong performance, but also because of its orthogonality to the existing scores.

ACKNOWLEDGMENTS

We wish to thank Mindaugas Margelevičius and Justas Dapkūnas for critically reading the manuscript. This work has been supported by European Social Fund under the Global Grant measure.

REFERENCES

1. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round x. Proteins 2014;82(Suppl 2):1–6.
2. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. Proteins 2016; 84:4–14.
3. Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A. Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins 2014;82(Suppl 2):112–126.
4. Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins 2015;84:349–369.
5. Ray A, Lindahl E, Wallner B. Improved model quality assessment using ProQ2. BMC Bioinformatics 2012;13:224.
6. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008;71: 261–277.
7. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
8. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
9. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
10. Shen M-y, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
11. Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One 2010;5:e15386.
12. Chae M-H, Krull F, Knapp E-W. Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction. Proteins 2015;83:881–890.
13. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008;72:793–803.
14. Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 2008;376:288–301.
15. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 2011;101:2043–2052.
16. Olechnovič K, Kulberkytė E, Venclovas Č. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81:149–162.
17. Zemla A, Venclovas Č, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;(Suppl 3):22–29.
18. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. Proc Natl Acad Sci U S A 2003;100:3215–3220.
19. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974; 82:1–14.
20. Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol 2004;14: 233–241. April,
21. Rother K, Hildebrand PW, Goede A, Gruening B, Preissner R. Voronoia: analyzing packing in protein structures. Nucleic Acids Res 2009;37:D393–D395.
22. Olechnovič K, Margelevičius M, Venclovas Č. Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure. Bioinformatics (Oxford, England) 2011; 27:723–724.
23. Esque J, L'eonard S, de Brevern AG, Oguey C. VLDP web server: a powerful geometric tool for analysing protein structures in their environment. Nucleic Acids Res 2013;41:W373–W378.
24. Zomorodian A, Guibas L, Koehl P. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. Comput Aided Geom Des 2006;23:531–544.
25. Mirzaie M, Sadeghi M. Delaunay-based nonlocal interactions are sufficient and accurate in protein fold recognition. Proteins 2014;82: 415–423. March,

26. Jafari R, Sadeghi M, Mirzaie M. Investigating the importance of Delaunay-based definition of atomic interactions in scoring of protein-protein docking results. J Mol Graph Model 2016;66:108–114. May,
27. Olechnovič K, Venclovas Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. J Comput Chem 2014;35:672–681.
28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
29. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction - Round VIII. Proteins 2009;77(Suppl 9):1–4.
30. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round IX. Proteins 2011;79(Suppl 10):1–5.
31. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591. August,
32. Zemla A. LGA: A method for finding 3d similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374. July,
33. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29:2722–2728.
34. Antczak PLM, Ratajczak T, Blazewicz J, Lukasiak P, Blazewicz J. SphereGrinder - reference structure-based tool for quality assessment of protein structural models. 665–668. IEEE, November, 2015.
35. Li J, Cao R, Cheng J. A large-scale conformation sampling and evaluation server for protein tertiary structure prediction and its assessment in CASP11. BMC Bioinformatics 2015;16:337.
36. Cao R, Cheng J. Protein single-model quality assessment by feature-based probability density functions. Sci Rep 2016;6:23990.
37. Uziela K, Wallner B. ProQ2: estimation of model accuracy implemented in Rosetta. Bioinformatics 2016;32:1411–1413.
38. Liu T, Wang Y, Eickholt J, Wang Z. Benchmarking deep networks for predicting residue-specific quality of individual protein models in CASP11. Sci Rep 2016;6:19301.
39. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77: 778–795.
40. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32:W526–W531.
41. Zhang Y. I-TASSER server for protein 3d structure prediction. BMC Bioinformatics 2008;9:40.
42. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715–1735.
43. Li W, Schaeffer RD, Otwinowski Z, Grishin NV. Estimation of uncertainties in the global distance test (GDT_TS) for CASP Models. PloS One 2016;11:e0154786.
44. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bull 1945;1:80.
45. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T. The Protein Model Portal–a comprehensive resource for protein structure and model information. Database 2013;2013:bat031.
46. Modi V, Xu Q, Adhikari S, Dunbrack RL. Assessment of template-based modeling of protein structure in CASP11. Proteins 2016; 84(Suppl 1):200–220.
47. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins Struct Funct Bioinformatics 2009;77:128–132.
48. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
49. Wallner B. ProQM-resample: improved model quality assessment for membrane proteins by limited conformational sampling. Bioinformatics (Oxford, England) 2014;30:2221–2223.