Supplementary Information

Smooth orientation-dependent scoring function for coarse-grained protein quality assessment

Mikhail Karasikov [1-3], Guillaume Pages [1], and Sergei Grudinin [1]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.
[2]Center for Energy Systems, Skolkovo Institute of Science and Technology, Moscow, 143026, Russia.
[3]Moscow Institute of Physics and Technology, Moscow, 141701, Russia.

A Literature Review

There are generally two types of QA methods. Consensus-model QA methods decide on the quality of individual protein models based on their statistics in the assessed set. On the contrary, single-model QA methods consider atoms of only the assessed protein structure without any additional information about other protein models and hence, these can be used for conformational sampling and structure refinement. Furthermore, performance of consensus-model QA methods usually depends on single-model QA methods involved in the conformational sampling used for generating a pool of protein models to be assessed. Also, single-model QA methods are proved to achieve better performance compared to consensus-model QA methods on unbalanced sets and in cases where the protein models within the assessed set are very similar (Ray *et al.*, 2012).

Among recently proposed single-model QA methods, there are generally two main approaches to design a scoring function, which are the physics-based and knowledge-based (data-driven) approaches (Faraggi and Kloczkowski, 2014; Liu *et al.*, 2014). Physics-based scoring functions are constructed according to some physical knowledge of the configuration and interactions in the system. This approach is based on the Gibbs free energy minimization principle, which states that all target protein structures minimize the Gibbs free energy over the whole conformational space. However, precise estimation of the Gibbs free energy requires exhaustive sampling of a huge number of conformational states (Cecchini *et al.*, 2009; Tyka *et al.*, 2006), which is computationally intractable in most practical cases. Thus, the physics-based approaches are aimed at constructing scoring functions (often called energy potentials or force-fields) that approximate only the enthalpic part of the Gibbs free energy and can be estimated efficiently. Usually, these potentials decompose the total energy into a sum of additive terms (contributions) that represent stretching of bonds and angles, dihedral potentials, electrostatic and van der Waals interactions, etc. Alongside with the physics-based approaches, there are so-called knowledge-based approaches that deduce the essential energies of molecular interactions from the structural and sequence databases assuming a certain distribution of conformations or minimizing a certain loss function. The corresponding scoring functions are typically derived either by machine learning or by estimating the probabilities of certain conformations (statistical QA methods) using the information from known protein structures found in structural databases.

There have been many QA methods proposed using either physics-based or knowledge-based approaches. However, recent QA methods that combine these two approaches appeared to be the most promising. Below, we briefly overview several commonly used QA methods involving the most representative techniques.

Statistical QA methods are derived according to the Boltzmann assumption, which states that the energy of a protein structure is proportional to the negative logarithm of the probability of its conformational state (Finkelstein *et al.*, 2004). Orientation-dependent statistical QA methods take into account the many-body interactions by describing both distance and relative orientation of atom groups involved in these interactions, and typically they outperform traditional distance-dependent-only methods. For example, the popular RWplus (Zhang and Zhang, 2010) scoring function combines pairwise distance-dependent and orientation-dependent contributions. For a given protein model, it computes the number $N_{\text{obs}}(\alpha, \beta, r)$ of atom pairs of types (α, β) at a distance in the interval from r to $r + \Delta r$ and divides it by the number of expected pairs $N_{\text{exp}}(\alpha, \beta, r)$ to estimate the likelihood of the assessed protein model. For each residue except glycines and alanines, Zhang and Zhang (2010) set a local frame defined by three side-chain atoms, which is centered at the location of one of them. The relative orientation of a pair of local frames is then represented by five parameters, which are two pairs of spherical angles and a torsion angle. Finally, the total orientation-dependent packing energy is calculated using the same technique for counting statistics as for the distance-dependent potentials.

ORDER_AVE (Liu *et al.*, 2014) is another orientation-dependent statistical potential, which assesses the quality of protein models using the joint probability distribution for four parameters to describe the geometric relationship between a pair of atoms (i, j) and connected to them heavy atoms (ir, jr). These parameters are three angles that describe the relative orientation of four atoms i, j, ir, jr, and the distance between atoms i and j. ORDER_AVE treats local (when two corresponding residues have a small sequence separation) and non-local interactions separately with the overall energy assigned as a weighted sum of these two.

The SELECTpro method (Randall and Baldi, 2008) is based on a potential consisting of physical and statistical terms as well as penalties inferred from structural predictions. Together with such conventional contributions as van der Waals, electrostatic, and side-chain hydrogen bonding interactions, Randall and Baldi (2008) proposed to also use β-strand pairing and introduced penalties for mismatches of observed and predicted structural features (secondary structure, solvent accessibility, contact map). The total energy is computed as a weighted sum of the introduced energy terms with weights that maximize the sum of the GDT-TS (Zhang and Skolnick, 2007) of the best-ranked models in the training set built from the CASP6 (Moult *et al.*, 2016) dataset.

Nowadays, more and more research is being devoted to techniques for building QA methods involving machine learning, especially to meta algorithms that combine several other knowledge-based scoring functions using their predictions as features. For example, the ProQ2 (Ray *et al.*, 2012) scoring function, which is one of the best QA methods according to experiments on the CASP11 (Moult *et al.*, 2016) dataset, is trained using support vector machine (SVM) in the space of structural and sequence-based features calculated from the model. As the structural features, ProQ2 uses contacts between 13 different atom types, residue-residue contacts, and the surface accessibility by aggregating amino acids into six different groups. The sequence-based features (secondary structure, surface accessibility, and sequence profiles) are derived using information predicted from the sequence. Another QA method by Faraggi and Kloczkowski (2014) uses physics-based electrostatic potentials and other knowledge-based scoring functions as the features and trains a neural network to predict the TM-score (Zhang and Skolnick, 2007) similarity measure between the protein models and the target structures. The Wang_SVM (Liu *et al.*, 2016) scoring function was trained by SVM using as features the protein sequences and several protein descriptors predicted

by external utilities. These are the secondary structure and the solvent accessibility predicted with SSPRO, residue-residue contact probabilities predicted with NNcon, and evolutionary information predicted with PSI-BLAST, for details see (Liu *et al.*, 2016). The Qprob (Cao and Cheng, 2016) scoring function was trained using structural features (compact score, surface score, the exposed mass, the exposed surface), features predicted from the protein sequence (secondary structure, solvent accessibility), and estimated distributions of features provided by other QA methods: RWplus by Zhang and Zhang (2010), ModelEvaluator by Wang *et al.* (2009), DOPE by Shen and Sali (2006), and RF_CB_SRS_OD by Rykunov and Fiser (2010). Following a similar strategy, Jing *et al.* (2016) applied a learning-to-rank technique (ranking SVM (Joachims, 2002)) to learn a scoring function using as features predictions by other QA methods: DFIRE (Zhou and Zhou, 2002), DOPE (Webb and Sali, 2014), GOAP (Zhou and Skolnick, 2011), RWplus (Zhang and Zhang, 2010), etc.

Although plenty of QA methods have been proposed, often they miss some meaningful contributions. For example, the solvation-related terms and terms related to hydrogen bonding interactions are very often neglected by many QA methods. However, these contributions are important and generally should be taken into account. For instance, hydrogen bonds (Hubbard and Kamran Haider, 2001) are stated to confer directionality and specificity to the intra-molecular interactions in structures. They provide structural organization of distinct protein folds because suitable interactions in the folded structure are typically achieved by the maximum number of hydrogen bonding groups. In addition, most of QA methods require all-atom protein models as input, and thus their performance critically depends on the accuracy of side-chain packing, that is, positions of the side-chain atoms, which can be modeled with the widely-used SCWRL4 method (Krivov *et al.*, 2009), as in (Cao and Cheng, 2016), or any other (Liang *et al.*, 2011). A simplified coarse-grained representation of amino acids, as in (Kmiecik *et al.*, 2016), overcomes this issue and also reduces the overall computational complexity. Another problem of many QA methods is that sequence-dependent features often introduce non-smooth terms that break the continuity and smoothness of the scoring function. Thus, these methods cannot be applied to gradient-based structure optimization.

In this paper, we propose a novel method for protein quality assessment, the Smooth Backbone-Reliant Orientation-Dependent (SBROD) scoring function. SBROD is a single-model QA method that scores protein models using only structural features along with the explicit representation of solvent generated on a regular grid. It requires only coordinates of the protein backbone, and thus it is insensitive to the side-chain conformations in protein models. In addition, the SBROD scoring function is continuous with respect to coordinates of the protein atoms. This makes it also applicable to be used in molecular mechanics or dynamics, for example.

B Methods Availability

For running comparison against SBROD, the following state-of-the-art methods were downloaded: RWplus, VoroMQA, ProQ2, ProQ2-refine, ProQ3, and ProQ3-repack.

The RWplus (Zhang and Zhang, 2010) QA method was downloaded from `https://zhanglab.ccmb.med.umich.edu/RW/calRWplus.tar.gz` (accessed Apr 20, 2018).

VoroMQA (Olechnovič and Venclovas, 2017) was downloaded from `https://bitbucket.org/kliment/voronota/downloads/voronota_1.18.1877.tar.gz` and installed according to instructions at `http://bioinformatics.lt/software/voromqa` (accessed Apr 20, 2018).

The ProQ2 (Uziela and Wallner, 2016) QA method was downloaded and installed according to instructions at `https://github.com/bjornwallner/ProQ_scripts` (accessed Apr 20, 2018). For scoring a protein model, the ProQ2-refine method repeats the exact framework of ProQ2 except for the fact that now ProQ2 method is used to assess quality of ten protein models with the same backbone and randomly repacked side-chains using Rosetta, as suggested in the instructions to ProQ2, and the highest score of the ten generated repacked protein models is used as prediction of ProQ2-refine for the assessed protein model.

The ProQ3 and ProQ3-repack (Uziela *et al.*, 2016) QA methods were downloaded and installed according to instructions at `https://bitbucket.org/ElofssonLab/proq3` (accessed Apr 20, 2018).

Rosetta 2018.12.60119 downloaded from the official website `https://www.rosettacommons.org/software/academic` was used to run the ProQ2 and ProQ3 methods.

C Data Preparation and Filtering

We were not able to get the ProQ3 score for the protein model `CASP12Stage2/T0912/Atome2_CBS_TS1` because ProQ3 was raising a segmentation fault when trying to assess that protein model. Therefore, we used scores predicted for the same protein model by the QA method ProQ3-repack. All other protein models were assessed directly by ProQ3.

D Performance on the CASP11 dataset

To compare the performance of SBROD with nine state-of-the-art QA methods, we first used the results obtained by Cao and Cheng (2016). They assessed the performance of several QA methods against the ground truth GDT-TS computed with the LGA utility (Zemla, 2003) for structures with side-chains repacked with SCWRL4 (Krivov *et al.*, 2009) on the CASP11 Stage1 and Stage2 datasets. Since the LGA utility (Zemla, 2003) is not open-source, we used the TM-score utility (Zhang and Skolnick, 2007) instead. Nonetheless, SBROD is not sensitive to the side-chains packing, and the difference between the GDT-TS computed by the TM-score and LGA utilities is negligible. Therefore, the measurements estimated by Cao and Cheng (2016) are consistent with ours, measured as described above, and all of these can be fairly compared to each other.

Tables S1a and S1b list the performance measures computed for the SBROD scoring function (trained on the CASP[5-10] data augmented with the generated NMA-based decoy models, with the CNDF smoothing parameters of $\sigma^a = \sigma^r = \sigma^h = \sigma^s = 0.187$ on the testing stage) and for nine other state-of-the-art methods on the CASP11 Stage1 and Stage2 datasets, correspondingly. It can be seen that our method outperforms all other methods on both stages of the CASP11 experiment if assessed by the mean score loss, and it is highly competitive to the other methods if assessed by the other performance measures.

E Additional Tests

We have assessed the performance of SBROD together with several other QA methods on the MOULDER dataset (Eramian *et al.*, 2006) downloaded from `https://salilab.org/decoys/moulder_decoys_scores.txt`. MOULDER is a conventional dataset, although outdated, for testing physics-based and statistical energy potentials. Table S2a lists the obtained results.

As we can see from this table, SVM_SCORE scoring function is a clear winner. However, Eramian *et al.* (2006) stated that SVMod (SVM_SCORE in Table S1) was derived by using a subsample of the MOULDER dataset. Hence, SVM_SCORE could not avoid overfitting, which can explain its prominent performance on the MOULDER dataset. Taking into account

Table S1. Performance of ten QA methods measured on the CASP11 dataset. Here and below the best values are highlighted in bold, results are sorted by the Spearman correlation, the reference measure is GDT-TS, and native protein structures were filtered out from the dataset.

(a) CASP11 Stage1.

QA Method	GDT-TS loss	Pearson	Spearman	Kendall
ProQ2-refine	0.093	0.653	**0.535**	**0.402**
Wang-SVM	0.109	**0.655**	**0.535**	0.401
SBROD (this study)	**0.083**	0.645	0.522	0.388
Qprob	0.097	0.631	0.517	0.389
ProQ2	0.090	0.643	0.506	0.379
ModelEvaluator	0.097	0.600	0.470	0.353
RWplus	0.135	0.536	0.433	0.433
VoroMQA	0.108	0.561	0.426	0.318
Dope	0.111	0.542	0.416	0.316
RF_CB_SRS_OD	0.162	0.486	0.357	0.357

(b) CASP11 Stage2.

QA Method	GDT-TS loss	Pearson	Spearman	Kendall
SBROD (this study)	**0.057**	**0.441**	**0.426**	**0.298**
Qprob	0.068	0.381	0.387	0.272
VoroMQA	0.069	0.401	0.386	0.269
ProQ2-refine	0.069	0.370	0.375	0.264
ProQ2	0.058	0.372	0.366	0.256
Wang-SVM	0.085	0.362	0.351	0.245
RF_CB_SRS_OD	0.097	0.360	0.350	0.243
Dope	0.077	0.304	0.324	0.228
RWplus	0.084	0.295	0.314	0.220
ModelEvaluator	0.072	0.324	0.305	0.212

Table S2. Performance of several QA methods on the MOULDER dataset. The SBROD scoring function is trained on the CASP[5-11] datasets. Results are sorted by the Spearman correlation.

(a) Metrics calculated with GDT-TS as the target scoring function.

QA Method	GDT-TS loss	Pearson	Spearman	Kendall
SVM_SCORE	0.023	0.938	0.935	0.778
SBROD (this study)	0.041	0.927	0.920	0.755
GA341	0.076	0.838	0.913	0.741
PSIPRED_WEIGHT	0.040	0.919	0.910	0.738
DOPE_AA	0.034	0.887	0.909	0.743
PSIPRED_PERCENT	0.050	0.910	0.900	0.723
PROSA_COMB	0.040	0.889	0.900	0.723
MP_COMBI	0.050	0.879	0.889	0.708
MODCHECK	0.052	0.888	0.887	0.708
DFIRE	0.040	0.846	0.883	0.709
ROSETTA	0.042	0.868	0.878	0.694
PROSA_SURF	0.065	0.873	0.873	0.692
RWplus	0.055	0.847	0.871	0.696
MP_SURF	0.071	0.855	0.864	0.677
Xd	0.077	0.838	0.835	0.656
DOPE_BB	0.053	0.798	0.826	0.634
PROSA_PAIR	0.070	0.799	0.818	0.626
MP_PAIR	0.080	0.793	0.812	0.615
SOLVX	0.100	0.815	0.810	0.614
GB	0.056	0.584	0.794	0.609
EEF1	0.054	0.535	0.791	0.605
FRST	0.100	0.761	0.773	0.588
Anolea_PUC	0.123	0.682	0.667	0.484
Anolea_Z	0.097	0.592	0.645	0.457
SIFT	0.220	0.226	0.275	0.187
Anolea_PE	0.453	0.093	0.127	0.091

(b) Metrics calculated with RMSD as the target scoring function.

QA Method	RMSD loss	Pearson	Spearman	Kendall
SVM_SCORE	0.601	0.874	0.881	0.696
DOPE_AA	0.675	0.870	0.872	0.690
SBROD (this study)	0.890	0.866	0.868	0.682
PSIPRED_WEIGHT	0.792	0.858	0.865	0.672
DFIRE	0.692	0.847	0.859	0.677
PSIPRED_PERCENT	0.919	0.847	0.855	0.661
RWplus	1.272	0.846	0.852	0.670
GA341	1.604	0.768	0.849	0.654
ROSETTA	0.868	0.846	0.843	0.652
PROSA_COMB	0.844	0.835	0.839	0.648
MODCHECK	1.045	0.806	0.827	0.634
MP_COMBI	1.231	0.816	0.823	0.627
PROSA_SURF	1.045	0.803	0.819	0.629
MP_SURF	1.811	0.783	0.809	0.615
DOPE_BB	1.119	0.783	0.787	0.589
Xd	1.741	0.753	0.770	0.585
SOLVX	2.431	0.753	0.762	0.566
EEF1	1.386	0.577	0.758	0.567
PROSA_PAIR	1.961	0.748	0.752	0.560
GB	1.562	0.620	0.751	0.562
MP_PAIR	1.770	0.730	0.739	0.544
FRST	2.081	0.693	0.699	0.515
Anolea_PUC	2.368	0.673	0.645	0.462
Anolea_Z	1.903	0.588	0.615	0.433
SIFT	6.052	0.203	0.257	0.175
Anolea_PE	10.748	0.136	0.159	0.115

that many conventional scoring functions were developed to predict the RMSD similarity, we repeated the same experiment using the RMSD similarity measure as the ground-truth. These results are listed in Table S1b. Again, SVM_SCORE shows the best performance. Also, the performance of DOPE_AA is higher than the performance of SBROD, since DOPE_AA was specifically optimized to predict the RMSD similarity measure as opposed to SBROD, which was trained to predict the GDT-TS.

References

Cao, R. and Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Scientific Reports*, **6**, 23990.

Cecchini, M., Krivov, S. V., Spichty, M., and Karplus, M. (2009). Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers. *The Journal of Physical Chemistry B*, **113**(29), 9728–9740.

Eramian, D., Shen, M.-y., Devos, D., Melo, F., Sali, A., and Marti-Renom, M. A. (2006). A composite score for predicting errors in protein structure models. *Protein Science : A Publication of the Protein Society*, **15**(7), 1653–1666.

Faraggi, E. and Kloczkowski, A. (2014). A global machine learning based scoring function for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, **82**(5), 752–759.

Finkelstein, A., Badretdinov, A., and Gutin, A. (2004). Why Do Protein Architectures Have Boltzmann-Like Statistics? *Proteins: Struct., Funct., Bioinf.*, **23**(2), 142–150.

Hubbard, R. E. and Kamran Haider, M. (2001). Hydrogen Bonds in Proteins: Role and Strength. In *eLS*. John Wiley & Sons, Ltd.

Jing, X., Wang, K., Lu, R., and Dong, Q. (2016). Sorting protein decoys by machine-learning-to-rank. *Scientific Reports*, **6**, 31571.

Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, **116**(14), 7898–7936.

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**(4), 778–795.

Liang, S., Zheng, D., Zhang, C., and Standley, D. M. (2011). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*, **27**(20), 2913–2914.

Liu, T., Wang, Y., Eickholt, J., and Wang, Z. (2016). Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11. *Scientific Reports*, **6**, 19301.

Liu, Y., Zeng, J., and Gong, H. (2014). Improving the orientation-dependent statistical potential using a reference state. *Proteins*, **82**(10), 2383–2393.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function, and Bioinformatics*, **84**, 4–14.

Olechnovič, K. and Venclovas, Č. (2017). Voromqa: Assessment of protein structure quality using interatomic contact areas. *Proteins: Structure, Function, and Bioinformatics*, **85**(6), 1131–1145.

Randall, A. and Baldi, P. (2008). SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Structural Biology*, **8**(1), 52.

Ray, A., Lindahl, E., and Wallner, B. (2012). Improved model quality assessment using ProQ2. *BMC Bioinformatics*, **13**(1), 224.

Rykunov, D. and Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, **11**(1), 128.

Shen, M.-y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, **15**(11), 2507–2524.

Tyka, M. D., Clarke, A. R., and Sessions, R. B. (2006). An Efficient, Path-Independent Method for Free-Energy Calculations. *The Journal of Physical Chemistry B*, **110**(34), 17212–17220.

Uziela, K. and Wallner, B. (2016). ProQ2: estimation of model accuracy implemented in Rosetta. *Bioinformatics*, **32**(9), 1411–1413.

Uziela, K., Shu, N., Wallner, B., and Elofsson, A. (2016). Proq3: Improved model quality assessments using rosetta energy terms. *Scientific Reports*, **6**, 33509 EP –.

Wang, Z., Tegge, A. N., and Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins*, **75**(3), 638–647.

Webb, B. and Sali, A. (2014). Comparative Protein Structure Modeling Using MODELLER. *Current protocols in bioinformatics*, **47**, 5.6.1–32.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, **31**(13), 3370–3374.

Zhang, J. and Zhang, Y. (2010). A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. *PLOS ONE*, **5**(10), e15386.

Zhang, Y. and Skolnick, J. (2007). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, **68**(4), 1020.

Zhou, H. and Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical Journal*, **101**(8), 2043–2052.

Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science : A Publication of the Protein Society*, **11**(11), 2714–2726.