MOLECULAR MODELLING QMEANDISCO—DISTANCE CONSTRAINTS APPLIED ON MODEL QUALITY ESTIMATION CTX-M TARGET PROTEIN RAMACHANDRAN PLOT IN URINARY TRACT INFECTION PATIENTS

Haseera. N[1], Baskaran. K [2*], Shalet Varghese[3], Nirmala Devi.N[4], Arifa.P. P[5].
[1 2 *,3,4,5] Department of Biochemistry, Sree Narayana Guru College, Coimbatore, Tamilnadu, India.

***Correspondence Author:** Dr. K. BASKARAN, MSC, M.Phil., Ph.D.
Assistant Professor Department of Biochemistry, Sree Narayana Guru College, Coimbatore-641105. Tamil Nādu, India.
E-mail: baskar.bio86@gmail.com, Cell number: 91+8760302579

ABSTRACT

3D protein structure model in absence of an experimental reference structure are crucial to determine a model's utility and potential applications. Single model methods assess individual models whereas consensus methods require an ensemble of models as input. In this work, we extend the single model composite score QMEAN that employs statistical potentials of mean force and agreement terms by introducing a consensus-based distance constraint (DisCo) score. DisCo exploits distance distributions from experimentally determined protein structures that are homologous to the model being assessed. Feed-forward neural networks are trained to adaptively weigh contributions by the multi-template DisCo score and classical single model QMEAN parameters. The result is the composite score QMEANDisCo, which combines the accuracy of consensus methods with the broad applicability of single model approaches. We also demonstrate that, despite being the de-facto standard for structure prediction benchmarking, CASP models are not the ideal data source to train predictive methods for model quality estimation. For performance assessment, QMEANDisCo is continuously benchmarked within the CAMEO project and participated inCASP13. For both, it ranks among the top performers and excels with low response times. Availability and implementation: QMEANDisCo is available as web-server at https://swissmodel.expasy.org/qmean.The source code can be downloaded from https://git.scicore.unibas.ch/schwede/QMEAN.

Keyword: QMEAN, DisCo) score, CAMEO

INTRODUCTION

The model of the CTX-M protein was constructed with SWISS-MODEL Server, available at https://swissmodel.expasy.org, and a suitable target was provided. The CTX-M gene of the sample was sequenced using Sanger Sequencing Method which was then used for homology modelling as the target and further analysis. The sequence was submitted (Accession No. OM965359) in National Center for Biotechnology Information (NCBI), which is located in Bethesda, MD, the United States available at *https://www.ncbi.nlm.nih.gov/*. Homology modelling depends on the evolutionary relationship between the target and template protein. The query for the SWISS-MODEL Server should be provided as amino-acid or protein sequence and ExPASy Translate tool was used for the purpose, available at *https://web.expasy.org/translate/*. SWISS-MODEL Server automated mode was followed since the alignment between the target and template sequences was showing high similarity, which is considered as the first step estimating the quality of the model. Target-template alignment was performed using a parallel search method with both BLAST and HH blits. Automated sequence alignments are generally considered acceptable when the target and template have more than 50% sequence identity. The target sequence was chosen and submitted on to the SWISS-MODEL workspace which was followed by template structures evaluation using the structurally conserved and changeable sections data using SWISS-MODEL homology modelling pipeline and thereby the best template was chosen. From the target–template alignment, the positioning of insertions and deletions was viewed in their structural context and changed was made accordingly.

The protein modelled structure assessment and model quality estimation was mainly performed using the Qualitative Model Energy Analysis (QMEANDisCo) (Studer et al., 2020) and Global Model Quality Estimation (GMQE) scores, which served in assessing the reliability of the modelled 3D structure. The Global Model Quality Estimate (GMQE), which takes into account of attributes such as target-template alignment and template structure. The GMQE score can goes from 0 to 1, with a higher score indicating more dependability (Waterhouse et al., 2018). Once the model was developed, the GMQE taken into consideration for the acquiring the model's QMEANDisCo global score to improve quality estimate reliability. Local model per residue score was accessed using the QMEANDisCo scoring function, which is a composite score for single model quality estimation. The QMEANDisCo global score value between 0 and 1, with higher number signifies better quality predicted models. QMEAN, a composite scoring function for model quality estimation, and DFIRE, an all-atom distance-dependent statistical potential is used in the SWISS-MODEL workspace. The QMEAN score was based on four statistical potentials of mean force and their linear combination but using Z-scores, all scores was compared which provided a comparison with experimentally determined structures of similar size. For global model quality estimations, the GMQE and QMEANDisCo global scores were used instead of the QMEAN Z-score analysis. A QMEAN Z-score around zero indicates that the projected structure is "native-like" structure and below -4.0 suggested a model of low quality (Benkert et al., 2011). A comparison plot was produced based on the number of standard deviations from the mean does the modelled structure fall, given a score distribution derived from a large number of empirically determined structures. The SWISS-MODEL Server was also used to provide the structural validation of the modelled target protein (CTX-M) for stereochemical quality and a Ramachandran plot using MolProbity. For an ideal case the MolProbity score will be as low as possible.

The homology modelled CTX-M target protein was downloaded from the SWISS-MODEL Server in protein databank file format (.pdb format). The target protein was prepared by removing the complexed ligands using UCSF Chimera v.1.16, followed by energy minimization using the Minimize Structure tool followed by Dock Prep tool.

MATERIALS AND METHODS

The single model scores from the current version of QMEAN (3) form the basis to obtain per-residue scores in QMEAN DisCo. They are suitable for assessing individual models and are summarized here with their respective statistical potential of mean force terms parameterized as further described in the Supplementary Materials:

- All-atom interaction potential: pairwise interactions are assessed between all chemically distinguishable heavy atoms. A sequence separation threshold has been introduced to reduce contributions from residues adjacent in sequence thereby focussing on long-range interactions and reduce the effect of local secondary structure.
- Cb interaction potential: this term assesses the overall fold by only considering pairwise interactions between Cb positions of the 20 default protein genic amino acids. In case of glycine, a representative of Cb is inferred from the N, Ca and C backbone atom positions. The same sequence separation as in the all-atom interaction potential is applied.
- Packing potential: Assesses the number of surrounding atoms around all chemically distinguishable heavy atoms not belonging to the assessed residue itself.
- Torsion potential: the central U/W angles of three consecutive amino acids are assessed based on the identities of the triplet using a grouping scheme described by Solis and Rackovsky (2006).
- Solvent accessibility agreement: binary classification whether solvent accessibility of a residue matches with the prediction from ACCpro (Cheng et al., 2005).QMEANDisCo (3) introduces four

additional terms. The first two of them are statistical potentials of mean force that are parametrized as described in the Supplementary Materials.

- Cb packing potential: Same concept as the packing potential, but only Cb atoms is considered. Glycine is treated the same way as in the Cb interaction potential.
- Reduced potential: assesses pairwise interactions between reduced representations of amino acids. The reduced representation is composed of the Ca position and a directional component constructed from backbone N, Ca and C positions with further details available in the Supplementary Materials. As in the two other interaction potentials, a sequence separation threshold is applied.
- Clash score: full-atomic clash score as defined for SCWRL3 (Canutescu et al., 2003).
- N: number of residues within 15 A ° by using Ca atoms as reference positions.

The scores are evaluated on a per-residue basis with full-atomic scores averaging their per-atom contributions. Before further processing, all per-residue scores except the number of residues (N) undergo a spherical smoothing (r ¼ 5 A°) as described for QMEAN Brane (Studer et al., 2014).

DisCo is derived from QMEAN DisCo, a quasi-single model method that participated in the CASP9 experiment as a global quality predictor (Biasini, 2013; Kryshtafovych et al., 2011). We revisited the approach of assessing the agreement of pairwise residue–residue distances with ensembles of constraints extracted from experimentally determined protein structures that are homologous to the assessed model. Instead of generating global quality estimates, DisCo aims to predict local per-residue quality estimates. After extracting the target sequence of the model to be assessed, homologues are identified using HH blits (Remmert et al., 2011, the used command line arguments are available in the Supplementary Materials). For each homologue k, all Ca positions are mapped onto the target sequence using the HH blits alignment. Gaussian distance constraints for residue pairs (i, j) are generated for all Ca–Ca distances.

The goal is to construct a pairwise scoring function sij(dij), that assesses the consistency of a particular pairwise Ca–Ca distance dij in the model with all corresponding constraints gijk(dij). In order to avoid biases towards overrepresented sequence families among all found homologues, they are clustered based on their pairwise sequence similarity as specified in the Supplementary Materials. Since the templates often do not cover the entire target sequence, some Ca–Ca pairs might not be represented in every template and consequently the number of templates nijc containing a Ca–Ca pair varies. Within a cluster for different (i, j). Only if a Ca–Ca pair is present in a cluster c, we construct a cluster scoring function.

To get the desired pairwise scoring function sij(dij) we combine hijc(dij) from each cluster c in a weighted manner as exemplified. Clusters expected to be closely related to the target sequence contribute more than others weights wc defined as exp[cSSc] and normalized, so that the weights of all clusters in which the Ca–Ca pair is present, sum up to one. SSc is the average normalized sequence similarity towards the target sequence of cluster c and c is a constant that controls how fast the influence of a cluster vanishes as a function of SSc. The default value for c is 70 and the effect of varying c is discussed in Supplementary Figure S3. The DisCo score of a single residue of the model at position i then is computed by averaging the outcome of all n pairwise scoring functions sij(dij) towards other residues j 6¼ i with their Ca positions within 15A ° . As the accuracy of DisCo depends on the underlying templates, features describing its reliability are required to optimally weigh DisCo with the single model scores in a subsequent machine learning step. For each residue i there are:

- ❖ Average number of clusters c of each evaluated pairwise function sij(dij).
- ❖ Average sequence similarity SS of each evaluated pairwise function sij(dij) with SS being defined as the maximum SSc of all underlying clusters c.
- ❖ Same as above but for sequence identities.
- ❖ Average variance v of each evaluated pairwise function sij(dij) with v being the variance of all observed distances in any of the underlying clusters c.

- Number of evaluated pairwise functions sij(dij).
- Total number of pairwise functions for residue.
- The fraction between the previous two items.
- Score combination fully connected feed-forward neural networks are used to learn complex interdependencies of scores described in the two previous sections. Furthermore, they adaptively weigh single model scores that are capable of scoring individual models and DisCo that depends on dynamically generated constraint data. Neural network training and validation relied on two datasets:
- CAMEO: All models submitted to the CAMEO QE category during 1 year (CAMEO weeks from July 1, 2017 to June 30, 2018) have been collected. These results in a set of _2.4 million per-residue data points from 9500 models built for 883 unique targets. DisCo scores for this set have been estimated from SWISS-MODEL HH blits template searches performed at the time of the CAMEO submission and thus do not contain the target structure.
- CASP12: the CASP12 EMA category (Kryshtafovych et al., 2018) submitted models for each target in two stages. ‘Stage 1’ was a selection of 20 models and ‘stage 2’ the 150 best models according to the Davis–EMA consensus baseline predictor. For each of the 70 finally evaluated targets, 101 of the 150 models submitted in ‘stage 2’ have randomly been selected. For the CASP12 dataset, these results in _1.9 million per-residue data points from 7070 models built for 70 unique targets. DisCo scores for this set have been estimated from HHblits template searches where every template with a release date after May 1, 2016 has been discarded.

Full-model scoring given the nature of lDDT, the average of accurate per-residue quality estimates can be expected to be a good approximation of the global overall quality. That is the definition of the QMEANDisCo global score. The expected error of the global score prediction is defined as the root mean square deviation of prediction and actual global lDDT on a large set of models. As this is derived from the global scoring evaluation, it will further be discussed in Section 3.5.

Blind test all data used for testing/benchmarking were obtained through regular blind predictions from QMEAN-Server instances registered to CAMEO and CASP13. For CASP13, we registered a private QMEAN-Server instance. The server initially deployed a development method called FaeNNz. FaeNNz is conceptually equivalent to QMEANDisCo 3 with the key difference of using less data from

The QMEAN-Server (https://swissmodel.expasy.org/qmean) makes QMEANDisCo accessible to non-expert users with the option to access it through an application programming interface. Alternatively, the underlying source code can be downloaded from https://git.sci core.unibas.ch/schwede/QMEAN under the permissive Apache v2.0 license. The software is based on the OpenStructure computational structural biology framework (Biasini et al., 2010). Computationally intensive tasks are implemented in Cþþ and exported to the Python scripting language to increase flexibility and speedup prototyping of new quality estimation algorithms.

RESULTS

Multiple sequence alignment of the target with the database sequence was performed for finding the template which resulted in 47 related templates. From the alignment result, CTX-M-15 in complex with FPI-1523 (PDB ID: 5FA7) showed 100% identity to the target sequence and was chosen as the template for modelling the protein. The quality estimation of the modelled 3D protein structure was assed using various parameters. The GMQE was found to be 0.97 for the modelled structure (Fig1, 2).

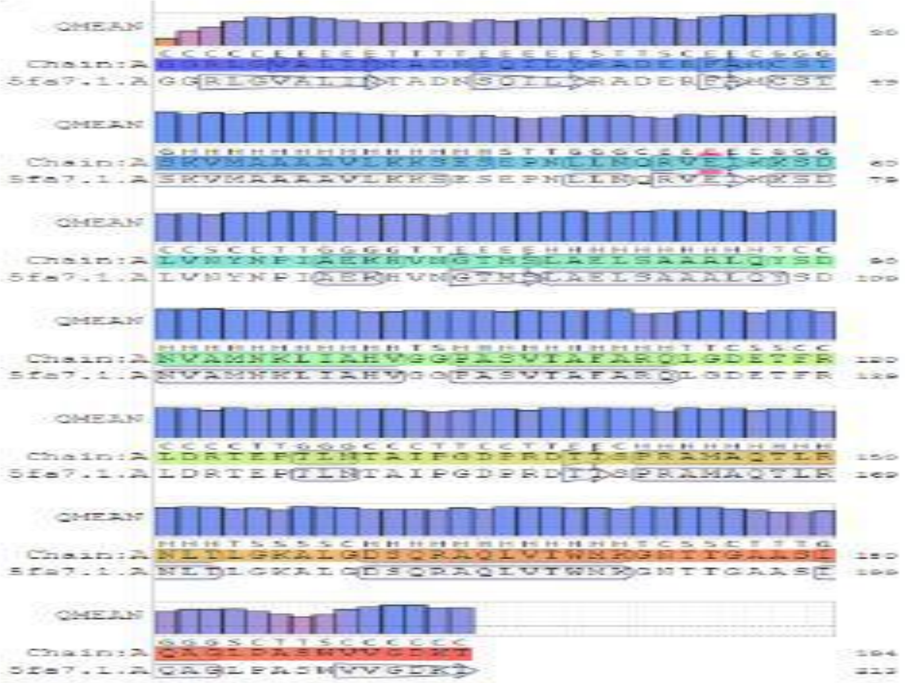
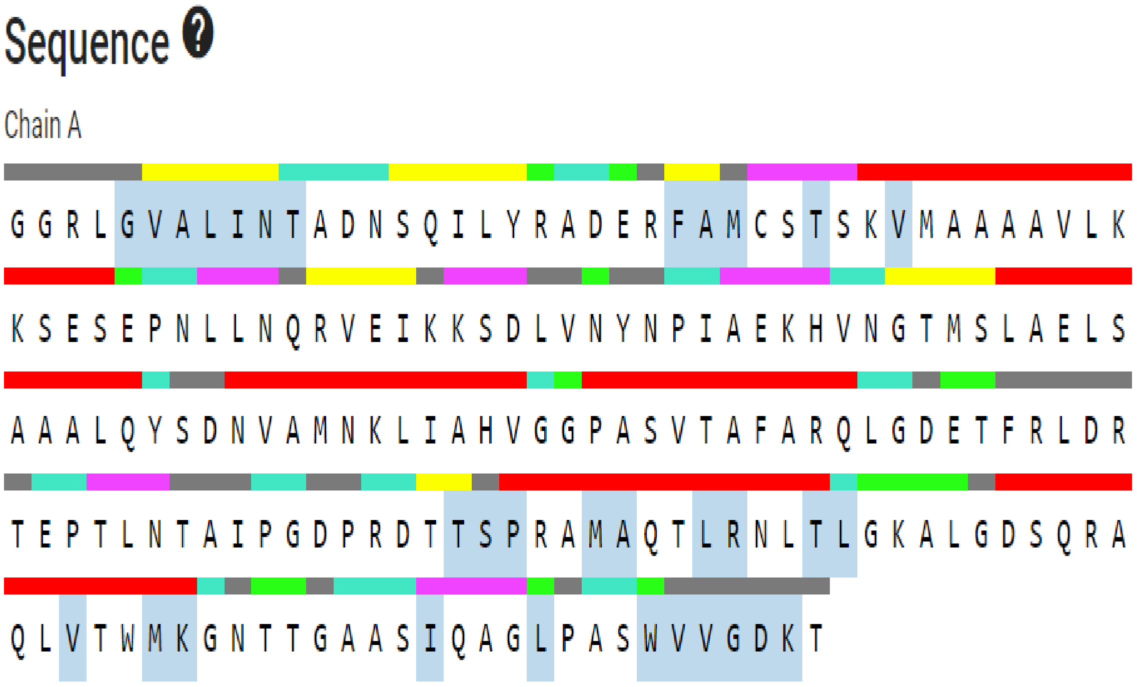


Fig.1. Showing the residue quality with alignment of target and template sequence from N-terminus to C-terminus



Sequence

Chain A

G G R L G V A L I N T A D N S Q I L Y R A D E R F A M C S T S K V M A A A A V L K

K S E S E P N L L N Q R V E I K K S D L V N Y N P I A E K H V N G T M S L A E L S

A A A L Q Y S D N V A M N K L I A H V G G P A S V T A F A R Q L G D E T F R L D R

T E P T L N T A I P G D P R D T T S P R A M A Q T L R N L T L G K A L G D S Q R A

Q L V T W M K G N T T G A A S I Q A G L P A S W V V G D K T

Fig.2. Showing the amino acid sequence (single letter representation) of the CTX-M (target protein). Amino acids in the active site have been highlighted blue

The Local Quality plot displayed the predicted similarity to the native structure (y-axis) for each model residue (stated on the x-axis). Residues with a score of less than 0.6 are usually considered low-quality. Different model chains are depicted in various colours. Comparing QMEAN and QMEANDisCo on CAMEO per-residue data reveals large improvements in overall AUC (0.87 versus

0.94) when adding the described scores and enhanced machine-learning techniques. This is consistent with the observed improvements during training. On both test sets, CAMEO and CASP13, QMEANDisCo, FaeNNz, respectively, have the highest overall AUC among all methods Fig 3.

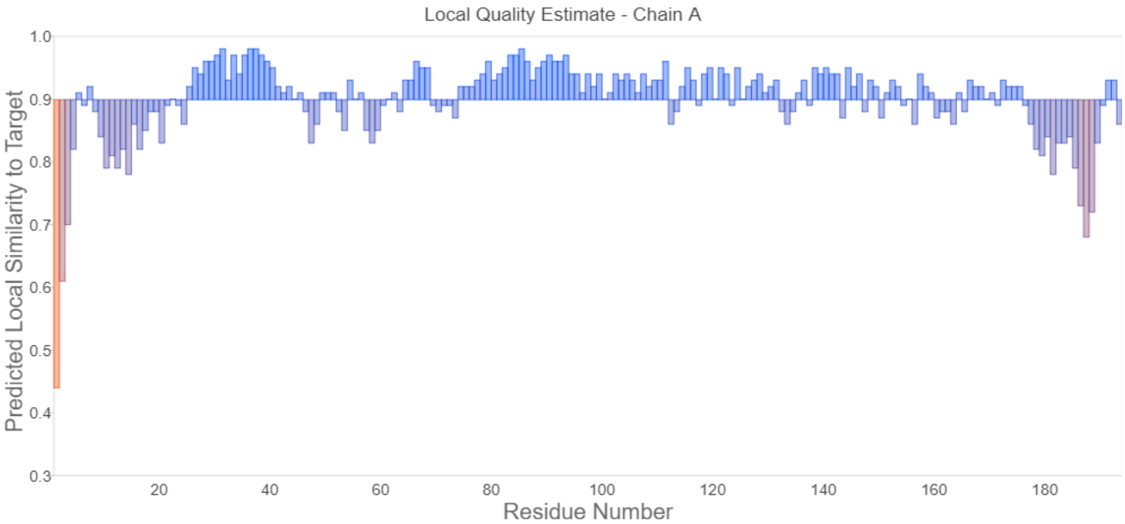


Fig.3. Showing the (QMEAN DisCo) Local Quality plot for modelled structure

QMEAN Z-Scores was also evaluated and a comparison plot was created. Experimental structures with a "QMEAN" score within 1 standard deviation of the mean (|Z-score| between 0 and 1) are black dots, whereas those with a |Z-score| between 1 and 2 are grey. Already during training, the substantially different target value distribution of the used training sets was a concern. Similar distributions can be observed for the test sets. CASP13 has many low quality data points largely originating from random coil models. This gives rise to the hypothesis that much of the overall AUC performance could already be retrieved by detecting those random coils and predicts all their residues to be of low quality.

To test this hypothesis, a naive predictor for CASP13 has been implemented. The global full- model score of the Davis–EMA consensus baseline predictor is blindly assigned to each residue of a model. Detecting random coils and scoring their residues accordingly is not necessarily a bad idea, but this implementation has the obvious flaw of not being able to discriminate correctly and wrongly modelled residues in one particular model. The naive predictor performs surprisingly well with an overall AUC value of 0.82 (Fig. 4). This observation suggests that a good performance in terms of overall AUC might not solely be the result of assigning meaningful per-residue scores but to some extent also a global effect. Consequently, we extended our evaluation to include per-model performance indicators and, for CASP13, repeated it on a subset composed of high quality models.

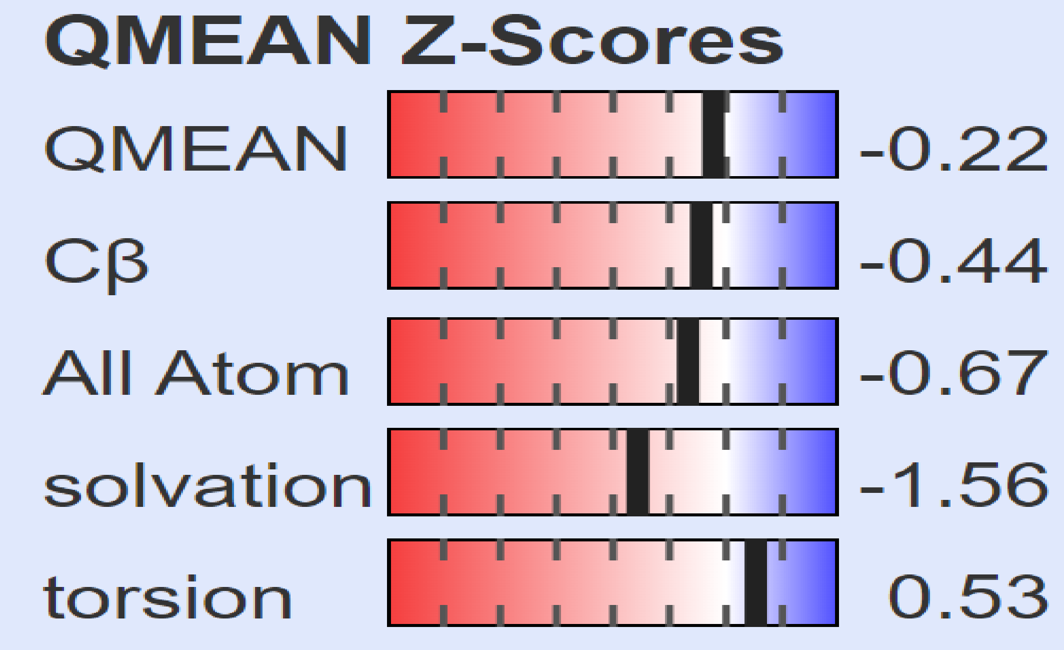


Fig.4. Showing QMEAN Z-scores for modelled structure

Ramachandran plot was assessed using MolProbity software for analysing the quality of the 3D model. The projected model's Ramachandran plot indicated that 98.44 percent of residues were in the most favourable zone, while 1.31 percent was in the permitted region and 0.52percent in the non-favourable region, indicating that the anticipated model is of excellent quality with a MolProbity score of 0.52 (Fig. 5).

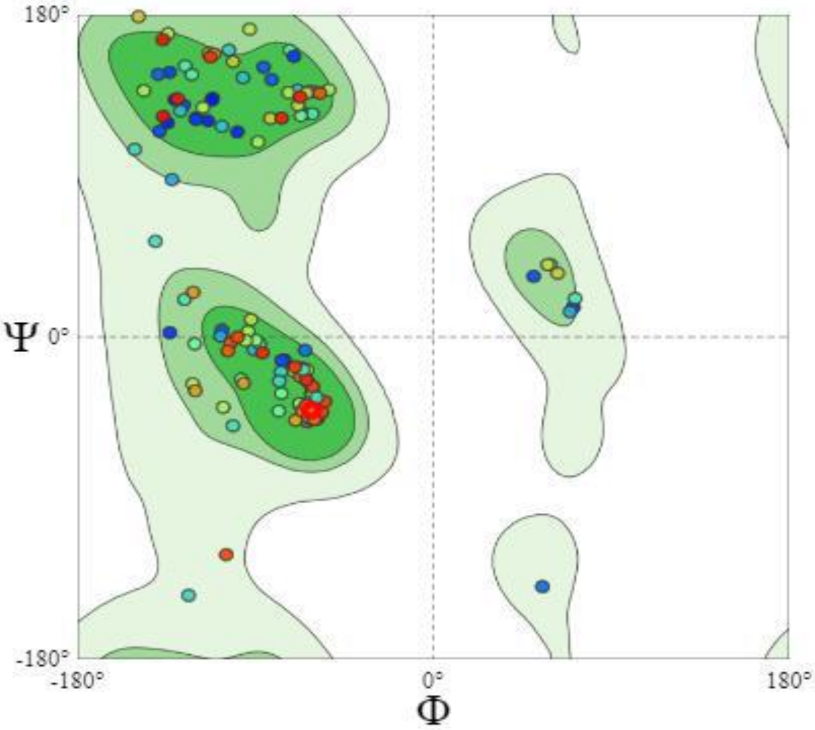


Fig. 5. Showing the Ramachandran plot of the modelled protein structure

QMEANDisCo Global for the modelled protein was 0.90 ± 0.06. Experimental structures farther from the mean represented by light grey and target modelled indicated using a red star. As for the overall AUC, QMEANDisCo performs best for the permodel AUC on CAMEO. On CASP13, CPClab (Mulnaes and Gohlke, 2018) slightly outperforms FaeNNz (per-model AUC of 0.79 versus 0.78). Also ModFOLD7_cor (Maghrabi and McGuffin, 2017), ProQ3D_LDDT and ProQ3D_CAD (Uziela et al., 2018) exhibit no significant difference in per-model AUC (permodel AUC of 0.77 for all three) (Fig. 6).

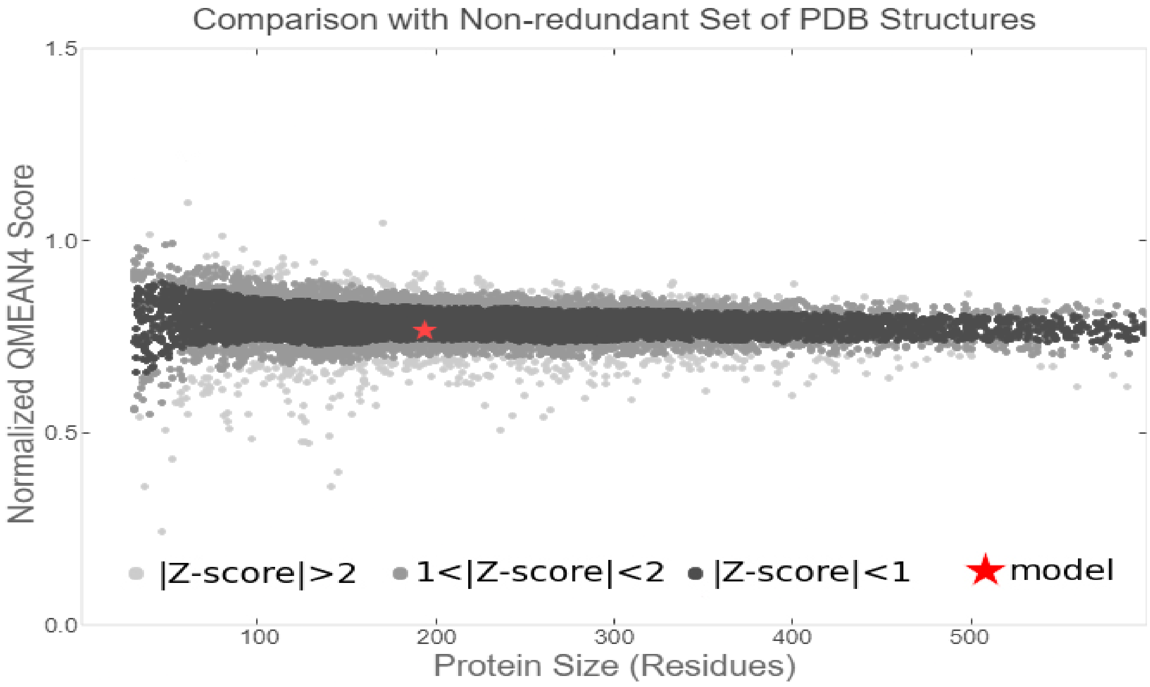


Fig. 6. Showing the comparison plot of QMEAN Z for modelled structure represented in red star

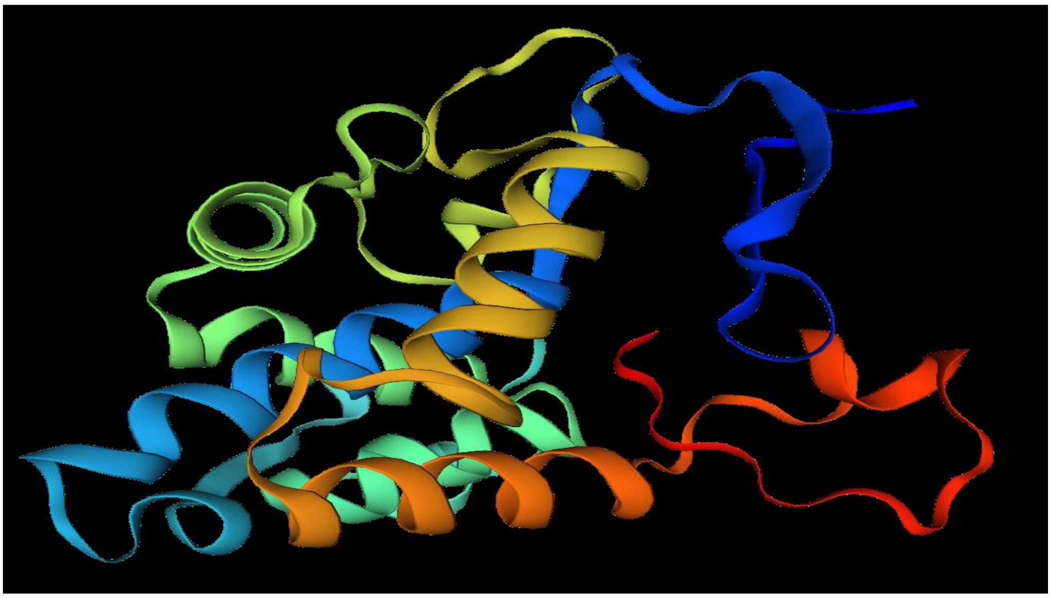


Fig.7. Showing the 3D representation of the modelled protein structure in rainbow colour scheme with N-terminus starting with violet and C-terminus with red

CONCLUSION

In this work, we describe the QMEANDisCo composite score for single model quality estimation. It employs single model scores suitable for assessing individual models, extended with a consensus component by additionally leveraging information from experimentally determined protein structures that are homologous to the model being assessed. By using the found homologues directly, QMEANDisCo avoids the requirement of an ensemble of models as input.

To find the optimal combination of scores, we did profit from recent developments in the machine-learning community providing computational tools that efficiently learn complex interdependencies in large amounts of training data. However, careful datapreparation and handling is crucial for optimal prediction performance.

QMEANDisCo has been developed with its application in the SWISS-MODEL homology modelling server in mind. A template search is the first step of any homology modelling pipeline. As this is the computationally most expensive step in QMEANDisCo, its integration into SWISS-MODEL comes at minimal additional computational cost. The low response times are also reflected in CAMEO where QMEANDisCo returns results within a few minutes with most of the time being spent in the template search step. We believe that we provide a valuable tool that can easily be accessed through the QMEAN-Server. We demonstrated state-ofthe- art performance in predicting lDDT scores with a focus on perresidue predictions. Prediction accuracy can expected to further increase given the growing number of experimentally determined protein structures.

REFERENCES

1. Studer,G. et al. (2014) Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinformatics, 30, i505–i511.
2. Waterhouse,A. et al. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res., 46, W296–W303.
3. Benkert,P. et al. (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics, 27, 343–350.
4. Solis,A.D. and Rackovsky,S. (2006) Improvement of statistical potentials and threading score functions using information maximization. Proteins, 62, 892–908.
5. Cheng,J. et al. (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res., 33, W72–W76.
6. Canutescu,A.A. et al. (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci., 12, 2001–2014.
7. Studer,G. et al. (2014) Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinformatics, 30, i505– i511.
8. Biasini,M. et al. (2013) OpenStructure: an integrated software framework for computational structural biology. Acta Crystallogr. D Biol. Crystallogr., 69, 701–709.
9. Kryshtafovych,A. et al. (2011) Evaluation of model quality predictions in CASP9. Proteins, 79 (Suppl. 10), 91–106.
10. Remmert,M. et al. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMMalignment. Nat. Methods, 9, 173–175.
11. Kryshtafovych,A. et al. (2018) Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. Proteins, 84 (Suppl. 1), 349–369.
12. Biasini,M. et al. (2010) OpenStructure: a flexible software framework for computational structural biology. Bioinformatics, 26, 2626–2628.
13. Mulnaes,D. and Gohlke,H. (2018) TopScore: using deep neural networks and large diverse data sets for accurate protein model quality assessment. J. Chem. Theory Comput., 14, 6117–6126.

14. Maghrabi,A.H.A. and McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models.Nucleic Acids Res., 45, W416–W421.
15. Uziela,K. et al. (2018) Improved protein model quality assessments by changing the target function. Proteins, 86, 654–663.