PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

QMEAN: A comprehensive scoring function for model quality assessment

Pascal Benkert,[1] Silvio C. E. Tosatto,[2] and Dietmar Schomburg[1,3]*

[1] Institute for Biochemistry, University of Cologne, 50674 Cologne, Germany
[2] Department of Biology and CRIBI Biotechnology Center, University of Padova, 35121 Padova, Italy
[3] Bioinformatics & Systems Biology, Technical University Braunschweig, 38106 Braunschweig, Germany

ABSTRACT

In protein structure prediction, a considerable number of alternative models are usually produced from which subsequently the final model has to be selected. Thus, a scoring function for the identification of the best model within an ensemble of alternative models is a key component of most protein structure prediction pipelines. QMEAN, which stands for Qualitative Model Energy ANalysis, is a composite scoring function describing the major geometrical aspects of protein structures. Five different structural descriptors are used. The local geometry is analyzed by a new kind of torsion angle potential over three consecutive amino acids. A secondary structure-specific distance-dependent pairwise residue-level potential is used to assess long-range interactions. A solvation potential describes the burial status of the residues. Two simple terms describing the agreement of predicted and calculated secondary structure and solvent accessibility, respectively, are also included. A variety of different implementations are investigated and several approaches to combine and optimize them are discussed. QMEAN was tested on several standard decoy sets including a molecular dynamics simulation decoy set as well as on a comprehensive data set of totally 22,420 models from server predictions for the 95 targets of CASP7. In a comparison to five well-established model quality assessment programs, QMEAN shows a statistically significant improvement over nearly all quality measures describing the ability of the scoring function to identify the native structure and to discriminate good from bad models. The three-residue torsion angle potential turned out to be very effective in recognizing the native fold.

Proteins 2008; 71:261–277.

Key words: *protein structure prediction; model quality assessment; comparative modeling; fold recognition; statistical potentials; torsion angle potential; scoring function; energy function.*

INTRODUCTION

Over the last two decades, large-scale sequencing projects of whole genomes have resulted in a vast amount of sequences. Of these, a considerable fraction has no annotated function or their mechanism of action is virtually unknown. On the basis of the fact that the biochemical function of a protein is determined by its structure, knowledge of the protein tertiary structure is of paramount importance. To close the gap between the number of known sequences and the fraction for which the structure is known, efficient methods for protein structure prediction are needed that complement current efforts in structural genomics to solve proteins with new folds experimentally.[1] Progress in techniques for protein structure prediction is regularly assessed at the biennial CASP experiment (Critical Assessment of techniques for protein Structure Prediction).[2] Two classes of protein structure prediction methods can be broadly distinguished. In the first class, which is called template-based modeling (TBM) in the terminology of the last CASP round, models are built by using information from experimentally solved structures having detectable similarity to the target sequence. Fold recognition and comparative modeling methods belong to this category. In contrast, the second class of methods, called *de novo* methods, does not rely on any similarity on the fold level (template-free modeling or free modeling). Particularly *de novo* methods, but increasingly also template-based approaches, usually produce a considerable amount of alternative models. Selecting the model being closest to the native conformation of a given protein out of an ensemble of models, independent of being produced during conformational search in a template-free approach[3,4] or on the basis of alternative alignments or different templates,[5–7] is a crucial step in protein structure prediction in general.

Scoring functions rely on the thermodynamic hypothesis stating that the native state of a protein lies in the free energy minimum

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.
Grant sponsors: German Federal Ministry for Education and Research (BMBF); Italian Ministry for University and Research (MIUR).
*Correspondence to: Dietmar Schomburg, Bioinformatics & Systems Biology, Technical University Braunschweig, Langer Kamp 19b, 38106 Braunschweig, Germany.
E-mail: d.schomburg@tu-bs.de
Received 24 April 2007; Revised 27 June 2007; Accepted 2 July 2007
Published online 11 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21715

under physiological conditions.[8] There are basically two categories of scoring functions: physics-based energy functions and knowledge-based statistical potentials. The former are true effective energy functions describing interactions observed in proteins and their parametrization is performed either by fitting experimental data or based on quantum chemical calculations.[9–11] The latter are energy functions derived from data of known protein structures and are usually formalized as either distance-dependent or -independent pairwise potentials of mean force.[12–20] Alternatively, statistical potentials have been derived for other structural features such as torsion angles[21–25] and solvent accessibility.[25–27] Knowledge-based energy functions are based on the inverse Boltzmann equation, although their physical basis has been questioned and is still not completely understood.[28–31] The Boltzmann equation describes a particular system in its thermodynamic equilibrium, whereas statistical potentials assume the system to be a database of protein structures in the free energy minimum. According to this assumption, structural elements such as pairwise distances or torsion angles obey a Boltzmann-type distribution based on a hypothetical reaction at equilibrium in which a unique structure consisting of averaged amino acids "mutates" to a unique sequence.[28,32] Regardless of their vague physical basis, statistical potentials have the advantage of being fast and simple to construct and they are widely used for various purposes among which are fold recognition,[33–37] identification of the native structure among decoys,[38,39] model quality assessment[25,40,41] or prediction of thermo stability.[42–45]

Combining several statistical potential terms covering different aspects of protein structures or models is a popular strategy and the combined potentials have been shown to outperform any single potential.[14,18,23,25,40,41] The composition and implementation of the combined scoring function is determined by its application. A scoring function designed and optimized for the identification of the native structure among decoys does not necessarily show a good performance in discriminating good from bad models. The composition also depends on the coarseness of the models and degree of nativeness of the structures. A hydrogen bonding term may improve the discrimination between near-native structures whereas in *de novo* structure prediction a term describing the compactness of the model or the burial of hydrophobic residues may have the biggest impact. A great variety of scoring functions has been proposed covering some aspects of protein structures and often being optimized for a specific task. There is no single energy function suitable for all tasks and success on one specific decoy set generated by a certain method does not guarantee good performance on another set.[46] The performance of energy functions on decoy sets depends on their similarity to the function used to generate the decoy set. On the other hand, model quality assessment programs (MQAPs) are used to assess models generated by various methods and the quality of the models range from very coarse *de novo* models often having a wrong fold to very accurate template-based models. Therefore, scoring functions consisting of several terms and being optimized on a diverse set of models will be more suitable for the task of discriminating good from bad models or for the identification of the most native-like structure. Model quality assessment programs have been tested the first time in a community-wide experiment in 2004 during CASP6 as part of Critical Assessment of Fully Automated Structure Prediction (CAFASP)[25,47] and only recently at CASP7.

An alternative strategy for model quality assessment is based on the use of consensus information. In consensus methods, the quality of a certain model is assessed by taking into account information contained in the ensemble of models regardless of being produced by a single method, some selected servers or based on the multitude of approaches present at CASP. The simplest consensus methods are purely based on geometric considerations, for example, 3DJury.[48] Alternatively, consensus information can be extracted by analyzing the density distribution in the model conformational space as described in self-RAPDF[49] or by biasing an energy function towards one of the most frequent models, for example, in the colony energy approach.[50] Several of the top performing methods at CASP7 integrated information of the ensemble of server models in some way. Nevertheless, no consensus information is used in this work for the following two reasons: first, the MQAP described in this work should also be applicable to single structures or small sets of structures and, second, a possible extension to a consensus approach can be reached, if required, with minor effort by integrating density information.

One common use of MQAP methods is to rerank the output of single methods,[51] for example, by deriving models from alternative alignments.[7] This is a different scenario than in CASP7, in that all models will be systematically biased towards the same method of generation. Using a clustering method will of course pick the most frequent model, thereby strengthening the systematic bias of the method rather than improving selection. As for integrating density information, one of us has previously shown this to be possible for generic energy functions.[50] However, this approach is comparable to using meta predictors in CASP: it obscures the added value of using more accurate energy functions, would bias the comparison to other "true" energy functions and therefore remains outside the scope of this work.

An early version of the MQAP described in this work has been successfully tested at the seventh round of CASP in summer 2006. The results motivated us to further improve the scoring function by optimizing the parameters of the different potentials and by investigating several alternative combination strategies. QMEAN, which stands for Qualitative Model Energy ANalysis, is a combined

scoring function consisting of three statistical potential terms and two additional terms describing the agreement of the predicted and observed secondary structure and solvent accessibility, respectively. The following aspects distinguish the present work from similar methods described before: QMEAN uses a new implementation of a torsion angle potential over three residues which appear to reflect the local geometry better than the ordinary single residue potential. A residue-level distance-dependent pairwise potential based on Cβ atoms was compiled in a secondary structure-specific manner and several alternative strategies to combine the five terms in a final score were investigated. Finally, the scoring function was compared with other widely-used methods using several decoy sets and the server models submitted to CASP7.

MATERIALS AND METHODS

Statistical potentials

All statistical potentials were extracted from a non-redundant set of high-resolution protein structures from the December 2006 version of the PDB.[52] The PISCES server[53] was used to select structures solved by X-ray crystallography sharing less than 30 sequence identity while having at least a resolution of 1.8 Å and a maximum *R*-value of 0.2. This resulted in an initial selection of 1801 chains. To reduce over-training of the potentials for structures subsequently used for training and testing, all target sequences of CASP6 and CASP7 were blasted against the 1801 chains and all detectable hits were removed resulting in 1688 structures. The following three filters were then applied: structures having less than 90% of the amino acids resolved were not included (171 chains removed), structures with a substantial part being flexible (more than 20% of the residues having an residue-averaged B-factor above two standard deviations) were removed (25 chains) as well as structures with missing backbone atoms or improper residue numeration (21 chains removed). For each of the remaining 1471 structures, DSSP[54] was executed in order to assign secondary structure and solvent accessibility.

Distance-dependent pairwise potential

The distance-dependent pairwise counts were extracted from the protein data set described earlier. To reduce the influence of sequentially local interactions, only contacting pairs separated by at least four residues were included. Alternatively, a sequential separation cutoff of 7 and an implementation without any cutoff have been investigated but resulted in worse performance (data not shown). Cα and Cβ atoms, respectively, have been investigated as possible interaction centers. In the secondary structure specific implementation of the residue-level pairwise potential, the potentials are calculated based on frequency counts extracted from residues of the same secondary structure state while ignoring the secondary structure state of the contacting residues. A distance range of 3–25 Å (bin size 0.5 Å) turned out to produce the best results. The final potential integrated in QMEAN is based on Cβ atoms and uses the secondary structure specific implementation. The calculation of the residue-level pairwise potentials has been carried out as described by Sippl[17] in analogy to the implementation of Melo and Feytmans.[14]

Solvation potential

The degree of residue burial was approximated by counting the number of interaction centers (Cβ atoms for QMEAN) within a sphere of 9 Å around the given amino acid in a similar way as described by Jones[33] and in FRST.[25] The cutoff of 9 Å used in this work resulted in a slightly better performance of the potential than other cutoffs tested. The relative accessibility was then calculated by dividing the counts by the maximum number of counts observed for the given amino acid type in the protein data set. The solvation potential reflects the propensity of a certain residue for a given solvent accessibility compared with any other residue. The potential has been implemented in analogy to Melo *et al.*, 2002.[55]

Torsion angle potential

The single residue torsion angle potential reflects the propensity of a certain residue for a given torsion compared with any other residue. The torsion angles were discretized in 10° bins.[25]

The three-residue torsion angle potential described here is a further development of the ordinary single residue torsion angle potential. The description of the local geometry for a certain residue was extended by including the torsion of the adjacent residues. The coarseness was increased by using 45° bins for the center residue and a bin size of 90° for the dihedral angles of the neighboring residues. Several alternative bin sizes have been investigated ranging from 30° to 90° (data not shown). The identity of the neighbors was not taken into account. The three-residue torsion angle potential is implemented as follows:

$$E^{a}_{\text{torsion}}\left(a,\Phi_{i-1},\Psi_{i-1},\Phi,\Psi,\Phi_{i+1},\Psi_{i+1}\right) = RT\ln\left(1+M_a\sigma\right) - RT\ln\left(1+M_a\sigma\,\frac{f_{\text{observed}}(a)}{f_{\text{reference}}}\right)$$

The weight given to each observation is usually set to $\sigma = 1/100$. M_a is the number of observations for residue type *a*.

$$M_a = \sum_{\Phi_{i-1}=1}^{4}\sum_{\Psi_{i-1}=1}^{4}\sum_{\Phi_i=1}^{8}\sum_{\Psi_i=1}^{8}\sum_{\Phi_{i+1}=1}^{4}\sum_{\Psi_{i+1}=1}^{4} \times f\left(a,\Phi_{i-1},\Psi_{i-1},\Phi,\Psi,\Phi_{i+1},\Psi_{i+1}\right)$$

f_{observed} describes the relative frequency of occurrence of the given local conformation as described by the 6 torsion angles for the amino acid type *a*.

$$f_{\text{observed}}(a) = \frac{f(a, \Phi_{i-1}, \Psi_{i-1}, \Phi, \Psi, \Phi_{i+1}, \Psi_{i+1})}{M_a}$$

$f_{\text{reference}}$ is the reference distribution and describes the relative frequency of occurrence of any residue type with the given local conformation.

$$f_{\text{reference}} = \frac{\sum_{a=1}^{20} f(a, \Phi_{i-1}, \Psi_{i-1}, \Phi, \Psi, \Phi_{i+1}, \Psi_{i+1})}{\sum_{a=1}^{20} M_a}$$

Secondary structure and solvent accessibility agreement

A term describing the agreement of the predicted secondary structure of the target sequence with the DSSP[54] derived secondary structure of the model was built. The DSSP output was converted into the three-state format (helix, sheet, coil) as used in EVA.[56] A consensus secondary structure prediction approach was investigated in an attempt to increase prediction accuracy. A consensus between PSIPRED,[57] SSpro,[58] and ProfSec[59] was build based on simple majority voting.[60] The fraction of residues with identical predicted and observed secondary structure states was used as a simple quality measure. In the final implementation of QMEAN, only PSIPRED was used since the consensus of the methods currently included did not lead to an improved performance.

A similar measure describing the agreement between the predicted binary burial status (buried/exposed) as provided by ACCpro and calculated solvent accessibility based on DSSP was implemented. The relative solvent accessibility was calculated by dividing the solvent accessibility extracted from DSSP by the maximum solvent accessibility for the given amino acid type observed in the training set. Afterwards, the relative solvent accessibility was transformed into the binary classification based on a cutoff of 25%. No consensus scheme was tested in this case.

Measures for the structural similarity between model and target

To evaluate the quality of the models in the two CASP decoy sets described below the GDT_TS score was used as an objective measure for the structural similarity between model and target.[61] The GDT_TS score was calculated using the TMscore software from Zhang and Skolnick.[62] GDT_TS is a well-established score used in the evaluation process of the last CASP rounds having the advantage of being less sensitive to local errors in models as the traditional RMSD (root mean square deviation). GDT (global distance test) describes the maximum percentage of residues which can be structurally aligned within a defined distance cut-off. In GDT_TS 4 increasing distance cut-offs are used ($x = 1$, 2, 4, and 8 Å) and the average of the percentage aligned residues p_x is calculated:

$$\text{GDT\_TS} = (p_1 + p_2 + p_4 + p_8) \div 4$$

For the decoy sets from the Decoys 'R' us website, the RMSD values as provided in the set have been used directly.

Data sets

CASP6 decoy set: Training

Parameter optimization as well as the evaluation of weighting factors for the combined energy function was performed on the CASP6[2] set. This set consists of all the models (TS and AL format) submitted to the 64 accepted targets of CASP6. To increase the quality of the data set and to reduce the influence of random predictions or very difficult targets, all models having a GDT_TS score of less than 0.2 are removed (11,475 models). The final dataset consists of 15,893 models.

Standard decoy sets: testing identification of native structures

The ability of a scoring function to identify the native structure among various decoy structures was investigated and compared with other state-of-the-art tools with the help of the following three frequently used decoy sets from the Decoys 'R' us website[63]: 4state_reduced,[38] lattice_ssfit[64] and LMDS.[65] The performance of the other methods on these decoy sets has not been recalculated here but the corresponding data were taken directly from a recent publication.[25] The two quality measures Z_{nat} and rank1 describe the Z-score of the native structure compared with the ensemble of decoys and the number of cases in which the native structure was ranked first in a given decoy set, respectively.

Molecular dynamics decoy set: testing on near-native solutions

The decoy set generated by Fogolari *et al.*[66] was used to estimate the performance on near-native structures. It consists of over 6000 snapshots from five independent molecular dynamics simulations. One simulation started from the native structure and the other four from minimized conformations of the thermo-stable sub-domain from the chicken villin headpiece consisting of 36 residues. The decoy set can also be downloaded from the Decoys 'R' us website and covers RMSD values from 2 to 12 Å.

CASP7 decoy set of server models: testing model quality assessment

The CASP7 server models for all 95 accepted targets were downloaded from the CASP website (URL: http://predictioncenter.org/casp7/). This is the same data basis used in the blind test for model quality assessment programs which was part of CASP7. Although all quality predictions submitted for the MQAP session were available on the CASP website, this data were not used here. Rather, predictions were recalculated with some well-established

model quality assessment programs (MQAPs) downloadable from the CAFASP4 website (URL: http://www.cs.bgu.ac.il/dfischer/CAFASP4/) or requested from the authors (DFIRE). This has the following reasons: First, many of the MQAPs joining CASP7 have not been published yet and from the abstracts submitted it was mostly impossible to understand how they work. Second, the top performing MQAPs all integrated consensus information in their calculation, which is not in the scope of this work. Third, the data is sometimes difficult to compare: some MQAPs fail to predict the model quality for many servers or have not submitted any predictions for some targets. The following MQAPs were used: FRST,[25] Modcheck,[51] ProQ,[41] RAPDF,[15] and DFIRE.[20] Only server models for which all of the five MQAPs were able to return a prediction were evaluated resulting in a total number of 22,420 models over all 95 targets. ProQ has been executed in two different modes either using secondary structure information (provided as a PSIPRED prediction) or not. Results are reported for ProQ-MX as this was found to give better results than ProQ-LG (data not shown).

The 95 targets were divided into the two categories free-modeling (FM) and template-based modeling (TBM) as introduced in the seventh round of CASP (see: http://predictioncenter.org/casp7/meeting_docs/difficulty.html). Since several targets are multi-domain structures and the domains can sometimes be assigned to different categories, multi-domain targets were assigned to the category of the most difficult domain they include (i.e. a target consisting of a FM domain and a TBM domain was assigned to the FM category). The final division is shown in Table SI in Supplementary Material.

Evaluation criteria

A variety of quality measures[25,49] have been used to compare the performance of the different methods. $\log P_{B1}$ and $\log P_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the 10 best-scoring models, respectively. Suppose the best scoring conformation *xi* has the GDT_TS rank of *Ri* in *n* decoy conformations, then the log probability is given by

$$\log P_{B1} = \log\left(\frac{R_i}{n}\right)$$

$$\text{for } \log P_{B10}: \quad R_i = \min\left[R_1, \ldots, R_{10}\right]$$

Fraction enrichment (FE) is the percentage of top 10% lowest RMSD conformations (Table V) or highest GDT_TS models (Table VI) among the top 10% best-scoring structures. In the FE curves[67] (Fig. 4) variable cutoffs are used ranging from 5 to 50%. The enrichment as defined in Tsai *et al.*[39] ($E_{15\%}$) is calculated by dividing the number of top 15% highest GDT_TS models found among the top 15% best predicted models divided by the number obtained in a random selection (15% × 15% × number of structures in the decoy set). Z_{nat} is the *Z*-score of the native structure as compared with the ensemble of models. Rank1 and rank10 are the number of targets in which the native structure (or the best model based on GDT_TS, excluding the native structure) was found on the first rank or among the top 10 predictions, respectively. *GDT_TS loss* is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set. Two kinds of regression coefficients have been used: Pearson's correlation coefficient r^2 and the Spearman's rank correlation coefficient ρ. The statistical significance of the difference between methods was assessed using the method described by Marti-Renom *et al.*[68] at the 95% confidence level.

Parameter optimization

Parameters for the statistical potentials (such as distance range, bin size, resolution, interaction center etc.) were optimized on the CASP6 set. To measure the ability of the potential to predict the model quality, the Pearson correlation coefficient between the predicted model energy and the measured quality in terms of GDT_TS was used. A variety of alternative implementations of the potentials were investigated and the best performing torsion angle potential, solvation potential, and pairwise potential are selected based on the correlation coefficients. Additionally, the weighting factors for the combined potential are evaluated by an exhaustive search strategy over reasonable ranges for the different weighting factors. The final combination is selected based on the maximum correlation coefficient.

Several alternative optimization strategies were investigated: Pearson's correlation coefficient vs. Spearman's rank correlation, energy vs. *Z*-scores compared with sequence-shuffled models. Parameters were optimized on a target-specific basis (i.e. regressions for all models of each target separately) or on a global basis by maximizing the regression over all models from all targets simultaneously.

The target-specific optimization was accomplished by averaging the Pearson's correlation coefficient over all targets provided that at least a suitable fraction (i.e. 150 models which are around 30%) have a GDT_TS higher than 0.2. In this way, 12 of the 64 accepted targets of CASP6 set were excluded from the target-specific evaluation. All but one belongs to the novel fold or fold recognition categories. The following targets were excluded in the target-specific optimization process (in brackets the number of models with GDT_TS > 20): T0202 (118), T0206 (94), T0228 (23), T0238 (129), T0242 (139), T0248 (5), T0262 (70), T0272 (4), T0273 (88), T0197 (51), T0198 (104), T0199 (12). This was done with the intent to reduce the influence of very difficult free modeling targets in which most of the groups failed to build a reasonable model. These targets are expected to add no value in the optimization process. In contrast to the

Pearson correlation, the Spearman rank correlation allows to investigate a relationship which does not have to be necessarily linear. As described in Pettitt *et al.*[51] *Z*-scores were built comparing the score of the model with the scores of models after sequence shuffling (1000 times in this work).

RESULTS

Generation of the potentials and training

All statistical potentials were extracted from a nonredundant protein data set of 1471 high-resolution structures using the PISCES server[53] and after having applied additional quality filters as described earlier. Parametrization of the different potentials and optimization of the weighting factors for the combined scoring function were both performed on the CASP6 decoy set by analyzing the regression between the GDT_TS score of the models and the predicted score provided by the energy function. For the three statistical potentials entering the QMEAN function a variety of alternative implementations has been investigated (detailed results will not be shown here). Table I provides a short description of all scoring function terms mentioned in this work and the different versions of QMEAN which were built to assess the influence of the two agreement terms. In the following, QMEAN, unless specified with an index, always indicates the original scoring function consisting of five terms (QMEAN5).

Table II contains regression coefficients achieved in a regression of the models GDT_TS scores and the QMEAN scores. Two different regression schemes were investigated: A direct correlation of the scores (Pearson's correlation coefficient) and a rank correlation (Spearman's ρ) in the hope of taking into account a possible nonlinear relationship. As an alternative, the scores are transformed into *Z*-scores by comparing the given model to 1000 other models with the same structure but randomly shuffled sequences. Shuffling the order of the residues has been shown[55] to work almost as good as randomizing the structure as originally proposed by Sippl.[18] Furthermore, two different strategies for the optimization of the weighting factors have been investigated: First, an optimization of the regression on a target-specific basis by maximizing the average of the regression coefficients achieved on the individual targets and second, a global approach in which the regression is optimized by using all models from all the targets at once. The regression coefficients achieved for the different scoring function terms and their combinations do not differ much between the six optimization strategies and all show the same tendency. QMEAN5, which is a linear combination of five terms (see Table I), consistently achieves the highest regression coefficients for all optimization strategies, directly followed by QMEAN4. QMEAN3, consisting only of statistical potential terms, shows a slightly worse correlation but is still better than any other single term. A Pearson correlation coefficient of 0.72 was observed for QMEAN5 in the global approach in which the regression is optimized over all models of all targets at once. The scatter plot in Figure 1(a) shows a clear trend but also the presence of some outliers. The weighting factors achieved in the two target-specific approaches (Spearman and Pearson) is quite similar to each other. In comparison to those in the global strategy, lower weight was assigned for the torsion and pairwise term (see Table SII in Supplementary Material). In any case, the performance differences when applying the weights of the six strategies to the decoy sets described in the next two sections are overall negligible. For the sake of simplicity, the weights of the global optimization strategy are used throughout:

Table I
Short Description of the Terms and Their Combinations Used in This Work

Scoring function	Description
Torsion single	Ordinary torsion potential based on Φ and Ψ propensities of single amino acids. Bin size: 10°
Torsion three-residue	Extended torsion potential over three consecutive residues. Bin sizes: 45° for the center residue, 90° for the two adjacent residues
Pairwise Cα/pairwise Cβ	Residue-specific pairwise distance-dependent potential using Cα or Cβ atoms, respectively, as interaction centers. Range 3–25 Å, step size: 0.5 Å
Pairwise Cβ/SSE	In analogy to pairwise Cβ, but a secondary structure specific implementation was used both for the derivation and application of the potential.
Solvation Cβ	Potential reflecting the propensity of a certain amino acid for the a certain degree of solvent exposure based on number of Cβ atoms within a sphere of 9 Å around the center Cβ.
SSE X	Agreement between the predicted secondary structure of the target sequence (using method X, or consensus of three methods) and the observed secondary structure of the model as calculated by DSSP. QMEAN uses X = PSIPRED
ACCpro	Agreement between the predicted relative solvent accessibility using ACCpro (two states buried/exposed) and the relative solvent accessibility derived from DSSP (>25% accessibility => exposed)
QMEAN3	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, salvation Cβ
QMEAN4	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED
QMEAN 5	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED, ACCpro

Table II
Absolute Values of the Regression Coefficients Received in a Regression of the GDT_TS Score Against the Predicted Score

	Pearson's correlation coefficient[a]				Spearman's correlation coefficient[b]	
Scoring function	Global[c]	Global/*Z*-score[c,d]	Target average[e]	Target average/*Z*-score[d,e]	Target average[d]	Target average/*Z*-score[d,e]
Torsion single	0.35	0.39	0.25	0.30	0.23	0.24
Torsion 3-residue	0.52	0.50	0.35	0.39	0.32	0.31
Pairwise Cα	0.54	0.57	0.42	0.54	0.37	0.42
Pairwise Cβ	0.57	0.59	0.47	0.56	0.43	0.46
Pairwise Cβ/SSE	0.61	0.60	0.49	0.58	0.45	0.48
Solvation Cβ	0.58	0.55	0.50	0.52	0.46	0.43
SSE PSIPRED	0.57	0.57	0.52	0.54	0.48	0.48
SSE ProfSec	0.53	0.53	0.49	0.52	0.45	0.45
SSE SSpro	0.56	0.56	0.50	0.52	0.45	0.45
SSE consensus	0.57	0.57	0.51	0.53	0.46	0.46
ACCpro	0.53	0.53	0.47	0.51	0.47	0.47
QMEAN 3terms	0.66	0.64	0.56	0.58	0.52	0.52
QMEAN 4terms	0.71	0.69	0.62	0.64	0.57	0.58
QMEAN 5terms	0.72	0.69	0.64	0.65	0.59	0.60

[a]weighting factors achieved by maximizing the Pearson's correlation coefficient.
[b]weighting factors achieved by maximizing the Spearman's rank correlation coefficient.
[c]Global: The regression is optimized on all models of all targets.
[d]*Z*-score: The *Z*-score of the model compared with 1000 sequence-shuffled models for the same structure is calculated.
[e]Target average: The average of the regression coefficients achieved on the models of the individual targets is optimized.

$$\begin{aligned}\text{QMEAN5} = {} & 0.3 \times \text{Score}_{\text{torsion 3-residue}} \\ & + 0.17 \times \text{Score}_{\text{pairwise C}\beta/\text{SSE}} + 0.7 \times \text{Score}_{\text{solvation C}\beta} \\ & + 80 \times \text{Score}_{\text{SSE PSIPRED}} + 45 \times \text{Score}_{\text{ACCpro}}\end{aligned}$$

The final implementation of the residue-level distance-dependent pairwise potential is based on Cβ atoms as interaction centers and the radial distribution between 3 and 25 Å (bin size 0.5 Å) was taken into consideration. As expected,[55,69] Table II shows that Cβ atoms are the better interaction centers as compared with Cα atoms. More interesting is the fact that compiling and applying the distance-dependent pairwise potential in a secondary structure specific manner (see Materials and Methods) results in a better regression as compared with the regular pairwise potential. Table III shows that the secondary structure-specific implementation does not have a significantly higher cross-correlation to any of the other terms.

A coarse-grained torsion angle potential using the Φ/Ψ angles of three consecutive residues was developed. The bin sizes are 45° for Φ and Ψ of the center residue and 90° for the neighboring torsion angles. Table II underlines the considerably better correlation of the three-residue torsion angle potentials with the GDT_TS score as compared with the regular single residue torsion angle potential. This goes together with a slight increase in the correlation to the other terms (Table III). Additionally, the three-residue torsion angle potential shows a remarkable performance in recognizing the native structure in the CASP7 decoy set (Table VI, last three columns) as well as in the Decoys'R'us decoy sets (Table IV) and often outperforms QMEAN in this task. The solvation potential shows a relatively high cross-correlation to the pairwise potentials which can be assigned to the similarity of their implementation (Table III). The correlation to the ACCpro term is lower than could be expected.

The SSE PSIPRED terms describing the agreement of the predicted secondary structure of the sequence provided by PSIPRED and calculated secondary structure as measured by DSSP results in an increase of the regression coefficient of at least 0.05 in all the approaches (Table III) while having no noticeable cross-correlation to any of the other terms and QMEAN3. A consensus secondary structure prediction[60] did not result in better regression coefficients as compared to PSIPRED. ACCpro, describing the agreement of the predicted and observed solvent accessibility, only leads to a minor increase of the regression coefficients of QMEAN5. ACCpro shows a cross-correlation around 0.6 to the distance-dependent potentials and the solvation potential and a comparison of the regression to QMEAN3 and QMEAN5 would suggests that ACCpro does not add any value to the combined score. However, Table VI proves that the opposite is true: ACCpro shows a very good performance according to the enrichment quality measures and is responsible for the constant improvement in all quality measures of QMEAN5 over QMEAN4.

According to Table III, a major part of the discriminatory power of QMEAN3 can be assigned to the pairwise Cβ/SSE and to the solvation potential. The correlation of the three-residue torsion angle potential is still rather high (regression coefficient 0.78). The secondary structure agreement term shows a significantly higher correlation to QMEAN5 than ACCpro.

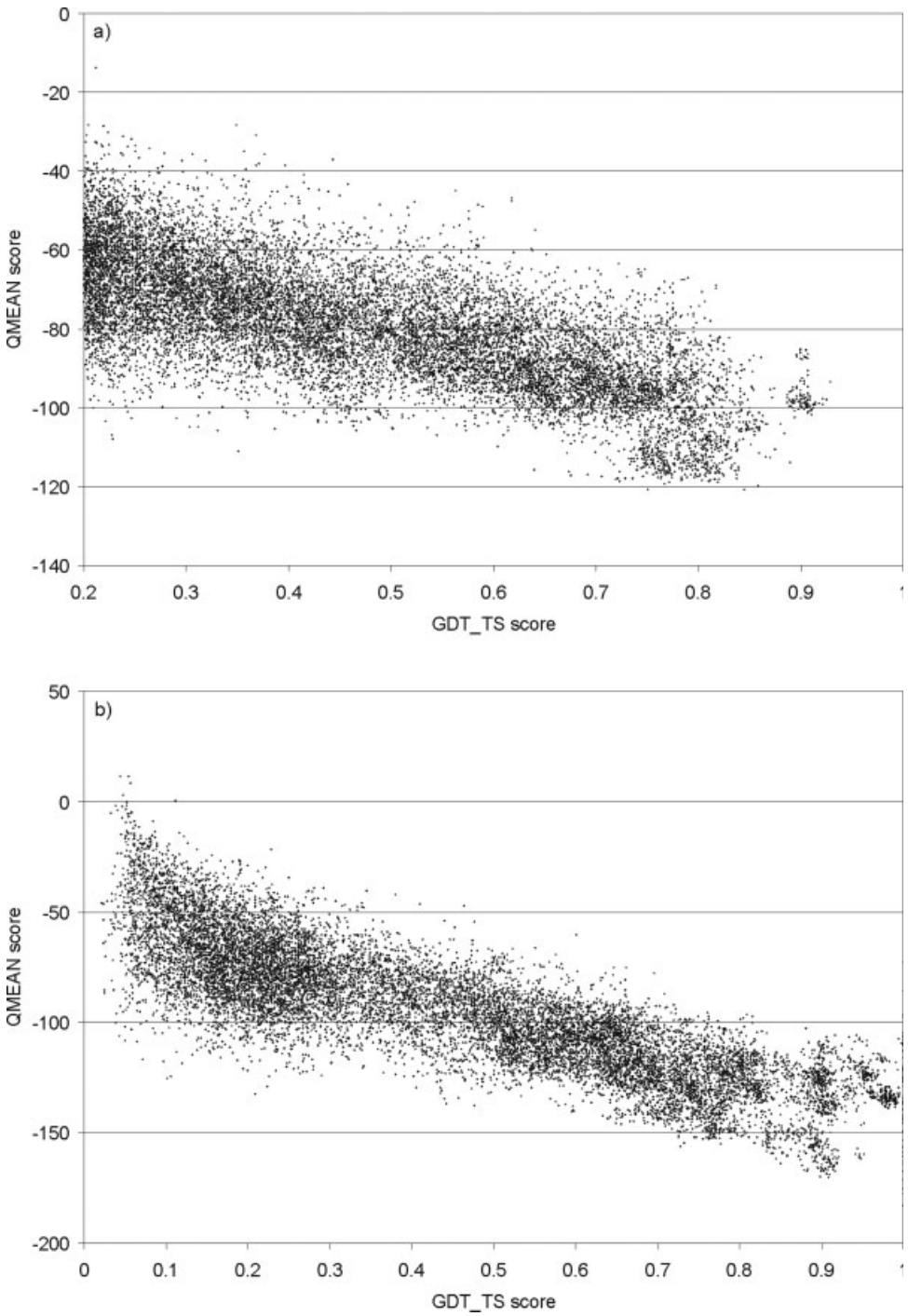

Figure 1

*Final correlation between GDT_TS and QMEAN score (based on 5 terms) for all models of the CASP6 training set (**a**) after optimization of the weighting factors and for all server predictions from the 95 accepted targets of CASP7 (**b**) applying the optimized weighting factors. Only models with GDT_TS > 0.2 have been use for training.*

Performance on three standard decoy sets

In order to compare the performance to several well-established scoring functions, QMEAN was tested on three standard decoy sets from Decoys 'R' Us. As can be seen from Table IV, the three-residue torsion angle potential shows the overall best performance in selecting the native structure and outperforms all other terms of QMEAN as well as all QMEAN versions. Except for the lattice_ssfit

Table III
Cross-Correlation Analysis of the Terms Entering the Combined Score (QMEAN) and Some Selected Scores for Comparison

	Torsion single	Torsion three-residue	Pairwise Cβ	Pairwise Cβ/SSE	Solvation	SSE PSIPRED	ACC Sspro	QMEAN 3terms	QMEAN 5terms	GDT_TS
Torsion single	1.00	0.81	0.41	0.43	0.34	0.35	0.31	0.59	0.54	−0.35
Torsion three-residue	0.81	1.00	0.58	0.60	0.50	0.48	0.41	0.78	0.73	−0.52
Pairwise Cβ	0.41	0.58	1.00	0.97	0.71	0.43	0.58	0.89	0.83	−0.57
Pairwise Cβ/SSE	0.43	0.60	0.97	1.00	0.72	0.44	0.62	0.92	0.85	−0.61
Solvation	0.34	0.50	0.71	0.72	1.00	0.48	0.62	0.87	0.81	−0.58
SSE PSIPRED	0.35	0.48	0.43	0.44	0.48	1.00	0.42	0.54	0.81	−0.57
ACCpro	0.31	0.41	0.58	0.62	0.62	0.42	1.00	0.65	0.64	−0.53
QMEAN3	0.59	0.78	0.89	0.92	0.87	0.54	0.65	1.00	0.93	−0.66
QMEAN5	0.54	0.73	0.83	0.85	0.81	0.81	0.64	0.93	1.00	−0.72
GDT_TS	−0.35	−0.52	−0.57	−0.61	−0.58	−0.57	−0.53	−0.66	−0.72	1.00

The Pearson's correlation coefficients are based on the "global" optimization strategy without *Z*-scores.

decoy set the torsion angle potential also produces the highest Z_{nat} scores. The pairwise potential performs comparably well on lattice_ssfit, shows a moderate performance on 4state_reduced and fails on LMDS. The solvation potential only produces reasonable *Z*-scores on the lattice_ssfit but fails completely on the other two sets. Comparing the performance of QMEAN5 on the three decoy sets, it seems that QMEAN5 performs best on lattice_ssfit. In general, the performance of QMEAN5 is comparable to the other methods taking into account the fact that QMEAN has been trained for model quality assessment and not specifically for the task of identifying native structures. The advantage of QMEAN5 as a combined scoring function over energy functions based on a single term is the decreased chance to fail on some decoy sets generated based on a specific method. Although the data basis is too sparse for well-founded conclusions, Table IV suggests that the performance of a certain scoring function is dependent on the decoy set. More precisely, how a given decoy set has been built appears to allow some terms to perform better on one decoy set than another.

Performance on a molecular dynamics simulation decoy set

The decoy set generated by Fogolari *et al.*[66] consists of 6255 snapshots from five different molecular dynamics simulations of the thermostable subdomain from the chicken villin headpiece. Since one simulation started from the native structure and the other four from alternative minimized conformation, this yields a wider range of RMSD values compared with the previously mentioned

Table IV
Comparison of QMEAN With Other Methods in the Performance of Selecting the Native Structure in Some Standard Decoy Sets From Decoys 'R' us

	4State_reduced		Lattice_ssfit		LMDS	
	Rank1	Z_{nat}	Rank1	Z_{nat}	Rank1	Z_{nat}
ProQ	5/7	4.1	7/8	12.1	4/10	3.7
Errat	1/7	2.5	3/8	5.1	5/10	3.1
ProsaII	5/7	2.7	8/8	5.6	6/10	2.5
Verify3D	4/7	2.6	7/8	4.5	2/10	1.4
SNAPP	3/7	2.6	5/8	3.5	2/10	1.1
AKBP	7/7	3.2	8/8	6.6	3/10	0.5
DFIRE	6/7	3.5	8/8	9.5	7/10	0.9
RAPDF	7/7	3.0	8/8	7.2	3/10	−0.5
FRST	7/7	4.4	8/8	6.7	6/10	3.5
Torsion three-residue	7/7	3.6	6/8	5.0	7/10	3.7
Pairwise Cβ/SSE	3/7	2.0	7/8	5.1	1/10	0.4
Solvation	0/7	1.6	3/8	3.1	0/10	1.1
SSE PSIPRED	0/7	1.6	7/8	5.4	2/10	1.3
ACCpro	1/7	2.0	5/8	3.7	3/10	1.9
QMEAN3	4/7	2.7	8/8	6.2	2/10	2.3
QMEAN4	3/7	2.4	8/8	7.5	4/10	2.3
QMEAN5	4/7	2.5	8/8	7.7	6/10	2.7

Rank1: Number of decoy set in with native structure was found on the first rank.
Z_{nat}: *Z*-score of the native structure compared with the ensemble of structure in the decoy set.

decoy sets. The other advantage is that it allows a direct comparison with molecular mechanics (MM) force fields.

As can be seen from Figure S1 in Supplementary Material, QMEAN consistently assigns low energies to the near-native conformations of the simulation starting from the native structure (colored in black). Especially, the decoys from the native simulation show a clear correlation between the RMSD and the score predicted by QMEAN5. Although the native structure was not predicted to have the lowest energy, several conformations around 2 Å RMSD get quite low energies. This is also reflected by the excellent $\log P_{B10}$ value of QMEAN5 as shown in Table V. The solvent accessibility agreement term seems to be quite good in identifying near-native structures and to a certain extent also the torsion angle potential over three residues, as reflected by the low $\log P_{B10}$ value and the high FE score. The secondary structure agreement term produces a FE of over 90% which indicates that there were no major changes in secondary structures during the simulation starting from the native structure. The RMSD values of the conformation with the lowest score are more or less the same for all three QMEAN versions whereas ACCpro is able to pick the second best conformation. The solvation potential produces bad results across all quality measures. In comparison to the three versions of MM energy functions, QMEAN shows comparable regression coefficients and $\log P_{B1}$ values but performs significantly better in the enrichment of near-native solutions.

Performance on the CASP7 decoy set

A different, and perhaps more realistic, test case is presented by the decoys from the CASP7. In Table VI QMEAN and its component scoring function terms are compared with five widely-used model quality assessment programs (MQAPs). The following executable programs could be downloaded from the CAFASP4 website (URL: http://www.cs.bgu.ac.il/dfischer/CAFASP4/): Modcheck,[51] RAPDF,[15] FRST,[25] and ProQ.[41] DFIRE[68] was requested from the author. ProQ was executed both with and without PSIPRED secondary structure prediction.

Table VI shows the average performance of the methods over all targets using different quality measures. Most of the quality measures have been previously introduced and described,[39,49] but a detailed definition can be found in Materials and Methods. The last three columns describe the scoring functions ability in identifying the native structure out of the ensemble of models for a specific target whereas all other measures describe different aspects of model quality assessment. The opposite algebraic sign of Modcheck and ProQ observed for the Pearson's correlation coefficients and for the Z_{nat} scores can be ascribed to the fact that these two tools use an inverse scaling compared with the other scoring function by assigning the highest scores to the best models. The statistical significance of these results was validated using the method described by Marti-Renom *et al.* at a 95% confidence level and the results are summarized in Figure 2.

In general, QMEAN5 consistently outperforms the other five MQAPs with respect to almost all tested quality measures on both categories (FM and TBM, see Supplementary Material Table SIII + SIV) and over all targets (see Table VI). On the two regression and enrichment quality measures, QMEAN5 performs significantly better than all other methods tested (see Fig. 2). For the task of identifying the native structure QMEAN3 is

Table V
Comparison of QMEAN and its Terms With Three Molecular Mechanics Energy Functions, a Contact Potential and FRST

Scoring function	$\log P_{B1}$[a]	$\log P_{B10}$[a]	FE[b]	Corr._coeff.[c]	RMSD (Å) best score
contact+pc	−1.08	−1.08	13.80	0.62	3.03
FRST	−1.38	−1.94	23.20	0.48	2.61
MM	−0.25	−1.39	10.60	0.21	7.45
MM/GBSA	−1.71	−2.02	29.60	0.66	2.40
MM/PBSA	−1.79	−2.02	23.20	0.58	2.35
QMEAN3	−1.50	−3.50	36.50	0.53	2.52
QMEAN4	−1.71	−2.80	90.20	0.56	2.40
QMEAN5	−1.51	−3.50	88.00	0.57	2.51
torsion 3-residue	−1.26	−2.80	58.40	0.57	2.71
pairwise Cb/SSE	−1.02	−1.41	35.50	0.64	3.34
solvation	−0.32	−0.98	6.10	0.20	7.15
SSE PSIPRED	−1.32	−1.32	91.20	0.55	2.58[d]
ACCpro	−3.50	−3.50	63.00	0.50	1.84

[a] $\log P_{B1}$ and $\log P_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the 10 best-scoring models, respectively.
[b] FE stands for fraction enrichment.
[c] Person's correlation coefficient.
[d] Since 56 structures with identical SSE PSIPRED scores were found, their average RMSD is shown.

Table VI
Performance of Different Scoring Functions in Predicting the Quality of the Server Models Submitted for all 95 Targets of CASP7. Comparison of QMEAN With Other Well-Known MQAPs

	Regression[a]		Enrichment[b]		Best predicted model[c]			Best model (GDT_TS)[d]			Native structure[d]		
Method	r^2	ρ	FE	$E_{15\%}$	Rank10	$\log P_{B1}$	$\log P_{B10}$	GDT_TS loss	Rank1	Rank10	Z_{nat}	Rank1	Rank10
Modcheck	0.64	0.59	0.33	2.70	17	−0.70	−1.67	−0.18	6	27	1.99	47	69
RAPDF	−0.50	0.50	0.31	2.44	17	−0.91	−1.67	−0.08	4	17	−2.09	55	77
DFIRE	−0.39	0.53	0.32	2.59	19	−0.93	−1.68	−0.08	5	18	−1.25	59	72
ProQ	0.36	0.26	0.13	1.22	5	−0.32	−0.99	−0.22	0	6	1.51	9	32
ProQ_SSE	0.54	0.43	0.19	1.71	8	−0.51	−1.21	−0.16	2	11	1.76	14	42
FRST	−0.57	0.53	0.30	2.36	21	−0.91	−1.74	−0.09	6	22	−2.41	56	72
QMEAN3	−0.65	0.58	0.33	2.57	16	−0.80	−1.83	−0.12	1	35	−2.27	59	75
QMEAN4	−0.71	0.63	0.38	2.76	28	−1.02	−1.90	−0.08	5	39	−1.86	55	69
QMEAN5	−0.72	0.65	0.40	2.90	30	−1.05	−1.94	−0.08	6	40	−1.89	56	71
Torsion single	−0.44	0.39	0.22	1.76	6	−0.60	−1.50	−0.13	0	13	−2.09	51	67
Torsion three-residue	−0.53	0.44	0.22	1.86	13	−0.76	−1.51	−0.11	1	10	−2.64	59	79
Pairwise Cβ	−0.58	0.51	0.30	2.51	17	−0.70	−1.70	−0.18	4	27	−1.96	39	69
Pairwise Cβ/SSE	−0.59	0.52	0.34	2.58	22	−0.84	−1.80	−0.13	5	36	−2.16	45	71
Solvation	−0.55	0.49	0.29	2.31	10	−0.55	−1.65	−0.24	2	27	−1.30	18	45
SSE PSIPRED	−0.65	0.52	0.24	2.03	9	−0.63	−1.43	−0.13	3	17	−0.89	7	25
ACCpro	−0.59	0.56	0.35	2.75	21	−0.85	−1.66	−0.11	6	33	−1.38	20	44

[a]Pearson's correlation coefficient r^2 and Spearmans's rank correlation coefficient ρ.
[b]*FE* stands for fraction enrichment and $E_{15\%}$ is the enrichment among the top 15% best predicted models as compared to a random selection.
[c]Rank10 are the number of targets for which the top-scoring models is among the top10 best models (based on GDT_TS). $\log P_{B1}$ and $\log P_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the 10 best-scoring models, respectively.
[d]GDT_TS loss is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set. Z_{nat} is the *Z*-score of the native structure as compared to the ensemble of models. Rank1 *and* rank10 are the number of targets in which the native structure (or the best model based on GDT_TS, excluding the native structure) was found on the first rank or among the top 10 predictions.

slightly better than QMEAN4 and QMEAN5 as a consequence of the inability of the secondary structure agreement term in recognizing the native structure which is reflected by the low *Z*-scores of the native structure and the rank measures (rank1 and rank10). RAPDF and, to a greater extend, FRST show a good performance over all quality measures but are especially apt in the task of recognizing the native structure as reflected by the good average *Z*-score of the native structures (Z_{nat}). DFIRE, together with QMEAN3 and the 3-residue torsion angle potential, identify to highest number of native structures whereas DFIRE has significantly worse Z_{nat} scores compared with all other methods (see Fig. 2). FRST produces better Z_{nat} scores than QMEAN3 but never better than the torsion angle potential over three residues which show an extraordinary good performance in recognizing the native structure. For the model quality assessment task described by the other quality measures, the three-residue torsion angle potential does mostly better than the ordinary single residue potential. Modcheck generates statistically significant better regression coefficients than the other methods except the three QMEAN functions. DFIRE and RAPDF show no significantly better or worse performance over most of the quality measures. DFIRE shows a worse Pearson's regression coefficient than RAPDF but is better based on Spearman's rank correlation coefficient. Consistently, over all quality measures (except for the Pearson's correlation coefficient), ProQ performs significantly worse than the other methods tested even after the integration of PSIPRED secondary structure prediction. The only exception is the good average Z_{nat} scores achieved on the free modeling targets which reflects the fact that ProQ has been trained specifically on fold recognition models (see Supplementary Material Table SIV). Furthermore, the comparably limited performance of ProQ can be tentatively ascribed to the original ProQ scoring function including several terms not directly accessible from the structure of the model alone (e.g. fraction of the protein modeled, information from the template). The secondary structure agreement term shows on average the highest Pearson correlation coefficient of all single terms and a reasonable performance on all the other model quality assessment measures. Compared with the other terms, SSE PSIPRED fails to recognize the native structure which is the reason why QMEAN3 is better than QMEAN4 and 5 in this specific task (see Discussion). The solvent accessibility agreement term on the other hand reaches the highest enrichment values and rank correlation coefficients and is very valuable for the selection of good models. In contrast to the secondary structure agreement term, the ACCpro score can help to identify the native structure in the case of free modeling targets where it recognizes 7 out of 18 native structures with an average *Z*-score of the native structure of more than 2 (see Supplementary Material). Over all quality measures and in both categories the secondary structure specific pairwise potential reaches significantly better scores than the regular one for the

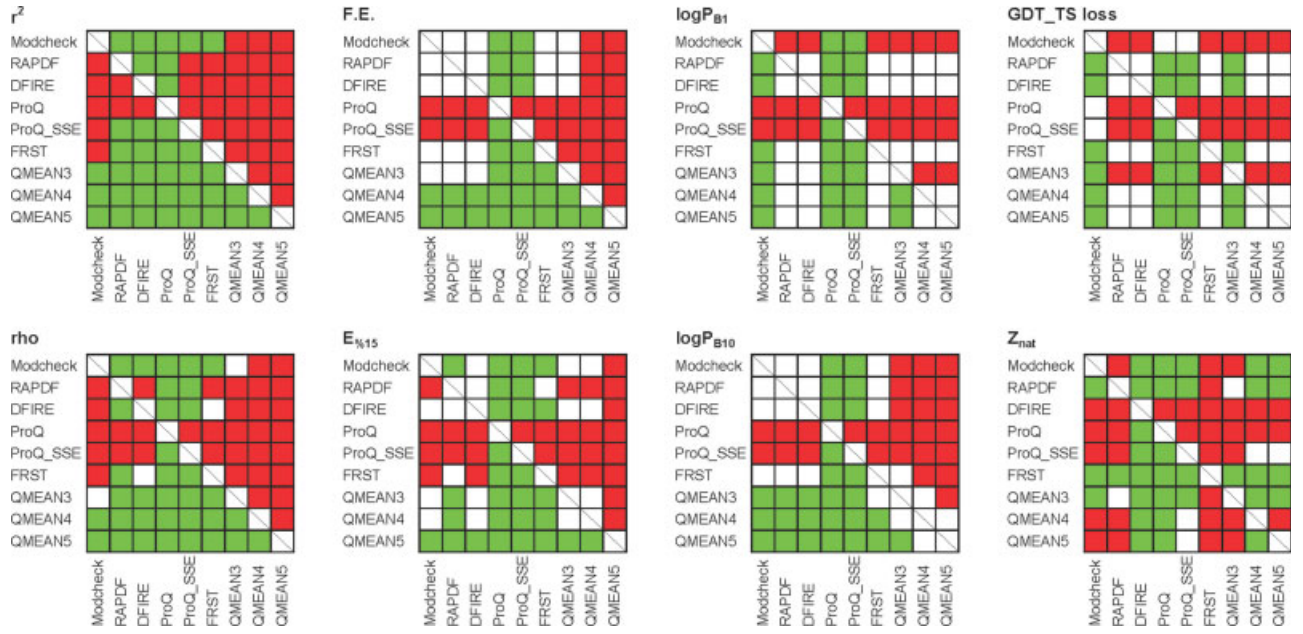

Figure 2

Statistical analysis of the performance differences between the methods at the confidence level of 95%. Green (red) squares indicate a significantly a better (worse) performance, white squares indicate no statistically significant difference in the performance of two methods.

model quality assessment task as well as in the identification of the native structure (see Supplementary Material Figure S2).

The differences in the results achieved for the free modeling and template-based modeling targets are frequently easy to explain but sometimes appear to be contra-intuitive. Figure 3 shows four selected targets belonging to the TBM and FM target category. The scatter plots on the left-hand side [Fig. 3(a,c)] represent two examples in which both the regression and the identification of the native structure went fine. As expected, the regression coefficients for TBM targets are on average higher than for FM targets. A slightly better enrichment is possible with FM targets, since the models in this category tend to be less similar to each other than, for example, in the high-accuracy TBM category in which a large fraction of the models can be more or less identical. Of the free modeling targets, the pairwise and solvation potentials as well as ACCpro all produce high enrichment values whereas on the TBM targets the performance of the solvation potential is significantly worse compared to the others over most quality measures. For the FM targets, the native structures are recognized with better *Z*-scores on average but, surprisingly, the relative number of native structures ranked as number one is lower (9 out of 18) as compared with the TBM targets (51 out of 77) (see Supplementary Material). Sometimes the native structure can be easily identified [target T0321, Fig. 3(c)] but sometimes the native structure is hidden among the bulk of the models [target T0300, Fig. 3(d)] even though the regression can be reasonably good. For most of the FM targets, no submitted model had a GDT_TS score of more than 50 and one should expect the native structure to be easy to identify. The enrichment for FM targets works rather well with enrichment values ($E_{15\%}$) on the order of factor 3 achieved on average.

Estimating overall performance

Fraction enrichment curves[67] are useful to compare and visualize the performance of different MQAPs in analogy to receiver operator characteristic (ROC) curves frequently used in benchmarks of fold recognition and alignment programs. They implicitly cover several quality measures used in Table VI, for example, enrichment and regression. Where ROC curves require the somewhat arbitrary definition of a threshold to distinguish good from bad models, fraction enrichment curves measure the added value of MQAPs in ranking different models. Figure 4 shows the fraction of best models (based on GDT_TS) found among a certain fraction of the top scoring models as predicted by the scoring function (FE). The calculations are performed on the server models of CASP7 after removing the native structure. The curves in Figure 4(a) reflect the ability of the scoring function to identify the best models among all models for a given target and are a measure for the scoring function's ability to predict the relative model quality. The steeper the pro-

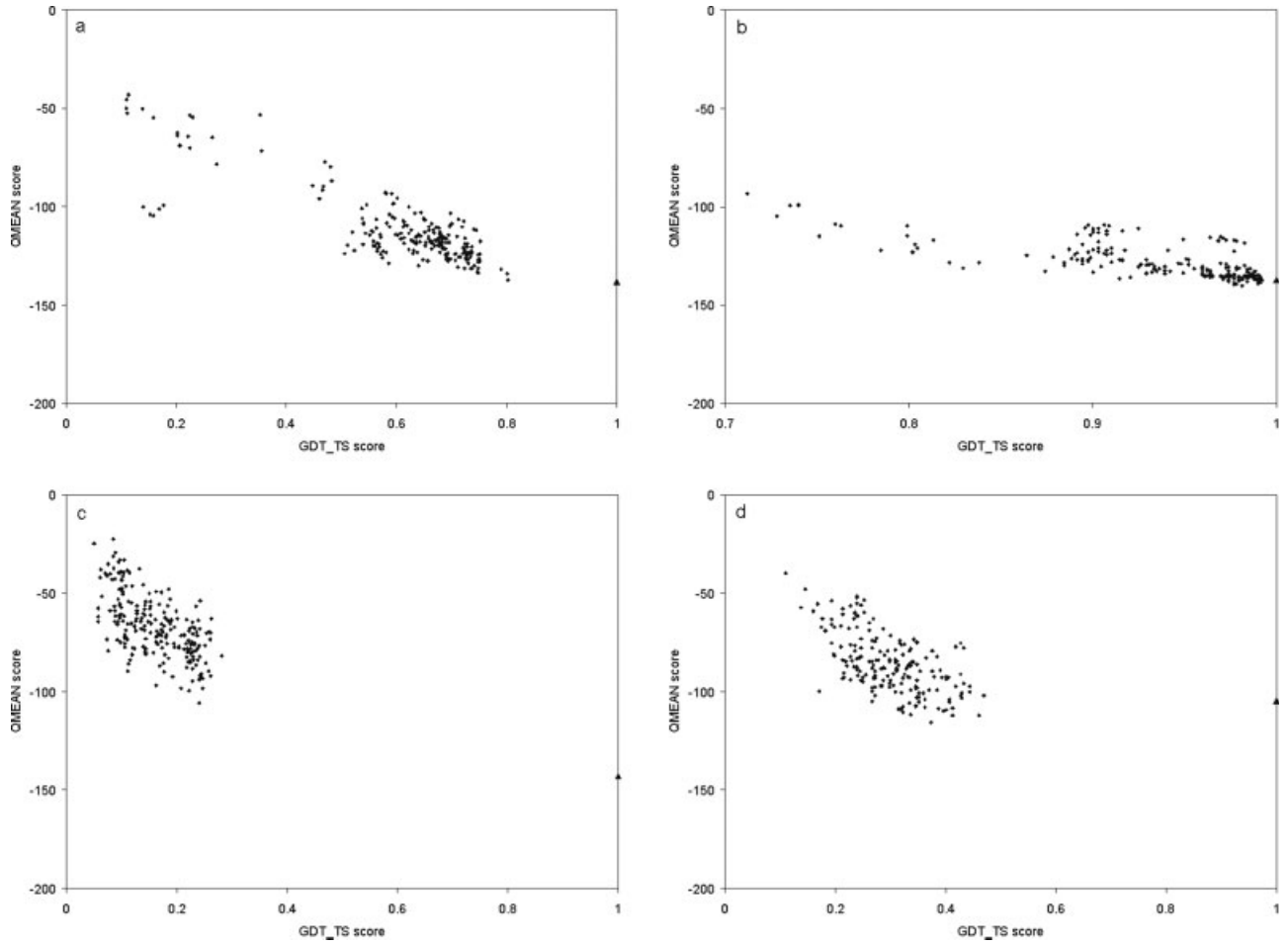

Figure 3

*Scatter plots showing the correlations between GDT_TS and QMEAN5 for four selected examples of the categories template-based modeling (target T0324, **a**), high accuracy template-based modeling (target T0290, **b**, mind the different scaling) and free modeling (targets T0321 and T0300, **c**,**d**). The native structure is represented as triangle at GDT_TS = 1.*

gression of the curve, and the larger the area under the curve, the better a scoring function agrees with the measured model quality. The average FE over the individual targets for cutoffs ranging from 5 to 50% is shown. QMEAN consistently shows the best performance over the whole range but especially between 5 and 15%, underlining its strength in recognizing the best models. Modcheck, RAPDF, DFIRE, and FRST show a quite similar behavior over the first three thresholds. Above 20%, the curve obtained for Modcheck and DFIRE are slightly higher which agrees with its good rank correlation coefficients and enrichment values in Table VI. ProQ performs significantly worse than the others. The global FE curves shown in Figure 4(b) are obtained by pooling together the models of all targets and calculating the FE over the whole set. In this way, the scoring function's ability to predict the absolute model quality (i.e. to estimate the degree of "nativeness" of a model) is investigated. In contrast to the results in Figure 4(a), the performance of RAPDF and especially DFIRE are strikingly low compared with Modcheck and FRST. FRST shows the best FE within the first 5% and appears to be good in recognizing native and native-like structures. This is also reflected by the low-average Z-scores of the native structure (Z_{nat}) shown in Table VI. In the global enrichment, ProQ shows a reasonable performance which can be mainly attributed to the secondary structure information included as the difference between ProQ and ProQ PSIPRED suggests. Above a fraction of 0.1, QMEAN consistently generates the highest FEs of all MQAPs tested. For example, among the 15% best QMEAN predictions more than 60% of the 15% best models are identified. The high enrichments are an evidence of a good global correlation between the QMEAN score and the effective model quality. The correlation between GDT_TS and QMEAN on the CASP7 test set is shown in Figure 1(b).

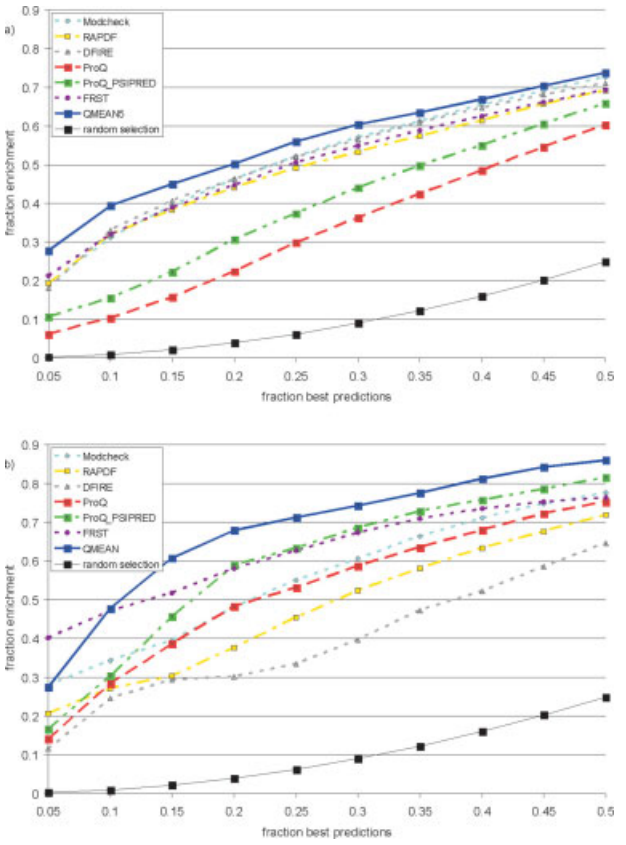

Figure 4

*Fraction enrichment curves describing the fraction on the x% best GDT_TS models observed among the x% best predictions. The ability of the scoring function to predict the relative model quality for models of the same target is assessed here by averaging the fraction enrichments obtained on the individual targets (**a**). The global enrichments are calculated by pooling together all models of all targets and reflect the ability to predict the absolute quality of models (**b**).*

DISCUSSION

General performance

The QMEAN scoring function has been shown to be a valuable tool for model quality assessment by distinguishing good from bad models and for the identification of the native structure among decoys generated by a variety of methods. On the comprehensive set of 22,420 server models of CASP7, QMEAN consistently outperforms the five model quality assessment programs over nearly all quality measures and model difficulty ranges.

Secondary structure and solvent accessibility agreement terms

Only in two decoy sets, lattice_ssfit and the molecular dynamics simulation set, did the integration of the secondary structure agreement term result in an improved ability of the combined scoring function in identifying the native structure compared with the statistical potential terms only (QMEAN3). This can be possibly attributed to the greater overall variability of the decoy structures in these sets. On the other two standard decoy sets, as well as on the CASP7 models, QMEAN3 performed better than QMEAN4 and 5. In contrast to this observation, the secondary structure agreement term turned out to be a valuable contributor to the good performance of QMEAN in the model quality assessment task. The different performance on these two tasks can, especially in the case of the CASP7 set, tentatively be ascribed to the fact that the secondary structure composition of the native structure can only be predicted with certain accuracy, typically around 76–80%. Consensus approaches integrating information of several secondary structure predictors are able to raise this threshold slightly.[60] A theoretical limit of prediction accuracy of 88% percent was proposed by Rost[70] arguing that minor variations in structures even between homologous proteins can result in different secondary structure assignments made by tools such as DSSP. It is therefore rather unlikely that the secondary structure agreement between PSIPRED and DSSP achieves 100 percent for the native structure and more likely that there is a tendency for models generated by methods taking implicitly advantage of predicted secondary structure information to receive better scores than the native structure. In analogy to the secondary structure agreement, a term describing the agreement between predicted and calculated solvent accessibility was implemented. The same argument given above holds for this term, although the effect seems to be less pronounced as reflected by the higher Z-scores of the native structure (Z_{nat}) achieved in the CASP7 decoy set. This might be explained by the significantly reduced sensitivity of this term toward minor differences in the structures, since it is based on a binary classification of solvent accessibility (buried/exposed) as provided by ACCpro. Thus, near-native structures would tend to have solvent accessibility agreement values (e.g. packing) similar to the native structure but bad models do not, which would explain the moderate Z_{nat} scores to some extent.

Torsion angle potential over three residues

The torsion angle potential over three residues turned out to be a very powerful term for the identification of the native structures out of a variety of decoy sets, suggesting that the three-residue torsion angle potential describes the propensity of a certain amino acid for a certain local geometry considerably better than the single residue torsion angle potential. The final bin sizes of 45° for Φ and Ψ for the center residue and 90° for the neighboring torsion angles are surprisingly coarse-grained, but can possibly be explained by reasonable binning of the Ramachandran plot[71] in 90° and 45° and how these values represent a trade-off between resolution

and number of states, reducing the danger of over-fitting. The resulting number of 327,680 ($= 20 \times (360/45)^2 \times (360/90)^2 \times (360/90)^2$) possible states is in the same order of magnitude as observed in some all-atom potentials. Betancourt and Skolnick[22] have shown that the dihedral angles of a residue are influenced by the identity and conformation of the adjacent residues. This effect is especially pronounced in loop regions and near the end of β-sheets. The three-residue torsion angle potential seems to capture this effect to a certain extend. In contrast to the potential introduced by Betancourt and Skolnick, the three-residue potential described in this work does not take into account the identity of the adjacent residues and is attractive in its simplicity. It basically reflects the propensity of a certain residue for a given local geometry (as described by six torsion angles) as compared with other residues.

Secondary structure specific pairwise potential

The secondary structure-specific implementation has shown to lead to a statistically significant improvement of the performance over all quality measures compared with the regular residue-level pairwise potential. Loops are primarily located at the protein surface and are to a larger extent influenced by nonlocal interactions in contrast to helices and sheets which are mainly determined by the local potential.[22] As loops have fewer contacts to the rest of the protein than helices and sheets, which are at least partially surrounded by more residues, it can be speculated that pairwise statistical potentials tend to be biased towards interaction patterns observed in the protein core. As a consequence, some motifs observed only in loop regions receive a slightly too high energy. A specialized potential compiled and applied in a secondary-specific manner may counteract this.

Training and evaluation Process

To reduce possible over-fitting of any of the potentials, all structures with detectable homology (based on a BLAST[72] search) to any of the structures of the two CASP decoy sets were removed from the protein data set used to build the potentials. In this way, several 100% sequence identity hits have been removed. Remarkably, comparing the results before and after adjusting the potentials, no considerable change has been observed even for the task of detecting the native fold (data not shown). This can be explained by the rather large number of structures used to compile the potentials, where the influence of one specific (even identical) structure is diminished by the others. In model quality assessment in particular, models with significant errors, not the actual structures, are evaluated, further reducing a possible bias from the presence of homologous structures in the data set.

Parameterizing and optimizing the single terms as well as the combined scoring function on CASP decoys[25] represents a reasonable approach since a variety of methods and the entire range of modeling difficulty is covered. The good performance of QMEAN on all decoy sets and the fact that the targets of two CASP rounds are completely different indicates that QMEAN has not been specifically trained to assess models produced by CASP participants but instead is applicable to the variety of methods. Although the strategy to derive the weighting factors for the composite score based on the regression coefficient represents a reasonable starting point (assuming a correlation between energy and degree of "nativeness"), this approach also has some disadvantages. Some terms showing a medium correlation to GDT_TS can still perform better on other quality measures and their discrimination power tends to be underestimated. A good example is the solvent accessibility agreement term which shows lower correlation to GDT_TS than the secondary structure agreement term (Table II) but performed consistently better in the CASP7 decoy set over a wide range of conditions (Table VI). A possible underestimation is also reflected by the low correlation to the QMEAN5 score as shown in Table III. The fact that some of the other terms show varying discrimination power depending on the modeling difficulty may warrant specialized versions of the scoring function, for example, for FM or TBM targets. In particular, it remains to be seen why decoys for certain free modeling targets have lower energy than the native structure.

Global and target-specific prediction of model quality

QMEAN shows a consistently better enrichment performance based on the FE curves shown in Figure 4 compared to other MQAPs for both the relative prediction of model quality for models of the same target as well as for global quality prediction over all targets. Since MQAPs are routinely used to assess ensemble of models for the same target, the target-averaged FE curves are probably of greater practical interest since they reflect the ability of the scoring function in discriminating good from bad models. On the other hand, the need for scoring functions predicting the absolute quality of a model has only recently been addressed at the CASP7 meeting.

CONCLUSIONS

We have presented a composite scoring function (QMEAN) consisting of three statistical potential terms covering the major aspects of protein stability and two additional terms describing the agreement of predicted and calculated secondary structure and solvent accessibility, respectively. QMEAN has been shown to be a valuable tool for model quality assessment in the process of structure prediction as well as for the task of recognizing

the native structure out of decoy structures. The different scoring function terms and their combination have been tested on several decoy sets containing models covering the whole quality range and a large number of quality measures have been used. Some of the scoring function terms turned out to be more specialized for a specific task or quality measure whereas others are more widely applicable. The results indicate that a combination of multiple terms increases the performance of the scoring function by taking advantage of the strengths of certain terms for a specific task while reducing a possibly negative contribution of other terms. The statistically significant improvement in performance of QMEAN over other well-established methods gets even more pronounced when taking into account that a rather simple approach was used to combine the different terms to the final scoring function. QMEAN represents a further step towards the prediction of the absolute quality of protein models.

ACKNOWLEDGMENTS

The authors are grateful to Dariusz Przybylski for help with the PROFsec software and to Yaoqi Zhou for providing the DFIRE software.

REFERENCES

1. Chance MR, Fiser A, Sali A, Pieper U, Eswar N, Xu G, Fajardo JE, Radhakannan T, Marinkovic N. High-throughput computational and experimental techniques in structural genomics. Genome Res 2004;14:2145–2154.
2. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round 6. Proteins 2005;61 (Suppl 7):3–7.
3. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
4. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
5. Contreras-Moreira B, Fitzjohn PW, Bates PA. In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. J Mol Biol 2003;328:593–608.
6. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
7. Sommer I, Toppo S, Sander O, Lengauer T, Tosatto SCE. Improving the quality of protein structure models by selecting from alignment alternatives. BMC Bioinformatics 2006;7:364.
8. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
9. Brooks B, RE B, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
10. Fogolari F, Brigo A, Molinari H. Protocol for MM/PBSA molecular dynamics simulations of proteins. Biophys J 2003;85:159–166.
11. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
12. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
13. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44:223–232.
14. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
15. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
16. Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
17. 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.
18. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput Aided Mol Des 1993;7:473–501.
19. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins 2000;41:40–46.
20. 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.
21. Albiero A, Tosatto SCE. Fine-grained statistical torsion angle potentials are effective in discriminating native protein structures. Curr Drug Discov Technol 2006;3:75–81.
22. Betancourt MR, Skolnick J. Local propensities and statistical potentials of backbone dihedral angles in proteins. J Mol Biol 2004; 342: 635–649.
23. Kocher JP, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
24. Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. Protein Sci 2002;11:18–26.
25. Tosatto SCE. The victor/FRST function for model quality estimation. J Comput Biol 2005;12:1316–1327.
26. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. J Mol Biol 1992;225:93–105.
27. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
28. Finkelstein AV, Badretdinov AY, Gutin AM. Why do protein architectures have Boltzmann-like statistics? Proteins 1995;23:142–150.
29. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
30. Rooman MJ, Wodak SJ. Are database-derived potentials valid for scoring both forward and inverted protein folding? Protein Eng 1995;8:849–858.
31. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
32. Shortle D. Propensities, probabilities, and the Boltzmann hypothesis. Protein Sci 2003;12:1298–1302.
33. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
34. Miyazawa S, Jernigan RL. Identifying sequence-structure pairs undetected by sequence alignments. Protein Eng 2000;13:459–475.
35. Reva BA, Finkelstein AV, Sanner MF, Olson AJ. Residue-residue mean-force potentials for protein structure recognition. Protein Eng 1997;10:865–876.
36. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 1992;13:258–271.
37. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
38. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
39. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53:76–87.

40. Eramian D, Shen M-y, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653–1666.
41. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
42. Gilis D, Rooman M. Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. J Mol Biol 1996;257:1112–1126.
43. Gilis D, Rooman M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J Mol Biol 1997;272:276–290.
44. Hoppe C, Schomburg D. Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. Protein Sci 2005;14:2682–2692.
45. Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. Proteins 2007;66:41–52.
46. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
47. Fischer D. Servers for protein structure prediction. Curr Opin Struct Biol 2006;6:178–182.
48. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
49. Wang K, Fain B, Levitt M, Samudrala R. Improved protein structure selection using decoy-dependent discriminatory functions. BMC Struct Biol 2004;4:8.
50. Fogolari F, Tosatto SCE. Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. Protein Sci 2005;14:889–901.
51. Pettitt CS, McGuffin LJ, Jones DT. Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 2005;21:3509–3515.
52. 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.
53. Wang G, Dunbrack RLJ. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
54. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
55. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
56. Eyrich VA, Martí-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
57. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
58. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33(Web Server issue):W72–W76.
59. Rost B. Prediction in 1D: secondary structure, membrane helices, and accessibility. Methods Biochem Anal 2003;44:559–587.
60. Albrecht M, Tosatto SCE, Lengauer TV, Giorgio. Simple consensus procedures are effective and sufficient in secondary structure prediction 2003;16:459–462.
61. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 2003;31:3370–3374.
62. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
63. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000; 9:1399–1401.
64. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
65. Keasar C, Levitt M. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 2003;329:159–174.
66. Fogolari F, Tosatto SCE, Colombo G. A decoy set for the thermostable subdomain from chicken villin headpiece, comparison of different free energy estimators. BMC Bioinformatics 2005;6:301.
67. Tosatto SC, Battistutta R. TAP score: torsion angle propensity normalization applied to local protein structure evaluation. BMC Bioinformatics 2007;8:155.
68. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure 2002;10:435–440.
69. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 2004;13:400–411.
70. Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
71. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–99.
72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.