QAcon: Single model quality assessment using protein structural and contact information with machine learning techniques

Renzhi Cao[1], Badri Adhikari[2], Debswapna Bhattacharya[3], Miao Sun[4], Jie Hou[2], Jianlin Cheng[2,5*]

[1]Department of Computer Science, Pacific Lutheran University, WA 98447, USA

[2]Department of Computer Science, University of Missouri, Columbia, Missouri 65211, USA.

[3]Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS 67260-0083, USA.

[4]Department of Electrical and Computer Engineering, University of Missouri, Columbia, Missouri 65211, USA.

[5]Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA.

Table S1. All score features used for our single model QA method.

Score name	Description
1. The RF_CB_SRS_OD score(Rykunov and Fiser, 2007)	A novel distance dependent residue-level potential energy score. The original RF_CB_SRS_OD score is normalized to the range of 0 and 1.
2. SS score	This score is calculated by the difference between secondary structure predicted by Spine X (Faraggi, et al., 2012) from the protein sequence and those of a model parsed by DSSP (Kabsch and Sander, 1983).
3. SP score	This score is calculated by the percentage of helix and sheet matching between secondary structure predicted and the on parsed from the model.
4. EC score	The Euclidian compact score is calculated by summation of pairwise Euclidean distance between amino acids divided by (N*N-1)*3.8, N is the total number of amino acids in the sequence.
5. SU score	This surface score is calculated by the total area of exposed nonpolar residues divided by the total area of all residues.

6. EM score	The exposed mass score is calculated as the total mass of nonpolar residues area divided by the total mass of exposed residue area.
7. ES score	The exposed surface score is calculated as the total exposed residue area divided by the total residue area.
8. SA score	The solvent accessibility score is calculated by the percentage of difference between the predicted solvent accessibility and the one parsed from the model.
9. RWplus score(Zhang and Zhang, 2010)	The RWplus score is from the energy score RWplus using pairwise distance-dependent atomic statistical potential function and side-chain orientation-dependent energy term, and normalized to the range of 0 and 1.
10. ModelEvaluator score(Wang, et al., 2009)	The ModelEvaluator score is predicted by the tool ModelEvaluator which is based on structural features using support vector machine.
11. Dope score(Shen and Sali, 2006)	A new statistical potential discrete optimized protein energy score. The Dope score is predicted and normalized to the range of 0 and 1.
12. Con score	The contact score is calculated by the satisfaction of contact predicted from the sequence and the one parsed from the model. PSI-COV(Jones, et al., 2012) is used for contact prediction, and the DNcon(Eickholt and Cheng, 2012) is used when PSI-COV fails to make predictions.

Table S2. The per-target average correlation, average loss for QAcon and several state-of-art (including quasi-single and clustering QA methods) on Stage1 and Stage2 of CASP11.

Server Name	Corr. Stage1	Loss. Stage 1	Corr. Stage 2	Loss. Stage 2
ProQ2	0.643	0.090	0.372	0.058
Qprob	0.631	0.097	0.382	0.068
ModFOLDclust2	0.737	0.047	0.560	0.069
ModFOLD5	0.748	0.033	0.501	0.079
DAVIS_consensus	0.798	0.052	0.570	0.073
MULTICOM-CONSTRUCT	0.670	0.073	0.536	0.050
QAcon	0.639	0.100	0.395	0.067

VoroMQA	0.561	0.108	0.401	0.069
Wang_SVM	0.655	0.109	0.362	0.085
Wang_deep_1	0.613	0.128	0.302	0.089

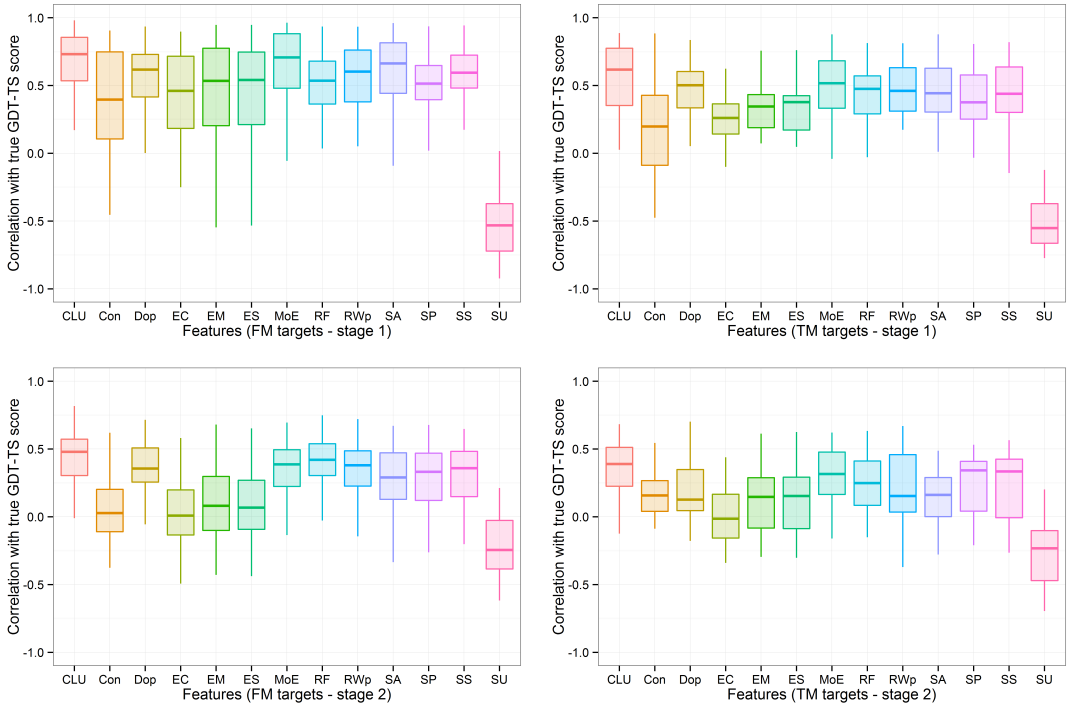

Figure S1. The performance of each feature on stage1 (plots in top row) and stage2 (plots in bottom row) of CASP11 free-modeling (plots in left column) and template-based modeling (plots in right-column) targets using the correlation between the feature values and the true GDT-TS of the models. The abbreviations CLU, RF, SS, SP, EC, SU, EM, ES, SA, RWp, MoE, Dop, and Con stand for RF_CB_SRS_OD score, secondary structure related score, helix and sheet related score, Euclidean compact score, surface score, exposed mass score, exposed surface score, solvent accessibility score, RWplus score, ModelEvaluator score, Dope score, and Contact score respectively. See Table S1 for the description of these features.

References:

Eickholt, J. and Cheng, J. (2012) Predicting protein residue–residue contacts using deep networks and boosting, *Bioinformatics*, **28**, 3066-3072.

Faraggi, E.*, et al.* (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles, *Journal of computational chemistry*, **33**, 259-267.
Jones, D.T.*, et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments, *Bioinformatics*, **28**, 184-190.
Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen‐bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.
Rykunov, D. and Fiser, A. (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance‐dependent statistical pair potentials, *Proteins: Structure, Function, and Bioinformatics*, **67**, 559-568.
Shen, M.y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures, *Protein Science*, **15**, 2507-2524.
Wang, Z., Tegge, A.N. and Cheng, J. (2009) Evaluating the absolute quality of a single protein model using structural features and support vector machines, *Proteins*, **75**, 638-647.
Zhang, J. and Zhang, Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction, *PloS one*, **5**, e15386.