OPUS-CSF: A C-atom-based Scoring Function for Ranking Protein Structural Models

Gang Xu[1,6], Tianqi Ma[2,3,6], Tianwu Zang[2,3], Qinghua Wang[4] & Jianpeng Ma[1,2,3,4,5,*]

[1]*School of Life Sciences*
Tsinghua University
Beijing, 100084, China

[2]*Applied Physics Program*
[3]*Department of Bioengineering*
Rice University
Houston, TX77005

[4]*Verna and Marrs Mclean Department of Biochemistry and Molecular Biology*
Baylor College of Medicine
One Baylor Plaza, BCM-125
Houston, TX77030

June 22, 2017

Keywords: Protein Structure Modeling, Protein Folding, Coarse-graining, Scoring Function, Decoy Recognition.

[5]Lead Contact.

[6]These authors contributed equally.

[*]Correspondence: jpma@bcm.edu.

Summary

We report a C-atom-based scoring function, named OPUS-CSF, for ranking protein structural models. Rather than using traditional Boltzmann formula, we built a scoring function (CSF score) based on the native distributions (analyzed through entire PDB) of coordinate components of mainchain C atoms on selected residues of peptide segments of 5, 7, 9, and 11 residues in length. In testing OPUS-CSF on decoy recognition, it maximally recognized 257 native structures out of 278 targets in 11 commonly used decoy sets, significantly more than other popular all-atom empirical potentials. The average correlation coefficient with TM-score was also comparable with those of other potentials. OPUS-CSF is a highly coarse-grained scoring function, which only requires input of partial mainchain information, and very fast. Thus it is suitable for applications at early stage of structural building.

Introduction

A potential function plays a central role in predicting protein structures. Generally, there are two kinds of potential functions: physics-based potentials and knowledge-based potentials. Physics-based potentials typically are the all-atom molecular mechanics force-fields (Arnautova et al., 2006; Brooks et al., 1983; Case et al., 2005; MacKerell Jr et al., 1998; Weiner et al., 1986), such as CHARMM (Brooks et al., 1983; MacKerell Jr et al., 1998) and AMBER (Case et al., 2005). They also include coarse-grained potentials such as MARTINI (Marrink et al., 2007), UNRES (Liwo et al., 1997a; Liwo et al., 1997b) and OPEP (Chebaro et al., 2012).

The knowledge-based potentials are derived from statistical analysis of known structures and are widely used in structural prediction (Buchete et al., 2004; DeBolt and Skolnick, 1996; Gilis et al., 2006; Gohlke and Klebe, 2001; Hendlich et al., 1990; Hoppe and Schomburg, 2005; Jernigan and Bahar, 1996; Jones et al., 1992; Koliński and Bujnicki, 2005; Lazaridis and Karplus, 2000; Lu and Skolnick, 2001; Lu et al., 2008; Ma, 2009; Miyazawa and Jernigan, 1985; Moult, 1997; Poole and Ranganathan, 2006; Russ and Ranganathan, 2002; Samudrala and Moult, 1998; Shen and Sali, 2006; Sippl, 1990, 1995; Skolnick, 2006; Skolnick et al., 2000; Tobi and Elber, 2000; Wu et al., 2007; Yang and Zhou, 2008; Zhang et al., 1997; Zhang and Zhang, 2010; Zhang et al., 2003; Zhou and Skolnick, 2011; Zhou and Zhou, 2002; Zhou et al., 2006). In general, knowledge-based potentials can be either constructed at coarse-grained residue level (Buchete et al., 2004; Gilis et al., 2006; Hendlich et al., 1990; Hoppe and Schomburg, 2005; Jones et al., 1992; Koliński and Bujnicki, 2005; Miyazawa and Jernigan, 1985; Sippl, 1990; Skolnick et al., 2000; Tobi and Elber, 2000; Wu et al., 2007; Zhang et al., 2003) or at atomic level (DeBolt and Skolnick, 1996; Lu and

Skolnick, 2001; Lu et al., 2008; Samudrala and Moult, 1998; Shen and Sali, 2006; Yang and Zhou, 2008; Zhang et al., 1997; Zhang and Zhang, 2010; Zhou and Skolnick, 2011; Zhou and Zhou, 2002). Although coarse-grained potentials may not be rigorous, it helps to focus on essential features and excludes less important details, thus reduces computational cost (Kmiecik et al., 2016; Noid, 2013). The performance of coarse-grained potential is highly related to how one designs the coarse-graining scheme. For example, OPUS-Ca potential (Wu et al., 2007) uses the positions of Ca atoms as input, calculate other atomic positions as pseudo-positions and significantly reduces the computing cost. Other applications of coarse-grained models using Ca positions are also reported in literature (Wu et al., 2005a; Wu et al., 2005b).

In this work, unlike traditional empirical potential functions using Boltzmann formula, we built a scoring function based on the native distributions of coordinate components of mainchain C atoms on a few selected residues of small peptide segments of 5, 7, 9 and 11 residues in length. A lookup table was first generated for native distributions of coordinate components by analyzing peptide segments in the entire Protein Data Bank (PDB). Then the scoring function was calculated for a particular test structure by comparing the information of its segments with the lookup table. The performance of OPUS-CSF was tested on 11 commonly used decoy sets, the results indicated that OPUS-CSF recognized significantly more native structures from their decoys than other empirical potentials. In terms of the correlation coefficients between CSF scores and TM-scores, they were comparable with those of the popular all-atom empirical potentials. Most importantly, OPUS-CSF achieved such performance despite its highly coarse-grained nature. That indicates the advantages of

OPUS-CSF in terms of its speed and its application in the early stage of structural modeling. This is vitally important for applications such as building structural models against intermediate resolution data from experimental techniques like cryogenic electro-microscopy (cryo-EM).

Results and Discussion

We compared the performance of OPUS-CSF on 11 commonly used decoy sets with that of popular all-atom potential functions. In Table 1, we listed the results of 5-residue segment case (OPUS-CSF5) and all-segment combined case (OPUS-CSF). For 5-residue segment case, OPUS-CSF5 successfully recognized 244 out of 278 native structures from their decoys and had the average Z-score (-3.56) comparable with that of GOAP (-3.57). For combined segment case, OPUS-CSF successfully recognized 257 out of 278 native structures from their decoys and had the average Z-score (-4.12) better than that of GOAP (-3.57). It is interesting that although OPUS-CSF is a highly coarse-grained scoring function, its performance is significantly better than other all-atom potentials.

We also calculated the Pearson's correlation coefficients between CSF score and TM-score (Zhang and Skolnick, 2004) in all decoy sets. The results are shown in Table 2. OPUS-CSF has comparable average correlation coefficient with those of GOAP and OPUS-PSP despite the fact that OPUS-CSF is highly coarse-grained and the other two are all-atom potentials.

Figure.1 shows the histogram of standard deviations of the coordinate components of C atoms of the 1st and 5th residues in the CND lookup table for 5-residue segment case. It is

clear that the distribution peaks at a very small value indicating that the coordinate components are clustered in a narrow distribution, i.e., the configurational distributions of the 5-residue peptide segments are narrow (Tang and Zhang, 2007), which provides a foundation for the success of OPUS-CSF. The narrow configurational distribution of small peptide fragments is also seen in other studies (Simons et al., 1997). In addition, the average value of the standard deviation is 1.20 Å, in a similar order of magnitude of a single chemical bond length.

It needs to be mentioned that, in the implementation of OPUS-CSF, we assume that the smaller the CSF score, the more likely the structure to be native. This is an approximation because even a native structure may not have a zero CSF score. However, the narrow distributions of standard deviations of the coordinate components of C atoms seem to be in favor of such an approximation. Figure 2 shows the distribution of frequency of sequence repeating in the CND lookup table. Half of the sequences repeat more than 26 times in the distribution. The largest value of X-axis is 29,618 with one sequence.

We examined OPUS-CSF with different length of residue segment. With the length of segment increases, the ratio of the number of segments that appear more than 5 times to the total number of segments in PDB decreases (Table 3). On the other hand, if the Coverage is defined as the ratio between the number of segments available in CND lookup table and the number of total segments of a test sequence, the average coverage of the 11 decoy sets (totally 278 targets) decreases as the length of segment increases. If a test sequence has less than 20% of its segments available in the CND lookup table, i.e., its coverage is less than 20%, it is regarded as Unknown, then the number of unknowns increases as the length of

segment increases. More details of OPUS-CSF on different segment length can be found in Supplemental Information.

Therefore, when working alone, the 5-residue case delivers the best performance in terms of decoy recognition (244 out 278 native recognition in Table 4). However, the Z-scores are better for longer-segment cases. This is probably because the longer segment preserves more sequence homology information.

The construction of CND lookup table ignores the secondary structural elements. We believe this to be an advantage as the prediction of secondary structural elements itself introduces additional uncertainty.

The advantages of OPUS-CSF are obvious. First, the CND lookup table is constructed from the entire PDB, and it contains the information of all allowed configurational information of the native segments (at least for the ones repeated more than 5 times in PDB). Second, the speed of OPUS-CSF is very fast, especially for longer polypeptide chains. This is because the entire chain is scanned once and linearly, it only requires partial mainchain atom coordinates to calculate the CSF score for a structure. Unlike in other potentials such as GOAP (Zhou and Skolnick, 2011) and OPUS-PSP (Lu et al., 2008), no inter-atomic distance needs to be calculated. We want to emphasize that, in modeling protein structures, an empirical potential function or a scoring function, should be fast and accurate. In early stage of modeling, it is advantageous that the scoring function requires minimal amount of structural information. In this regard, OPUS-CSF seems to be a good choice.

Methods

Scanning through the polypeptide chain with a step size of one residue, we collected small peptide segments with sequence length of 5, 7, 9, and 11 residues and searched for their configurations in the entire PDB. Totally, we downloaded 130,054 PDB structures on June 7, 2017 via ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb. The sequences that appeared less than 5 times in PDB were discarded.

Here we use 5-residue segment case as an example to illustrate the detail procedure. The ratio of segments that appear more than 5 times to all segments in PDB is 75.1%, which means we can utilize 75.1% information in the whole PDB using 5-residue segments (also see Table 3 in Results and Discussion).

A local molecular coordinate system was defined for every segment using the positions of three main-chain atoms in the middle residue. The origin was set at the Ca atom, the X-axis was defined along the line connecting Ca and C atoms, Y-axis was in the Ca-C-O plane, parallel to component of C-O vector that was perpendicular to the X-axis, and the Z-axis was defined correspondingly (Figure 3).

For a 5-residue segment with a specific sequence, we saved the C atom coordinates of the 1st and 5th residue in the local coordinate system, denoted as (x_1, y_1, z_1) and (x_5, y_5, z_5). And under our assumption, we treated $x_1, y_1, z_1, x_5, y_5, z_5$ as six independent variables. By scanning through the entire PDB, we generated six independent distributions of these variables, called configurational native distributions (CNDs) of 5-residue segments. We then calculated the means and standard deviations of the distributions and they were kept as the CND lookup table.

For a test structure, we scanned through its sequence with 5-residude-segment. With each

segment and its sequence, we looked for the Z-scores of the six independent variables in the CND lookup table. At the end, we added up all the absolute values of Z-scores of all variables for all segments, and it was called CSF score. We assume the polypeptide structure with smallest CSF score has the largest likelihood to be the native structure.

The segments of varying lengths are denoted as 5(1, 3, 5), 7(2, 4, 6), 9(1, 3, 5, 7, 9) and 11(2, 4, 6, 8, 10). Here, in segments with the form of 5(1, 3, 5), for example, the first number 5 is the segment length, 1,5 in the parenthesis are the residues that we record C atom positional distributions in local coordinate system, 3 is the residue on which the local coordinate system is defined. For 9(1, 3, 5, 7, 9) and 11(2, 4, 6, 8, 10), four atoms are used for recording C atom positional distributions, thus totally 12 independent variables are used.

The CSF score can be calculated either based on one particular segment length or by combining all segment length together. In the case of combined segment length, final CSF score is a linear sum of all CSF scores of different segment length. No weighting function is introduced for the contribution of different segment length.

The 11 commonly used decoy sets we used to test OPUS-CSF are the same as those used in GOAP (Zhou and Skolnick, 2011), including decoy sets of 4state_reduced (Park and Levitt, 1996), fisa (Simons et al., 1997), fisa_casp3 (Simons et al., 1997). hg_structal, ig_structal and ig_structal_hires (R. Samudrala, E. Huang, and M. Levitt, unpublished). I-TASSER (Zhang and Zhang, 2010), lattice_ssfit (Samudrala et al., 1999; Xia et al., 2000), lmds (Keasar and Levitt, 2003), MOULDER (John and Sali, 2003) and ROSETTA (Tsai et al., 2003).

Author Contributions

GX designed the initial concept of OPUS-CSF potential. GX and TM implemented the computational details and thus they made equal contributions to the project. GX and TM also participated in the writing of paper. TZ participated in discussion and contributed to the understanding of the potential function. QW contributed to the understanding of the potential function and participated in the writing of paper. JM supervised the entire project and participated in the writing of paper.

Acknowledgements

JM thanks support from the National Institutes of Health (R01-GM067801, R01-GM116280), and the Welch Foundation (Q-1512). QW thanks support from the National Institutes of Health (R01-AI067839, R01-GM116280), the Gillson-Longenbaugh Foundation, and The Welch Foundation (Q-1826).

Figure Captions

Figure 1. The histogram of standard deviations of the coordinate components in the CND lookup table for 5-residue segment case. The distribution peaks at a very small value of standard deviation indicating that the coordinate components of the 1st and 5th C atoms are clustered in a narrow distribution, i.e., the configurational distributions of the 5-residue peptide segments are narrow. In addition, the average value of the standard deviation is 1.20 Å, in a similar order of magnitude of a single chemical bond length.

Figure 2. The distribution of frequency of sequence repeating in the CND lookup table. The X-axis is the repeating frequency, and the Y-axis is the number of sequences with particular repeating frequency. Sequences that repeat less than 5 times were omitted in our study. Analysis of this distribution indicates that half of the sequences repeat more than 26 times. The largest value of X-axis is 29,618 with one sequence, but not shown for the purpose of clarity.

Figure 3. Local molecular coordinate system in OPUS-CSF defined by the mainchain atoms of the 3rd residues. The origin is on Ca atom. The X-axis is along the Ca-C line. Y-axis is in the plan of Ca-C-O atoms, and parallel to the orthogonal projection of C-O vector. Z-axis is defined accordingly.

Reference

Arnautova, Y.A., Jagielska, A., and Scheraga, H.A. (2006). A new force field (ECEPP-05) for peptides, proteins, and organic molecules. J Phys Chem B *110*, 5025-5044.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem *4*, 187-217.

Buchete, N., Straub, J., and Thirumalai, D. (2004). Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol *14*, 225-232.

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J Comput Chem *26*, 1668-1688.

Chebaro, Y., Pasquali, S., and Derreumaux, P. (2012). The coarse-grained OPEP force field for non-amyloid and amyloid proteins. J Phys Chem B *116*, 8741-8752.

DeBolt, S.E., and Skolnick, J. (1996). Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Eng *9*, 637-655.

Gilis, D., Biot, C., Buisine, E., Dehouck, Y., and Rooman, M. (2006). Development of Novel Statistical Potentials Describing Cation− π Interactions in Proteins and Comparison with Semiempirical and Quantum Chemistry Approaches. J Chem Inf Model *46*, 884-893.

Gohlke, H., and Klebe, G. (2001). Statistical potentials and scoring functions applied to protein–ligand binding. Curr Opin Struct Biol *11*, 231-235.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models: the calculation of low energy conformations from potentials of mean force. J Mol Biol *216*, 167-180.

Hoppe, C., and Schomburg, D. (2005). Prediction of protein thermostability with a direction‐and distance‐dependent knowledge‐based potential. Protein Sci *14*, 2682-2692.

Jernigan, R.L., and Bahar, I. (1996). Structure-derived potentials and protein simulations. Curr Opin Struct Biol *6*, 195-209.

John, B., and Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res *31*, 3982-3992.

Jones, D.T., Taylort, W., and Thornton, J.M. (1992). A new approach to protein fold recognition. Nature *358*, 86-89.

Keasar, C., and Levitt, M. (2003). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol *329*, 159-174.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A.E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. Chem Rev *116*, 7898-7936.

Koliński, A., and Bujnicki, J.M. (2005). Generalized protein structure prediction based on combination of fold‐recognition with de novo folding and evaluation of models. Proteins: Structure, Function, and Bioinformatics *61*, 84-90.

Lazaridis, T., and Karplus, M. (2000). Effective energy functions for protein structure prediction. Curr Opin Struct Biol *10*, 139-145.

Liwo, A., Ołdziej, S., Pincus, M.R., Wawak, R.J., Rackovsky, S., and Scheraga, H.A. (1997a). A united ‐ residue force field for off ‐ lattice protein ‐ structure simulations. I. Functional forms and parameters of long ‐ range side ‐ chain interaction potentials from protein crystal data. J Comput Chem *18*, 849-873.

Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S., Ołdziej, S., and Scheraga, H.A. (1997b). A united ‐ residue force field for off ‐ lattice protein ‐ structure simulations. II. Parameterization of short ‐ range interactions and determination of weights of energy terms by Z ‐ score optimization. J Comput Chem *18*, 874-887.

Lu, H., and Skolnick, J. (2001). A distance ‐ dependent atomic knowledge ‐ based potential for improved protein structure selection. Proteins: Structure, Function, and Bioinformatics *44*, 223-232.

Lu, M., Dousis, A.D., and Ma, J. (2008). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol *376*, 288-301.

Ma, J. (2009). Explicit orientation dependence in empirical potentials and its significance to side-chain modeling. Acc Chem Res *42*, 1087-1096.

MacKerell Jr, A.D., Bashford, D., Bellott, M., Dunbrack Jr, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., and Ha, S. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins†. J Phys Chem B *102*, 3586-3616.

Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., and De Vries, A.H. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. J Phys Chem B *111*, 7812-7824.

Miyazawa, S., and Jernigan, R.L. (1985). Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures - Quasi-Chemical Approximation. Macromolecules *18*, 534-552.

Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol *7*, 194-199.

Noid, W. (2013). Perspective: Coarse-grained models for biomolecular systems. J Chem Phys *139*, 09B201_201.

Park, B., and Levitt, M. (1996). Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol *258*, 367-392.

Poole, A.M., and Ranganathan, R. (2006). Knowledge-based potentials in protein design. Curr Opin Struct Biol *16*, 508-513.

Russ, W.P., and Ranganathan, R. (2002). Knowledge-based potential functions in protein design. Curr Opin Struct Biol *12*, 447-452.

Samudrala, R., and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol *275*, 895-916.

Samudrala, R., Xia, Y., Levitt, M., and Huang, E. (1999). A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Pacific Symposium on Biocomputing 505-516.

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

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol *268*, 209-225.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol *213*, 859-883.
Sippl, M.J. (1995). Knowledge-based potentials for proteins. Curr Opin Struct Biol *5*, 229-235.
Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol *16*, 166-171.
Skolnick, J., Kolinski, A., and Ortiz, A. (2000). Derivation of protein‐specific pair potentials based on weak sequence fragment similarity. Proteins: Structure, Function, and Bioinformatics *38*, 3-16.
Tang, H.-Y., and Zhang, Z.-G. (2007). Using C′ deviation to study structures of central amino acids in peptide fragments. Amino Acids *33*, 689-693.
Tobi, D., and Elber, R. (2000). Distance‐dependent, pair potential for protein folding: Results from linear optimization. Proteins: Structure, Function, and Bioinformatics *41*, 40-46.
Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. Proteins: Structure, Function, and Bioinformatics *53*, 76-87.
Weiner, S.J., Kollman, P.A., Nguyen, D.T., and Case, D.A. (1986). An all atom force field for simulations of proteins and nucleic acids. J Comput Chem *7*, 230-252.
Wu, Y., Chen, M., Lu, M., Wang, Q., and Ma, J. (2005a). Determining protein topology from skeletons of secondary structures. J Mol Biol *350*, 571-586.
Wu, Y., Lu, M., Chen, M., Li, J., and Ma, J. (2007). OPUS‐Ca: A knowledge‐based potential function requiring only Cα positions. Protein Sci *16*, 1449-1463.
Wu, Y., Tian, X., Lu, M., Chen, M., Wang, Q., and Ma, J. (2005b). Folding of small helical proteins assisted by small-angle X-ray scattering profiles. Structure *13*, 1587-1597.
Xia, Y., Huang, E.S., Levitt, M., and Samudrala, R. (2000). Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol *300*, 171-185.
Yang, Y., and Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins: Structure, Function, and Bioinformatics *72*, 793-803.
Zhang, C., Vasmatzis, G., Cornette, J.L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol *267*, 707-726.
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.
Zhang, Y., Kolinski, A., and Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J *85*, 1145-1164.
Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics *57*, 702-710.
Zhou, H., and Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J *101*, 2043-2052.
Zhou, H., and Zhou, Y. (2002). Distance‐scaled, finite ideal‐gas reference state improves structure‐derived potentials of mean force for structure selection and stability prediction. Protein Sci *11*, 2714-2726.

Zhou, Y., Zhou, H., Zhang, C., and Liu, S. (2006). What is a desirable statistical energy functions for proteins and how can it be obtained? Cell Biochem Biophys *46*, 165-174.

Decoy sets	Total # of targets	DFIRE	RWplus	dDFIRE	OPUS-PSP	GOAP	OPUS-CSF5	OPUS-CSF
4state_reduced	7	6(-3.48)	6(3.51)	7(-4.15)	**7(-4.49)**	7(-4.38)	7(-3.38)	7(-3.31)
fisa	4	**3(-4.87)**	3(-4.79)	3(-3.80)	3(-4.24)	3(-3.97)	2(-2.31)	2(-2.55)
fisa_casp3	5	4(-4.80)	4(-5.17)	4(-4.83)	**5(-6.33)**	5(-5.27)	4(-4.38)	4(-6.72)
hg_structal	29	12(-1.97)	12(-1.74)	16(-1.33)	18(1.87)	22(-2.73)	**23(-2.07)**	23(-2.06)
ig_structal	61	0(0.92)	0(1.11)	26(-1.02)	20(0.69)	47(-1.62)	49(-2.03)	**56(-2.14)**
ig_structal_hires	20	0(0.17)	0(0.32)	16(-2.05)	14(-0.77)	18(-2.35)	19(-2.19)	**20(-2.08)**
I-TASSER	56	49(-4.02)	56(-5.77)	48(-5.03)	55(-7.43)	45(-5.36)	55(-5.32)	**56(-6.39)**
lattice_ssfit	8	8(-9.44)	8(-8.85)	8(-10.12)	8(-6.75)	8(-8.38)	8(-9.56)	**8(-11.79)**
lmds	10	7(-0.88)	7(-1.03)	6(-2.44)	8(-5.63)	7(-4.07)	8(-5.47)	**8(-6.80)**
MOULDER	20	19(-2.97)	19(-2.84)	18(-2.74)	19(-4.84)	19(-3.58)	**20(-3.18)**	20(-3.16)
ROSETTA	58	20(-1.82)	20(-1.47)	12(-0.83)	39(-3.00)	45(-3.70)	49(-3.68)	**53(-4.53)**
Total	278	128(-1.94)	135(-2.13)	164(-2.52)	196(-2.86)	226(-3.57)	244(-3.56)	**257(-4.12)**

Table 1. The results of OPUS-CSF5 (5-residue segment) and OPUS-CSF (combined segment length) on 11 decoys sets compared with different potentials. The results of other potentials come from GOAP paper. The numbers of targets, with their native structures successfully recognized by various potentials, are listed in the table. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score, the better. Out of totally 278 targets in 11 decoy sets, OPUS-CSF5 (5-residue segment) recognized 244 and OPUS-CSF (combined segment length) recognizes 257 native structures from their decoys. The bold number in each row indicates the best one among all the potential functions for that particular decoy set (if the numbers of targets are the same, the bold face is on the one with the better Z-score).

Decoy sets	OPUS-PSP	GOAP	OPUS-CSF
4state_reduced	-0.589	**-0.694**	-0.667
fisa	-0.282	-0.347	**-0.552**
fisa_casp3	-0.095	-0.221	**-0.333**
hg_structal	-0.752	**-0.825**	-0.803
ig_structal	-0.779	-0.865	**-0.882**
ig_structal_hires	-0.832	-0.885	**-0.901**
I-TASSER	-0.284	**-0.477**	-0.452
lattice_ssfit	-0.051	-0.058	**-0.151**
lmds	-0.091	-0.146	**-0.342**
MOULDER	-0.802	**-0.886**	-0.863
ROSETTA	-0.343	**-0.476**	-0.391
Average	-0.521	-0.632	-0.624

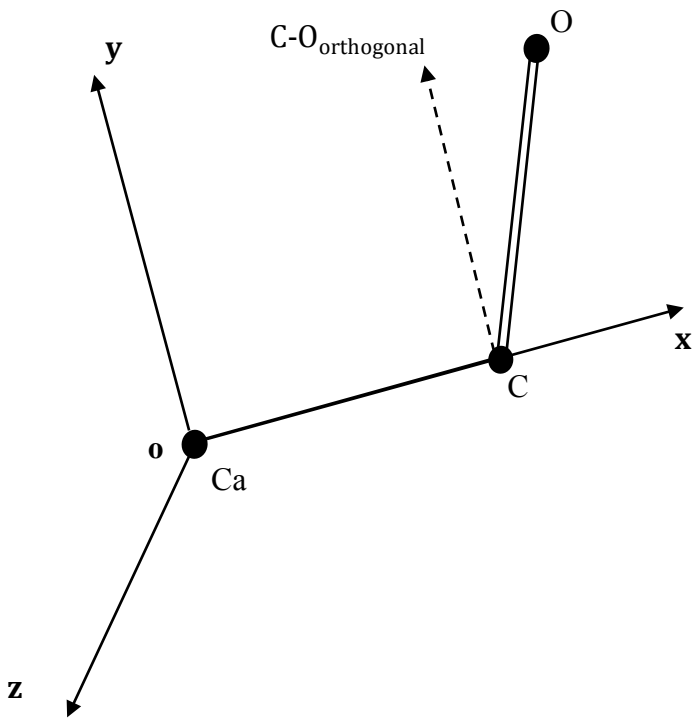
Table 2. Average Pearson's correlation coefficients of CSF scores with TM-scores. The correlation coefficient of a decoy set is the average coefficient of all targets in that decoy set. In calculating the correlation coefficients, the native structure was excluded. OPUS-CSF has comparable average correlation coefficient with other two potentials. The bold number in each row indicates the best one among the three potential functions for that particular decoy set. For OPUS-CSF, only those results for the combined segment case are listed.

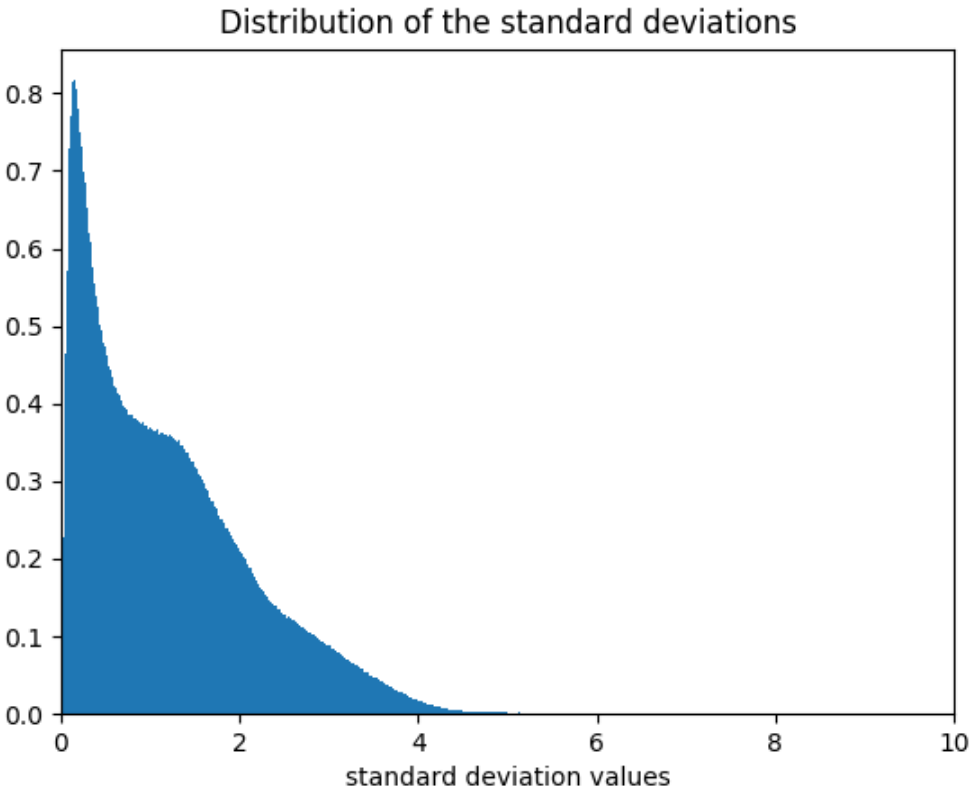
	Num_above5	Num_all	Num_above5/Num_all
5-residues	1766273	2350969	0.751
7-residues	3736778	9544858	0.391
9-residues	3713506	10262243	0.362
11-residues	3743204	10698802	0.350

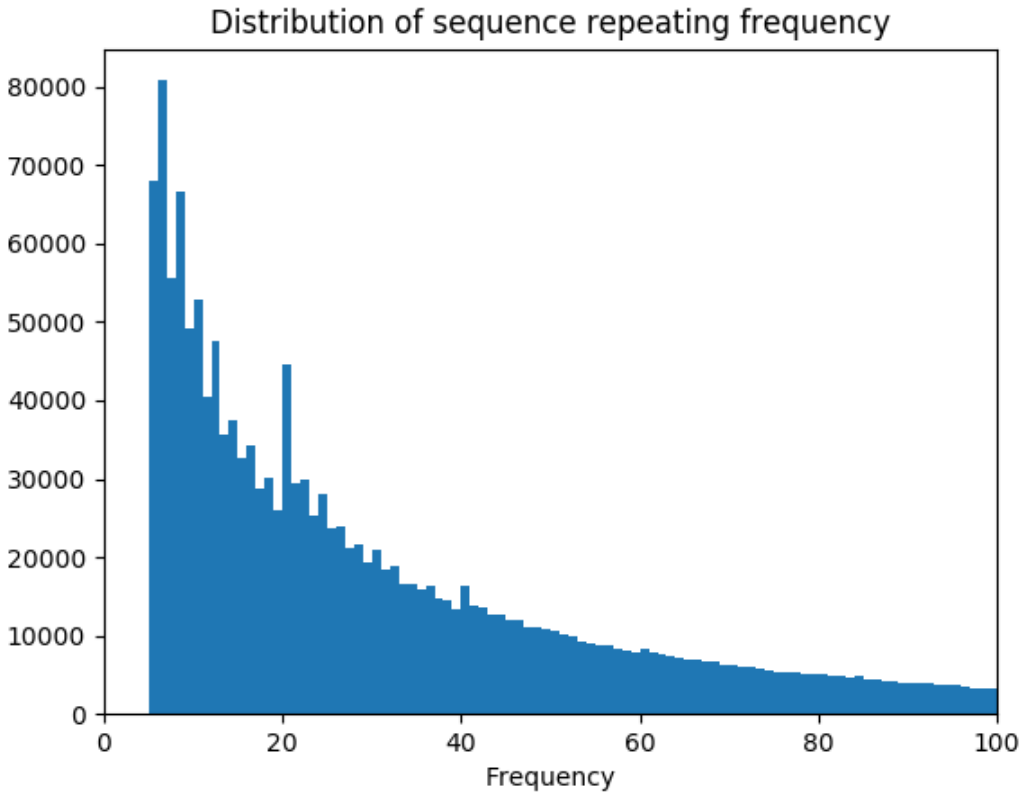
Table 3. The result of OPUS-CSF built by different length of residue segments. Num_above5 is the number of sequence segments which occur at least five times in PDB. Num_all shows the total number of sequence segments in PDB. The ratio decreases as the length of segments increases.

	5-residues	7-residues	9-residues	11-residues
Success numbers	244 (278)	218 (278)	220 (278)	219 (278)
Z-scores	-3.556	-4.546	-4.616	-4.569
Average Coverage	0.971	0.749	0.712	0.683
Unknowns	0	41	45	46

Table 4. The performance of OPUS-CSF based on different length of residue segments on 11 decoys sets. Success numbers are the numbers of native structures that OPUS-CSF correctly recognized from the decoys. Numbers in parentheses (278) is the total number of native structures (or targets) in 11 decoy sets. The Z-scores are the ones calculated based on the CSF scores of the native structures with respect to their decoys. Coverage means the ratio between the number of segments available in CND lookup table and the number of total segments of a target sequence. The table shows the average coverage among 278 targets in 11 decoy sets. Unknowns are the numbers of target sequences that have less than 20% of coverage. For these sequences, OPUS-CSF is not applicable. Note, 5-residue case does not have sequence classified as unknown, while 7-residue case, for example, has 41 out of 278 sequences not applicable for OPUS-CSF. The number of unknown increases slightly as the length of segment increases. Note, in the combined segment case, the longer segments may make no contribution to the CSF score if they are regarded as unknowns. Since 5-residue segment case has no unknown, it guarantees OPUS-CSF applicable to all target sequences even in rare ones that all longer segments are regarded as unknown.


y
C-O orthogonal
O
x
C
o
Ca
z


Distribution of the standard deviations
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
standard deviation values


Distribution of sequence repeating frequency
80000
70000
60000
50000
40000
30000
20000
10000
0
0
20
40
60
80
100
Frequency