SHORT RESEARCH ARTICLE

REVISED Protein structure quality assessment based on the distance profiles of consecutive backbone Cα atoms [version 3; referees: 2 approved]

Sandeep Chakraborty[1], Ravindra Venkatramani[2], Basuthkar J. Rao[1], Bjarni Asgeirsson[3], Abhaya M. Dandekar[4]

[1]Department of Biological Sciences, Tata Institute of Fundamental Research, Mumbai, 400 005, India
[2]Department of Chemical Sciences, Tata Institute of Fundamental Research, Mumbai, 400 005, India
[3]Science Institute, Department of Biochemistry, University of Iceland, Reykjavik, IS-107, Iceland
[4]Plant Sciences Department, University of California, Davis, CA 95616, USA

v3 **First published:** 10 Oct 2013, **2**:211 (doi: 10.12688/f1000research.2-211.v1)
Second version: 21 Nov 2013, **2**:211 (doi: 10.12688/f1000research.2-211.v2)
Latest published: 17 Dec 2013, **2**:211 (doi: 10.12688/f1000research.2-211.v3)

Abstract

Predicting the three dimensional native state structure of a protein from its primary sequence is an unsolved grand challenge in molecular biology. Two main computational approaches have evolved to obtain the structure from the protein sequence - *ab initio/de novo* methods and template-based modeling - both of which typically generate multiple possible native state structures. Model quality assessment programs (MQAP) validate these predicted structures in order to identify the correct native state structure. Here, we propose a MQAP for assessing the quality of protein structures based on the distances of consecutive Cα atoms. We hypothesize that the root-mean-square deviation of the distance of consecutive Cα (RDCC) atoms from the ideal value of 3.8 Å, derived from a statistical analysis of high quality protein structures (top100H database), is minimized in native structures. Based on tests with the top100H set, we propose a RDCC cutoff value of 0.012 Å, above which a structure can be filtered out as a non-native structure. We applied the RDCC discriminator on decoy sets from the Decoys 'R' Us database to show that the native structures in all decoy sets tested have RDCC below the 0.012 Å cutoff. While most decoy sets were either indistinguishable using this discriminator or had very few violations, all the decoy structures in the fisa decoy set were discriminated by applying the RDCC criterion. This highlights the physical non-viability of the fisa decoy set, and possible issues in benchmarking other methods using this set. The source code and manual is made available at https://github.com/sanchak/mqap and permanently available on 10.5281/zenodo.7134.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	 2
REVISED version 3 published 17 Dec 2013	✓	✓
REVISED version 2 published 21 Nov 2013	↑	↑
version 1 published 10 Oct 2013	✓ report	✓ report

1 **Bairong Shen**, Soochow University China

2 **Simon Lovell**, University of Manchester UK

Discuss this article

Comments (2)

Associated Research Article

Chakraborty S, Venkatramani R, Rao BJ *et al.* » The electrostatic profile of consecutive Cβ atoms applied to protein structure quality assessment, *F1000Research* 2014, **2**:243 (doi: 10.12688/f1000research.2-243.v3)

Corresponding author: Sandeep Chakraborty (sanchak@gmail.com)

How to cite this article: Chakraborty S, Venkatramani R, Rao BJ *et al.* **Protein structure quality assessment based on the distance profiles of consecutive backbone Cα atoms [version 3; referees: 2 approved]** *F1000Research* 2013, **2**:211 (doi: 10.12688/f1000research.2-211.v3)

Grant information: BJ and RV acknowledge financial support from Tata Institute of Fundamental Research (Department of Atomic Energy). Additionally, BJR is thankful to the Department of Science and Technology for the JC Bose Award Grant. BA extends gratitude to the University of Iceland Research Foundation for supporting the project financially. AMD wishes to acknowledge grant #12-0130-SA from California Department of Food and Agriculture CDFA PD/GWSS Board.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 10 Oct 2013, **2**:211 (doi: 10.12688/f1000research.2-211.v1)

REVISED **Amendments from Version 2**

We have implemented the following changes to add data in accordance with comments from Dr Rafael Najmanovich:

1). We obtained a set of PDB structures from the PISCES database (http://dunbrack.fccc.edu/PISCES.php) - they have a precompiled set of structures below a certain resolution and with a certain homology cut-off.

2). We have binned the structures based on resolution into different sets.

3). We have plotted the frequency distribution of the RDCC for each of these sets and display them in a new figure (Figure 1e).

Incidentally, and as expected, we could not detect any correlation based on RDCC and the resolution of the protein structure.

See referee reports

Introduction

The structure of a protein is a veritable source of information about its physiological relevance in the cellular context[1]. In spite of rapid technical advances in crystallization techniques, the number of protein sequences known far exceeds the known structures. There are essentially two different computational approaches to predict protein structures from its primary sequence: 1) Template based methods (TBM) which are based on features obtained from the database of known protein structures[2–4] and 2) *ab initio* or *de novo* methods which are based on the intrinsic laws governing atomic interactions and are applicable in the absence of a template structure with significant sequence homology[5,6]. While at present TBM methods fare much better than the *de novo* approaches, the requirement of a known template protein can sometimes be a constraining factor. Both these methods typically generate multiple possibilities for the native structure of a given sequence. Selecting the best candidate from the set of putative structures is an essential aspect that is performed by model quality assessment programs (MQAP).

MQAPs can be classified as energy based, consensus based or knowledge based. The refinement of structures based on modeling of atomic interactions in energy based methods, such as molecular dynamic simulations, are subject to limited sampling of possible conformations due to large run times, and force field inaccuracies due to the approximations involved in describing the dynamics of large multi-atomic systems[7–10]. Consensus methods are based on the principle that sub-structures of the native structure are likely to feature frequently in a set of near-native structures[11–14]. These methods are currently the best performing amongst MQAPs[13], but are prone to be computationally intensive due structure-to-structure comparison of all models[14], and are of limited use when the number of possible structures is small[15]. Knowledge based methods rely on the assignment of an empirical potential (also known as statistical potential) from the frequency of residue contacts in the known structures of native proteins[16,17]. In statistical physics, for a system in thermodynamic equilibrium, the accessible states are populated with a frequency which depends on the free energy of the state and is given by the Boltzmann distribution. The Boltzmann hypothesis states that if the database of known native protein structures is assumed to be a statistical system in thermodynamic equilibrium, specific structural features would be populated based on the free energy of the protein conformational state. Sippl argued using a converse logic that the frequencies of occurrence of structural features such as interatomic distances in the database of known protein structures could determine a free energy (potential of mean force) for a given protein conformation, and thus be used to discriminate the native structure[18,19]. A crucial aspect in applying statistical potentials is the proper characterization of the reference state[20]. The application of such empirical energy functions to predict and assess protein structures, while quite popular, are vigorously debated[21,22], and several approaches for using statistical potentials for protein structure prediction been described to date[20,23–26].

Here, we propose a new statistical potential based MQAP for assessing the quality of protein structures based on the distances of consecutive Cα atoms - Protein structure quality assessing based on Distance profile of backbone atoms (PROQUAD). We first propose a statistical potential based on the distance of consecutive Cα distances. In a set of high quality protein structures (top100H[27]), we demonstrate that the distance between consecutive Cα atoms are distributed normally with a mean of 3.8 Å and standard deviation of 0.04 Å. Based on this observation, the reference state for our statistical potential calculations is defined as one where all consecutive Cα atoms are 3.8 Å apart. We propose a scoring function which measures the deviation of consecutive Cα atoms from 3.8 Å, and hypothesize that this score is minimized in native structures. Based on the top100H database, we chose a cutoff of 0.012 Å for this scoring function to identify non-native states. We show that all the decoy structures from the fisa decoy set taken from the Decoys 'R' Us database[28] are distinguished using this discriminator. It has been previously proposed that native structures have constrained interatomic distances[29]. Interatomic distances, and other metrics, have been combined in several such methods - Molprobity (http://molprobity.biochem.duke.edu/), PROSA (https://prosa.services.came.sbg.ac.at/prosa.php) and the WHATIF server (http://swift.cmbi.ru.nl/whatif)[30–32]. These identify possible anomalies in a given protein structure. While Molprobity and WHATIF identified steric clashes in the decoy structure in fisa, distance checks between consecutive Cα are not part of checks in these methods, and they failed to detect the consecutive Cα atoms anomaly in the fisa decoy set.

Thus, we propose a simple and fast discriminator for protein structure quality based on the distance profiles of consecutive backbone Cα atoms that identifies decoy structures that are physically nonviable.

Results and discussion

The frequency distribution of the distance of consecutive Cα atoms in ~100 proteins in the top100H database (a database consisting of high quality structures)[27] shows that the distance between consecutive Cα atoms are distributed normally with a mean of 3.8 Å and standard deviation of 0.04 Å (Figure 1a). Out of 16,162 pairs of consecutive Cα atom distances, 14,281 (88%) were spaced 3.8 Å apart, 1297 (8%) were spaced 3.9 Å apart and 553 (3%) were spaced 3.7 Å apart. Only 31 (0.1%) pairs of consecutive Cα atom distances had values different than these (highest being 4 Å and the lowest being 2.9 Å). It would be interesting to correlate these distance deviant residue pairs to structural or functional aspects of the protein - *It is well worth examining every outlier and either correcting it if possible, giving up gracefully if it really cant be improved (more*

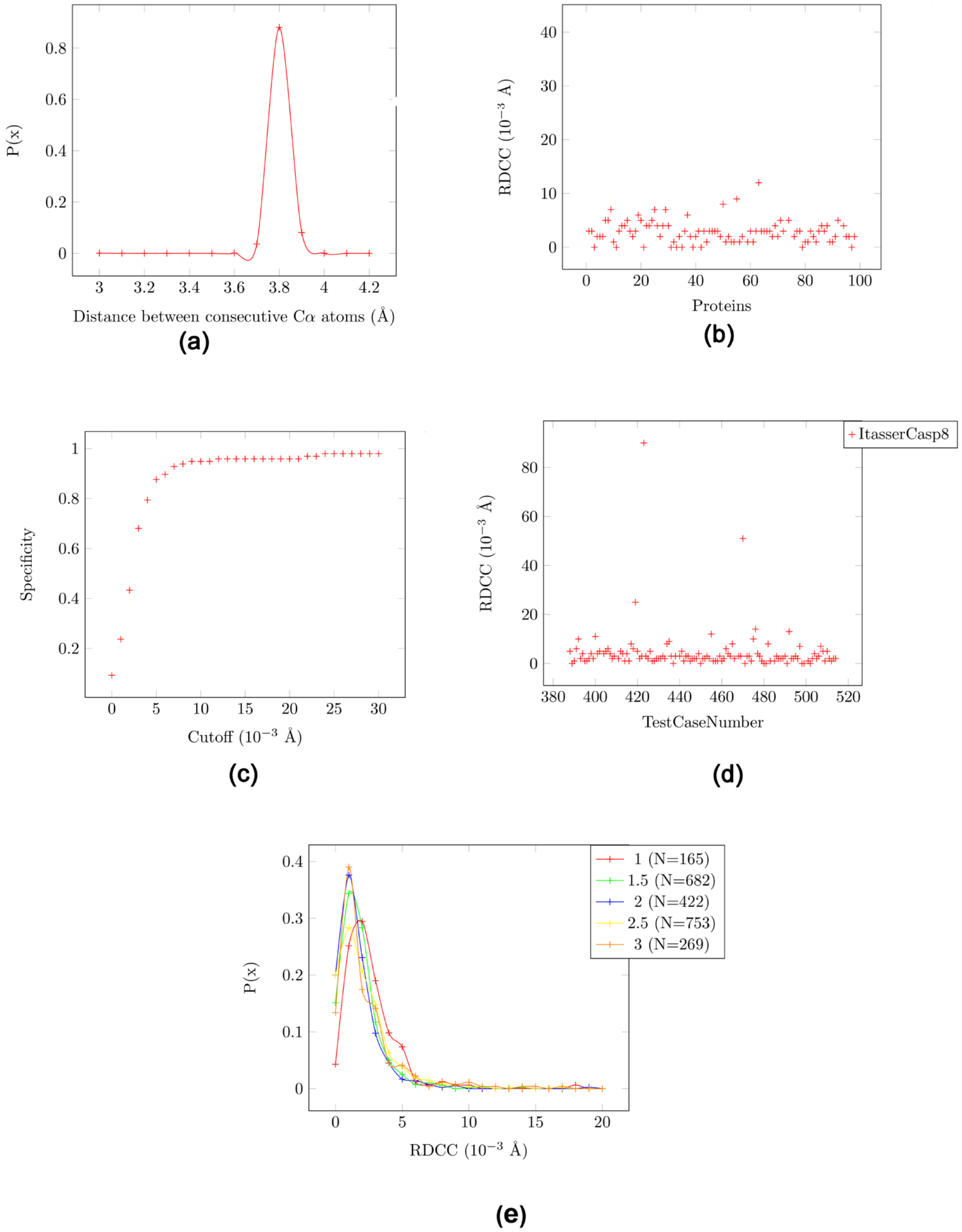

Figure 1. Root-mean-square deviation of the distance of consecutive Cα (RDCC) atoms from the ideal value of 3.8 Å. (**a**) Probability distribution (P(x)) for the distance of consecutive Cα in ~100 proteins in the top100H database. (**b**) RDCC in ~100 high quality structures from the top100H database. (**c**) Variation in specificity based on the cutoff value. We choose 0.012 Å as the cutoff for filtering out non-native structures. (**d**) RDCC in I-TASSER CASP8 decoy suite. (**e**) RDCC for protein structures based on the resolution.

often true at low resolution), or celebrating the significance of why it is being held in an unfavorable conformation[33].

The *cis* confirmations of peptide bonds are mostly responsible for these deviations. For example, in the protein concanavalin B (PDBID:1CNV), there are four violation of the 3.8 Å constraint: Ile33/Ser34 - 4 Å, Ser34/Phe35 - 3 Å, Pro56/Ser57 - 4 Å and Trp265/Asn266 - 3.4 Å. These all these deviations are noted in the PDB file as footnotes, mentioning that ‘peptide bond deviates significantly from *trans* conformation’[34]. Another example is the Glu223-Asp24 violation in PBDid:1ADS, which is between two *cis* prolines (as noted in the PDB file)[35]. However, these conformations are rare and not expected to occur frequently in a protein structure.

Figure 1b plots the root-mean-square deviation of the distance of consecutive Cα (RDCC) for these ~100 proteins. All structures in the top100H database have low RDCC values, barring three proteins (PDBids: 2ER7, 1XSO and 4PTP), which had multiple conformations for some residues, and were excluded from the processing. This validates our hypothesis that RDCC is minimized in native structures. Hence, structures that have a RDCC value more than a user specified threshold can be pruned out as structures with low quality or non-native structures.

We evaluated the results using the measures of specificity (the ability of a test to identify negative results) which is defined as:

$$specificity = \frac{TN}{FP + TN} \quad (1)$$

(TN = true negatives, FP = false positives). The specificity variation with the cutoff chosen is shown in Figure 1c. We choose 0.012 Å as the cutoff value for RDCC, which has a specificity of 1. We also plot the RDCC of the 121 testcases (Figure 1d) in the I-TASSER decoy set - http://zhanglab.ccmb.med.umich.edu/casp8/decoys[36]. Only five sets have RDCC values above the 0.012 Å threshold: T0492 - 0.013 Å, T0476 - 0.014 Å, T0419 - 0.025 Å, T0470 - 0.051 Å, T0423 - 0.09 Å. Some of these are the result of erroneous residue numbering in the CASP8 I-TASSER decoy set. For example, Ala24 is mistakenly numbered as Ala19 in T0423 (PDBid:3D01, identified by doing a BLAST search). Correcting this numbering results in a RDCC of 0.002 Å. Similarly, T0470 (PDBid:3DJB) has a correct RDCC of 0.001 Å, since Ser112 is mistakenly numbered as Ser101.

Figure 1e plots the frequency distribution of RDCC values of protein structures based on their resolution. The RDCC values are much lower than the 0.012 Å cutoff proposed. The non homologous structures (20% identity cutoff) are obtained from the PISCES database (http://dunbrack.fccc.edu/PISCES.php). Certain outliers have been removed - for example, PDBid:2JLI mentions a 'cleaved peptide bond between N263 and P264'. The distance between the Cα atoms of N263 and P264 in this protein is 9.4 Å. Table 1 shows the mean and standard deviation for these sets, and demonstrates that the RDCC values are independent of the resolution of the structure under consideration.

We have applied this cutoff on decoy sets from the Decoys 'R' Us database[28]. The first protein (the native structure) in all decoy sets has RDCC below the 0.012 Å cutoff (Figure 1c). Figure 2 shows the RDCC for the hg_structral and fisa decoy sets from the Decoys 'R' Us database. All 500 decoy structures in each protein structure in the fisa decoy set are discriminated by applying the RDCC criterion. Figure 3 shows the superimposition of the native structure and the first decoy structure (AXPROA00-MIN) for a protein (PDBid:1FC2) taken from the fisa decoy set. The distance between Ile12/Cα and Leu13/Cα atoms is 3.8 Å and 4.1 Å in the native and the decoy structures, respectively. According to our hypothesis, a 4.1 Å distance between consecutive Cα atoms is typically unfeasible in protein structures, and their occurrence should be relatively rare. The presence of such deviations throughout the protein structures categorizes it as a non-native structure. MolProbity[30] and ProSA[31] are two programs used as a pre-processing step for structures used in CASP[38]. MolProbity was able to discriminate the decoy structure (AXPROA00-MIN) from the native structure (PDBid:1FC2) using a metric called the ClashScore (the number of serious steric overlaps) and the Cβ deviations[39]. PROSA was unable to discriminate between the decoy and the native structures, reporting equivalent Zscores of -4.12 and -5.28, respectively. The WHATIF server report also reports steric clashes in the decoy structures (Data File 1). None of the above mentioned methods use a metric similar to the RDCC proposed in this paper, and thus did not report the abnormal distance between consecutive Cα atoms in the decoy structure.

Table 1. Mean and standard deviation (SD) of RDCC values for structures based on resolution. The number N signifies the number of protein structures analzyed that have resolution less than the specified number, but more than the previous one. For example, there are 165 protein with less than 1 Å resolution, and 682 proteins which have more than 1 Å but less than 1.5 Å resolution, and so on.

Resolution(Å)	N Mean	RDCC(10-3 Å)	SD(10-3 Å)
1.0	165	3.0	4.3
1.5	682	1.9	2.7
2.5	422	2.2	2.6
2.0	753	1.8	2.8
3.0	269	2.2	2.5

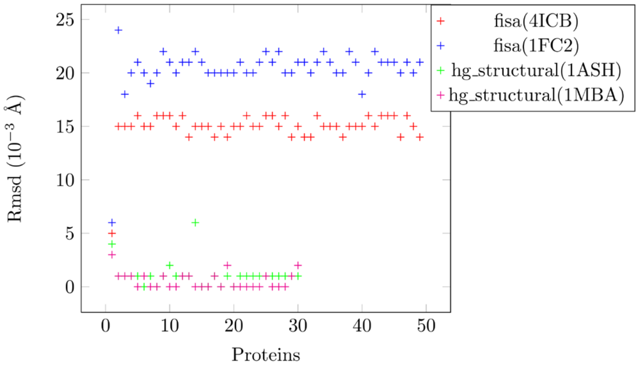

Figure 2. Root-mean-square deviation (RMSD) of the distance of consecutive Cα (RDCC) atoms from the ideal value of 3.8 Å in decoy sets. The hg_structal and misfold decoy sets are indistinguishable using the distance discriminator, unlike the fisa decoy set. We have shown ~25 decoy structures from the fisa set, but the values apply to all the decoys (more than 500). The first protein (the native structure) in each set has RDCC below the 0.012 Å cutoff.

The hg_structal and misfold decoy sets are indistinguishable using this distance discriminator, while only a few decoy structures failed in the 4state_reduced decoy set. This relationship between RDCC and proteins structure quality is therefore not an equivalence relationship. In propositional calculus, a relationship is equivalent if 'A' implies 'B' and 'B' implies 'A'. A high RDCC implies a low quality structure, but a low quality structure does not necessitate a high RDCC. We therefore suggest the usage of the RDCC measure as a first pass to rule out the non-native contacts prior to applying other discriminators.

The model quality assessment program (MQAP) used to choose the best structure from the multiple closely related structures generated

$$\Phi_{proteins} = \{P_1, P_2 \ldots P_M\} \quad (2)$$

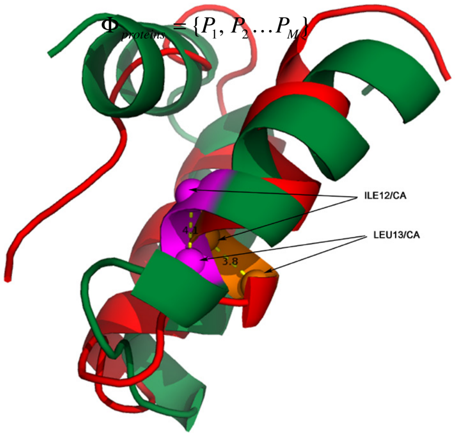

Figure 3. Superimposition of the native structure and a decoy structure (AXPROA00-MIN) for a protein (PDBid:1FC2) taken from the fisa decoy set. The native structure is in red, and the decoy structure is in green. The structures are superimposed using MUSTANG[58]. The distance between Ile12/Cα and Leu13/Cα atoms is 3.8 Å and 4.1 Å in the native and the decoy structures, respectively.

by structure prediction programs is of critical importance. We have in the past used electrostatic congruence to detect a promiscuous serine protease scaffold in alkaline phosphatases[40] and a phosphoinositide-specific phospholipase C from *Bacillus cereus*[41], and a scaffold recognizing a β-lactam (imipenem) in a cold-active *Vibrio* alkaline phosphatase[42,43]. However, continuum models[44] that compute potential differences and pK_a values from charge interactions in proteins[45] are sensitive to the spatial arrangement of the atoms in the structure. Thus, an incorrect model will generate an inaccurate electrostatic profile of the peptide[46]. It is thus possible to functionally characterize a protein from its sequence by applying such *in silico* tools subsequent to the protein structure prediction and MQAPs tools[47].

The estimation of the model quality by MQAPs is achieved by formalizing a scoring function[48], referred to as a knowledge-based or statistical potential, constructed from the database of known structures, assuming that the distribution of the structural features obtained from these structures follows the Boltzmann distribution[20,23,24,26]. The validity of statistical potential and the method to choose a proper reference frame in such models are still widely debated[21,22]. Methods that use consensus values from numerous models outperform other MQAP methods[11–14], and are 'very useful for structural meta-predictors'[49]. It has been shown that many of the MQAP programs perform considerably better when different statistical metrics are combined[50–53]. The state of the art methods for predicting structures[54] and MQAPs[38,49,55] are evaluated by researchers every two years.

Here, we propose a discriminator (RDCC) based on the distance of consecutive Cα atoms in the peptide structure. The discriminator is independent of the database of structures[56], and is thus an absolute discriminator. Our proposed RDCC criterion is satisfied in high quality protein structures taken from the top100H database. As a specific application, we show that all decoy structures in the fisa decoy set from the Decoys 'R' Us database Cα atoms do not satisfy this criterion. It has been previously shown that the fisa decoy set violates the van der Waals term[57]. We propose a fast complementary method to identify this transgression. It is also an interesting fact that most consensus methods will fare poorly in the fisa decoy set, since the majority of sub-structures are incorrect in all the decoy sets. Therefore, the fisa decoy set consists of physically nonviable structures and one should exercise caution when benchmarking other methods using this decoy set[58].

Data File 1: WHATIF server report

1 Data File

http://dx.doi.org/10.6084/m9.figshare.813320

Materials and methods

The set of proteins $\Phi_{proteins}$ consists of the native structure P_1 and M-1 decoys structures (Equation 2). We ignore the first x=*IgnoreNTerm* and last y=*IgnoreCTerm* pairs of residues in the protein structure to exclude the terminals (Equation 3). For every consecutive pair of residues in the structure we calculate the distance between the consecutive Cα atoms ($Res_n(C\alpha)$ and $Res_{n+1}(C\alpha)$), and its deviation from the ideal value of 3.8 Å. The square of the summation of these deviations is then normalized based on the number of pairs processed, and results in the *CADistScore*. We hypothesize that $CADistScore^{P1}$ is minimum in a native structure (Equation 4). Algorithm 1 shows the pseudocode for the function that generates the *CADistScore*.

Algorithm 1: AssessCADist()

```
Input: $P_1$ : Protein under consideration
Input: IgnoreNTerm: Ignore this number of residues in the N Terminal
Input: IgnoreCTerm: Ignore this number of residues in the C Terminal
Output: CADistScore: Score indicating deviation of successive Cα atoms from 3.8 Å
begin
    CADistScore = 0 ; NumberCompared = 0 ; N = NumberOfResidues($P_1$);
    for p ← IgnoreNTerm to N – IgnoreCTerm do
        q = p + 1 ;
        CADist = Distance(p, q, Cα, Cα)
        NumberCompared = NumberCompared + 1 ;
        diff = absolute(CADist – 3.8 Å) ;
        CADistScore = CADistScore + diff ∗ diff;
    end
    /* Normalize */
    CADistScore = sqrt(CADistScore/( NumberCompared ∗ NumberCompared));
    return (CADistScore);
end
```

$$\Phi_{proteins} = \{P_1, P_2 \ldots P_M\} \quad (2)$$

$$CADistScore^{Pi} = \sqrt{\frac{\sum_{n=1+x}^{N-y-1}(dist(Res_{n}(C\alpha), Res_{n+1}(C\alpha)) - 3.8)^{2}}{(N-y-x-2)^{2}}} \quad (3)$$

$$[\forall i = 2 \ldots M](CADistScore^{P1} < CADistScore^{Pi}) \quad (4)$$

In order to validate our hypothesis on known structures, we applied our discriminator to the top100H database (a database consisting of high quality structures)[27] - http://kinemage.biochem.duke.edu/databases/top100.php. In order to benchmark model quality assessment programs, we used decoy sets from the Decoys 'R' Us database[28] - http://dd.compbio.washington.edu/. Each set has several structures that are supposed to be ranked worse than the native structure.

Structural superimposition has been done using MUSTANG[59]. Protein structures were rendered by PyMol (http://www.pymol.org/).

The source code and manual is made available at https://github.com/sanchak/mqap.

Author contributions

Conceived and performed the experiments: SC. Analyzed the data, and improved experiments: SC BA AMD BJR RV. Wrote the manuscript: SC BA AMD BJR RV.

Competing interests

No competing interests were disclosed.

Grant information

BJ and RV acknowledge financial support from Tata Institute of Fundamental Research (Department of Atomic Energy). Additionally, BJR is thankful to the Department of Science and Technology for the JC Bose Award Grant. BA extends gratitude to the University of Iceland Research Found for supporting the project financially. AMD wishes to acknowledge grant #12-0130-SA from California Department of Food and Agriculture CDFA PD/GWSS Board.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Wise EL, Rayment I: **Understanding the importance of protein structure to nature's routes for divergent evolution in TIM barrel enzymes.** *Acc Chem Res.* 2004; **37**(3): 149–158.
PubMed Abstract | Publisher Full Text

2. Soding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics.* 2005; **21**(7): 951–960.
PubMed Abstract | Publisher Full Text

3. Peng J, Xu J: **RaptorX: exploiting structure information for protein alignment by statistical inference.** *Proteins.* 2011; **79**(Suppl 10): 161–171.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Zhang Y: **Template-based modeling and free modeling by I-TASSER in CASP7.** *Proteins.* 2007; **69**(Suppl 8): 108–117.
PubMed Abstract | Publisher Full Text

5. Wu S, Skolnick J, Zhang Y: **Ab initio modeling of small proteins by iterative TASSER simulations.** *BMC Biol.* 2007; **5**: 17.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Rohl CA, Strauss CE, Misura KM, *et al.*: **Protein structure prediction using Rosetta.** *Methods Enzymol.* 2004; **383**: 66–93.
PubMed Abstract | Publisher Full Text

7. Chen J, Brooks CL: **Can molecular dynamics simulations provide high-resolution refinement of protein structure?** *Proteins.* 2007; **67**(4): 922–930.
PubMed Abstract | Publisher Full Text

8. Zhu J, Fan H, Periole X, *et al.*: **Refining homology models by combining replica-exchange molecular dynamics and statistical potentials.** *Proteins.* 2008; **72**(4): 1171–1188.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Raval A, Piana S, Eastwood MP, *et al.*: **Refinement of protein structure homology models via long, all-atom molecular dynamics simulations.** *Proteins.* 2012; **80**(8): 2071–2079.
PubMed Abstract | Publisher Full Text

10. Lee MR, Tsai J, Baker D, *et al.*: **Molecular dynamics in the endgame of protein structure prediction.** *J Mol Biol.* 2001; **313**(2): 417–430.
PubMed Abstract | Publisher Full Text

11. Ginalski K, Elofsson A, Fischer D, *et al.*: **3D-Jury: a simple approach to improve protein structure predictions.** *Bioinformatics.* 2003; **19**(8): 1015–1018.
PubMed Abstract | Publisher Full Text

12. Terashi G, Oosawa M, Nakamura Y, *et al.*: **United3D: a protein model quality assessment program that uses two consensus based methods.** *Chem Pharm Bull (Tokyo).* 2012; **60**(11): 1359–1365.
PubMed Abstract | Publisher Full Text

13. Wallner B, Elofsson A: **Prediction of global and local model quality in CASP7 using Pcons and ProQ.** *Proteins.* 2007; **69**(Suppl 8): 184–193.
PubMed Abstract | Publisher Full Text

14. Adamczak R, Pillardy J, Vallat BK, *et al.*: **Fast geometric consensus approach for protein model quality assessment.** *J Comput Biol.* 2011; **18**(12): 1807–1818.
PubMed Abstract | Publisher Full Text

15. McGuffin LJ: **Benchmarking consensus model quality assessment for protein fold recognition.** *BMC Bioinformatics.* 2007; **8**: 345.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Tanaka S, Scheraga HA: **Model of protein folding: inclusion of short-, medium-, and long-range interactions.** *Proc Natl Acad Sci U S A.* 1975; **72**(10): 3802–3806.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules.* 1985; **18**(3): 534–552.
Publisher Full Text

18. 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**(4): 859–883.
PubMed Abstract | Publisher Full Text

19. Sippl MJ: **Knowledge-based potentials for proteins.** *Curr Opin Struct Biol.* 1995; **5**(2): 229–235.
PubMed Abstract | Publisher Full Text

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**(11): 2714–2726.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Thomas PD, Dill KA: **Statistical potentials extracted from protein structures: how accurate are they?** *J Mol Biol.* 1996; **257**(2): 457–469.
PubMed Abstract | Publisher Full Text

22. Hamelryck T, Borg M, Paluszewski M, *et al.*: **Potentials of mean force for protein**

structure prediction vindicated, formalized and generalized. *PLoS ONE.* 2010; **5**(11): e13714.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Samudrala R, Moult J: **An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.** *J Mol Biol.* 1998; **275**(5): 895–916.
PubMed Abstract | Publisher Full Text

24. Shen MY, Sali A: **Statistical potential for assessment and prediction of protein structures.** *Protein Sci.* 2006; **15**(11): 2507–2524.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Rajgaria R, McAllister SR, Floudas CA: **A novel high resolution Calpha-Calpha distance dependent force field based on a high quality decoy set.** *Proteins.* 2006; **65**(3): 726–741.
PubMed Abstract | Publisher Full Text

26. Lu H, Skolnick J: **A distance-dependent atomic knowledge-based potential for improved protein structure selection.** *Proteins.* 2001; **44**(3): 223–232.
PubMed Abstract | Publisher Full Text

27. Word JM, Lovell SC, LaBean TH, *et al.*: **Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms.** *J Mol Biol.* 1999; **285**(4): 1711–1733.
PubMed Abstract | Publisher Full Text

28. Samudrala R, Levitt M: **Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction.** *Protein Sci.* 2000; **9**(7): 1399–1401.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Engh RA, Huber R: **Accurate bond and angle parameters for X-ray protein structure refinement.** *Acta Crystallographica Section A.* 1991; **47**: 392–400.
Publisher Full Text

30. Chen VB, Arendall WB, Headd JJ, *et al.*: **MolProbity: all-atom structure validation for macromolecular crystallography.** *Acta Crystallogr D Biol Crystallogr.* 2010; **66**(Pt 1): 12–21.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Wiederstein M, Sippl MJ: **ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins.** *Nucleic Acids Res.* 2007; **35**(Web Server issue): W407–410.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Vriend G: **WHAT IF: a molecular modeling and drug design program.** *J Mol Graph.* 1990; **8**(1): 52–56.
PubMed Abstract | Publisher Full Text

33. Richardson J, Richardson D: **The zen of model anomalies – correct most of them. Treasure the meaningful valid few. Live serenely with the rest!** *Advancing Methods for Biomolecular Crystallography*, Springer Netherlands. 2013; 1–10.
Publisher Full Text

34. Hennig M, Jansonius JN, Terwisscha van Scheltinga AC, *et al.*: **Crystal structure of concanavalin B at 1.65 A resolution. An "inactivated" chitinase from seeds of Canavalia ensiformis.** *J Mol Biol.* 1995; **254**(2): 237–246.
PubMed Abstract | Publisher Full Text

35. Wilson DK, Bohren KM, Gabbay KH, *et al.*: **An unlikely sugar substrate site in the 1.65 A structure of the human aldose reductase holoenzyme implicated in diabetic complications.** *Science.* 1992; **257**(5066): 81–84.
PubMed Abstract | Publisher Full Text

36. Zhang Y: **I-TASSER: fully automated protein structure prediction in CASP8.** *Proteins.* 2009; **77**(Suppl 9): 100–13.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Wang G, Dunbrack RL Jr: **PISCES: a protein sequence culling server.** *Bioinformatics.* 2003; **19**(12): 1589–91.
PubMed Abstract | Publisher Full Text

38. Kryshtafovych A, Monastyrskyy B, Fidelis K: **Casp prediction center infrastructure and evaluation measures in casp10 and casp roll.** *Proteins: Structure, Function, and Bioinformatics.* 2013.
PubMed Abstract | Publisher Full Text

39. Lovell SC, Davis IW, Arendall WB, *et al.*: **Structure validation by Calpha geometry: phi, psi, and Cbeta deviation.** *Proteins.* 2003; **50**(3): 437–450.
PubMed Abstract | Publisher Full Text

40. Chakraborty S, Minda R, Salaye L, *et al.*: **Active site detection by spatial conformity and electrostatic analysis–unravelling a proteolytic function in shrimp alkaline phosphatase.** *PLoS ONE.* 2011; **6**(12): e28470.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Rendon-Ramirez A, Shukla M, Oda M, *et al.*: **A Computational Module Assembled from Different Protease Family Motifs Identifies PI PLC from *Bacillus cereus* as a Putative Prolyl Peptidase with a Serine Protease Scaffold.** *PLoS One.* 2013; **8**(8): e70923.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Helland R, Larsen RL, Asgeirsson B: **The 1.4 Å crystal structure of the large and cold-active *Vibrio.* sp. alkaline phosphatase.** *Biochim Biophys Acta.* 2009; **1794**(2): 297–308.
PubMed Abstract | Publisher Full Text

43. Chakraborty S, Asgeirsson B, Minda R, *et al.*: **Inhibition of a cold-active alkaline phosphatase by imipenem revealed by in silico modeling of metallo-*β*-lactamase active sites.** *FEBS Lett.* 2012; **586**(20): 3710–3715.
PubMed Abstract | Publisher Full Text

44. Honig B, Nicholls A: **Classical electrostatics in biology and chemistry.** *Science.* 1995; **268**(5214): 1144–1149.
PubMed Abstract | Publisher Full Text

45. Bashford D, Karplus M: **pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model.** *Biochemistry.* 1990; **29**(44): 10219–10225.
PubMed Abstract | Publisher Full Text

46. Baker NA, Sept D, Joseph S, *et al.*: **Electrostatics of nanosystems: application to microtubules and the ribosome.** *Proc Natl Acad Sci U S A.* 2001; **98**(18): 10037–10041.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Chakraborty S, Rao BJ: **A measure of the promiscuity of proteins and characteristics of residues in the vicinity of the catalytic site that regulate promiscuity.** *PLoS One.* 2012; **7**(2): e32011.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Shen MY, Sali A: **Statistical potential for assessment and prediction of protein structures.** *Protein Sci.* 2006; **15**(11): 2507–2524.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Kryshtafovych A, Fidelis K, Tramontano A: **Evaluation of model quality predictions in CASP9.** *Proteins.* 2011; **79**(Suppl 10): 91–106.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Benkert P, Tosatto SC, Schomburg D: **QMEAN: A comprehensive scoring function for model quality assessment.** *Proteins.* 2008; **71**(1): 261–277.
PubMed Abstract | Publisher Full Text

51. Tosatto SC: **The victor/FRST function for model quality estimation.** *J Comput Biol.* 2005; **12**(10): 1316–1327.
PubMed Abstract | Publisher Full Text

52. Bagaria A, Jaravine V, Huang YJ, *et al.*: **Protein structure validation by generalized linear model root-mean-square deviation prediction.** *Protein Sci.* 2012; **21**(2): 229–238.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Zhou J, Yan W, Hu G, *et al.*: **Svr caf: An integrated score function for detecting native protein structures among decoys.** *Proteins: Structure, Function, and Bioinformatics.* 2013; In press.
PubMed Abstract | Publisher Full Text

54. Moult J: **A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.** *Curr Opin Struct Biol.* 2005; **15**(3): 285–289.
PubMed Abstract | Publisher Full Text

55. Kryshtafovych A, Barbato A, Fidelis K, *et al.*: **Assessment of the assessment: Evaluation of the model quality estimates in CASP10.** *Proteins.* 2013.
PubMed Abstract | Publisher Full Text

56. Benkert P, Biasini M, Schwede T: **Toward the estimation of the absolute quality of individual protein structure models.** *Bioinformatics.* 2011; **27**(3): 343–350.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Handl J, Knowles J, Lovell SC: **Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction.** *Bioinformatics.* 2009; **25**(10): 1271–1279.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Chakraborty S, Rao BJ, Asgeirsson B, *et al.*: **The electrostatic profile of consecutive cβ atoms applied to protein structure quality assessment [v2; ref status: awaiting peer review, http://f1000r.es/2cf].** *F1000Research.* 2013; **2**: 243.
Publisher Full Text

59. Konagurthu AS, Whisstock JC, Stuckey PJ, *et al.*: **MUSTANG: a multiple structural alignment algorithm.** *Proteins.* 2006; **64**(3): 559–574.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ✓ ✓

Version 3

Referee Report 11 March 2014

doi:10.5256/f1000research.3328.r4038

✓ **Bairong Shen**
Center for Systems Biology, Soochow University, Suzhou, China

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 05 March 2014

doi:10.5256/f1000research.3328.r3986

✓ **Simon Lovell**
Faculty of Life Sciences, University of Manchester, Manchester, UK

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 06 November 2013

doi:10.5256/f1000research.2356.r2086

✓ **Simon Lovell**
Faculty of Life Sciences, University of Manchester, Manchester, UK

This is a nice, straightforward analysis of features of protein decoy sets. The authors find that a simple, novel measure of protein geometry is sufficient to distinguish native structures from decoys in the popular fisa decoy set. A rational response to this work would be to include measures of geometry such as that which are used here in the generation of the decoys. This would make the measure less useful, but would improve decoy sets.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 18 Nov 2013

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

Dear Dr Lovell,

Thank you for your encouraging comments on our manuscript. We took a lot of inspiration from your work "Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics.* 2009" (ref 52), while writing our manuscript.

Additionally, we have revised our manuscript adding some additional results we have obtained on the CASP8 I-TASSER decoy set.

Warm regards,

Sandeep

Competing Interests: No competing interests were disclosed.

Referee Report 30 October 2013

doi:10.5256/f1000research.2356.r2091

✓ **Bairong Shen**

Center for Systems Biology, Soochow University, Suzhou, China

This paper proposes an absolute discriminator to identify non-native protein structures based on backbone C_α atom distance. Interestingly the authors found that most methods performed poorly in the fisa decoy set from the Decoys “R” Us database and remind researchers to be cautious of using this decoy set since most of the sub-structures in the fisa decoy set are incorrect. This work provides a simple and fast assessment for protein structure quality.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 18 Nov 2013

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

Dear Dr Shen,

Thank you for taking the time to review our manuscript. We appreciate your encouraging comments. We have revised our manuscript with some additional results from the CASP8 I-TASSER decoy set, and also cited some recent manuscripts which had been published since the version 1 went online. We hope you find the revised version improved.

Warm regards,

Sandeep

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 3

Author Response 18 Dec 2013

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

Dear Dr Najmanovich,

Thank you for taking the time to read our Ms and for suggesting relevant enhancements.
We hope you are satisfied with this version.

Best regards,

Sandeep

Competing Interests: No competing interests were disclosed.

Version 2

Reader Comment (*F1000Research Advisory Board Member*) 03 Dec 2013

Rafael Najmanovich, Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montreal, Canada

I would have appreciated to see a bootstrapped distribution of RDCC on non redundant samples of the entire PDB database as a function of resolution.

Competing Interests: none