SHORT RESEARCH ARTICLE

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

Sandeep Chakraborty¹, Ravindra Venkatramani², Basuthkar J. Rao¹, Bjarni Asgeirsson³, Abhaya M. Dandekar⁴

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

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.
Changes from Version 1

The main changes incorporated in this version are summarized below.

1. Investigated the reason behind the root-mean-square deviation of the distance of consecutive Cα (RDCC) violation in the 31 anomalous Ca-Cα distances in the top100H database. These violations are the result of the cis conformations of the peptide bond, which are known to be unfavorable compared to the trans conformations. As we have mentioned, such occurrences are typically unfeasible in protein structures, and their occurrence should be relatively rare.

2. Added results from the Iassar CASP8 dataset - we have seen five outliers. 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'. We expect the others to have a similar numbering problem.

See referee reports

Introduction

The structure of a protein is a veritable source of information about its physiological relevance in the cellular context. 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 and are based on the intrinsic laws governing atomic interactions and are applicable in the absence of a template structure with significant sequence homology. 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. 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. These methods are currently the best performing amongst MQAPs, but are prone to be computationally intensive due structure-to-structure comparison of all models, and are of limited use when the number of possible structures is small. 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. 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. A crucial aspect in applying statistical potentials is the proper characterization of the reference state. The application of such empirical energy functions to predict and assess protein structures, while quite popular, are vigorously debated, and several approaches for using statistical potentials for protein structure prediction been described to date.

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), 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 are distinguished using this discriminator. It has been previously proposed that native structures have constrained interatomic distances. Interatomic distances, and other metrics, have been combined in several such methods - Molprobity and WHATIF server (http://swift.cmbi.ru.nl/whatif/) and the WHATIF server (http://swift.cmbi.ru.nl/whatif/) which 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) 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
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:

$$\text{specificity} = \frac{TN}{FP + TN}$$

where \(TN\) = 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 for filtering out non-native structures.

Figure 1b plots the root-mean-square deviation of the distance of consecutive Ca (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:

$$\text{specificity} = \frac{TN}{FP + TN}$$

Figure 1b plots the root-mean-square deviation of the distance of consecutive Ca (RDCC) atoms from the ideal value of 3.8 Å. (a) Probability distribution (P(x)) for the distance of consecutive Ca 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.
a correct RDCC of 0.001 Å, since Ser112 is mistakenly numbered as Ser101.

We have applied this cutoff on decoy sets from the Decoys ‘R’ Us database\[^{26}\]. 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_structural 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\[^{97}\] and ProSA\[^{10}\] are two programs used as a pre-processing step for structures used in CASP\[^{10}\]. 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 CB deviations\[^{10}\]. 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.

The hg_structural 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 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 phos-pha-tases\[^{10}\] and a phosphoinositide-specific phospholipase C from Bacillus cereus\[^{10}\], and a scaffold recognizing a β-lactam (imipenem) in a cold-active Vibrio alkaline phosphatase\[^{10}\]. However, continuum models\[^{10}\] that compute potential differences and pK\(_a\) values from charge interactions in proteins\[^{10}\] 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\[^{10}\]. 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\[^{10}\].

The estimation of the model quality by MQAPs is achieved by formalizing a scoring function\[^{10}\], 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\[^{10-14,26}\]. The validity of statistical potential and the method to choose a proper reference frame in such models are still widely debated\[^{10}\]. Methods that use consensus values from numerous models outperform other MQAP methods\[^{10-14,26}\], and are ‘very useful for structural meta-predictors’\[^{10}\]. It has been shown that many of the MQAP programs perform considerably better when different statistical metrics are combined\[^{10-14,26}\]. The state of the art methods for predicting structures\[^{10}\] and MQAPs\[^{10,14,26}\] 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\[^{10}\], 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 Ca atoms do not satisfy this criterion. It has been previously shown that the fisa decoy set violates the van der Waals term\textsuperscript{16}. 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.

Materials and methods

The set of proteins $\Phi_{\text{proteo}}$ consists of the native structure $P_i$ and M-1 decoys structures \textit{(Equation 2)}. We ignore the first $x=$\textit{IgnoreNTerm} and last $y=$\textit{IgnoreCTerm} pairs of residues in the protein structure to exclude the terminals (\textit{Equation 3}). For every consecutive pair of residues in the structure we calculate the distance between the consecutive Ca atoms $(\text{Res}_n, (\text{Ca})_n$ and $\text{Res}_{n+1}, (\text{Ca})_{n+1}$), 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 $\text{CADistScore}$. We hypothesize that $\text{CADistScore}^{P_i}$ is minimum in a native structure \textit{(Equation 4)}. Algorithm 1 shows the pseudocode for the function that generates the $\text{CADistScore}$.

$\Phi_{\text{proteo}} = \{P_1, P_2, \ldots P_M\}$ \hfill (2)

$\text{CADistScore}^{P_i} = \sqrt{\sum_{n=1}^{N-m-1} (\text{dist}((\text{Ca})_n, \text{Res}_n), (\text{Ca})_{n+1}, \text{Res}_{n+1} - 3.8)^2} / (N - y - x - 2)^2$ \hfill (3)

$[\forall i = 2 \ldots M] (\text{CADistScore}^{P_i} < \text{CADistScore}^{P_j})$ \hfill (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)\textsuperscript{27} - \textit{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\textsuperscript{28} - \textit{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\textsuperscript{31}. Protein structures were rendered by PyMol (\textit{http://www.pymol.org/}). The source code and manual is made available at \textit{https://github.com/sanchak/mqap}.

Algorithm 1: AssessCADist()

\begin{verbatim}
Input: $P_i$ : 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 Ca atoms from 3.8 Å

\begin{verbatim}
begin
    CADistScore = 0; NumberCompared = 0; N = NumberOfResidues($P_i$);
    for p ← IgnoreNTerm to N − IgnoreCTerm do
        q = p + 1;
        CADist = Distance(p, q, Ca, Ca)
        NumberCompared = NumberCompared + 1;
        diff = absolute(CADist − 3.8 Å)
        CADistScore = CADistScore + diff * diff;
    end
    /* Normalize */
    CADistScore = sqrt(CADistScore(NumberCompared * NumberCompared));
    return (CADistScore);
end
\end{verbatim}
\end{verbatim}

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 Fund 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.


**Open Peer Review**

**Current Referee Status:** ✅ ✅

---

**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.
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\textalpha{} 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 2

Reader Comment (F1000Research Advisory Board Member) 03 Dec 2013
Rafael Najmanovich, Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montréal, 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