The electrostatic profile of consecutive Cβ atoms applied to protein structure quality assessment

Permalink
https://escholarship.org/uc/item/63z0p8p6

Authors
Chakraborty, Sandeep
Venkatramani, Ravindra
Rao, Basuthkar J
et al.

Publication Date
2013

DOI
10.12688/f1000research.2-243.v3

Peer reviewed
The electrostatic profile of consecutive Cβ atoms applied to protein structure quality assessment [v3; ref status: indexed, http://f1000r.es/4dq]

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, IS-107 Reykjavik, Iceland
⁴Plant Sciences Department, University of California, Davis, CA, 95616, USA

Abstract
The structure of a protein provides insight into its physiological interactions with other components of the cellular soup. Methods that predict putative structures from sequences typically yield multiple, closely-ranked possibilities. A critical component in the process is the model quality assessing program (MQAP), which selects the best candidate from this pool of structures. Here, we present a novel MQAP based on the physical properties of sidechain atoms. We propose a method for assessing the quality of protein structures based on the electrostatic potential difference (EPD) of Cβ atoms in consecutive residues. We demonstrate that the EPDs of Cβ atoms on consecutive residues provide unique signatures of the amino acid types. The EPD of Cβ atoms are learnt from a set of 1000 non-homologous protein structures with a resolution cutoff of 1.6 Å obtained from the PISCES database. Based on the Boltzmann hypothesis that lower energy conformations are proportionately sampled more, and on Annsen's thermodynamic hypothesis that the native structure of a protein is the minimum free energy state, we hypothesize that the deviation of observed EPD values from the mean values obtained in the learning phase is minimized in the native structure. We achieved an average specificity of 0.91, 0.94 and 0.93 on hg_structal, 4state_reduced and ig_structal decoy sets, respectively, taken from the Decoys ‘R’ Us database. The source code and manual is made available at https://github.com/sanchak/mqap and permanently available on 10.5281/zenodo.7134.
Associated Short Research Article

Chakraborty S, Venkatramani R, Rao BJ et al. *Protein structure quality assessment based on the distance profiles of consecutive backbone Ca atoms*, F1000Research 2013, 2:211 (doi: 10.12688/f1000research.2-211.v3)

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

How to cite this article: Chakraborty S, Venkatramani R, Rao BJ et al. *The electrostatic profile of consecutive Cβ atoms applied to protein structure quality assessment* [v3; ref status: indexed, http://f1000r.es/4dq] F1000Research 2014, 2:243 (doi: 10.12688/f1000research.2-243.v3)

Copyright: © 2014 Chakraborty S et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

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.

Competing interests: No competing interests were disclosed.

First published: 13 Nov 2013, 2:243 (doi: 10.12688/f1000research.2-243.v1)

First indexed: 18 Jul 2014, 2:243 (doi: 10.12688/f1000research.2-243.v2)
Introduction

The challenge of deriving the native structure of a protein from its sequence has intrigued researchers for decades. Methods that predict putative structures from sequences are based either on features from databases of known structures (template-based methods) or use first principles of atomic interactions (ab initio or de novo methods). Typically, these methods yield multiple, closely-ranked possibilities. Model quality assessment programs (MQAP) that validate accuracy of these predicted structures are used to select the best candidate from the set of predicted structures.

MQAPs can be classified as energy, consensus or knowledge based. Two major sources of errors in energy based methods used for refining or discriminating protein structures are inaccuracies in the force field due to the inherent approximations in equations that model multi-atomic configurations, and inadequate sampling of the conformational space. Consensus based methods are based on the principle that structural features that are frequently observed in a population of structures are more likely to be present in the native structure. These clustering methods outperform other MQAP methods and are “very useful for structural meta-predictors.” However, they 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 proceed by deriving an empirical potential (also known as statistical potential) from the frequency of residue contacts in the known structures of native proteins. For a system in thermodynamic equilibrium, statistical physics hypotheses that 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. Applying a converse logic, Sippl reasoned that the frequencies of occurrence of structural features such as interatomic distances in the database of known protein structures could be used to assign a free energy (potential of mean force) for a given protein conformation. Furthermore, this statistical potential can be used to discriminate the native structure. The proper characterization of the reference state is a critical aspect in applying statistical potentials. In spite of their popularity, the application of such empirical energy functions to predict and assess protein structures is vigorously debated. Many MQAP programs perform better when multiple statistical metrics are combined. The paramount importance of obtaining high quality protein structures from sequences using in silico methods can be estimated by the effort invested by researchers every two years to evaluate both structure prediction tools and MQAPs.

Here, we propose a novel statistical potential to assess the quality of protein structures based on the electrostatic potential difference (EPD) of Cβ atoms in consecutive residues - EPD profile of side-chain atoms used in assessment of protein structures (ESCAPIST). Previously, we have established that the EPD is conserved in cognate pairs of active site residues in proteins with the same function. The ability of finite difference methods to quickly obtain consistent electrostatic properties from peptide structures provides an invaluable tool for investigating other innate properties of protein structures. We plot the EPD profiles for different atom types (Cα atoms, Cβ atoms and the C-N bond) in consecutive residues from a set of non-homologous protein structures obtained from the PISCES database. We proceed to show that the EPD between Cβ atoms in consecutive residues can be used to generate a scoring function that assesses the quality of protein structures. This EPD scoring function is then applied to standard decoy sets from the Decoys ‘R’ Us database to establish the validity of our method.

Results

Electrostatic potential difference (EPD) based discrimination

To extract feature values we chose a set of 1000 proteins from the PISCES database with percentage identity cutoff of 20%, resolution cutoff of 1.6 Å and a R-factor cutoff of 0.25 (SI Table 1).

Invariance of the EPD in the C-N peptide bond and between Cα atoms of consecutive residues

Adaptive Poisson-Boltzmann Solve (APBS) writes out the electrostatic potential in dimensionless units of kT/e where k is Boltzmann’s constant, T is the temperature in K and e is the charge of an electron. The units of EPD are same as that of the electrostatic potential. The EPD of the C-N peptide bond has a Gaussian distribution with mean = 420 EPD units and SD = 55 EPD units (Figure 1). In the probability distribution for four pairs of amino acids the mean of all pairs of amino acids are the same (Figure 1a). Figure 1b shows the scatter plot for the mean and standard deviation (SD). Thus, the amino acids are indistinguishable using the profile of the EPD of the C-N peptide bond across all protein structures since they have identical mean values and a large variance (SD=±50).

The probability distribution for four pairs of amino acids for the EPD between the Cα atoms of consecutive residues (Figure 2a) have means that are slightly more varied than those for the C-N bond (Figure 1a). In the scatter plot for the mean and SD of all pairs (Figure 2b) the outliers are pairs that include proline, which have a higher mean, although the magnitude of SD is the same (Table 1).

Distinctive EPD between Cβ atoms of consecutive residues for certain amino acid pairs

In contrast to the results described above, the EPD between the Cβ atoms in consecutive residues in the peptide structure can be used to discriminate different amino acid pairs in the protein structure. The mean EPD of all amino acid pairs are much more varied (Figure 2a). These pairs do not include glycine, which lacks a sidechain. In the scatter plot for the mean and SD, the outliers are pairs that include
Figure 1. Electrostatic potential differences (PD) for the C-N peptide bond. AA: Alanine/Alanine, AC: Alanine/Cysteine, HS: Histidine/Serine and DF: Aspartic-acid/Phenylalanine. (a) Probability distribution for four pairs of amino acids. (b) Scatter plot for all pairs of amino acids. It can be seen that the mean and SD for all pairs of amino acids are the same. Further, the variance is large (SD=−50), indicating that this feature is not tightly constrained in peptide structures.

Figure 2. Electrostatic potential differences (PD) for consecutive residue pairs for Ca atoms. A: Alanine/Alanine, AC: Alanine/Cysteine, HS: Histidine/Serine, DF: Aspartic-acid/Phenylalanine. (a) Probability distribution for four pairs of amino acids. (b) Scatter plot for all pairs of amino acids. It is seen that pairs of amino acids which include proline have a higher mean, although the magnitude of SD is the same. Cysteine (Figure 3b), which have a much higher SD (≈90) as compared to other pairs (SD=−35) (Table 2), and thus cannot be used for discriminatory purposes.

These values are used as a discriminator when choosing the native structure from a set of possible candidates (Table 3). To establish the non-triviality of these values, we also show that the variance of the EPD between these pairs increases with increasing sequence distance. Thus, the EPD between the pairs ‘DF’ and ‘HS’ has lesser correlation as the sequence distance between them increases (sample size for each sequence distance is > 30) (Figure 4). The SD for distance 1 (i.e. consecutive residues) is 29.8 EPD units and 31.8 EPD units for ‘DF’ and ‘HS’, respectively - and rises to around 60 EPD units with increasing sequence distance.

Validating using decoy sets
We obtained the score (PDScore) of any given protein structure by comparing the electrostatics of the Cβ atoms based on Table 3. To benchmark model quality assessment programs, we used decoy sets from the Decoys ’R’ Us database⁴³. We detail our results from some of these decoy sets. Each set has several structures that are supposed to be ranked worse than the native structure.
Figure 3. Electrostatic potential differences (PD) for consecutive residue pairs for Cβ atoms. AA: Alanine/Alanine, AD: Alanine/Aspartic-acid, AE: Alanine/Glutamic-acid, DF: Aspartic-acid/Phenylalanine, DY: Aspartic-acid/Tyrosine, HT: Histidine/Threonine, HS: Histidine/Serine. (a) Probability distribution for seven pairs of amino acids. (b) Scatter plot for all pairs of amino acids. The pairs which include cysteine have a high standard deviation. It is seen that the mean is much more varied than the electrostatic potential difference (EPD) for Cα and the C-N peptide bond.

Table 1. Electrostatic potential differences (EPD) for consecutive residue pairs for Cα atoms for residue pairs that include proline. While these pairs have for a low standard deviation (SD) like all other pairs, the absolute value of their mean is different (higher) than any pair that does not include a proline. This also highlights the unique nature of proline in protein structures.

| Pair | Mean EPD | SD  | Number of samples |
|------|----------|-----|-------------------|
| AP   | -167.3   | 28.6| 328               |
| CP   | -153     | 30.3| 45                |
| DP   | -184.5   | 29.5| 290               |
| EP   | -176.6   | 27.3| 346               |
| FP   | -160.4   | 25.3| 173               |
| GP   | -165.3   | 29.2| 339               |
| HP   | -162.7   | 34.6| 92                |
| IP   | -161.9   | 27.2| 175               |
| KP   | -156.6   | 29.6| 203               |
| LP   | -165.2   | 28.3| 323               |
| MP   | -161.3   | 29.5| 70                |
| NP   | -159.6   | 27.6| 168               |
| PQ   | 168.5    | 26.1| 131               |
| PR   | 156.2    | 31.5| 184               |
| PS   | 172.3    | 25.9| 269               |
| PT   | 170.8    | 27.6| 218               |
| PV   | 164.4    | 30.4| 299               |
| PW   | 158.5    | 29.7| 70                |
| PY   | 155.5    | 29  | 141               |

Table 2. Electrostatic potential differences (EPD) for consecutive residue pairs for Cβ atoms for residue pairs that has one cysteine. These pairs have a random values for the mean and a high standard deviation (SD), with the exception of the pair ‘CC’ (not the disulfide bond) which has a low mean value and SD. Consequently, these values can not discriminate between pairs of amino acids.

| Pair | Mean EPD | SD  | Number of samples |
|------|----------|-----|-------------------|
| AC   | -53.7    | 86.9| 178               |
| CC   | -7.1     | 30.4| 36                |
| CD   | 103.8    | 92.7| 154               |
| CE   | 96.8     | 94.7| 121               |
| CF   | -21.4    | 84.2| 85                |
| CH   | -12      | 93.3| 97                |
| CI   | 32.8     | 80.9| 136               |
| CK   | 50.2     | 93.9| 131               |
| CL   | 42.8     | 90.6| 224               |
| CM   | 61.9     | 100 | 39                |
| CN   | 63.7     | 96.1| 115               |
| CP   | 24.9     | 88  | 45                |
| CQ   | 66.7     | 92.1| 95                |
| CR   | 35.4     | 95.1| 144               |
| CS   | 106.3    | 98.1| 184               |
| CT   | 109.9    | 97.5| 173               |
| CV   | 54.9     | 90.7| 183               |
| CW   | -0.2     | 85.9| 43                |
| CY   | 8.5      | 91.5| 96                |
The misfold decoy set has ~20 protein structures, each of which has a correct and an incorrect structure specified (three structures have two incorrect structures: we randomly chose the first). The PDScore of these proteins were computed (Table 4). Barring three structures (PDBids: 1CBH, 1FDX and 2SSI), the PDScore of the incorrect structure is higher than that of the correct structures.

The hg_structal set has about ~30 proteins. Each protein has 30 structures (including the native structure). Table 5A shows specificity obtained for structures in this decoy set. The average specificity obtained for this decoy set is 0.91 (Table 5A). The decoy set 4state_reduced has ~600 structures for each of the seven proteins. We obtain an average specificity of 0.94 for this decoy set (Table 5B). Similarly, for the ig_structal decoy set we obtain a specificity of 0.93 (Supplementary Table 1).

Table 4. Misfold decoy set. This decoy set has ~20 protein structures - each of which has a correct and an incorrect structure specified. The PDBs are sorted based on the number of residues in the structure (NRes). Three of the structures (1CBH, 1FDX and 2SSI) have a lower PDScore for the incorrect structure.

| PDB | NRes | Correct PDScore | Incorrect PDScore | Specificity |
|-----|------|-----------------|-------------------|-------------|
| 1CBH | 36   | 18.7            | 12.6              | 0           |
| 1PPT | 36   | 18              | 33.5              | 1           |
| 1FDX | 54   | 33              | 30.9              | 0           |
| 5RXN | 54   | 25.1            | 35                | 1           |
| 1SN3 | 65   | 20.7            | 30.3              | 1           |
| 2CI2 | 65   | 19.9            | 35.2              | 1           |
| 2CRO | 65   | 26.7            | 43.4              | 1           |
| 1HIP | 85   | 19.1            | 36.8              | 1           |
| 2BS5 | 85   | 22.1            | 34.3              | 1           |
| 2CDV | 107  | 17.4            | 40.9              | 1           |
| 2SSI | 107  | 22.6            | 20                | 0           |
| 1BP2 | 123  | 21              | 44.1              | 1           |
| 2PAZ | 123  | 19.3            | 27.3              | 1           |
| 1P2P | 124  | 28.6            | 29                | 1           |
| 1RN3 | 124  | 20.7            | 28.8              | 1           |
| 1LH1 | 153  | 18              | 26.2              | 1           |
| 211B | 153  | 19              | 27.6              | 1           |
| 1REI | 212  | 16.5            | 21.8              | 1           |
| 5PAD | 212  | 18.8            | 32                | 1           |
| 1RHD | 293  | 23.2            | 31.9              | 1           |
| 2CYP | 293  | 21.2            | 35.8              | 1           |
| 2TMN | 316  | 27.2            | 32.2              | 1           |
| 2TS1 | 317  | 21.3            | 28.2              | 1           |

Table 3. Electrostatic potential differences (EPD) in a sample of consecutive residue pairs of Cβ atoms. These pairs are used for discriminating predicted structures in order to obtain the native structure. The complete set is available at https://github.com/sanchak/mqap.

| Pair | Mean EPD | SD | Number of samples |
|------|----------|----|-------------------|
| DF   | -108.9   | 29.5 | 481               |
| DY   | -107.4   | 30.7 | 442               |
| DH   | -105.2   | 33.5 | 242               |
| DW   | -104.1   | 27.7 | 209               |
| EH   | -98.5    | 28.5 | 200               |
| EY   | -96.5    | 28   | 378               |
| EW   | -94.2    | 29.8 | 184               |
| SY   | -93.5    | 27.5 | 403               |
| EF   | -93.1    | 27.6 | 439               |
| TY   | -93      | 28.6 | 384               |
| TW   | -90.8    | 28.7 | 144               |
| SW   | -89.2    | 27.7 | 169               |
| FT   | 89.2     | 26.8 | 436               |
| FS   | 92.3     | 28.4 | 453               |
| HS   | 93.7     | 31.8 | 235               |
| HT   | 95.1     | 31.5 | 235               |

Figure 4. Standard deviation (SD) of the electrostatic potential difference between Cβ atoms increases with increasing sequence distance for amino acid pairs. Each sequence distance has at least 30 sample points. DF: Aspartic-acid/Phenylalanine, HS: Histidine/Serine. As expected, there is lesser correlation in the EPD values between the shown amino acid pairs ‘DF’ and ‘HS’ as the sequence distance between the residues increases. The SD for distance 1 (i.e. consecutive residues) is 29.8 EPD units and 31.8 EPD units for ‘DF’ and ‘HS’, respectively - and rises to around 60 EPD units with increasing sequence distance.
Table 5. hg_structal and 4state_reduced decoy sets. The PDBs are sorted based on specificity. (A) The hg_structal decoy set has ~30 protein structures - each of which has 30 structures. The average specificity obtained for the set is 0.91. (B) The 4state_reduced decoy set has 7 protein structures - each of which has ~600 structures. The average specificity obtained for the set is 0.94. (C) The fisa set has 4 protein structures - each of which has 500 structures. The electrostatic discriminator has low specificities in this case. We have previously demonstrated that this decoy set can be discriminated by a distance based criterion. It consists of physically nonviable structures, thus rendering an electrostatic analysis meaningless. NRes = number of residues, NStructures = number of structures in the decoy set.

| PDB    | NRes | NStructures | Specificity |
|--------|------|-------------|-------------|
| 2PGHA  | 141  | 30          | 0.2         |
| 1MBS   | 153  | 30          | 0.5         |
| 2DHBA  | 141  | 30          | 0.6         |
| 1HDAB  | 145  | 30          | 0.9         |
| 1MYT   | 146  | 30          | 0.9         |
| 1HLM   | 158  | 30          | 0.9         |
| 1HSY   | 153  | 30          | 0.9         |
| 1MBA   | 146  | 30          | 0.9         |
| 1MYGA  | 153  | 30          | 0.9         |
| 1MYJA  | 153  | 30          | 0.9         |
| 1ASH   | 147  | 30          | 1           |
| 1BABB  | 146  | 30          | 1           |
| 1COLA  | 197  | 30          | 1           |
| 1CPCA  | 162  | 30          | 1           |
| 1ECD   | 136  | 30          | 1           |
| 1EMY   | 153  | 30          | 1           |
| 1FLP   | 142  | 30          | 1           |
| 1GDM   | 153  | 30          | 1           |
| 1HBG   | 147  | 30          | 1           |
| 1HBHA  | 142  | 30          | 1           |
| 1HBHB  | 146  | 30          | 1           |
| 1HDA   | 141  | 30          | 1           |
| 1HLD   | 153  | 30          | 1           |
| 1ITHA  | 141  | 30          | 1           |
| 1LHT   | 153  | 30          | 1           |
| 2DHBB  | 146  | 30          | 1           |
| 2LHB   | 149  | 30          | 1           |
| 2PGHB  | 146  | 30          | 1           |
| 4SDHA  | 145  | 30          | 1           |
| 2CRO   | 65   | 675         | 0.8         |
| 3ICB   | 75   | 654         | 0.9         |
| 4RXN   | 54   | 677         | 0.9         |
| 4PTI   | 118  | 688         | 1           |
| 1CTF   | 131  | 631         | 1           |
| 1R69   | 97   | 676         | 1           |
| 1SN3   | 65   | 661         | 1           |
| 4ICB   | 76   | 501         | 0           |
| 1FC2   | 44   | 501         | 0.4         |
| 1HDDC  | 57   | 501         | 0.1         |
| 2CRO   | 65   | 501         | 0.7         |
Discussion
The functional characterization of a protein from its sequence using in silico methods based on the ‘sequence to structure to function’ paradigm is contingent upon the availability of its 3D-structure. The rapidly developing field of next generation sequencing has exacerbated the bottleneck of obtaining structural data using crystallization techniques. This ever-widening gap has been filled by methods that predict structures from sequences, based either on features from databases of known structures or from first principles of atomic interactions.

The various sources of error in protein structure prediction have been previously discussed in detail. An incorrect model of a protein structure will result in an inaccurate analysis of its properties. For example, continuum models that compute potential differences and pK values from charge interactions in proteins are sensitive to the spatial arrangement of the atoms in the structure. It must be pointed out that other detailed methods using quantum mechanical/molecular mechanical (QM/MM) techniques and doing extensive conformational sampling have been able to determine the side chain pK values with high accuracy. Accurate structural information is indispensable for in silico methods that extract the electrostatic profile of atoms in the peptide structure, and for other methods widely used in pharmacology for drug discovery.

Model quality assessment programs (MQAP) that validate the accuracy of predicted structures are thus a critical aspect in the process of modeling a protein structure from its sequence. MQAPs can be classified as energy- or knowledge-based (statistical potential). The state of the art methods for predicting structures and MQAPs are evaluated by researchers every two years.

Previously, we hypothesized and demonstrated that the electrostatic potential difference (EPD) in cognate pairs in the active site are conserved in proteins with the same functionality, even when evolution has converged to the same catalytic from completely different sequences. This similarity is observed in structures solved independently over many years and establishes the reliability of the APBS and PDB2PQR implementations. We focused on unraveling other electrostatic features that are innate to protein structures. Here, we first demonstrate that the EPD between the C-N peptide bond and the Ca atoms of consecutive residues are independent of the amino acid type. This is expected, since the distance between these atoms is almost invariant across all structures. The EPD of the C-N bond has a high variance, implying that the backbone accommodates relatively large variations while seeking energetically minimized structures.

The true source of the chemical and structural diversity in protein structures is the side chain atoms. Every amino acid, except glycine, has a Cβ atom that hosts a unique moiety of atoms. Although the reactive groups are different for amino acids, we show that this difference is encapsulated in the backbone Cβ atoms. We first show that different pairs of amino acids have significantly different mean EPD values in side chain Cβ atoms (Figure 3), unlike the EPD of the C-N peptide bond (Figure 1) or the EPD between consecutive Ca atoms (Figure 2). Further, the variance is much less than in the EPD of the C-N bond. These facts suggested that the EPD between Cβ atoms of consecutive residues in the peptide structure might act as a discriminator. Our hypothesis is based on the insightful Boltzmann law that lower energy conformations are disproportionately sampled, on the thermodynamic hypothesis that the native structure has minimal energy, and the hypothesis that statistical derived features in known protein structures have a Gaussian distribution.

We apply our discriminator to standard decoy sets from the Decoys ‘R’ Us database to establish the validity of the method. Our work also highlights the unique properties of proline in the protein structure. This is evident from the different magnitude of EPD in consecutive Cα atoms involving proline (Table 1). Another noteworthy aspect is the large variation in EPD in consecutive Cβ atoms involving cysteine (Table 2), demonstrating the unique role of cysteine plays in providing flexibility to protein structures, a critical element in the evolution of complex organisms. The discrimination of Cβ atoms also provides a uniform basis for methods that require a single-atom representation of a residue. Such methods depend on a correct parameterization of the reactive atoms, a task further complicated by amino acids such as histidine which has two reactive atoms. For example, the EPD between the negatively charged E and D with respect to the aromatic phenylalanine is -108 and -93 EPD units, in spite of the difference in their reactive atom. Similarly, the EPD between alanine and the three aromatic amino acids (F, W and Y) are -67, -66 and -63 EPD units respectively.

We achieved an average specificity of 0.91, 0.94 and 0.93 on hg_structal, 4state_reduced and ig_structal decoy sets, respectively, taken from the Decoys ‘R’ Us database. We have previously demonstrated that the fisa decoy set can be discriminated by a distance based discriminator. ESCAPIST does not discriminate the native structure in this decoy set (Table 5C). The physical implication of ESCAPIST results on the fisa decoy set, which has significant RMSD for backbone Cα atoms, needs further elaboration. The input to a finite difference Poisson-Boltzmann (FDPB) analysis is a charge distribution that might be unfeasible due to energy functions other than electrostatics. For example, van der Waals force or the elastic bond length force components might prevent two atoms from being in close proximity. However, if such a physically impossible configuration were presented to a FDPB-based analysis tool, such as APBS, it would still generate an electrostatic potential landscape. Inferences based on this potential landscape would be incorrect due to its physical non-viability. Thus, before invoking the EPD constraints specified here, it is imperative that other spatial constraints that are rarely violated in structures are checked. Possibly for this same reason, MQAPs that combine many features in input to a finite difference Poisson-Boltzmann discriminator are prone to biases and possible errors. In fact, the fisa decoy set has been shown to violate the van der Waals term. To summarize, we present a novel discriminating feature in protein structures based on the electrostatic properties of the side chain atoms. We validated this discrimination in several decoy sets taken from the Decoys ‘R’ Us database, and achieved high specificities in most decoy sets.

Methods
Our proposed method has two phases. In the learning phase, we process multiple structures to extract the feature values - mean values...
of electrostatic potential difference (EPD) for each amino acid pair. These feature values are applied on query proteins to obtain a score (PDscore) that indicates the deviation of the feature values in the given structure from the ‘ideal’ values. Thus, a lower PDscore indicates a better candidate.

**Learning phase**

Algorithm 1 shows the procedure LearnFeatures() that extracts the feature values from a set of proteins \( \Phi_{\text{proteins}} \) (Equation 1). We ignore the first \( x=\text{IgnoreNTerm} \) and last \( y=\text{IgnoreCTerm} \) pairs of residues in the protein structure to exclude the terminals. For every consecutive pair of residues in the structure, we calculate the EPD (see below for method) between two specified atoms (\( \text{atomP} \) and \( \text{atomQ} \)). Both \( \text{atomP} \) and \( \text{atomQ} \) are set to \( C_{\beta} \) to obtain EPD values for \( C_{\beta} \) atoms, while we set \( \text{atomP} \) to ‘C’ and \( \text{atomQ} \) to ‘N’ in order to obtain the C-N peptide bond EPD values. The mean (Mean learnt value - MLV) and standard deviation (SD) are computed for each amino acid type pair (AAType) in protein (Equation 2), and the mean computed for the globals set of proteins (MLV(AAType')) for each pair of amino acid types (Equation 3). Pairs whose EPD have a SD greater than a threshold value (sdThresh, 50 by default) are ignored. Means for all significant pairs (\( \phi_{\text{pairMean}} \)) are noted to a file, which is the input to the quality assessment phase.

\[
\Phi_{\text{Learning Phase}}^{\text{proteins}} = \{P_1, P_2, \ldots, P_M\}
\]

\[
MLV(AAType^{Res_a}, AAType^{Res_{a+1}})_{Pi} = \sum_{n=1}^{N-y-1} \frac{EPD(Res_n(\text{atomP}), Res_{n+1}(\text{atomQ}))}{(N - y - x - 2)}
\]

\[
MLV(AAType', AAType') = \left\{ \forall i = 1 \ldots M, \text{and AAType'} \text{and AAType'} \right\} \frac{\sum_{i=1}^{M} (MLV(AAType', AAType')_{Pi})}{M}
\]

\[
\Phi_{\text{Assessment Phase}}^{\text{proteins}} = \{P_1, P_2, \ldots, P_N\}
\]

\[
PDscore^{P_i} = \sum_{n=x}^{N-y-1} \frac{Abs(EPD(Res_n(\text{atomP}), Res_{n+1}(\text{atomQ})) - MLV(AAType^{Res_a}, AAType^{Res_{a+1}}))}{(N - y - x - 2)}
\]

\[
[\forall i = 2 \ldots N] (PDscore^{P_1} < PDscore^{P_i})
\]
Algorithm 1. LearnFeatures(): extract electrostatic potential difference (EPD) values from a given pair of amino acids

Input: $\phi_{\text{proteins}} = \{P_1, \ldots, P_M\}$: M Proteins in the learning set
Input: IgnoreNTerm: Ignore this number of residues in the N Terminal
Input: IgnoreCTerm: Ignore this number of residues in the C Terminal
Input: atomP: Atom type in first residue
Input: atomQ: Atom type in second residue
Input: $sd\text{Thresh}$: Threshold for standard deviation of the EPD
Output: $\phi_{\text{pairMean}} = \{\text{meanPDC}_1, \ldots, \text{meanPDC}_K\}$: Mean values of EPD between specified atoms of successive residues, there being K such significant pairs

begin
/*K pairs of amino acid type (sorted: AC and CA are equivalent)*/
/*Each set is initialize to be the null set*/
$\phi_{\text{pair}} = \{\phi_{1\text{PDC}}, \ldots, \phi_{K\text{PDC}}\}$:
foreach $P_i$ in $\phi_{\text{proteins}}$ do
    $N = \text{NumberOfResidues}(P_i)$;
    for $p \leftarrow \text{IgnoreNTerm}$ to $(N - \text{IgnoreCTerm})$ do
        $q = p + 1$;
        /* Amino acid pairs are order independent */
        ResiduePairTypeString = ResidueTypeString($p$) + ResidueTypeString($q$);
        ResiduePairTypeStringSorted = Sort(ResiduePairTypeString);
        /* Reverse sign of potential difference accordingly */
        multfactor = 1;
        if (ResiduePairTypeStringSorted $\neq$ ResiduePairTypeString) then
            multfactor = -1;
        end
        $PD = \text{PotentialDifference}(p, q, \text{atomP}, \text{atomQ}) \times \text{multfactor}$;
        /* Let the amino acid pair be the kth in the set $\phi_{\text{pair}}$ */
        InsertInSet($PD, \phi_{k\text{PDC}}$);
    end
end
/* Compute Mean and SD of each set - ignore pairs with SD greater than $sd\text{Thresh}$*/
$\phi_{\text{pairMean}} = \emptyset$;
foreach $\phi_{\text{pair}}$ in $\phi_{\text{pair}}$ do
    $(\text{Mean}, SD) = \text{MeanAndSD}(\phi_{\text{pair}})$;
    if $(SD > sd\text{Thresh})$ then
        Add($\text{Mean}, \phi_{\text{pairMean}}$);
    end
end
return ($\phi_{\text{pairMean}}$);
Algorithm 2. AssessEPDQuality()

Inputs:
- $P_1$: Protein under consideration
- IgnoreNTerm: Ignore this number of residues in the N Terminal
- IgnoreCTerm: Ignore this number of residues in the C Terminal
- atomP: Atom type in first residue
- atomQ: Atom type in second residue
- $\phi_{pairMean} = \{\text{meanPDC}_1, \ldots, \text{meanPDC}_M\}$: Mean values of EPD between specified atoms of successive residues

Output: $PDscore$: Score indicating the normalized distance of the observed values from the (mean) learnt values from native structures

begin

PDscore = 0; NumberCompared = 0; N = NumberOfResidues($P_1$);
for $p \leftarrow$ IgnoreNTerm to $(N - \text{IgnoreCTerm})$ do

  $q = p + 1$;
  /* Amino acid pairs are order independent */
  ResiduePairTypeString = ResidueTypeString($p$) + ResidueTypeString($q$);
  ResiduePairTypeStringSorted = Sort(ResiduePairTypeString);
  /* Reverse sign of potential difference accordingly */
  multfactor = 1;
  if (ResiduePairTypeStringSorted != ResiduePairTypeString) then
    multfactor = -1;
  end

  /* Let the amino acid pair be the $k$th in the set $\phi_{pair}$ */
  PD = PotentialDifference($p$, $q$, atomP, atomQ) * multfactor;
  if ($\exists$meanPDC) then
    NumberCompared = NumberCompared + 1;
    diff = absolute(PD - meanPDC);
    PDscore = PDscore + diff;
  end
end

/* Normalize */
PDscore = PDscore/NumberCompared;
return (PDscore);
end
The top level procedure ESCAPIST() is shown in Algorithm 3. It invokes the function LearnFeatures() once, and applies the learnt values to assess the quality of structures based on the feature values obtained.

**Algorithm 3. ESCAPIST(): Top level function**

| Input | Description |
|-------|-------------|
| $\phi_{\text{proteins}}$ | Learning set |
| $P_1$ | Protein to be scored |
| IgnoreNTerm | Ignore this number of residues in the N Terminal |
| IgnoreCTerm | Ignore this number of residues in the C Terminal |
| atomP | Atom type in first residue |
| atomQ | Atom type in second residue |
| sdThresh | Threshold for standard deviation of the EPD |

| Output | Description |
|--------|-------------|
| PDscore | Score indicating the normalized distance of the observed values from the (mean) learnt values from native structures |

```
begin
/* This is invoked once*/ $\phi_{\text{pairMean}} =$
LearnFeatures($\phi_{\text{proteins}}$, IgnoreNTerm, IgnoreCTerm, atomP, atomQ, sdThresh);;
PDscore = AssessEPDQuality($P_1$, IgnoreNTerm, IgnoreCTerm, atomP, atomQ, $\phi_{\text{pairMean}}$);
return (PDscore);
end
```

Computing electrostatic potentials
Adaptive Poisson-Boltzmann Solver (APBS) and the PDB2PQR package was used to calculate the potential difference between the reactive atoms of the corresponding proteins. The APBS parameters are set as follows - solute dielectric: 2, solvent dielectric: 78, solvent probe radius: 1.4 Å, Temperature: 298 K and 0 ionic strength. APBS writes out the electrostatic potential in dimensionless units of $kT/e$ where $k$ is Boltzmann’s constant, $T$ is the temperature in K and $e$ is the charge of an electron.

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.

Acknowledgements
We are grateful to Mary Lou Mendum critical reading of the manuscript.
### Table S1. Proteins from the PISCES database used for learning values.

Set of proteins from the PISCES database with percentage identity cutoff of 20%, resolution cutoff of 1.6 Å, R-factor cutoff of 0.25, and a RDCC cutoff of 0.012 Å used to learn feature values.

| Protein ID | Chain | PDB Code |
|------------|-------|----------|
| 1A62       | 1A7H  | 1A4E     |
| 1A9H       | 1A9H  | 1ABF     |
| 1B5E       | 1B5E  | 1B8K     |
| 1B7X       | 1B7X  | 1C1K     |
| 1C4Q       | 1C4Q  | 1C5E     |
| 1C75       | 1C75  | 1C8C     |
| 1C8W       | 1C8W  | 1C8W     |
| 1D07       | 1D07  | 1D07     |
| 1D57       | 1D57  | 1D57     |
| 1E05       | 1E05  | 1E2W     |
| 1E5B       | 1E5B  | 1E5K     |
| 1E7L       | 1E7L  | 1E9Q     |
| 1E9W       | 1E9W  | 1E9W     |
| 1E9Z       | 1E9Z  | 1F1E     |
| 1F46       | 1F46  | 1F6B     |
| 1F8E       | 1F8E  | 1F94     |
| 1F94       | 1F94  | 1F94     |
| 1FK5       | 1FK5  | 1FSG     |
| 1FSG       | 1FSG  | 1FX2     |
| 1FX2       | 1FX2  | 1G2R     |
| 1G5A       | 1G5A  | 1G61     |
| 1GBQ       | 1GBQ  | 1GBQ     |
| 1GCI       | 1GCI  | 1GCI     |
| 1GJ7       | 1GJ7  | 1GJ7     |
| 1GK1       | 1GK1  | 1GK1     |
| 1GMU       | 1GMU  | 1GMU     |
| 1GXP       | 1GXP  | 1GXP     |
| 1GPQ       | 1GPQ  | 1GPQ     |
| 1GQ1       | 1GQ1  | 1GQ1     |
| 1GQ2       | 1GQ2  | 1GQ2     |
| 1GQ9       | 1GQ9  | 1GQ9     |
| 1GQI       | 1GQI  | 1GQI     |
| 1GRI       | 1GRI  | 1GRI     |
| 1GRF       | 1GRF  | 1GRF     |
| 1GSA       | 1GSA  | 1GSA     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSF       | 1GSF  | 1GSF     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
| 1GSG       | 1GSG  | 1GSG     |
| 1GSR       | 1GSR  | 1GSR     |
| 1GSI       | 1GSI  | 1GSI     |
Table S2. ig_structal decoy set. The PDBs are sorted based on specificity: The ig_structal decoy set has ~61 protein structures - each of which has 61 structures. The average specificity obtained for the set is 0.97. NRes =: number of residues, NStructures =: number of structures in the decoy set.

| PDB  | NRes | NStructures | Specificity |
|------|------|-------------|-------------|
| 1FPT | 11   | 61          | 0           |
| 1IKF | 233  | 61          | 0.5         |
| 1IGM | 115  | 61          | 0.9         |
| 1ACY | 211  | 61          | 1           |
| 1BAF | 214  | 61          | 1           |
| 1BBD | 219  | 61          | 1           |
| 1BBJ | 211  | 61          | 1           |
| 1DFB | 212  | 61          | 1           |
| 1DVF | 107  | 61          | 1           |
| 1EAP | 213  | 61          | 1           |
| 1FAI | 214  | 61          | 1           |
| 1FBI | 214  | 61          | 1           |
| 1FGV | 107  | 61          | 1           |
| 1FIG | 214  | 61          | 1           |
| 1FLR | 219  | 61          | 1           |
| 1FOR | 210  | 61          | 1           |
| 1FRG | 217  | 61          | 1           |
| 1FVC | 109  | 61          | 1           |
| 1FVD | 214  | 61          | 1           |
| 1GAF | 214  | 61          | 1           |
| 1GGI | 211  | 61          | 1           |
| 1GIG | 210  | 61          | 1           |
| 1HL  | 211  | 61          | 1           |
| 1HKL | 214  | 61          | 1           |
| 1IAI | 214  | 61          | 1           |
| 1IBG | 213  | 61          | 1           |
| 1IGC | 213  | 61          | 1           |

References

1. Zhang Y: Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008; 18(3): 342–348. PubMed Abstract | Publisher Full Text | Free 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. Roth CA, Strauss CE, Misura KM, et al.: Protein structure prediction using Rosetta. Methods Enzymol. 2004; 383: 66–93. PubMed Abstract | Publisher Full Text
7. Karplus K, Karchin R, Draper J, et al.: Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins. 2003; 53(Suppl 6): 491–496. PubMed Abstract | Publisher Full Text
8. Mirjalili V, Noyes K, Feig M: Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins. 2013. Publisher Full Text
10. 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.

11. Raval A, Piana S, Eastwood MP, et al.: Refinement of protein structure models via long, all-atom molecular dynamics simulations. Proteins. 2012; 80(6): 2071–2079.

12. 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.

13. Ginalska K, Elsösser A, Fischer D, et al.: 3D-Jury: a simple approach to improve force field predictions. Proteins. 2003; 51(3): 534–552.

14. Tersas G, Osowa 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): 1399–1395.

15. Walther B, Elsösser A: Prediction of global and local model quality in CASP7 using Poons and ProQ. Proteins. 2007; 69(Suppl 8): 184–193.

16. 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.

17. Krzyżtofowich A, Fidelis K, Tramontano A: Evaluation of model quality predictions in CASP9. Proteins. 2011; 79(Suppl 10): 91–106.

18. McQuillen L: Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics. 2007; 8: 345.

19. Tanaka S, Scharinger HA: Model of protein folding: Inclusion of short-, medium-, and long-range interactions. Proc Natl Acad Sci U S A. 1970; 72(10): 3805–3806.

20. Miyazawa S, Jemigan RL: Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 1985; 18(3): 534–552.

21. 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(3): 859–883.

22. Sippl MJ: Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995; 5(2): 229–235.

23. Zhou H, Zhou Y: Using Pcons and ProQ. Proteins. 2007; 65(3): 726–741.

24. Lu H, Skolnick J: A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001; 44(3): 223–232.

25. Thomas PD, Dill KA: Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol. 1996; 257(2): 457–469.

26. 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.

27. Benkert P, Tosatto SC, Schomburg D: GMEM: A comprehensive scoring function for model quality assessment. Proteins. 2008; 71(1): 261–277.

28. Tosatto SC: The victor/FRST function for model quality estimation. J Comput Biol. 2005; 12(10): 1316–1327.

29. Archie JG, Paluszewski M, Kulpke K: Applying Underiker to quality assessment. Proteins. 2009; 77(Suppl 9): 191–195.

30. Zhu J, Yan W, Hu G, et al.: SVR, CAF: an integrated score function for detecting native protein structures among decoys. Proteins. 2014; 82(4): 556–64.

31. Krzyżtofowich A, Monastyrsky B, Fidelis K: CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLLS. Proteins. 2013; 82(Suppl 2): 7–17.

32. Moult J: A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 2005; 15(3): 285–289.

33. Krzyżtofowich A, Barbato B, Fidelis K, et al.: Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins. 2014; 82(Suppl 2): 112–126.

34. Chakraborty S, Mindra R, Saliay 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.

35. Chakraborty S, Asgeirsson B, Mindra 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: 3710–3715.

36. Chakraborty S, Rao BJ, Baker N, et al.: Structural phylogeny by profile extraction and multiple superimposition using electrostatic congruence as a discriminator. In Silico Disord Proteins. 2013; 3: e25463.

37. Rendón-Ramírez A, Shukla M, Osa 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.

38. 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): 10937–10941.

39. Holm L, Sander C: Evaluation of protein models by atomic solvation preference. J Mol Biol. 2002; 319(1): 93–105.

40. Metzker ML: Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11(1): 31–46.

41. Lewis TE, Stillo E, Andreeva A, et al.: Genom3D: A UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic Acids Res. 2013; 41(Database issue): D498–507.

42. Homig B, Nichols A: Classical electrostatics in biology and chemistry. Science. 1995; 268(5214): 1144–1149.

43. Kamerlin SC, Haranczyk M, Warshel A: Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of pKa, redox reactions and solvation free energies. J Phys Chem B. 2009; 113(9): 1253–72.

44. Holl D, Fersht A: A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. J Comput Chem. 1991; 12(4): 435–445.

45. Ekins S, Mestre J, Testa B: In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol. 2007; 152(1): 21–37.

46. Helland R, Larsen RL, Asgiersson B: The 1.4 Å crystal structure of the large Vibrio sp. alkaline phosphatase. Biochim Biophys Acta. 2009; 1792(2): 297–308.

47. Rawlings ND, Barrett AJ: Evolutionary families of peptidases. Biochem J. 1993; 290(Pt 1): 205–218.

48. Dolinsky TJ, Nielsen JE, McCammon JA, et al.: PDB2PQR: an automated pipeline...
for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 2004; 32(Web Server issue): W665–667.

57. Anfinsen CB: Principles that govern the folding of protein chains. Science. 1973; 181(4096): 223–230. PubMed Abstract | Publisher Full Text

58. MacArthur MW, Thornton JM: Influence of proline residues on protein conformation. J Mol Biol. 1991; 218(2): 397–412. PubMed Abstract | Publisher Full Text

59. Miseta A, Csutora P: Relationship between the occurrence of cysteine in proteins and the complexity of organisms. Mol Biol Evol. 2000; 17(8): 1232–1239. PubMed Abstract

60. Chakraborty S, Rao BJ, Asgeirsson B, et al.: Protein structure quality assessment based on the distance profiles of consecutive backbone Cα atoms. [v1; ref status: awaiting peer review. http://f1000r.es/1tg]. F1000 Research. 2013; 2: 211. Publisher Full Text

61. 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
Open Peer Review

Current Referee Status: ✔ ✔

Version 2

Referee Report 18 July 2014

doi:10.5256/f1000research.3039.r5376

Patricia C Weber
Implex LLC, Bristol, PA, USA

The quality of protein structure models is assessed by the geometry of adjacent C beta atoms. The approach successfully distinguishes properly folded proteins in most cases. It adds a new way to assess protein models and could be included in the protein model assessment toolbox.

Just a few comments:

a) Table 4 lists three of 20 structures where the incorrect one has a lower score. A few comments about structural features of those examples that lead unexpected scores would be useful.

b) It might be preferable to note in the title that the method is applied to computational models of protein structure as a way to distinguish the manuscript from those that deal with quality assessment of experimentally determined structures.

c) I did not test the available source code.

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 15 Sep 2014

Sandeep Chakraborty, Tata Institute of Fundamental Research, India

We would like to thank you for taking the time and reviewing our paper, and deeply appreciate your positive comments. We do not have the expertise to comment on the structural characteristics that lead to the 3 negative results in Table 4 - these require insights into crystal structures that we do not possess at the present time. Also, we believe that this method could be applied to any structure - computational or experimentally obtained - and thus are leaving the title unchanged.

Competing Interests: No competing interests were disclosed.
Shina Caroline Lynn Kamerlin  
Computational and Systems Biology, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden  

This is an interesting idea that uses the physical (electrostatic) properties of amino acid side chains in order to predict secondary structure from sequence, and thus assess (and rank) the quality of protein structures. The manuscript is well-written, and the authors provide comprehensive information to allow others to follow the study-design and methodology. The focus on electrostatics is important as this has been repeatedly shown by rigorous theoretical studies (work by Warshel and others) to be the primary driving force in determining protein function and most likely folding stability as well (whether directly or indirectly). As the specific methodology the authors use is slightly further from my area of expertise I cannot directly comment on this, however, importantly the source code has been made Open Access and freely available through Github.

My only comment is on the second paragraph of the Discussion, which comments on the pitfalls when using an incorrect model of a protein structure, particularly when trying to calculate $pK_a$s using continuum models which are dependent on the initial conformation. While the authors highlight a very important challenge, I would like to point out that it can to some extent be resolved by extensive conformational sampling (particularly the $pK_a$ problem, as the $pK_a$ is an average property over all possible protein conformations), which we discussed at length in a review a few years ago (Kamerlin et al., 2009). Otherwise this is a nice manuscript and a valuable contribution to the literature.

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.