



# A simple and efficient statistical potential for scoring ensembles of protein structures

Pilar Cossio<sup>1</sup>, Daniele Granata<sup>1</sup>, Alessandro Laio<sup>1</sup>, Flavio Seno<sup>2</sup> & Antonio Trovato<sup>2</sup>

<sup>1</sup>SISSA - Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 256, I-34136 Trieste, Italy, <sup>2</sup>Dipartimento di Fisica e Astronomia-Universita' di Padova e CNISM-Unita' di Padova, Via Marzolo 8, I-35131 Padova.

SUBJECT AREAS:

BIOINFORMATICS

COMPUTATIONAL BIOLOGY

STRUCTURAL BIOLOGY

PROTEINS

Received  
5 January 2012

Accepted  
14 March 2012

Published  
3 April 2012

Correspondence and  
requests for materials  
should be addressed to  
A.L. (laio@sissa.it)

In protein structure prediction it is essential to score quickly and reliably large sets of models by selecting the ones that are closest to the native state. We here present a novel statistical potential constructed by Bayesian analysis measuring a few structural observables on a set of 500 experimental protein structures. Even though employing much less parameters than current state-of-the-art methods, our potential is capable of discriminating with an unprecedented reliability the native state in large sets of misfolded models of the same protein. We also introduce the new idea that thermal fluctuations cannot be neglected for scoring models that are very similar to each other. In these cases, the best structure can be recognized only by comparing the *probability distributions* of our potential over short finite temperature molecular dynamics simulations starting from the competing models.

Knowledge based potentials (KBP)<sup>1–4</sup> are energy functions derived from databases of known protein conformations that empirically aim to capture the most relevant aspects of the physical chemistry of protein structure and function. They are derived by measuring the probability of an observable in an ensemble of experimental structures relative to a reference state<sup>2,6,16–18</sup>. The conversion of the probability into an energy function is normally done employing Boltzmann's law<sup>3</sup>. Theory of conditional probabilities<sup>5,6</sup>, linear and quadratic programming<sup>7</sup> and information theory<sup>8</sup> have been invoked to justify the approach. The simplest observable<sup>1,2</sup> that one can use to characterize a structure is the presence of a contact between two specific residues. This procedure has been generalized to include more and more complex observables<sup>3,6,7,9–15</sup>, making the KBPs more and more accurate. Because of their excellent balance between accuracy and computational efficiency, KBPs are widely used in protein design<sup>19</sup>, in simulation of protein folding<sup>20</sup>, binding<sup>21</sup> and aggregation<sup>22</sup>, in protein structure prediction<sup>23</sup> and in fold recognition<sup>24</sup>.

The quality of knowledge-based potentials is normally benchmarked on their capability of recognizing the folded state and/or the best model in a set of decoys<sup>5,25–27</sup>. It is now believed that decoy sets composed by models submitted during the biennial CASP (Critical Assessment in Structure Prediction) competition<sup>28</sup> are the most challenging<sup>29</sup>. Although it is not trivial to definitely assess their absolute efficiency, many KBP<sup>7,15,17,30–34</sup> perform quite well. Rosetta<sup>10,27</sup>, a scoring function derived using an elegant Bayesian analysis, the composite scoring function QMEAN6<sup>35,36</sup> and the potential RF\_CB\_SRS\_OD introduced by Rykunov and Fiser<sup>18</sup> are particularly successful, even when tested on CASP targets<sup>18,28</sup>. However, even the best performing KBP is not capable of distinguishing the folded state in all the decoy sets, indicating that improvements are still possible. Moreover, the state-of-the-art KBPs exploit many complex observables, such as the distance between pair of residues or atoms, which significantly boosts the number of parameters. The weights of the various terms (which may include correlated observables, such as value of the torsions and the presence of secondary structure elements) are *optimized* on the decoy sets in order to obtain the best performance<sup>10,35</sup>. This may affect the robustness of the potentials and the possibility to use them for different purposes other than fold recognition.

We here introduce a statistical potential that we call BACH (Bayesian Analysis Conformation Hunt). Its definition relies on a few binary structural observables, such as the presence of short range contacts between pair of residues, in the spirit of the original works<sup>1,2</sup> on KBPs. As a consequence, the energy function depends on only 1091 parameters, much less than other state-of-the-art potentials. The parameters are derived from a relatively small set of experimental folded structures (TOP500) but do not vary significantly when computed over a larger dataset, revealing a remarkable robustness. Moreover (and possibly more importantly) the parameters are not optimized on any decoy set. This potential, despite of its relative simplicity, has a remarkable



capability in recognizing the folded state in the CASP sets. By assigning to it the lowest energy value, BACH discriminates the native conformation in 58 % of the sets we considered. In all the other cases the native structure is among the best. In the worst case, it ranks #35 in a decoy set with 248 structures. This performance is significantly better than that of other potentials, and remains remarkably unaffected for different sets of alternative BACH parameters, derived either from a subset of TOP500 consisting of only 50 randomly chosen structures, or from TOP500 with a reduced 9 class representation of aminoacids<sup>37</sup>. We also found that BACH performs marginally better than other competing potentials in finding a model as close as possible to the native state within the pool of the few lowest energy decoys.

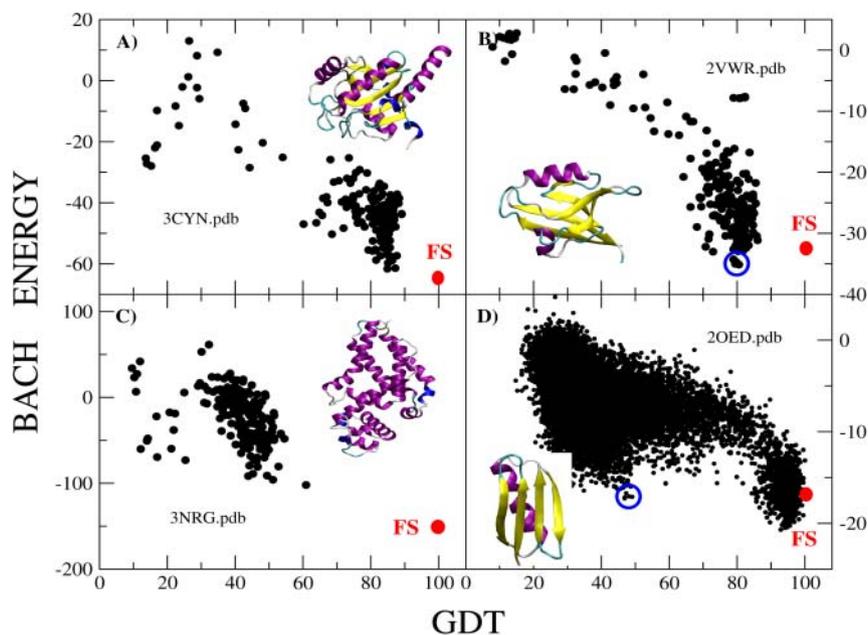
We finally investigated the reason of the poorer performance of BACH on some decoys sets. By analyzing ensembles generated by molecular dynamics from a single model we find that the BACH energy of thermodynamically equivalent structures is affected by large fluctuations. In fact, fluctuations of similar amplitude are also present among different models of the same protein obtained by NMR. Based on these observations, we propose that computing a scoring function on single structures may not be always adequate for discriminating the best. A more reliable quality measure is the *probability distribution* of a scoring function computed in a finite temperature molecular dynamics run, rather than a single value. Applying this idea we are able to discriminate the native state also in cases where evaluating the scoring function on single structures leads to ambiguous results.

## Results

The main goal of this work was to develop an efficient and physically inspired knowledge based potential which is able to discriminate the native conformation from a set of different structures with the same primary sequence. This potential is constructed by analyzing a set of 500 experimentally resolved protein structures (see Methods), monitoring the probability of the event that single residues or residue pairs are observed in one of the structural classes defined as follows. Each residue is classified as solvent exposed or not by computing its

solvent accessible surface<sup>38</sup>. Every pair of residues is classified in one out of five mutually exclusive states: (i) as forming an  $\alpha$ -bridge; (ii) as forming an anti-parallel  $\beta$ -bridge; (iii) as forming a parallel  $\beta$ -bridge; (iv) as forming a contact through their side chain atoms; (v) as not realizing any of the previous four conditions. All classifications are verified considering the full atomic configuration of the system (see Methods). Following a Bayesian analysis procedure, the 1090 energy parameters associated to event occurrences are estimated from their observed probabilities. An extra parameter,  $p$ , fixes the relative weight of the two contributions, the solvation and the pairwise. This parameter is determined ( $p = 0.6$ ) by monitoring the fluctuations of the two components on the TOP500 dataset as described in Methods. These 1091 parameters define the BACH statistical potential. Moreover, we have validated the robustness of our approach by checking that the performance remains remarkably unaffected for different sets of alternative BACH parameters (see Methods).

**Discriminating the native conformation.** We have calculated the BACH energy over various decoy sets, and checked its ability of assigning the lowest energy value to the native conformation. In figure 1 we show the BACH energy versus the distance in GDT<sup>39</sup> to the native structure for four decoy sets: A) T0388 from CASP8 (pdb code 3CYN), B) T0488 from CASP8 (pdb code 2VWR), C) T0575 from CASP 9 (pdb code 3NRG), D) 15561 decoy structures of protein GB3 generated by molecular dynamics and bias exchange metadynamics (details of the simulations are presented in Methods). One can see that in cases A) and C) BACH is able to discriminate the native structure from the other structures by assigning to it the lowest energy value. Case B) is the CASP8-9 decoy set in which BACH performance is the worst. The native conformation is ranked #35 out of 248 structures (14%). In the case of the structures generated by molecular dynamics (figure 1-D) the native conformation is in the best 0.5%, ranked #86 out of 15561 structures. In all cases, the closer a decoy structure is to the native state (as quantified by a large GDT), the lower is, on average, its BACH energy, thus showing a funnel-like behaviour. This is a first qualitative indication that BACH might be a rather powerful tool for protein structure discrimination. In the following, we will provide quantitative



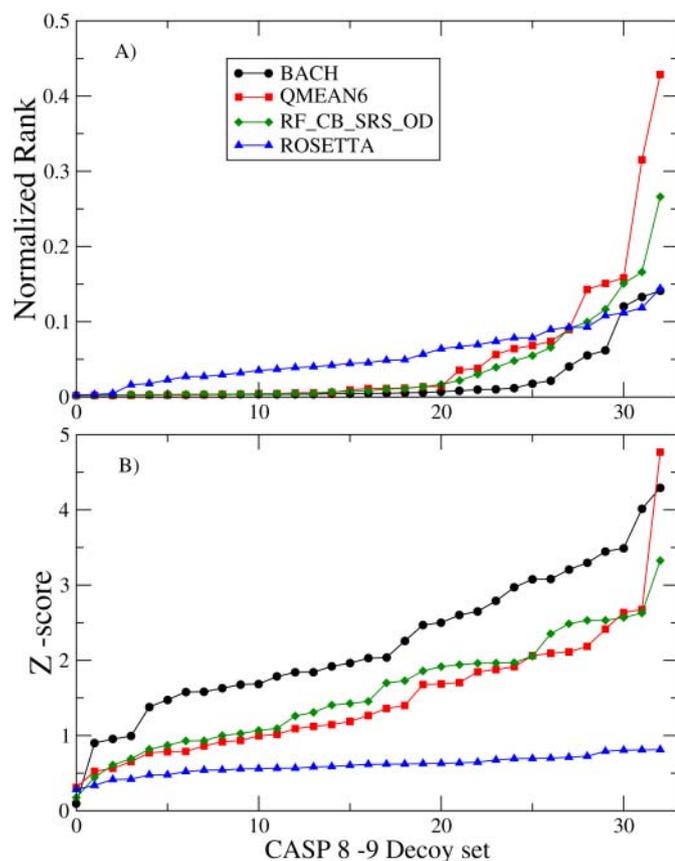
**Figure 1** | The BACH energy as a function of the GDT with respect to the native structure, for three decoy sets of CASP8-9: A) T0388 (native pdb code 3CYN), B) T0488 (pdb code 2VWR), C) T0575 (pdb code 3NRG), and D) for a decoy set of 15561 structures of protein GB3 (pdb code 2OED) generated by molecular dynamics simulations and bias exchange metadynamics. Red points: native conformation. Blue circles: structures with low GDT (distant from the folded state) and with a BACH energy lower than the native state.



evidence supporting this claim and compare the performance of BACH with other knowledge based potentials.

**Comparison with other knowledge-based potentials.** We compare the performance of BACH with QMEAN6<sup>35,36</sup>, ROSETTA<sup>5,10,27</sup> and RF\_CB\_SRS\_OD<sup>18</sup>. These potentials have been shown to perform extremely well in decoy discrimination<sup>18,28</sup>. The comparison is performed on a subset of CASP8-9 targets (see Methods), as it has been previously shown that these decoys are the most challenging<sup>29</sup>. We used only sets in which the decoy structures have the same sequence as the native structure, and that contain the coordinates of all the heavy atoms (see Table S1). This greatly reduces the number of targets, but allows comparing structures that are strictly chemically equivalent, removing a source of systematic errors.

The first measure of performance we use is the *normalized rank*, defined as the rank of the native structure divided by the total number of structures in the decoy set. For example, if the native structure has the lowest (resp. highest) possible energy among a set of 100 structures, its rank will be 0.01 (resp. 1). In figure 2-A we show the normalized rank for the decoy sets in CASP 8/9 (see Methods) for BACH, QMEAN6, RF\_CB\_SRS\_OD and ROSETTA. Strikingly, BACH has the lowest sorted rank, namely the best performance in discriminating the native structure from the decoys. BACH ranks the native within the best 5% for 28 decoy sets, whilst QMEAN6 does it for 23, RF\_CB\_SRS\_OD for 25, and ROSETTA for 19 out of a total of 33 sets. Moreover, for 19 sets BACH ranks the native structure as the first, whilst QMEAN6 does this for 14 cases, RF\_CB\_SRS\_OD for 13 and ROSETTA for 3. These results show that BACH is able to discriminate accurately the native structure, assigning to it, in many cases, the lowest energy value.

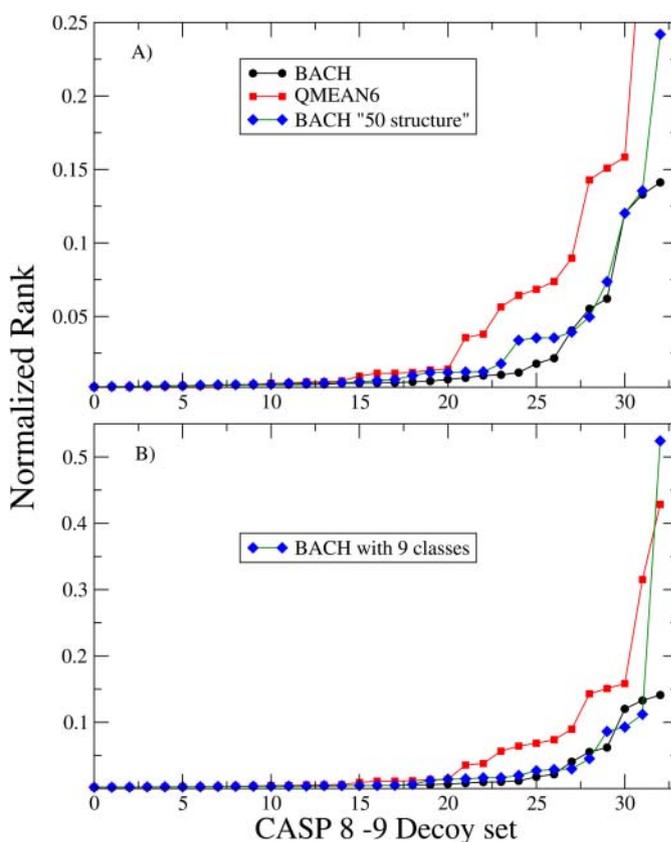


**Figure 2** | A) Normalized rank and B) Z-score sorted for the decoy sets in CASP 8–9, and calculated for the BACH, QMEAN6, RF\_CB\_SRS\_OD and ROSETTA scoring functions.

Another standard measure to characterize the performance of a scoring function is the Z-score, defined as the distance, measured in standard deviations, of the energy of the native state from the mean energy of the set. The larger the Z-score, the better the potential is in discriminating the native structure. In figure 2-B, the Z-scores for the decoy sets in CASP 8/9 are shown for each potential. As one can see, BACH has, in almost all cases, the largest Z-score values. Also for the Z-score, QMEAN6 has the second best performance, while RF\_CB\_SRS\_OD and ROSETTA perform more poorly.

BACH performance in native state discrimination is largely unaffected by varying the relative weight  $p$  of the solvation term with respect to the pairwise term by a factor of 2 (see Eq. (1)). For both  $p = 0.3$  and  $p = 1.2$  we still rank the native structure first in 17 cases out of 33. We performed even more stringent tests on the robustness of the BACH potential. First, we rederived the 1090 BACH energy parameters on a much smaller subset of the TOP500 dataset, consisting of 50 randomly chosen structures. Secondly, we derived 243 reduced BACH energy parameters, from the TOP500 dataset, implementing a 9 classes residue alphabet<sup>37</sup> (see Methods). In both cases we kept  $p = 0.6$ . As is shown in Fig. 3, the performance in properly ranking native structures is strikingly comparable, even though slightly worse, to the one obtained with the BACH energy parameters constructed with 20 aminoacid classes from the full TOP500 dataset.

In order to quantify the “funnelness” we also considered the Pearson correlation coefficient between the score and the GDT with respect to the native structure. This quantity measures how large are the fluctuations of a set of points with respect to a linear fit of the data. Its absolute value for the 33 decoy sets and for the four scoring functions we considered is shown in Fig. 4-A. For this specific quantity, QMEAN6 performs better than BACH, namely it produces on



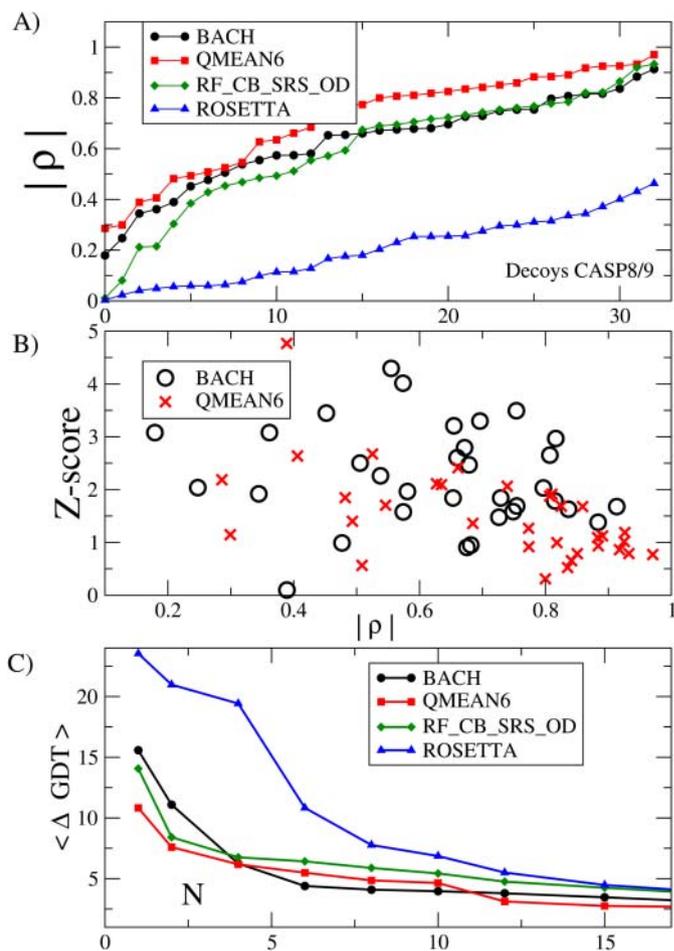
**Figure 3** | Normalized rank sorted for the decoy sets in CASP 8–9, calculated for the BACH, QMEAN6, and for alternative implementations of the BACH scoring functions: A) BACH parameters derived from 50 randomly chosen structures of TOP500 B) reduced BACH parameters for 9 aminoacid classes. The black curve is the same as in Fig. 2-A.



average a more linear dependence of the scoring function on GDT. However, the highest Pearson coefficient is observed in the decoy sets in which the native state is poorly discriminated. This is shown in figure 4-B, where we plot the absolute value of the Pearson correlation coefficient versus the Z-score for BACH and QMEAN6 for the 33 decoy sets of CASP 8/9. Strikingly, in almost all the cases in which the Pearson coefficient is higher than 0.8 the Z-score is below 1.5, indicating that a scoring function producing a good linear correlation is normally less capable of distinguishing the native state. This could indicate that the profile of the energy score as a function of the GDT to the native is not necessarily best fitted to a linear curve for a funnel-shaped landscape. This idea is consistent with previous theoretical arguments<sup>40</sup>.

We also benchmarked the capability of BACH in recognizing models that are close to the native structure. This is relevant, as KBPs are normally used in computational procedures aimed at iteratively finding conformations that are closer and closer to an unknown folded state. As we will show quantitatively later, estimates of a scoring function on a single structure are unavoidably affected by fluctuations, that hinder their statistical significance. If in an algorithm one always selects only the lowest energy structure, there would be high chances of missing a relevant “branch” of native-like conformations. In order to quantify this effect, we consider the  $N$  lowest energy structures, and we select among them the one with the largest GDT. We compute the difference  $\Delta_N\text{GDT}$  between this value and the GDT of the best model in all the decoy set, namely the decoy with the largest GDT. If the best model is one of the  $N$  lowest energy structures  $\Delta_N\text{GDT} = 0$ . In figure 4-C, we report  $\Delta_N\text{GDT}$  as a function of  $N$  averaged over the 33 CASP8-9 decoy sets. For  $N=1$ , only the lowest energy structure is selected, providing a quality measure similar to the one used in ref.<sup>18</sup>. For BACH and QMEAN6 the average  $\Delta_1\text{GDT}$  is 16 and 11 GDT units, respectively. This indicates that the top scoring QMEAN6 model is on average 5 GDT units closer to the best model in the set. For increasing values of  $N$ , the average  $\Delta_N\text{GDT}$  decreases rapidly down to  $\sim 5$  for all scoring functions, indicating a significant gain in information achieved by considering 5–10 alternative structures. In this range of  $N$ , the performance of BACH becomes better than that of the other potentials. Adding more and more structures reduces only marginally the average  $\Delta_N\text{GDT}$ , but with the drawback that in a hypothetical algorithm one would have to consider a very large number of branches which would increase the computational cost.

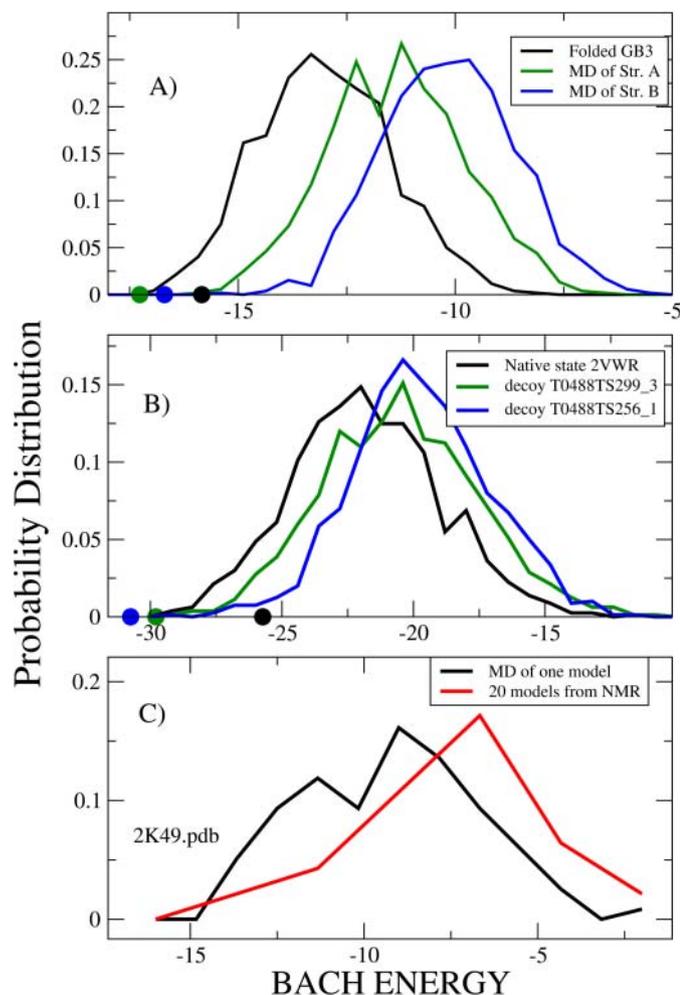
**The performance of BACH on traditional decoy sets.** We also benchmarked the performance of BACH over decoy sets that are considered the standard ones to test the scoring functions: semfold<sup>26</sup>, 4state<sup>25</sup>, fisa<sup>5</sup> and RosettaAll<sup>27</sup>. Since these decoy sets include a large number of targets satisfying the aforementioned condition of chemical equivalence, we use these sets to assess how BACH’s performance is affected by the presence of bound partners (e.g. ligands or other protein chains) in the native conformation. As supplementary information we present the normalized rank of the folded state, and compare BACH’s performance with the other potentials. In figure S1-B we have selected, from these sets, only the decoys that have a monomeric native structure with no ligands. The performance of BACH remains marginally better than the other KBPs also in this case. In particular, the normalized rank of BACH is smaller than 0.2 in all the cases. For 12 sets BACH ranks the native structure as first, similarly to QMEAN6, whilst RF\_CB\_SRS\_OD does it for 2 cases and ROSETTA for 8 out of a total of 19 sets. In figure S1-A, we plot the normalized ranking over all the sets, also those including target structures that are multimeric or have ligands. Even if these decoy sets include cases in which the native state is not monomeric, BACH still performs at least as good as other competitors. We also see that the other scoring functions, in



**Figure 4** | A) Absolute value of the Pearson correlation coefficient sorted for the decoy sets in CASP8-9, for the BACH, QMEAN6, RF\_CB\_SRS\_OD and ROSETTA scoring functions. B) Absolute value of the Pearson correlation coefficient versus the Z-score for BACH and QMEAN6 for the decoys in CASP 8-9. C) Difference in GDT between the structure in each set which is closest to the folded state and the low energy representative structure (the decoy that is ranked within the first  $N$ , according to each potential, and that has the highest GDT).

particular ROSETTA, perform much better with these decoys than with the CASP decoy sets, consistently with what is found in Ref.<sup>29</sup>.

**The role of fluctuations.** The results presented in Fig. 1 and 2 demonstrate the excellent capability of BACH in distinguishing the folded state. However, in the decoy sets we considered there are still several cases in which the native structure does not have the lowest BACH energy. Two examples are shown in Fig. 1-B and 1-D. For instance, for the decoys of GB3 generated by molecular dynamics and bias exchange metadynamics there are several structures that have high GDT and a BACH energy lower than the native structure. Even some structures of low GDT have a remarkably low BACH energy (blue circle in Fig. 1-D). These decoys have a different tertiary arrangement with respect to the native structure. One wonders whether these structures are a signal of a flaw in the BACH potential, that it is not able to recognize them as misfolded states. In order to investigate this point, we selected two structures enclosed in the blue circle in Figure 1-D and we performed for each of them a finite temperature molecular dynamics (MD) simulation in explicit solvent (see Methods). We compare the results with an ensemble of structures obtained with a similar run performed starting from the native state. As found in Fig. 5 all the MD simulations show high



**Figure 5** | A) BACH energy distributions for sets of structures generated with molecular dynamics simulations for GB3 protein (see Methods): the folded state in explicit solvent (black line), two structures with initial low BACH energy (green and blue lines, respectively). Points: initial conformations. B) 2VWR protein (set T0488 of CASP8) 1ns of MD simulations for: the native conformation (black line), decoy structure T0488TS256\_1 (blue line) and decoy structure T0488TS299\_3 (green line). Points: initial conformations. C) BACH energy distribution for the 20 NMR models of protein 2K49 (set T0472 of CASP8), and for the structures obtained by 1ns of MD simulations for model #1.

thermal fluctuations in BACH energy, of the order of 7 or more. Remarkably, the structures generated from the folded state show a BACH energy distribution (Fig. 5-A) clearly different from those obtained for the other ensembles. This demonstrates that the low BACH energy structures enclosed by the blue circle in Figure 1-D do not correspond to meaningful low energy conformations.

To confirm and exploit further this idea we considered the CASP8-9 decoy set in which BACH has the worst performance (T0488, where the native was ranked #35 out of 248 structures, see Fig. 1-B). Following a procedure similar to the one described above, we generated by MD simulations three sets of structures (see Methods for details) starting from the native state and from two decoy structures with lower energy than the native state (T0488TS256\_1 and T0488TS299\_3, ranked according to BACH #1 and #4, respectively). These structures have a high GDT score (GDT > 80.0) and therefore a high topological similarity with the folded conformation. In Fig 5-B we compare the BACH energies distributions for the three sets of structures. The ability of BACH

in distinguishing the native state from both decoys is confirmed; the discrimination is less sharp than in the case of GB3, because of the higher similarity of decoys to the native state. In conclusion, the failure of BACH in pinpointing the native state in the original decoy set can be removed by looking at the appropriate energy distribution over a finite temperature run.

The fluctuations we observe in Fig. 5-A/B are not artifacts of the simulation, but are also present in experimental data. This is shown in Fig 5-C, where we compare the probability distribution of BACH energy observed for the 20 native NMR models of the T0472 CASP8 decoy set (pdb code 2K49) with that computed over a finite temperature run starting from one of such models. These results confirm what we discussed previously, namely that the energy value of a single structure is not always significant, as it is affected by thermal fluctuations. A more reliable quality measure of a model is the probability distribution of the BACH energy computed on an ensemble of structures that can be obtained by means of a finite temperature molecular dynamics run.

## Discussion

We have developed a knowledge based potential, named BACH, that employs in its definition only a few binary structural observables, monitoring the presence of contacts, like originally proposed more than thirty years ago<sup>1,2</sup>. A novel ingredient in BACH is splitting residue-residue contacts, within the same Bayesian framework, in those present within  $\alpha$ -helices or  $\beta$ -sheets or instead in those that are not involved in secondary structure motifs. The former are monitored by verifying the presence of hydrogen bonds between backbone atoms, the latter by assessing the proximity between side chain heavy atoms. This last choice of defining non-secondary structure contacts by ignoring backbone atoms has been already recognized to improve crucially the performance of a scoring function<sup>41</sup>. A second innovative aspect of BACH is the evaluation, again within a Bayesian framework, of the propensities of single residues to be buried within the interior of the protein globule or to stay exposed on its surface.

BACH's performance has been benchmarked against other state-of-the-art potentials on the most challenging decoy sets. We used indicators that assess either how the native structure is recognized or how different models are sorted based on their similarity to the native state. BACH performs better than other scoring functions in discriminating the native structure and in assigning to it the largest energy gap with respect to the mean energy of the decoy set. This is a striking result since the parameters employed by BACH are fewer in comparison with other KBPs, and no optimization on decoy sets was used to boost the performance. The latter is moreover maintained if the number of parameters is further reduced by employing a reduced 9 classes aminoacid alphabet<sup>37</sup> or if the parameters are derived from just 50 structures, demonstrating the remarkable robustness of our potential.

We then benchmarked the capability of BACH in producing “funnel-shaped” distributions. QMEAN6 performs better than BACH in generating a linear correlation of the score with the GDT from the native state, as quantified by the Pearson coefficient. However, the highest Pearson coefficient is observed in the decoy sets in which the native state is poorly discriminated.

We have also shown that the 5-10 lowest energy decoys selected by BACH are likely to include better models of the native structure than other competing potentials. The procedure of searching for the best native-like model in a pool of low energy conformations is justified by the observation that the rankings may be affected by the presence of thermal fluctuations. Therefore, considering only the lowest energy structure might lead to inaccurate modeling. In fact, we have shown that thermal fluctuations are present with the same magnitude both among the different NMR models of the same protein and within a structural ensemble obtained by finite temperature molecular dynamics simulations. As a consequence, we propose that



an appropriate strategy to rank two structures requires to run a short MD simulation starting from each structure, in order to produce ensembles of thermodynamically equivalent conformations. We have shown that the probability distribution of the BACH energy estimated on the two ensembles allows discriminating correctly the two models. This procedure can be applied successfully not only for models that are very dissimilar from the native state but also for those which are close to it. This suggests that BACH might be a valuable tool for the refinement of protein structures roughly determined by experimental or numerical methods.

## Methods

**BACH parameters.** The BACH energy function is based on two terms

$$E_{\text{Bach}} = pE_{\text{PAIR}} + E_{\text{SOLV}}. \quad (1)$$

Both are statistical potentials derived from a dataset of native PDB structures. They take respectively into account effective pairwise residue-residue interactions and single residue solvation properties.  $p$  is a parameter that fixes the relative units of the two energy terms.

The pairwise statistical potential  $E_{\text{PAIR}}$  is based on classifying all residue pairs within a protein structure in five different structural classes, labeled by  $x$ : two residues may form a  $\alpha$ -helical bridge ( $x = 1$ ), or an anti-parallel  $\beta$ -bridge ( $x = 2$ ), or a parallel  $\beta$ -bridge ( $x = 3$ ), or may be in contact with each other through side chain atoms ( $x = 4$ ), or may not realize any of the previous four conditions ( $x = 5$ ). Pairs of consecutive residues along the protein chain are included in the above classification.  $\alpha$ - and  $\beta$ -bridges are detected by using a modified version of the DSSP algorithm<sup>42</sup> that employs a more stringent energy threshold ( $-1$  Kcal/mol in place of the original  $-0.5$  Kcal/mol, as already done in<sup>43</sup>) to assess hydrogen bond formation. A residue pair is assigned to the side chain - side chain contact class ( $x = 4$ ) if it is not assigned to any of the previous classes ( $x < 4$ ) and if any inter-residue pair of side chain heavy atoms is found at a distance lower than  $4.5\text{\AA}$ . If none of the above conditions are verified, the residue pair is assigned to the no-interaction class ( $x = 5$ ). The pairwise statistical potential  $E_{\text{PAIR}}$  requires five distinct symmetric matrices  $\epsilon_{ab}^x$ , where  $a$  and  $b$  vary among the 20 aminoacid types, for overall 1050 parameters:

$$E_{\text{PAIR}} = \sum_{i < j} \epsilon_{a_i a_j}^{x_{ij}} \quad (2)$$

$a_i$  is the aminoacid type of residue at position  $i$ , and  $x_{ij}$  is the structural class to which residue pair  $i - j$  is assigned.

The five interaction matrices  $\epsilon_{ab}^x$  are determined from a dataset of native protein structures employing the ensemble of all residue pairs from the dataset as the reference state<sup>6</sup>:

$$\epsilon_{ab}^x = -\ln \left[ \frac{\sum_x n_{ab}^x}{\sum_x \sum_{a,b} n_{ab}^x} \right] \quad (3)$$

where  $n_{ab}^x$  is the total number of residue pairs of type  $a$  and  $b$  found in the structural class  $x$  within the dataset. In order to treat coherently as a unique physical system the full sequence which determines a given protein structure, residues that are not seen in the PDB structure, even if they are part of the sequence expressed by the crystallographer, are considered as disordered ones and included in the countings, so that any pair involving a disordered residue is classified in the no-interaction class ( $x = 5$ ).

The solvation statistical potential  $E_{\text{SOLV}}$  is based on classifying all residues in two different environmental classes, either buried ( $b$ ) or solvent exposed ( $s$ ). The environmental class is defined based on the evaluation of the solvent accessible surface area (SASA) performed by the SURF tool of VMD graphic software<sup>44,45</sup>. The SASA is computed by SURF for all heavy atoms of the protein chain by rolling a probe sphere (representing a water molecule) on the surface of the set of spheres centered at heavy atom coordinates. We input to SURF the same value ( $1.8\text{\AA}$ ) for the radii of all atom types and the radius of the probe sphere. The latter is higher than what is employed in VMD ( $1.4\text{\AA}$ ) because we want to avoid considering internal cavities as areas exposed to the solvent. The output of SURF is the number of triangle vertices associated to each atom of the protein. These vertices are used in the triangulated representation of the protein surface employed by VMD, and the area associated with each vertex is  $\sim 0.15\text{\AA}^2$ . By summing over all atoms of a given residue, we obtain the number of vertices  $t$  associated to that residue, which is proportional to its SASA. The distribution of the values of  $t$  observed in the dataset for alanine, valine and arginine residues is shown as supplementary information in Fig. S2; the observed behaviour is typical of all residue types. The presence of a sharp peak at  $t = 0$ , well separated by a broader peak at larger values of  $t$ , allows a clearcut definition of residue environment as either buried ( $t \leq t^*$ ) or solvent exposed ( $t > t^*$ ), using the same threshold  $t^* = 10$  for all residues. The single residue statistical potential  $E_{\text{SOL}}$  requires two separate parameter sets  $\lambda_{a_i}^e$ , for overall 40 parameters:

$$E_{\text{SOLV}} = \sum_i \lambda_{a_i}^{e_i} \quad (4)$$

$e_i = b$  or  $s$  is the environmental class of residue at position  $i$ . The two parameter sets  $\lambda_{a_i}^e$

are determined employing the ensemble of all residues from the dataset as the reference state<sup>6</sup>:

$$\lambda_a^e = -\ln \left[ \frac{\sum_e m_a^e}{\sum_e \sum_a m_a^e} \right] \quad (5)$$

where  $m_a^e$  is the total number of residues of type  $a$  found in the environment class  $e$  within the dataset. Residues that are not seen in the PDB structure, even if they are part of the sequence expressed by the crystallographer, are included in the counting as solvent exposed ( $e = s$ ).

An alternative implementation of BACH was derived using a reduced aminoacid alphabet consisting of 9 classes<sup>37</sup>: small hydrophobic (ALA,VAL,ILE,LEU,MET), large hydrophobic (TYR,TRP,PHE), small polar (SER,THR), large polar (ASN,GLN,HIS), positively charged (ARG,LYS), negatively charged (ASP,GLU), and finally GLY, PRO, CYS separately on their own. In this way we obtain just 225 energy parameters for the pairwise term and 18 energy parameters for the solvation term.

The parameter  $p$  entering in Eq. 1 and fixing the relative weight of the two contributions  $E_{\text{PAIR}}$  and  $E_{\text{SOLV}}$  is chosen in such a way that the energy per residue of the two terms has approximately the same standard deviation over the dataset. This criterion gives  $p = 0.6$ . Therefore, at variance with other composite methods<sup>10,35</sup> no optimization criteria is introduced in BACH to fix the relative weights of different physical terms. We have also checked that the value of  $p$  does not affect significantly the performance of BACH in native state discrimination.

**PDB dataset.** The BACH parameters have been derived using the TOP500 database<sup>46</sup>. This set includes 500 non redundant protein domain conformations, extracted from both monomeric and multimeric PDB protein structures, between 30–840 amino acids long, which have been solved with resolution better than  $1.8\text{\AA}$  by X-ray crystallography (no NMR). The structures in the set include disordered and not resolved regions. We count the contributions of these amino acids as no-interaction ones. We have checked that the choice of the fold library does not significantly affect parameters. In figures S3 and S4, we present the correlation between the BACH parameters calculated using either TOP500 or an ensemble of 8000 structures found in the CATH database<sup>47</sup>. Correlation is excellent especially for parameters corresponding to favourable interactions, that are by definition highly represented in the datasets. We have checked that the BACH energy calculated for a decoy set using the parameters obtained with the two different databases, have an excellent correlation as well (see figure S5). We have also derived a different alternative set of the 1090 energy parameters from a subset of TOP500 consisting of only 50 randomly chosen structures (these structures are listed in Table S2).

**Scoring function evaluation.** The scoring functions used in this work were evaluated on native and decoy structures as follows: online submission at the server <http://swissmodel.expasy.org/qmean/cgi/index.cgi> for the composite function QMEAN6; a linux executable downloaded from <http://www.fiserlab.org/potentials> was used for the RF\_CB\_SRS\_OD potential and the 3.2 ROSETTA scoring function was downloaded from <http://boinc.bakerlab.org/>.

**CASP decoy sets.** The performance of BACH has been assessed on selected decoy sets from CASP 8–9. A decoy set was used if the sequence of the experimentally solved structure is the same as the sequence given initially in the CASP competition, so that the majority of decoy structures have the same sequence as the native structure. The list of the decoy sets used is presented in Table S1 (33 decoys sets). The structures in each decoy set were used if they had the same length and sequence as the native structure, and had all the side-chain and backbone atoms. If for any reason a structure could not be analyzed by one of the scoring functions with whom we compare the performance of BACH, the structure was taken out of the set. The native conformations given by NMR were cut to have the same number of residues as the decoys of the set, and since their folded structure is given in the form of 20 different models, we defined as native, the pdb model that had the lowest energy (for each potential). So, for two scoring functions the native structure for a decoy set could be a different model out of the 20 presented in the NMR structures.

**Decoy set generated by Molecular Dynamics.** We also generated decoy sets, using molecular dynamics simulations, either in explicit solvent (MD) or in combination with bias-exchange metadynamics<sup>48</sup> (BE) to produce realistic structures, for three proteins: GB3 (pdb code 2OED), protein T0488 from CASP8 (pdb code 2VWR) and protein T0472 from CASP8 (pdbcode 2K49). For protein GB3 we ran three different simulations: i) MD starting from the native state at  $T = 330\text{ K}$  for 10 ns, ii) MD starting from two selected structures (blue circle in Fig. 1-D) at  $330\text{ K}$  for 1ns, iii) BE of 20 ns at  $400\text{ K}$ . For protein 2VWR we ran MD of 1ns at  $T = 300\text{ K}$  starting from i) the native conformation, and two decoys: ii) T0488TS256\_1 and iii) T0488TS299\_3. For protein 2K49 we ran MD of 1ns at  $T = 300\text{ K}$  starting from NMR model #1. MD simulations were performed using the GROMACS 4.5.3 package<sup>49</sup>, employing the AMBER99-ILDN<sup>50</sup> force field, using TIP3P water model<sup>51</sup> for the explicit solvent. BE together with the OBC<sup>52</sup> implicit solvent model and  $T = 400\text{ K}$  were used to enhance the conformational searching<sup>53</sup>. See Supplementary Text for details on MD and BE simulations.



**Similarity measure between structures: GDT.** As a similarity measure between structures, we use the Global Distance Test (GDT)<sup>39</sup>, defined as:

$$GDT = (1/4)(F_{\leq 1} + F_{\leq 2} + F_{\leq 4} + F_{\leq 8}) \quad (6)$$

where  $F_{\leq x}$  denotes the percentage of residues under a distance cutoff of  $X\text{\AA}$  after the two molecules have been optimally superimposed. To detect the optimal superposition, we used the MaxSub algorithm<sup>54</sup>, in the implementation given within the MaxCluster tool ([www.sbg.bio.ic.ac.uk/maxcluster/index.html#MaxSub](http://www.sbg.bio.ic.ac.uk/maxcluster/index.html#MaxSub)).

- Tanaka, S. & Scheraga, H. A. Medium and long range interaction parameters between amino acids for predicting three dimensional structures of proteins. *Macromolecules* **9**, 945–950 (1976).
- Miyazawa, S., Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures. *Macromolecules* **18**, 534–552 (1985).
- Sippl, M. J. 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.* **213**, 859–883 (1990).
- Lazaridis, T. & Karplus, M. Effective energy functions for protein structure predictions. *Curr. Opin. Struct. Biol.* **10**, 139–145 (1996).
- Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
- Samudrala, R. & Moulton, J. An all atom distance dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895–916 (1998).
- Rajgaria, R., McAllister, S. R. & Floudas, C. A. Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins* **70**, 950–970 (2008).
- Solis, A. D. & Rackovsky, S. Information and discrimination in pairwise contact potentials. *Proteins* **71**, 1071–1087 (2008).
- Jones, D. T., Taylor, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).
- Simons, K. T. *et al.* Improved Recognition of Native like Protein structures using a combination of sequence dependent and sequence independent features of proteins. *Proteins* **34**, 82–95 (1999).
- Tobi, D. & Elber, R. Distance dependent, pair potential for protein folding: results from linear optimization. *Protein* **41**, 40–56 (2000).
- Lu, H. & Skolnick, J. A distance dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44**, 223–232 (2001).
- Zhou, H. & Zhou, Y. Single body knowledge based potentials based energy score combined with sequence profile and secondary structure information for fold recognition. *Protein* **55**, 1005–1013 (2004).
- Betancourt, M. R. & Skolnick, J. Local propensities and statistical potentials of backbone and dihedral angles in proteins. *J. Mol. Biol.* **235**, 1598–1613 (2004).
- Fang, Q. & Shortle, D. Protein refolding in silico with atom based statistical potentials and conformational search using a simple genetic algorithm. *J. Mol. Biol.* **395**, 1456 (2006).
- Skolnick, J., Jaszczewski, L., Kolinski, A. & Godzik, A. Derivation and testing of pair potentials of mean force for structure selection and stability. *Protein Science* **6**, 676 (1997).
- Shen, M. Y. & Sali, A. Statistical potentials for assessment and prediction of protein models and a survey of energy functions. *Protein Science* **15**, 2507–2524 (2006).
- Rykunov, D. & Fiser, A. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics* **11**, 128 (2010).
- Boas, F. E. & Harbury, P. B. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* **17**, 199–204 (2007).
- Lee, J., Liwo, A. & Scheraga, H. A. Energy based de novo protein folding by conformational space annealing and off-lattice united-residue force field: Application to the 10–55 fragment of staphylococcal protein A and to apo calbindin D9k. *Proc. Natl. Acad. Sci. USA* **96**, 2025–2030 (1999).
- Turjanski, A. G., Gutkind, J. S., Best, R. B. & Hummer, G. Binding-induced folding of a natively unstructured transcription factor. *Plos. Computational Biology* **4**, e1000060 (2008).
- Bereau, T. & Deserno, M. Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **130**, 235106 (2009).
- Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **16**, 166–171 (2006).
- Buchete, N. V., Straub, Thirumalai, D. Development of novel statistical potentials for protein folding recognition. *Curr. Opin. Struct. Biol.* **14**, 225–232 (2006).
- Park, B. & Levitt, M. Energy functions that discriminate X-ray and near native folds from well constructed decoys. *J. Mol. Biol.* **258**, 367–392 (1996).
- Samudrala, R. & Levitt, M. Decoy 'R' Us: a database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.* **9**, 1399–1401 (1998).
- Tsai, J., Bonneau, R., Morozov, A. V., Kuhlman, R., Rohl, C. A. & Baker, D. An Improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **52**, 76–87 (2003).
- Cozzetto, D., Kryshchafovich, A. & Tramontano, A. Evaluation of CASP8 model quality predictions. *Proteins* **77**, 157–166 (2009).
- Handl, J., Knowles, J. & Lovell, S. C. Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics* **25**, 1271–1279 (2009).
- 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 Science* **11**, 2714 (2002).
- McConkey, B. J., Sobolev, Edelman, M. Discrimination of native protein structures using atomatom contact scoring. *Proc. Natl. Acad. Sci. USA* **100**, 3215 (2003).
- Dong, Q., Wang, X. & Lin, L. Novel knowledge based mean force potential at the profile level. *BMC Bioinformatics* **7**: 324 (2006).
- Ferrada, E., Vergara, I. A. & Melo, F. A knowledge based potential with an accurate description of local interactions improves discrimination between native and near native protein conformations. *Cell Biochem Biophys.* **49**, 111–24 (2007).
- Lou, M., Dousis, A. D. & Ma, J. OPUS-PSP: An orientation dependent Statistical All-atom Potential derived from Side chain packing. *J. Mol. Biol.* **376**, 288–301 (2008).
- Benkert, P., Tosatto, S. C. E. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins* **71**, 267–277 (2008).
- Benkert, P., Kunzli, M. & Schwede, T. QMEAN server for protein quality estimation. *Nucleic Acids Res.* **37**, 510–514 (2009).
- Buchete, N.-V., Straub, J. E. & Thirumalai, D. Dissecting contact potentials for proteins: Relative contributions of individual amino acids. *Proteins* **70**, 119–130 (2008).
- Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713 (1983).
- Zemla, A. LGA: a method for finding 3d similarities in protein structures. *Nucl. Ac. Res.* **31**, 3370–3374 (2003).
- Bogatyeva, N. S. & Finkelstein AV. Cuning simplicity of protein folding landscapes. *Prot. Eng.* **14**, 521–523 (2001).
- Berrera, M., Molinari, H. & Fogolari, F. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* **4**, 8 (2003).
- Kabsch, W. & Sander, C. Dictionary of protein secondary structure pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2257–2637 (1983).
- Trovato, A., Chiti, F., Maritan, A. & Seno, F. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *Plos Comp. Biol.* **2**, 1602–1618 (2006).
- Varshney, A., Brooks, F. P. & Wright, W. V. Computing smooth molecular-surfaces. *IEEE Computer Graphics and applications.* **14**, 19–25 (1994).
- Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Jour. Mol. Gra.* **14**, 33 (1996).
- Lovell, S. C. *et al.* Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437–50 (2003).
- Orengo, C. *et al.* CATH- A Hierarchic Classification of Protein Domain Structures. *Structure*, **5**, 1093–1108 (1997).
- Piana, S. & Laio, A. A bias-exchange approach to protein folding. *J. Phys. Chem. B.* **111**, 4553–4559 (2007).
- Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Mod.* **7**, 306–317 (2001).
- Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **78**, 1950–1958 (2010).
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
- Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **55**, 383–394 (2004).
- Cossio, P. *et al.* Exploring the universe of protein structures beyond the protein data bank. *Plos Comput. Biol.* **6**, 11 (2010).
- Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16**, 776–85 (2000).

## Acknowledgments

We thank Marco Baiesi, Federico Fogolari and Fabio Pietrucci for stimulating discussion, and an anonymous referee for drawing our attention to the possibility of using a reduced number of amino acid classes. We acknowledge financial support from University of Padua through Progetto di Ateneo n. CPDA083702 and from the grant AIRC 5 per mille, Rif. 12214.

## Author contributions

PC, DG, AL, FS and AT performed research. PC, AL, FS and AT designed research. PC, DG, AL, FS and AT wrote the manuscript.

## Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.



**License:** This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

**How to cite this article:** Cossio, P., Granata, D., Laio, A., Seno, F. & Trovato, A. A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci. Rep.* **2**, 351; DOI:10.1038/srep00351 (2012).