Published in final edited form as:
Proteomics. 2013 January ; 13(2): 230–238. doi:10.1002/pmic.201200330.

An Artificial Neural Network Approach to Improving the Correlation Between Protein Energetics and the Backbone Structure

Timothy M Fawcett[1], **Stephanie J Irausquin**[1], **Mikhail Simin**[1], and **Homayoun Valafar**[1]
[1]Computer Science & Engineering, University of South Carolina, Columbia, SC 29208, USA

Abstract

Computational approaches to modeling protein structures have made significant advances over the past decade. However, the current limitation in modeling protein structures is to produce protein structures consistently below the limit of 6Å compared to their native structure. Therefore improvement of protein structures consistently below the 6Å limit using simulation of biophysical forces is of significant interest. Current protein force fields such as those implemented in CHARMM, AMBER, and NAMD have been deemed complete, yet their use in *ab initio* approaches to protein structure determination has been unsuccessful. Here we introduce a new approach in evaluation of protein structures based on analysis of energy profiles produced by the SCOPE software package. The latest version of SCOPE produces a hydrogen bond profile that is substantially more informative than a single hydrogen bond energy value. We demonstrate how analysis of SCOPE's energy profile by an Artificial Neural Network (ANN) shows a significant improvement compared to the traditional force-based approaches to evaluation of structures. The ANN based analysis of SCOPE's energy profile showed identification of structures to within 1.5-3.0Å of the native structure. These results have been obtained by testing structures in the same Homology, Topology, Architecture, or Class of the CATH family.

Keywords

Protein Structure Prediction; Protein Structure Refinement; Artificial Neural Network; Protein Energetics; Hydrogen Bonding

1. Introduction

Currently, there are speculated to be over 10,000 distinct fold families for proteins [1]. However, the Protein Data Bank (PDB) [2] contains approximately 1,500 fold families as reported by CATH [3] or SCOP [4]. This suggests that homology modeling, the conventional strategy for protein structure prediction, may have a limited rate of success. In contrast, methods based purely on known physical forces and computational modeling have far less limitations, since proteins are not suspended in a specific state or in a specific medium. This promotes better modeling and comprehension of internal dynamics for proteins especially for those which exist in different conformations, as exemplified by the hTS protein [5], and provides a more accurate structure in turn. Similarly, physics-based folding of proteins will prove useful with regard to protein/protein and protein/ligand

Corresponding Author: Homayoun Valafar, Associate Professor, Computer Science & Engineering, University of South Carolina, 3A01 Swearingen Engineering Center, homayoun@cec.sc.edu*Phone:* 1 803 777 2880 *Fax:* 1 803 777 3767.

Conflict of Interest Statement
No competing financial or commercial interests exist.

interactions, further contributing to our understanding of diseases at the molecular level and potentially leading to effective treatments or therapies. Therefore, the concept of an energy landscape becomes fundamentally important in understanding the mechanism by which proteins fold [6,7] into their native conformation.

Previously we introduced the program SCOPE [8], as an alternative tool for structure evaluation based on energetics alone. In contrast to the current protein force fields of other similar programs, SCOPE contains a number of novel approaches. For example, its representation of proteins in rotamer space [9] provides a more manageable energy landscape; while a reduction in the energy terms used, translates to a simplification of the force field. Moreover, explicit inclusion of a hydrogen bond energy term has traditionally been excluded from molecular force fields, based on the argument that it can be represented as a combination of the Van der Waals and electrostatic terms. However, results obtained from recent works [10,11] have argued the importance of including a formal hydrogen bond energy term and provided the motivation for its calculation in our most recent version of SCOPE [12].

Here we utilize SCOPE [12] to further establish the possibility of improving protein structure evaluations purely from energetics by use of Artificial Neural Networks (ANN). The summary of our work reveals the importance of directly including a hydrogen bond energy term, as well as the potential impact of utilizing the extended energy profile that is produced by SCOPE. Our general strategy consists of training an ANN on existing proteins and testing it on unknown proteins. Using the CATH classification of proteins, provides a mechanism to investigate the structural similarity needed between the known and unknown protein structures for successful deployment of the presented method. Our examination begins with the lowest and most similar level (Homologous superfamily) between structures and ascends to the level with the least amount of structural similarity (Class).

2. Materials and Methods

2.1. Existing Models and Limitations

CHARMM, AMBER, NAMD, and Xplor-NIH [13-17] utilize a similar formulation of the force field that governs the energetic stability of proteins (Equation 1). In this equation the terms E_{bonds}, E_{angles}, $E_{dihedrals}$, $E_{impropers}$, E_{vdw}, and E_{elec} refer to the potential energy contribution of bond lengths, bond angles, dihedral geometry, planarity of certain atomic groups, Van der Waals and electrostatic interactions, respectively. SCOPE adopts the same set of calculations for these potential energies as the standard force field defined by CHARMM [13,14]. Although this force field should be sufficient for energy based assessment of a given structure, it often fails for various reasons, including the following two primary reasons:

1. Inadequacies in global optimization techniques fall victim to the complexity of protein energy landscape, especially given the high-dimensionality of the search space;
2. An insufficient or incomplete representation of the energy forces leads to the ineffective folding of proteins. For example, it is clear that hydrophobicity/hydrophilicity of residues, cooperative hydrogen bonding, and other interactions with water molecules can play an important role in the folding of proteins. Therefore a more complete representation of these forces may reduce the complexity of the energy landscape.

$$U((R))=\sum E_{\text{bonds}}+\sum E_{\text{angles}}+\sum E_{\text{dihedrals}}+\sum E_{\text{impropers}}+\sum E_{\text{vdw}}+\sum E_{\text{elec}} \quad (1)$$

2.2. SCOPE

SCOPE, whose general implementation details have been presented previously [8] and is available for download at http://ifestos.cse.sc.edu, is an open-source software program with a true Object Oriented implementation in C++. One unique feature of SCOPE is its representation of protein structures in rotamer space, as opposed to the more traditional Cartesian space. This translates to a reduction in the number of parameters used to create and calculate the energies for a given protein structure, which helps to mitigate the first noted impediment listed in Section 2.1. For example, the Cartesian representation of a single Alanine residue requires 30 degrees of freedom (10 atoms × 3 coordinates per atom). In contrast, representation of the same residue in the rotamer space requires a total of 4 parameters: φ, ψ, ω and the χ angles. Therefore, SCOPE boasts a significant decrease in the number of parameters needed to reconstruct an entire protein. In addition, use of the rotamer space preserves the perfect peptide geometries; thereby eliminating the need to calculate bonded energy terms such as: E_{bonds}, E_{angles}, $E_{dihedrals}$, and $E_{impropers}$ [8]. This has allowed SCOPE to focus strictly on calculations of the non-bonded energy terms (E_{vdw} and E_{elec}), which in so doing has simplified the force field formulation from 6 terms to two.

SCOPE's rendition of the standard peptide geometry is based on CHARMM's force field and its calculation of non-bonded energy terms has been confirmed to be identical to that of CHARMM [8]. More recently we have modified SCOPE to include a hydrogen bond energy term, as well as a hydrogen bond profile (discussed next), to aid with an improved protein structure evaluation.

2.2.1. Extended Force Field by Inclusion of an Explicit Hydrogen Bond Term—

Among the strongest non-bonded interactions [18], hydrogen bonds are critical in both the formation and stabilization of protein secondary structures (α-helices and β-sheets). Yet, the hydrogen bond energy term has not been specifically included in traditional force fields. This exclusion is based on the argument that a combination of electrostatic and Van der Waals interactions should implicitly encapsulate the hydrogen bond term. It can be speculated however, that its implicit inclusion does not allow for the proper evaluation of hydrogen bond energies, since hydrogen bonds consist of both distance and orientational components. Therefore, in the case that distance requirements are satisfied but orientational requirements are not, it is possible for a hydrogen bond to not be formed [18]. Furthermore, consecutive hydrogen bonds have been shown to exhibit a cooperativity phenomenon, where the total potential energy is greater than the sum of its individual components [18]. This has motivated the explicit calculation of hydrogen bonds within SCOPE [12], as well as other approaches [10,11]. Our implementation of SCOPE's hydrogen bond calculation, the details of which have been published previously [8,12], is designed specifically to be consistent with that of the DSSP program [19].

The current version of SCOPE limits its calculation of hydrogen bonds to only the backbone-backbone interactions that stabilize secondary structural elements. Because secondary structures are the result of consecutive hydrogen bonds in the primary sequence, it may be useful to consider such information for a given protein structure. Consequently, the resulting SCOPE output includes the total number of hydrogen bonds and a hydrogen bonding profile, in addition to the calculated hydrogen bond energy. The hydrogen bond profile is determined based on the number of ungapped and consecutive hydrogen bonds that are formed along the backbone of the protein. The least number of sections per protein could be zero (no hydrogen bonding), while the maximum number of sections may be n/2 (n is the protein size in number of residues). Figure 1 exemplifies the SCOPE energy profile created for each protein: the first number output, 24 in our example, identifies the total number of sections with consecutive hydrogen bonds; this is followed by the number of consecutive

hydrogen bonds in each section (for example, section 1 contains 2 consecutive hydrogen bonds while section 2 contains 1); the last line of contains the output file name, the calculated energies (total, Van der Waals, electrostatic and hydrogen bond, respectively) and the total number of hydrogen bonds in the protein (102 in our example).

2.3. Artificial Neural Network

The evaluation of protein structures has conventionally been performed by assessing the calculated total energy term for the entire protein. While the result of this approach is easy to utilize, it may lack the ability to identify severe localized problems. Therefore the applicability of the conventional energy based approaches may be limited. Figure 2A illustrates the results of energy-based evaluation of an ensemble of 1000 protein structures as a function of their similarity to the native structure measured as the BackBone Root Mean Squared Deviation (BBRMSD). The lack of any significant correlation between potential energy of a protein structure and its structural viability provides the motivation for additional investigations and establishes a reference for improvement by any new mechanism.

Here we demonstrate a different approach, which incorporates the energy profiles produced by SCOPE. SCOPE's energy profile clearly provides more information regarding the fitness of a structure in contrast to its overall potential energy. Although several approaches can be used for the analysis of multidimensional data, here we utilize an Artificial Neural Network (ANN) [20] that accepts a SCOPE profile in order to quantify the viability of the corresponding structure compared to its native structure. The ANN architecture that was utilized in our work consisted of a two stage, feed forward, standard Multilayer perceptron. The ANN was trained by using a Levenberg-Marquardt [21] based back propagation algorithm. It is important to note that the presented strength of our work is not in the use of artificial neural networks, but in the utility of representing local energies of a protein structure in a novel way. The use of an ANN is motivated by its ease of use, robustness, and automatic evaluation of the relative importance of each parameter. For example, inclusion of an explicit hydrogen bond term in addition to the traditional Van der Waals and electrostatic terms may present redundant information since Van der Waals and electrostatic terms can potentially represent hydrogen bond energies in some instances. Therefore the proper use of these terms will require elimination of the mutual information content. ANN's can automatically make the appropriate adjustments of the input parameters during the process of their training. Training of ANN consists of modifying its operation parameters (weights) such that the proper relationship between the presented input and desired output is established. During ANN training sessions, slightly modified SCOPE energy profiles are presented as input with the corresponding desired output. These modifications consist of scaling both the Van der Waals and electrostatic terms to the range of 0-10, in order to reduce the dominance of these two terms compared to other energy terms. Scaling of the input parameters is a standard practice in the training of ANN's since it simplifies the training process. The exact details on training, design and preparation of the input to the ANN has been previously described [12].

2.4. Testing Strategy

Our initial investigations with SCOPE [12] established a better correlation between the energetics based evaluation of protein structures and their corresponding BBRMSD with respect to the native structure. The previous work demonstrated how the existence of such an approach would be instrumental in improving the quality of protein structures produced by various computational modeling techniques [8]. Additional investigations, using output produced by SCOPE, consisted of establishing both the effect of directly including a hydrogen bond term for potential energy calculations of protein structures and the

improvements that can be observed through analysis of a hydrogen bond profile by the use of ANN [12]. Although this work was instrumental in establishing the proof of concept, it lacked practical implication. The training and testing of the previous work was conducted on the same protein, leading to the potential problem of memorization where ANN may potentially fail for the purposes of inference. The practical deployment of this technology necessitates a training mechanism that is independent of the structure of the unknown protein.

To investigate the ANN's generalization ability, we have modified our training procedure to be independent of the unknown protein. Therefore the general procedure consists of training the ANN on a protein other than the unknown protein, and testing its performance on the unknown protein. The main question that arises in relation to this procedure is the structural difference that can exist between the training structures and the testing structure (the unknown structure). To address this question we have utilized the protein structure similarity as reported by CATH [3]. The training and testing structures will be selected to represent increasing degrees of structural dissimilarity. Experiments will begin at the CATH level which corresponds to the highest degree of structural similarity, and proceed up the CATH tree where structural similarity is gradually diminished. The Homologous superfamily (denoted by H), is the lowest level of the CATH tree and corresponds to protein structures with the highest degree of evolutionary relationship. The next level Topology (T), are proteins that share structural features. Moving up, Architecture (A) corresponds to the structures that have high similarity but no evidence of structural homology. Finally, the last level is Class (C), which describes the secondary structure composition of a protein.

To further establish the general applicability of our approach to diverse classes of proteins, our studies will include two α-helical, two β-sheet, and two α/β mix proteins. Table 1 lists the proteins that are used in our experiments and their relevant information. One thousand derivative structures were created for each of the proteins by randomly altering their backbone dihedral angles to sample structural variation up to 7Å BBRMSD. This process produced a total of 1001 structures (including the original structure) that can be used accordingly during the training and testing of ANN. The exact details of the training and testing have been described in our previous work [12].

3. Results

Our strategy in testing the general applicability of our approach is to explore the distance between the training and testing structures. To that end, our exploration spans selection of the training and testing proteins exhibiting the following distance criteria: sequence homology of higher than 60%, sequence homology of less than 40%, structural similarity at the Topology level of CATH, structural similarity at the Architecture level of CATH, and finally structural similarity at the Class level of CATH. The categories selected for percent sequence homology have been chosen based on the thresholds reported in the literature. Proteins with high sequence homology (greater than 60%) generally produce similar structures and imply similar biological function, while those with less than 40% sequence homology suggest regions of structural similarity [1,22]. In the interest of brevity, we only present results for the latter four categories which exemplify the most challenging cases in the following subsections, and defer the results of the former case to the Supporting Information (Figure S1).

3.1. Evaluation of Protein Structures Using Conventional Force Fields

It is important to report the performance of traditional force-based approaches in evaluating protein structures in order to compare the improvements that are gained from the new approach. Figure 2A shows the correlation plot of 4997 derivative structures for the protein

1A1Z. In this plot, the total energy of each structure is plotted versus the BBRMSD of the structure with respect to the known native structure of this protein. The lack of any correlation between the total energy and structural viability is clear and can be used in order to evaluate the contribution of our presented work.

To establish the importance of utilizing SCOPE's energy profile, we have utilized an optimized linear regression in analysis of structural viability of the same set of 4997 structures using the conventional total energy. In this exercise the Van der Waals and electrostatic energy terms were used in order to calculate the optimal coefficients of a linear model. Figure 2B shows the results of this exercise and confirms the absence of any notable correlation between the conventional energies and structural viability. Finally, we repeat the same exercise after inclusion of the total hydrogen bond term in order to establish a point of reference for the results that are reported in the following sections. Furthermore, comparison of Figure 2B and Figure 2C establishes the potential usefulness of information presented by an explicit hydrogen bond, since the latter exhibits an improved correlation between total energy and BBRMSD of structures.

3.2. Evaluation of Proteins in CATH's Level H: Homologous Superfamily with Less Than 40% Sequence Identity

Experiments were first conducted using training and testing proteins in the same Homologous superfamily with at most 40% sequence similarity. Figure 3A illustrates the results of the evaluation prediction of the ANN for the protein 3CZZ, with the training conducted on the protein 1N02. Results shown in this figure clearly illustrate a linear correlation between the ANN's output and the actual BBRMSD of the unknown protein. The top 10% of the ranked structures in the case of 3CZZ exhibit a BBRMSD between 0 – 0.996Å, which indicates the success of the ANN in selecting an accurate structure with BBRMSD less than 1Å.

The prediction of 3CUO trained on 2D2W has a very strong linear correlation (Supporting Information Figure S2A). The top 10% of the ranked structures in the case of 3CUO have a BBRMSD between 0.006-1.267Å. This again demonstrates the success of our approach in selecting a highly accurate structure with great confidence. This is especially noteworthy considering that these two proteins share less than 30% sequence identity.

The prediction of 2JXT trained on 1PGX has a moderate linear correlation (Supporting Information Figure S2B) and shows the existence of the funneling effect despite the diminished performance. Although the best predicted structure was around 2.2Å, the original structure was predicted to be eleventh. The top 10% of the ranked structures for 2JXT exhibited a BBRMSD range of 0-3.447Å. While not as compelling as the previous two cases, it is important to note that deployment of the presented work can potentially improve the quality of a starting structure from 6.559Å to at least 3.447Å.

3.3. Evaluation of Proteins in CATH's Level T: Topology

In the Topology level of CATH proteins do not demonstrate significant sequence similarity and the measure of structural similarity in terms of BBRMSD begins to lose its meaning. Results of the ANN evaluation of the protein 3FB9 with training based on 3CQT is shown in Figure 3B. These results clearly illustrate the clustering of the ANN with direct correlation to their structural quality. The top 10% ranked results of the protein 3FB9 have a BBRMSD between 0 – 1.512Å. This is a substantial selection of proteins with as high as 6.427Å BBRMSD with respect to the native structure. This level of performance is on par with previous experiments from the homologous superfamily.

The prediction of 1SAU trained on 3CUO (shown in Supporting Information Figure S3A) demonstrates the existence of the desired funneling effect. The top 10% structures have a BBRMSD between 0-3.235Å indicating the possibility of improving computed structures to within 3.235Å of the native structure.

The prediction of 1VJK trained on 1LXD (Supporting Information Figure S3B) also shows a funneling effect. The top 10% of the ranked structures of 1VJK have a BBRMSD between 0-1.827Å. Once again this shows a remarkable ability to identify the more viable structures below 2Å.

3.4. Evaluation of Proteins in CATH's Level A: Architecture

The prediction of 3HAK trained on 3CUO exhibits a strong linear correlation (Figure 3C) with a remarkable funneling effect. The top 10% of the ranked structures of 3HAK have BBRMSDs between 0.022-1.864Å. This provides evidence for the success of our approach within the Architecture level of the CATH tree.

The prediction of 3CQT trained on 2H3L (Supporting Information Figure S4A) shows similar results with the top 10% of the results demonstrating a BBRMSD between 0-5.939Å to the native structure. The expanded BBRMSD range in this instance compared to the previous results is due to one anomalous structure with a BBRMSD of 5.939Å. Elimination of this one anomalous structure would produce results spanning the range of 0-2.079Å BBRMSD, which is similar to the previous outcomes.

The prediction of 20ZT trained on 1VJK (Supporting Information Figure S4B) reveals a cluster of structures between 5Å and 6Å that appear as good structures in the graph. The top 10% of the ranked structures of 2Z0T had a BBRMSD between 0-5.701Å. In this instance, five anomalous structures were the culprits with BBRMSD's ranging between 5.205-5.701Å. Elimination of these five structures would result in BBRMSDs ranging between 0-1.131Å.

3.5. Evaluation of Proteins in CATH's Level C: Class

The prediction of 2EPI trained on 1VJK also demonstrates a strong linear correlation (Figure 3D) and an extraordinary funneling effect, considering that structures at this level display the least amount of structural similarity. In this exercise, the original structure was predicted to be eighty-ninth. The original structure was not in the top thirty but the BBRMSD of the top 10% was between 0.107-1.890Å. Even though the original structure was not predicted in the top thirty structures, the selected structures exhibited significant structural similarity to the native structure (under 2Å).

The prediction of 1TQG trained on 3CUO shows a very strong linear correlation (Supporting Information Figure S5A). However, a number of structures ranging from 0Å to over 4Å have been ranked highly by ANN. Therefore the top 10% of structures for 1TQG have a BBRMSD between 0.076-4.316Å and demonstrate a potential difficulty in predicting structures which share structural similarity only within the Class level of CATH.

The prediction of 1SMX trained on 3CQT shows a strong linear correlation (Supporting Information S5B) and illustrates a strong funneling effect. The original structure was predicted to be twelfth overall. These results are notable with the two structures only sharing mostly β secondary structures. 1SMX's top thirty structures have a BBRMSD that range between 0-3.338Å. This would refine a protein from 6.933Å down to 3.338Å.

4. Discussion

Our first experiments demonstrated the consistent ability in utilizing ANN based evaluation of SCOPE's energy profile in application to structures that exhibit a strong evolutionary relationship. During the generalization exercises, we demonstrated the practical application of our presented work to identify viable structures to within 1.5-3.0Å of the native structure of the unknown protein. Our presented experiments utilized the CATH classification of a protein to establish the required structural similarity between the training and testing proteins. To that end we have presented results spanning the most similar and related proteins (Homologous superfamily) and conclude with studies using the most dissimilar proteins (Class). The Homologous superfamily experiments show the existence of the funneling effect at this level, and demonstrate the ability to identify the native protein structure within the top 10% of the ranked structures. The experiments also present a potential approach to improve the structural quality of a protein by over 3Å; in fact, in most cases the protein could be improved to less than 2Å. The testing of proteins with the same Topology, but different Homologous superfamiles, demonstrated that an ANN was able to predict proteins with similar accuracies to the Homologous superfamily level. The experiments in this section displayed results that were able to identify a protein structure to within 3.235Å in one instance, and to under 2Å in the other instances. It is important to note, that the funneling effect still exists in these experiments, however some anomalous behaviors are starting to be observed. Also of note, the linear correlation coefficient is not associated with the prediction of the original structure; this is summarized in Table 2 which demonstrates that the experiment with the worst linear correlation coefficient did not have the worst prediction. The next two sets of experiments, using proteins with the same Architecture and the same Class respectively, also reveal surprisingly promising results with respect to the ANN's ability to predict the original structure. The examples examining protein structural similarity in the Architecture level still show that the native structure was predicted to be in the top 14% or better. The funneling effect can still be seen but the predicted anomalous structures are becoming more frequent at this level; if the anomalous structures are removed, however, then an ANN has the ability to identify structures of the unknown protein to less than 2.1Å. When experiments involving the same Class of proteins are conducted, there is a long line of structures that are being predicted by the ANN to be good structures - with the original structure being predicted out of the top thirty best structures in most cases.

When combined with an ANN, the energy profile produced by SCOPE has demonstrated the potential to improve the structural quality of an unknown protein from 7Å to approximately 2Å. Our presented method of evaluating protein structures purely based on biophysical energies clearly shows a strong funneling effect that is otherwise missing from traditional force fields.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

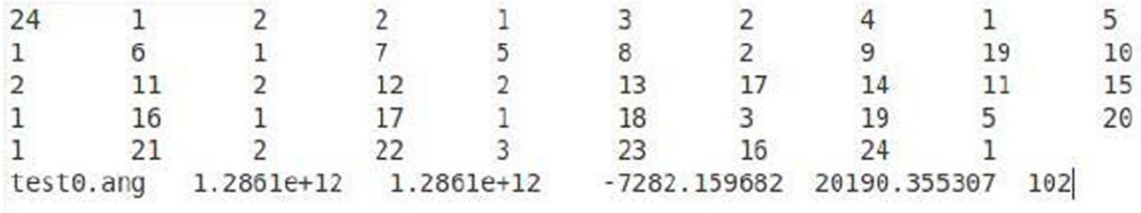
Acknowledgments

We would like to thank Dalton Brown for helping with source code development. This work was supported by grant No. NIH-1R01GM081793 and the SC INBRE grant No. P20 RR-016461.

References

1. Holm L, Sander C. Mapping the protein universe. Science. 1996; 273:595–602. [PubMed: 8662544]

2. Berman HM, Westbrook J, Feng Z, Gilliland G, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
3. Orengo CA, Michie AD, Jones S, Jones DT, et al. CATH - a hierarchic classification of protein domain structures. Structure. 1997; 5:1093–1108. [PubMed: 9309224]
4. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP - A Structural Classification Of Proteins Database For The Investigation Of Sequences And Structures. J Mol Biol. 1995; 247:536–540. [PubMed: 7723011]
5. Berger SH, Berger FG, Lebioda L. Effects of ligand binding and conformational switching on intracellular stability of human thymidylate synthase. Biochimica et biophysica acta. 2004; 1696:15–22. [PubMed: 14726200]
6. Dobson CM. Protein folding and misfolding. Nature. 2003; 426:884–90. [PubMed: 14685248]
7. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973; 181:223–230. [PubMed: 4124164]
8. Fawcett, TM.; Irausquin, S.; Simin, M.; Valafar, H. SCOPE: An Open-Source, C++ Implementation for Calculation of Protein Energetics from First Principles. Proceedings of the International Conference on Bioinformatics & Computational Biology; 2011.
9. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of molecular biology. 1997; 273:283–98. [PubMed: 9367762]
10. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. Journal of molecular biology. 2003; 326:1239–59. [PubMed: 12589766]
11. Brylinski M, Gao M, Skolnick J. Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function. Physical chemistry chemical physics : PCCP. 2011; 13:17044–55. [PubMed: 21655593]
12. Fawcett, TM.; Irausquin, S.; Simin, M.; Valafar, H. An Artificial Neural Network Based Approach for Identification of Native Protein Structures Using an Extended ForceField. Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine; 2011.
13. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry. 1983; 4:187–217.
14. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, et al. CHARMM: the biomolecular simulation program. Journal of computational chemistry. 2009; 30:1545–1614. [PubMed: 19444816]
15. Case DA, Cheatham TE, Darden T, Gohlke H, et al. The Amber biomolecular simulation programs. Journal Of Computational Chemistry. 2005; 26:1668–1688. [PubMed: 16200636]
16. Phillips JC, Braun R, Wang W, Gumbart J, et al. Scalable molecular dynamics with NAMD. Journal Of Computational Chemistry. 2005; 26:1781–1802. [PubMed: 16222654]
17. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. Journal of Magnetic Resonance. 2003; 160:65–73. [PubMed: 12565051]
18. Desiraju, GR.; Steiner, T. The Weak Hydrogen Bond in Structural Chemistry and Biology. Oxford University Press Inc; New York: 1999.
19. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
20. Morris R, Willshaw D. Parallel Distributed-Processing - Explorations in the Microstructures of Cognition, Vol 1, Foundations, Vol 2, Psychological and Biological Models - Rumelhart,De, Mcclelland,Jl. Nature. 1987; 327:469–470.
21. Ranganathan A. The Levenberg-Marquardt Algorithm. 2004
22. Sierk ML, Smoot ME, Bass EJ, Pearson WR. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. BMC bioinformatics. 2010; 11:146. [PubMed: 20307279]



24	1	2	2	1	3	2	4	1	5
1	6	1	7	5	8	2	9	19	10
2	11	2	12	2	13	17	14	11	15
1	16	1	17	1	18	3	19	5	20
1	21	2	22	3	23	16	24	1	
test0.ang	1.2861e+12	1.2861e+12	-7282.159682	20190.355307	102				

Figure 1.
An example of the SCOPE output.

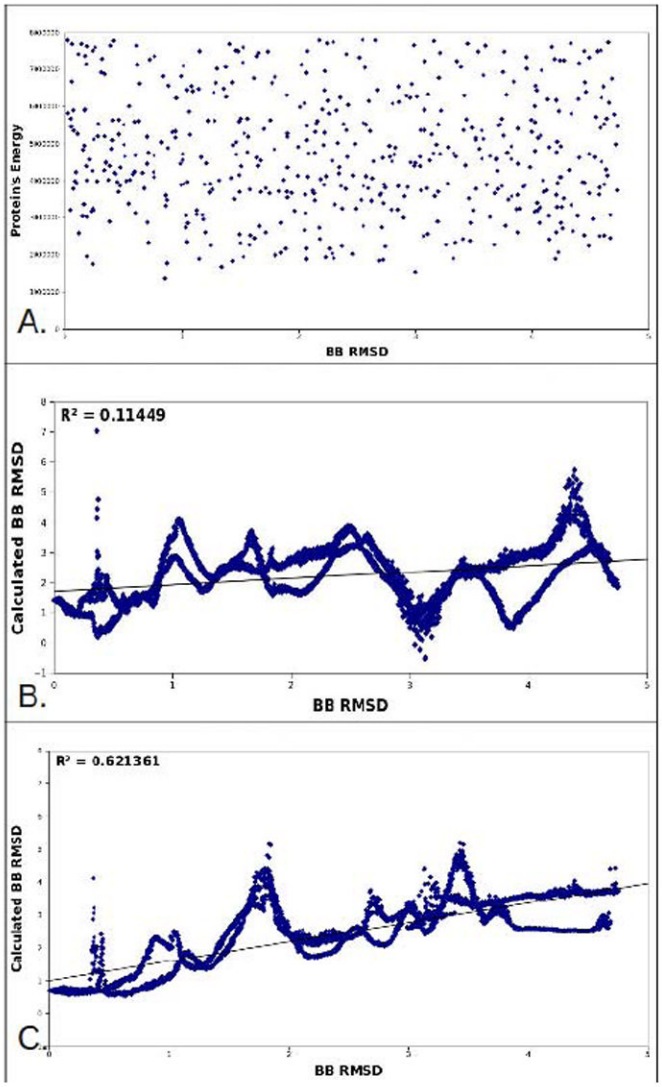


Figure 2.

Using the same set of 4997 derivative structures of 1A1Z: A) energy-based evaluation of the derivative structures as a function of their similarity to the native structure measured by BBRMSD; B) linear regression analysis of structural viability utilizing SCOPE's energy profile (scaled Van der Waal and scaled electrostatic energy terms); C) linear regression analysis of structural viability utilizing SCOPE's energy profile and the inclusion of a hydrogen bond energy term.

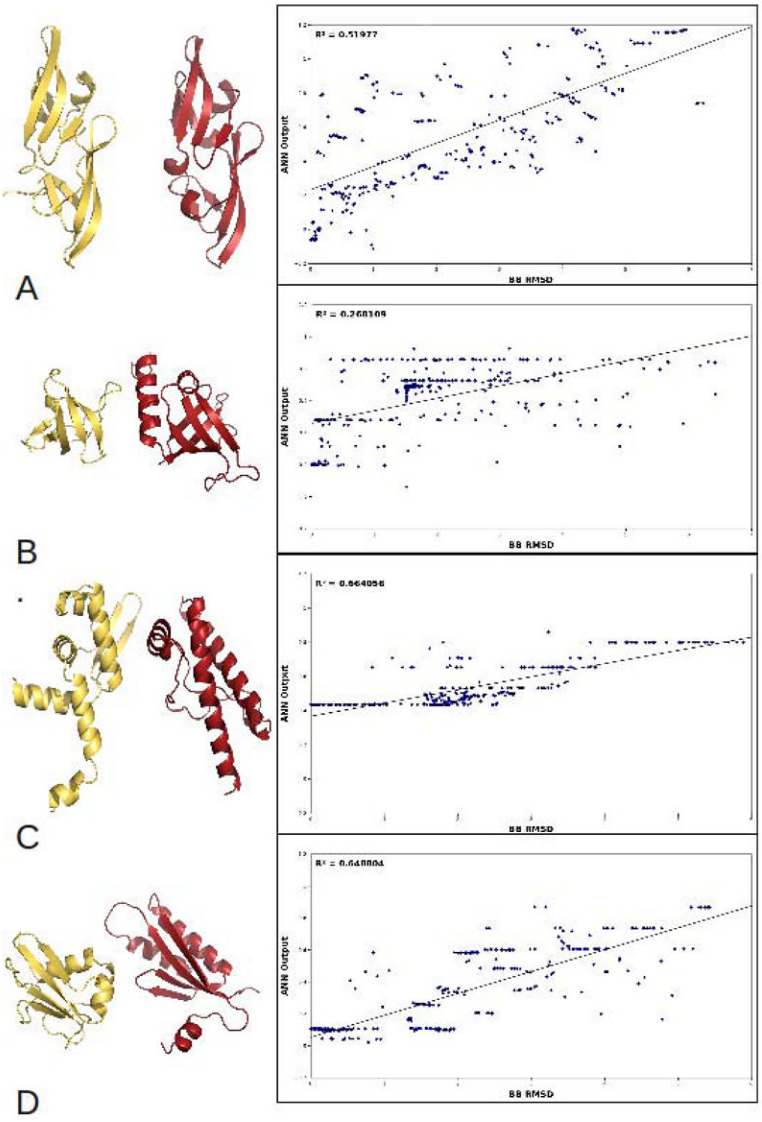


Figure 3.

ANN Refinement of Proteins in CATH's A) Level H: Homologous superfamily, using 1N02 for training (yellow) and 3CZZ for testing (red) with 19 hidden neurons; B) Level T: Topology, using 3CQT for training (yellow) and 3FB9 for testing (red) with 16 hidden neurons; C) Level A: Architecture, using 3CUO for training (yellow) and 3HAK for testing (red) with 11 hidden neurons; and D) Level C: Class, using 1VJK for training (yellow) and 2EPI for testing (red) also with 11 hidden neurons.

Table 1

Summary of all proteins used for ANN Testing.

	Training				Testing				
CATH Level	**PDB ID**	**CATH Classification**	**Class**	**Length (Residues)**	**PDB ID**	**CATH Classification**	**Class**	**Length (Residues)**	**BBRMSD (Å)**
Homology (≥60%)	1A1W	1.10.533.10	α	83	1A1Z	1.10.533.10	α	83	1.61
	2EZM	2.30.60.10	β	101	2EZN	2.30.60.10	β	101	2.34
	1P7E	3.10.20.10	α/β	56	1P7F	3.10.20.10	α/β	56	1.15
Homology (≤40%)	1N02	2.30.60.10	β	102	3CZZ	2.30.60.10	β	101	9.429
	2D2W	1.10.10.10	α	103	3CUO	1.10.10.10	α	98	12.060
	1PGX	3.10.20.10	α/β	69	2JXT	3.10.20.10	α/β	86	12.032
Topology	3CQT	2.30.30.40	β	79	3FB9	2.30.30.100	β	90	N/A
	3CUO	1.10.10.10	α	98	1SAU	1.10.10.370	α	115	N/A
	1LXD	3.10.20.90	α/β	100	1VJK	3.10.20.30	α/β	98	N/A
Architecture	3CUO	1.10.10.10	α	98	3HAK	1.10.790.10	α	103	N/A
	2H3L	2.30.42.10	β	103	3CQT	2.30.30.40	β	79	N/A
	1VJK	3.10.20.30	α/β	98	20ZT	3.10.480.10	α/β	109	N/A
Class	1VJK	3.10.20.30	α/β	98	2EPI	3.30.70.930	α/β	100	N/A
	3CUO	1.10.10.10	α	98	1TQG	1.120.120.160	α	105	N/A
	3CQT	2.30.30.40	β	79	1SMX	2.40.50.140	β	100	N/A

Table 2

Results obtained for ANN Testing.

CATH Level	Training PDB ID	Testing PDB ID	Correlation Coefficient	Predicted Ranking of the Native Structure (Out of 334)	BBRMSD of Top 10% Structures (Å)
Homology (≥60%)	1A1W	1A1Z	0.8654	12	0 – 0.640
	2EZM	2EZN	0.5507	33	0.019 – 1.805
	1P7E	1P7F	0.7198	22	0 – 2.018
Homology (≤40%)	1N02	3CZZ	0.7209	4	0 – 0.996
	2D2W	3CUO	0.9327	37	0.006 – 1.267
	1PGX	2JXT	0.6346	11	0 – 3.347
Topology	3CQT	3FB9	0.5178	25	0 – 1.512
	3CUO	1SAU	0.7070	20	0 – 3.235
	1LXD	1VJK	0.6340	25	0 – 1.827
Architecture	3CUO	3HAK	0.8149	44	0.022 – 1.864
	2H3L	3CQT	0.7348	8	0 – 5.939
	1VJK	20ZT	0.7348	23	0 – 5.701
Class	1VJK	2EPI	0.8055	89	0.107 – 1.860
	3CUO	1TQG	0.9011	56	0.076 – 4.316
	3CQT	1SMX	0.7901	12	0 – 3.338