Protein Science (2000), *9*:1399–1401. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Decoys ‘R’ Us: A database of incorrect conformations to improve protein structure prediction

RAM SAMUDRALA AND MICHAEL LEVITT
Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305

(Received October 19, 1999; Final Revision May 4, 2000; Accepted May 5, 2000)

Abstract: The development of an energy or scoring function for protein structure prediction is greatly enhanced by testing the function on a set of computer-generated conformations (decoys) to determine whether it can readily distinguish native-like conformations from nonnative ones. We have created “Decoys ‘R’ Us,” a database containing many such sets of conformations, to provide a resource that allows scoring functions to be improved.

Keywords: refinement; scoring/energy functions

What are decoys: Predicting the structure of protein using the amino acid sequence information alone is one of the fundamental unsolved problems in computational molecular biology (Richards, 1991). Any algorithm that attempts to predict protein structure requires a scoring or discriminatory function that can distinguish between correct and incorrect conformations. A major issue in developing any discriminatory function for work with proteins is deciding how to test its performance.

We introduce a database, Decoys ‘R’ Us ⟨http://dd.stanford.edu⟩, that contains a wide variety of decoys generated by different methods with the aim of fooling scoring functions. Decoys are computer-generated conformations of protein sequences that possess some characteristics of native proteins, but are not biologically real. Decoys have been based on discrete-state models (Park & Levitt, 1996), molecular dynamics trajectories (Wang et al., 1995; Huang et al., 1996), crystal structures of different resolutions (Subramaniam et al., 1996), conformations with different loops (Samudrala & Moult, 1998), and amino acid sequences mounted on radically different folds (Novotny et al., 1984; Holm & Sander, 1992).

World Wide Web sites have been established to provide decoy test sets for fold recognition functions ⟨http://fold.doe-mbi.ucla.edu⟩ and for general protein structure prediction functions ⟨http://prostar.carb.nist.gov⟩. These sites are useful because most functions generally are tested on only one or two types of decoy because generation of decoy sets is a time-consuming task. Using only a few types of decoy, discrimination may be achieved by some specific artifacts of the decoys, such as noncompactness or systematic distortion of detailed features like hydrogen bond length (Park et al., 1997; Samudrala & Moult, 1998). Multiple decoy sets are essential to not only measure the “orthogonality” (i.e., the ability to succeed on many different sets) between a discriminatory function and a method for generating decoys, but also the “complementarity” between a method for exploring the conformational space of proteins and a given scoring function.

It is difficult to generate high-quality decoy sets that can readily fool discriminatory functions. Because the Web has been useful for

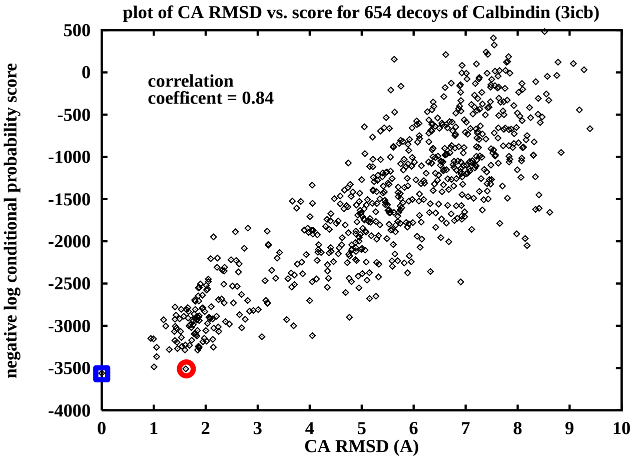

Fig. 1. C_α RMSD vs. score for 653 decoys of Calbindin (Protein Data Bank code 3icb) from the 4state_reduced decoy set. The scores are calculated using an all-atom distance-dependent conditional probability discriminatory function (Samudrala & Moult, 1998). The RMSD for the lowest scoring structure (circled) is 1.63 Å, and the $\log_{10}$ odds of picking out this conformation by chance is -1.40 ($p = 0.04$). The correlation coefficient of the score and RMSD is 0.84. The zero RMSD structure is the experimental conformation (boxed), which is shown here for informative purposes; in a bona fide prediction scenario this conformation is not likely to be seen in the sample space. Similar results are observed with this function for all the proteins in the 4state_reduced set. This example, although idealized, indicates that if a method that can sample the conformational space in an ab initio manner to produce the distribution of RMSDs depicted is available, then the function is able to select conformations around 2.0 Å RMSD.

Reprint requests to: Ram Samudrala, Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305; e-mail: ram@csb.stanford.edu.

testing scoring functions, and because many decoy sets generated by us and other research groups already exist, a wide variety of decoys have been placed on the Decoys ‘R’ Us Web site, with the hope of aiding developers of scoring functions in finding challenging tests for their work. The goal of this effort is to present sets of decoys in a readily available and usable manner, complementing the other existing efforts in this area. The focus is on diversity and volume-collecting data for many different proteins, and providing a large number of decoys per protein so that a giving scoring function can be tested exhaustively. Data sets generated in an ab initio manner by different search algorithms are also provided (e.g., the lattice_ssfit decoy set). A scoring function that does well in these types of sets will most likely do well in a blind bona fide prediction scenario, such as the one provided by the Critical Assessment of Protein Structure Prediction Methods (CASP) conference, which in turn, can lead to elucidation of function using a combination of theory and experiment based on predicted structure (Wei et al., 1999; Samudrala et al., 2000).

How are decoys used: How does one evaluate the performance of a scoring function in a manner that enables different scoring func-

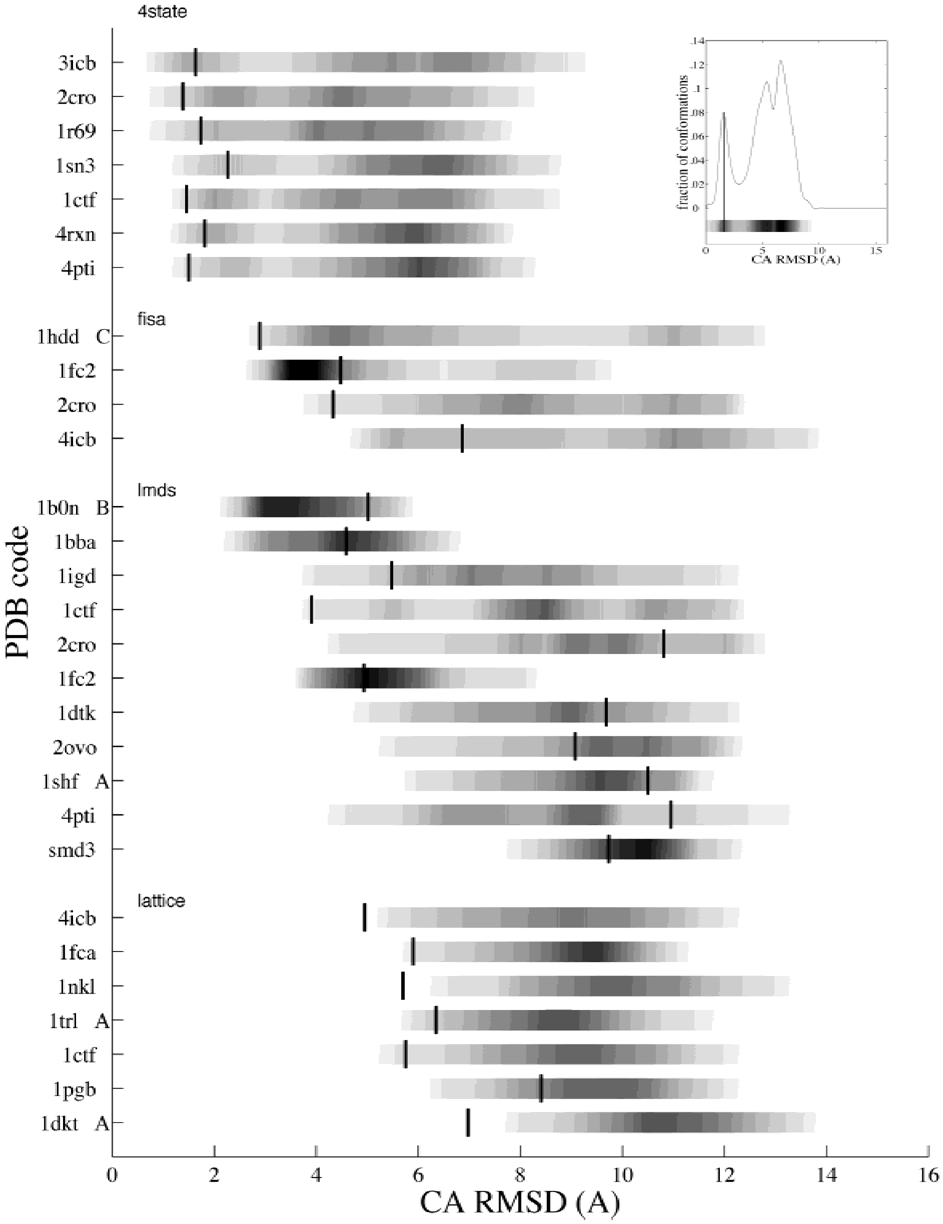

Fig. 2. Performance of our discriminatory function on four different decoy sets. The distribution of C_α RMSDs for each protein in each of the decoy sets is represented by a shaded density bar, where the density of the shading at a given RMSD range (horizontal axis) is an indicator of the fraction of conformations present. The thick bar indicates the RMSD of conformation selected by our scoring function. The inset image shows how the density bar maps to the fraction of conformations for a given RMSD range. The scoring function we use performs well across a variety of decoy sets, doing worst in the case of the lmds decoy set and doing best when there are clear native-like conformations present (4state_reduced decoy set).

tions to be compared to each other between different decoy sets? For decoy sets with one correct and one incorrect conformation, we use two primary measures: the percentage (or fraction) of cases where the correct/experimental conformation has a better score than the incorrect conformation (the higher the percentage, the better the discrimination), and the discrimination ratio between the score of the incorrect conformation and the correct conformation, averaged over all correct/incorrect pairs in the particular set.

For decoy sets with one correct and many incorrect conformations, we begin by plotting the score vs. the root-mean-square deviation (RMSD) of the C_α atoms between the native conformation and each decoy. An example of this is illustrated in Figure 1 using an all-atom distance-dependent conditional probability discriminatory function (Samudrala & Moult, 1998), one of the many scoring functions that have been published in the literature. The RMSD of the lowest scoring conformation (excluding the experimental structure) is one measure of how well the function performs, but this is an extremely noisy quality. One can also estimate the probability of selecting the conformation by chance (RMSD rank of the conformation/divided by the total number of conformations). This is also a noisy estimate. A reasonably robust measure we have found to work in practice is the correlation coefficient of the RMSDs and the scores, because it incorporates information about all the conformations produced by a particular decoy-generation method (Fig. 1).

We have also developed a new method for examining the discriminatory power of a scoring function within the distribution of conformations in a graphical manner using the gel-like plots in Figure 2. This allows us to both qualitatively and quantitatively assess the performance of a scoring function for a large number of different types of decoy sets at a glance. In these plots, not only is the final selection information present, but the distribution of conformations is represented by shading density. This allows us to visualize and compare the performance of a function across many decoy sets.

For example, from Figure 2 we can glean that the discriminatory function we use for our simulations performs fairly well across a wide variety of decoys. However, in certain cases such as for the 1fc2 protein in the fisa decoy set, even though the selection of 4.480 Å RMSD seems reasonable, it is actually a poor one as it falls to the right of the mean of the distribution, which has a lower RMSD. Likewise, the function does not always succeed on the lmds decoy set even though in some cases near-native conformations are present in the set with reasonably high density. We note that it performs consistently well when very native-like conformations are present in the set, as is the case with the 4state_reduced set.

Utility of Decoys 'R' Us and future plans: The Decoys 'R' Us database has been available in preliminary form for about a year. During that time, a number of groups have made use of these decoys in published and unpublished works, using them to evaluate and improve scoring function performance for predicting structure (Samudrala et al., 1999; Simons et al., 1999a, 1999b) to elucidate the physical nature of protein–protein interactions (Lazaridis & Karplus, 1999), and to assess the degree to which biologically relevant functional sites are preserved in predicted structures (Wei et al., 1999). In addition, there are over 50 unique downloads of at least one decoy set each month.

Besides maintaining the database and adding more decoys as scoring functions start performing better on these decoys, we also will have software that can test the ability of a scoring function to drive a conformation toward the native structure. A suite of programs will be made available that will help create decoy sets using some of the methods described above for different proteins and evaluating them using different scoring functions. A preliminary version of this set of programs is available at ⟨http://www.ram.org/computing/ramp/⟩. Programs to visualize the decoy set data and discriminatory function performance will also be made available. Finally, this database will also serve to validate sampling efforts by different methods by collecting conformations produced in a bona fide manner (i.e., "blind prediction") by those methods.

A detailed description of the organization of the database and the format of the conformations, examples of decoys, and usage guidelines is available on the Decoys 'R' Us Web site: ⟨http://dd.stanford.edu⟩.

References

Holm L, Sander C. 1992. Evaluation of protein models by atomic solvation preference. *J Mol Biol 225*:93–105.

Huang E, Subbiah S, Tsai J, Levitt M. 1996. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J Mol Biol 257*:716–725.

Lazaridis T, Karplus M. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol 288*:477–487.

Novotny J, Bruccoleri R, Karplus M. 1984. An analysis of incorrectly folded protein models. Implications for structure predictions. *J Mol Biol 177*:787–818.

Park B, Huang E, Levitt M. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol 266*:831–846.

Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol 258*:367–392.

Samudrala R, Moult J. 1998. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol 275*:895–916.

Samudrala R, Xia Y, Huang E, Levitt M. 1999. *Ab initio* protein structure prediction using a combined hierarchical approach. *Proteins Struct Funct Genet S3*:194–198.

Samudrala R, Xia Y, Levitt M, Cotton N, Huang E, Davis R. 2000. Probing structure–function relationships of the dna polymerase alpha-associated zinc-finger protein using computational approaches. In: Altman R, Dunker A, Hunter L, Klein T, Lauderdale K, eds. *Proceedings of the Pacific Symposium on Biocomputing*. Singapore: World Scientific Press. pp 179–189.

Simons K, Bonneau R, Ruczinski I, Baker D. 1999a. Ab initio structure prediction of CASP3 targets using ROSETTA. *Proteins Struct Funct Genet S3*:171–176.

Simons K, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. 1999b. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins Struct Funct Genet 34*:82–95.

Subramaniam S, Tcheng DK, Fenton J. 1996. A knowledge-based method for protein structure refinement and prediction. In: States D, Agarwal P, Gaasterland T, Hunter L, Smith R, eds. *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology*. Menlo Park, California: AAAI Press. pp 218–229.

Wang Y, Zhang H, Scott R. 1995. Discriminating compact non-native structures from the native structure of globular proteins. *Proc Natl Acad Sci USA 92*:709–713.

Wei L, Huang E, Altman R. 1999. Are predicted structures good enough to preserve functional sites? *Structure 7*:643–650.