PROTEINS: Structure, Function, and Bioinformatics 53:76–87 (2003)

An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction

Jerry Tsai,[1]* Richard Bonneau,[2] Alexandre V. Morozov,[3] Brian Kuhlman,[4] Carol A. Rohl,[5] and David Baker[6]
[1]*Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas*
[2]*Institute for Systems Biology, Seattle, Washington*
[3]*Department of Physics, University of Washington, Seattle*
[4]*Department of Biochemistry, University of Washington, Seattle*
[5]*Department of Biomolecular Engineering, University of California, Santa Cruz, California*
[6]*Department of Biochemistry, University of Washington, Seattle*

ABSTRACT **We have improved the original Rosetta centroid/backbone decoy set by increasing the number of proteins and frequency of near native models and by building on sidechains and minimizing clashes. The new set consists of 1,400 model structures for 78 different and diverse protein targets and provides a challenging set for the testing and evaluation of scoring functions. We evaluated the extent to which a variety of all-atom energy functions could identify the native and close-to-native structures in the new decoy sets. Of various implicit solvent models, we found that a solvent-accessible surface area–based solvation provided the best enrichment and discrimination of close-to-native decoys. The combination of this solvation treatment with Lennard Jones terms and the original Rosetta energy provided better enrichment and discrimination than any of the individual terms. The results also highlight the differences in accuracy of NMR and X-ray crystal structures: a large energy gap was observed between native and non-native conformations for X-ray structures but not for NMR structures. Proteins 2003;53:76–87.**

Key words: scoring functions; Rosetta method and decoys; protein structure prediction

INTRODUCTION

The development and evaluation of new energy functions is critical to the accurate modeling of the properties of biological macromolecules. Because the native structure of a protein must be low in free energy relative to almost all other conformations of the chain in order to be almost exclusively populated in solution, a stringent test of energy functions is the extent to which they attribute lower energies to native and near native conformations than to non-native conformations. Indeed, "decoy discrimination" tests have become a widely used approach for testing and validating alternative energy models.[1–3]

An optimal decoy set should (1) contain conformations for a wide variety of different proteins to avoid over-fitting; (2) contain conformations close (<4Å) to the native structure because structures more distant from the native structure may not be in the native structure's energy basin and thus impossible to recognize; (3) consist of conformations that are at least near local minima of a reasonable scoring function, so they are not trivially excludable based on obviously non protein like features; and (4) be produced by a relatively unbiased procedure that does not use information from the native structure during the conformational search. If (4) is the case, then a method that performs well on the decoy set can immediately be used for structure prediction.

The Rosetta algorithm developed by our group over the past several years has shown a degree of success at *de novo* structure prediction.[4,5] Reducing the representation of a protein to only the main chain atoms and a side chain centroid,[6] Rosetta can generate reasonable low-resolution structures much of the time, but cannot reliably identify the most native-like model.[4] Here we use Rosetta to generate a large and improved decoy set for testing energy functions that satisfies the four above criteria better than previously described sets. In the development of this augmented decoy set, we add sidechains to the centroid/backbone models and refine the structures to remove steric clashes. Next, we evaluate the capability of different energy/scoring functions, including a number of different implicit solvent models, to recognize the near-native structures in the decoy set. We then develop a combined scoring function that exhibits an enhanced performance over a variety of folds and assess the performance of this hybrid scoring function in simulated as well as real tests of structure prediction.

RESULTS

Decoy Set

We set out to create a decoy set that satisfied the four criteria listed in the introduction. To satisfy requirement

---

The Rosetta All-atom Decoy Set may be downloaded from http://depts.washington.edu/bakerpg/decoys/ using the link "Download the all atom decoys" used by Tsai et al. (pdbs).

Grant sponsor: Howard Hughes Medical Institute.

*Correspondence to: Department of Biochemistry and Biophysics, Texas A&M University, 2128 TAMU, Room 111, College Station, TX 77843. E-mail: jerrytsai@tamu.edu

Received 18 November 2002; Accepted 18 February 2003

TABLE I. Statistics on the Rosetta All Atom Decoy Set

Number	2°	Residues	PDB code	Experiment	Relative contact order	Lowest CαRMSD	Number within CαRMSD bins				
							<3	3 to 4	4 to 5	5 to 6	6<
1	α	35	1res	NMR	0.11	1.24	1,280	111	8	0	0
2	α	43	1uxd	NMR	0.12	1.31	760	235	189	70	145
3	α	43	2pdd	NMR	0.11	2.65	36	367	212	289	496
4	α	45	1uba	NMR	0.09	3.00	1	78	290	376	655
5	α	47	1gab	NMR	0.11	1.93	438	351	171	87	353
6	α	56	1bw6	NMR	0.10	2.29	163	203	245	248	541
7	α	56	2hp8	NMR	0.09	3.13	0	25	368	105	901
8	α	57	1am3	X-RAY	0.09	1.66	345	123	154	137	640
9	α	61	1r69	X-RAY	0.12	1.58	232	101	205	212	649
10	α	62	1c5a	NMR	0.11	3.18	0	67	231	191	910
11	α	62	1utg	X-RAY	0.08	3.68	0	2	176	225	996
12	α	65	1a32	X-RAY	0.09	1.20	321	118	87	102	772
13	α	65	2ezh	NMR	0.10	2.43	2	170	150	100	977
14	α	66	1nre	NMR	0.09	1.75	91	65	82	97	1,064
15	α	67	1ail	X-RAY	0.10	2.82	3	15	106	120	1,155
16	α	68	1hp8	NMR	0.08	4.02	0	0	117	220	1,062
17	α	69	1lfb	X-RAY	0.11	2.73	1	49	106	79	1,164
18	α	70	1nkl	NMR	0.09	2.80	1	122	245	204	827
19	α	70	1pou	NMR	0.11	2.70	4	88	125	80	1,102
20	α	71	1mzm	X-RAY	0.10	2.67	3	134	198	135	972
21	α	73	1acp	NMR	0.10	3.69	0	15	339	96	949
22	α	74	1jvr	NMR	0.09	3.85	0	3	95	171	1,130
23	α	74	1kjs	NMR	0.11	3.30	0	144	272	139	845
24	α	74	1ner	NMR	0.08	3.53	0	23	167	140	1,070
25	α	75	1hyp	X-RAY	0.08	4.41	0	0	14	102	1,284
26	α	76	1adr	NMR	0.11	4.04	0	0	27	88	1,285
27	α	76	1cc5	X-RAY	0.10	4.29	0	0	26	188	1,185
28	α	77	2 pac	NMR	0.12	4.24	0	0	92	170	1,137
29	α	81	1coo	NMR	0.11	4.21	0	0	16	132	1,252
30	α	83	1a1z	NMR	0.09	3.68	0	2	6	21	1,371
31	α	85	1cei	X-RAY	0.11	4.63	0	0	2	25	1,373
32	α	85	1ngr	NMR	0.11	3.28	0	9	75	127	1,189
33	α	86	1aca	NMR	0.14	3.77	0	3	71	84	1,242
34	α	86	2af8	NMR	0.09	3.33	0	8	78	86	1,228
35	α	87	1a6s	NMR	0.11	4.12	0	0	47	165	1,188
36	α	87	1ddf	NMR	0.11	3.95	0	1	2	41	1,355
37	αβ	25	5znf	NMR	0.17	0.78	537	213	293	199	157
38	αβ	43	1ptq	X-RAY	0.21	5.25	0	0	0	36	1,363
39	αβ	52	1ap0	NMR	0.14	5.75	0	0	0	8	1,392
40	αβ	52	1bor	NMR	0.17	4.79	0	0	1	9	1,389
41	αβ	56	1aa3	NMR	0.12	2.21	23	131	133	111	1,001
42	αβ	56	1orc	X-RAY	0.09	3.05	0	69	230	347	753
43	αβ	57	1pgx	X-RAY	0.17	2.22	76	182	185	308	648
44	αβ	59	1tif	X-RAY	0.16	2.64	1	68	268	382	680
45	αβ	60	2ptl	NMR	0.18	2.53	43	304	186	78	788
46	αβ	62	1dol	X-RAY	0.14	3.91	0	1	2	7	1,390
47	αβ	63	1leb	NMR	0.12	2.42	12	60	126	181	1,020
48	αβ	65	1tnt	NMR	0.14	3.57	0	1	18	42	1,338
49	αβ	65	2fmr	NMR	0.16	3.34	0	26	146	110	1,118
50	αβ	66	1fwp	NMR	0.19	5.10	0	0	0	24	1,375
51	αβ	66	1sap	NMR	0.10	3.33	0	5	14	34	2,346
52	αβ	66	2fow	NMR	0.13	3.02	0	133	215	174	877
53	αβ	67	1ctf	X-RAY	0.18	2.91	1	195	250	140	867
54	αβ	68	1stu	NMR	0.12	3.69	0	2	19	70	1,309
55	αβ	69	2bby	NMR	0.10	3.79	0	6	158	141	1,372
56	αβ	69	4ull	NMR	0.15	5.08	0	0	0	28	1,372
57	αβ	71	1bb8	NMR	0.09	6.37	0	0	0	0	1,400
58	αβ	71	1vig	NMR	0.14	3.70	0	2	18	12	1,368
59	αβ	72	1afi	NMR	0.19	2.20	11	163	105	65	1,056
60	αβ	72	1lea	NMR	0.14	3.65	0	4	98	118	1,180

TABLE I. (Continued)

Number	2°	Residues	PDB code	Experiment	Relative contact order	Lowest CαRMSD	Number within CαRMSD bins				
							<3	3 to 4	4 to 5	5 to 6	6<
61	αβ	72	5icb	X-RAY	0.11	3.15	0	70	227	156	946
62	αβ	76	2ula	NMR	0.17	4.16	0	0	7	13	1,380
63	αβ	77	1vcc	X-RAY	0.11	3.76	0	1	4	57	1,338
64	αβ	78	1aoy	NMR	0.11	4.17	0	0	39	103	1,258
65	αβ	81	2fxb	X-RAY	0.16	5.50	0	0	0	9	1,391
66	β	42	1qyp	NMR	0.19	3.14	0	78	187	100	1,034
67	β	48	1vif	X-RAY	0.20	0.61	225	26	79	67	1,002
68	β	53	1bq9	X-RAY	0.18	2.86	1	18	96	112	1,173
69	β	54	2cdx	NMR	0.18	6.77	0	0	0	0	1,399
70	β	55	1ark	NMR	0.19	3.25	0	12	129	134	1,124
71	β	55	5pti	X-RAY	0.17	4.05	0	0	19	149	1,231
72	β	60	1msi	X-RAY	0.19	5.59	0	0	0	12	1,387
73	β	61	1tuc	X-RAY	0.20	4.54	0	0	14	152	1,234
74	β	62	1aiw	NMR	0.15	6.56	0	0	0	0	1,400
75	β	64	1csp	X-RAY	0.16	3.35	0	33	196	102	1,068
76	β	65	1kde	NMR	0.17	5.92	0	0	0	2	1,398
77	β	66	1sro	NMR	0.14	2.60	2	103	198	112	984
78	β	69	1pse	NMR	0.17	5.81	0	0	0	1	1,399

(1), we sought to produce a decoy set using Rosetta for as large a set of proteins as possible. We started with a previously defined set[7] and augmented it with proteins from a set selected by another group.[8] The final set comprises 78 proteins, which are listed in Table I, and includes proteins that range from 25 to 81 residues in length and from 0.08 to 0.21 in relative contact order.[9] Based on the native structures, we loosely group the sets into one of three categories: 36 all α-helical, 29 mixed α-helical and β-sheet, and 13 all β-sheet (see Table I). For each of these proteins, Rosetta simulations were used to produce 1,000 independent conformations following the original protocol and using the energy functions described previously (see Methods).[7] Side chains were added to these models using the energy function and Monte Carlo search procedure described in Kuhlman and Baker[31]. To satisfy requirement (2) that the decoy set contain structures within the native energy basin, large numbers of additional simulations were carried out and only conformations close in CαRMSD to the native structure were saved. These structures totaled ~400 per protein and were added to the initial sets of 1,000 conformations. The average CαRMSD of the most native-like decoys is 3.5 Å. The number of structures within 3, 4, 5, and 6 Å CαRMSD for each protein are listed in Table I. Requirement (3) is satisfied because all structures are the result of Rosetta conformational searches, which usually produce structures with quite protein-like properties. Requirement (4) is satisfied for the initial sets of 1,000 conformations, but is broken to some extent by the inclusion of the additional low CαRMSD decoys. Each individual simulation was ignorant of the native structure, however, and thus the sets are quite different from previous sets such as those generated using molecular dynamics starting from a native structure[10] or built up using information derived from the native structure.[11] Rather than perturbing the native state, each simulation begins from an extended chain; the native structure is only used in the selection of a subset of the decoys. These steps insure that the resulting models populate minima spread throughout conformational space. The new, all-atom Rosetta decoy set is diverse, well populated with near native conformations, and well suited for evaluation of scoring functions.

We attempted to relax structures following addition of the side chains by explicitly minimizing the all-atom energy. Through this minimization, we hoped to push the structures towards the native structure and to obtain better discrimination. Extensive testing of this approach using a number of procedures for backbone relaxation and side chain selection did not prove fruitful (see Methods). We did find that we could improve structures whose backbones were less than 2 Å CαRMSD to the native (Fig. 1), which is similar to what was found to be the limit for side chain packing.[12] Most of our starting models, however, were well beyond 3 Å CαRMSD to the true structure (Table I). While the relaxation procedure did not make the models more native-like as measured by CαRMSD, it does make the models more physically consistent by removing bad clashes, and hence the decoy set available for distribution and used in the tests below consists of the relaxed structures.

Evaluation of Scoring Functions

An accurate energy function should on average attribute lower energies to native-like conformations than to non-native conformations. A useful measure that captures the extent to which a function has this property is the enrichment: the fraction of low CαRMSD models in a low-energy subset of the total decoy population, divided by the fraction of low CαRMSD models in the total population (see Methods). Enrichment values greater than one indicate that the function enriches for lower CαRMSD structures, and vice versa. An alternative measure is the native Z score: the number of standard deviations separating the

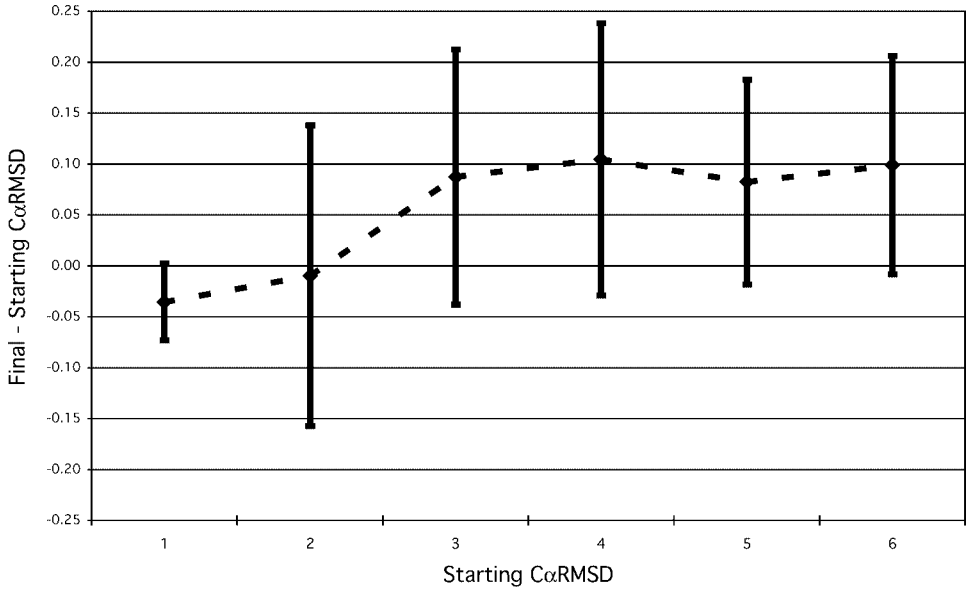

Fig. 1. Change in CαRMSD resulting from decoy refinement. The change in CαRMSD (final-starting) is binned according to starting CαRMSD values. Negative values indicate models moved closer to the native structure after refinement. We only show the bins from 0 to 6 Å, although bins up to a 15 Å are populated.

native structure energy from the average energy of the decoys. For each of the energies described below, we calculated the enrichment of the decoys and the Z-score of the native structure. For those energies involving side chains, we also calculated the Z-score for a native structure with side chains repacked using the same protocol as the decoys. We show the overall average for this repacked Z-score, as well as contrast the averages for structures solved by X-ray crystallography vs. those solved by NMR. Negative Z-scores indicate the native structure is lower in energy than the average value for decoys, whereas positive Z-scores indicate the native energy is higher than this average. We will break up our discussion loosely upon types of energies as organized in Table II.

Original Rosetta energies

These energies are primarily knowledge-based, probabilistic distribution functions used in the initial generation of the models. Since all of these energies use the reduced representation, repacking the side chains has no effect. The decoys are heavily minimized on the residue-environment and residue-residue energies and, as a result, the Z-scores of their native structures are poor. Even so, the energies exhibit enrichment for more native-like decoys.

Van der Waals interactions

Close packing of side chains is a characteristic feature of globular proteins.[13,14] We separately analyzed the attractive and repulsive parts of the Lennard Jones (LJ) interactions. Two different repulsive terms were considered: one with a truncated $1/r^{12}$ dependence, and one with a reduced linear dependence (see Methods). Table II shows the enrichments and Z-scores for the attractive and repulsive terms separately, and in combination. All the different LJ terms exhibit good Z-scores to the native structure, which decrease for the repacked native structure. A striking result is the much poorer repacked Z-score from both LJ total terms for NMR structures compared to X-ray crystal structures, which reflects the greater deviation of sidechian conformations in NMR structures from the canonical rotamer conformations used in the repacking calculations.

Implicit solvent models

The large energy decrease associated with desolvating non-polar atoms provides much of the driving force for protein folding. Explicit solvent models, which treat the solvent in atomistic detail, clearly are the most physically realistic, but are computationally prohibitive. Therefore, many groups have developed implicit solvent models, which can be readily tested using our decoy set. The near-native decoy enrichments and the native Z-scores are listed in Table II.

The Generalized Born (GB) model is an implicit solvent model that takes into account charge-charge interactions in vacuum screened by polarization on the solvent-solute boundary, the desolvation penalty of bringing a charge inside a protein cavity (charge self-energy), and the cost of making a solute cavity in solvent.[15] We implemented a version of the GB model compatible with the AMBER force field[16] developed previously by other groups[17–19] that is known to reproduce fairly well the electrostatic energies obtained through a solution of the Poisson-Boltzman equation. The GB model uses one dielectric constant for the solvent and one for the solute (we used a dielectric constant of one for the protein interior and 80 for the solvent surrounding the molecule), even though the dielectric constant is not well defined in the protein interior where heterogeneous and nonuniform distributions of polar and nonpolar atoms may have quite different mobilities depending on the

TABLE II. Evaluation of Energy Functions

Energy/score	Enrichment (15 × 15%)	Z score (native)	Z score (native repacked)	Z score XRAY (native repacked)	Z score NMR (repacked native)
Section I: Centroid/backbone					
Residue-environment (structural)	1.22	1.22			
Residue-residue (pair)	1.33	1.14			
Hard sphere repulsion	0.98	−0.53			
Strand assembly in sheets	0.99	−0.18			
Strand orientation	1.41	−1.38			
Strand packing	1.38	−0.98			
Helix-strand packing	1.04	0.45			
Section II: All atom					
LJ attractive	1.40	−1.48	−0.98	−0.97	−0.98
LJ attractive side chain only	1.35	−1.47	−0.67	−0.85	−0.58
LJ repulsive, capped	0.85	−1.09	4.37	1.19	5.87
LJ repulsive, linear	0.78	−1.44	3.10	−0.02	4.57
LJ repulsive, linear, side chain only	0.78	−1.48	2.41	−0.73	3.89
LJ total, capped	0.92	−1.19	4.38	1.15	5.90
LJ total, linear	1.14	−2.48	2.83	−0.66	4.47
LJ total, linear, side chain only	1.26	−2.86	1.56	−1.57	3.04
Coulomb	1.14	−1.52	−0.68	−1.09	−0.49
Screened Coulomb	0.87	−0.96	−0.26	−1.56	0.05
GB desolvation	0.63	1.51	1.10	0.16	1.55
GB SA	1.61	−1.29	−0.97	−1.16	−0.89
GB total	0.63	1.08	0.91	1.00	1.56
SASA–ASP	1.53	−1.60	−0.97	−0.99	−0.95
Effective solvent	0.93	1.77	0.61	0.55	0.65
Main chain hydrogen bonding	1.01	−1.16	−1.16	−3.11	−0.24
Side chain hydrogen bonding	0.97	−2.05	−0.36	−0.69	−0.20
Section III: Combined					
Centroid/backbone	1.60	0.82			
All atom	1.86	−4.48	−0.92		
All atom α	1.53	−7.77	−1.26		
All atom αβ	2.08	−1.94	−0.44		
All atom β	2.30	−1.04	−1.07		

rigidity of the structural element/side chain to which they are attached. We found that Coulomb energies, screened Coulomb energies, and GB surface area (cavity) terms produce reasonable enrichments and Z-scores. The total GB energy is worse because the atomic desolvation penalties (see GB Desolvation in Table II) tend to disfavor native structures. A more rigorous examination of electrostatic models has been done recently using Rosetta.[33]

A scaled solvent accessible surface area term SASA-ASP[20] exhibits the second best enrichment and a good Z-score. Even for the repacked native structures, the Z-score is good and consistent between X-ray and NMR structures. Exhibiting an enrichment less than one and a positive Z-score, the Lazaridis and Karplus effective solvation model[21] does not work well in isolation, probably because relaxed decoys tend to have larger surface areas than the native structures, and therefore the desolvation penalty for burying polar atoms is smaller in decoys. This phenomenon also occurs when the GB model[15] is used to compute desolvation penalties (see Table II). On the other hand, the GB SA term,[22] which approximates the cost of making an empty solute-sized cavity in the solvent, exhibits behavior similar to that of the SASA-ASP model. This model's relative success in discriminating native and near-native structures likely results from larger decoy surface areas relative to native proteins. We also experimented with an SASA-based term that penalized buried polar atoms,[23] but found that it did not actually help in discrimination and in some cases caused the structures to unfold during the relaxation protocol in order to avoid the penalty.

Hydrogen bonding

We used an empirical, orientation-dependent hydrogen bonding potential developed by Gordon and Mayo.[24] With enrichment values around one, no hydrogen bonding term shows enrichment for native-like decoys. As shown by the Z-score, main-chain hydrogen bonding distinguishes the native structure from decoys, and as expected, the Z-score for the native repacked is the same because the backbone has not changed. This good discrimination results primarily from X-ray structures with a Z-score of −3.11, whereas the NMR structures have a poor Z-score of −0.24. For side chain hydrogen bonds, native structures have a good Z-score to decoys, but the Z-score from structures repacked

TABLE III. Rosetta All-Atom Energy: Weights and Contribution

	Weights			% Contribution		
Score	α	αβ	β	α	αβ	β
Residue-environment (structural)	1.75E-02	2.15E-02	1.55E-02	0.01	0.02	0.01
Residue-residue (pair)	1.75E-02	2.15E-02	1.55E-02	0.01	0.02	0.01
Helix-strand packing	—	1.27E-02	—	—	0.00	—
Strand packing	—	3.32E-02	—	—	0.03	—
Strand orientation	—	1.15E-01	2.60E-01	—	0.01	0.01
Side chain hydrogen bonding	1.56E-02	—	1.13E-02	0.01	—	0.00
LJ repulsive	4.65E-04	2.39E-04	5.24E-04	0.82	0.59	0.70
LJ attractive	1.01E-02	—	—	0.03	—	—
SASA · ASP	1.21E-02	2.39E-02	3.60E-02	0.12	0.33	0.27
Dunbrack	1.70E-02	2.14E-02	4.88E-03	0.01	0.01	0.00

Dashes indicate the term produced negative correlations from the fitting procedure (see Methods) and were not used (set to zero) in the energy function.

on native backbones is not as good at −0.36. In this case, the X-ray structures are only slightly better than NMR structures, but both groups exhibit poor discrimination with Z-scores larger than −1. The current version of Rosetta uses a more accurate orientation dependent hydrogen bonding potential consistent with both the distributions of hydrogen bond geometries in high resolution protein structures and with quantum mechanical calculations.[33,34]

Combined energies

In Table II (Section III), we evaluate combinations of scoring terms. The first is the centroid/backbone energy used in generating the starting decoys, which provides a good enrichment of near native decoys, but is poor at discriminating the decoys from the native structure. To combine the LJ terms with the implicit solvation and backbone/centroid based terms, we used logistic regression (see Methods) to obtain relative weights, which are shown in Table III. The combined all-atom energy provides on average better discrimination and enrichment. Breaking the numbers down based on type of secondary structure, we see that the enrichment primarily results from improvements for proteins with β-sheets, but the all-helical structures have a favorable Z-score. We also see a modest improvement in enrichment after refinement, but the Z-score decreases because the refinement moves the structures to more native-like LJ repulsive energies.

Use of Potential to Improve Rosetta Predictions

The enrichments for the scoring function developed above are significant. Given the failure of the refinement procedure to consistently improve the CαRMSD, we turned to the strategy used in CASP4[4] to normalize the contact order distribution[25] and cluster the models.[26] We first applied this procedure to the augmented decoy set described above [Fig. 2(A)] and then to a more objective test of 100,000 unbiased decoys generated by Rosetta (see below) [Fig. 2(B)]. In both plots, the y-axes are the number of sets and the x-axes are CαRMSD to the native structure. Each selection scheme makes a set of 78 predictions of the native structure. Figure 2 plots the number of predictions that are within the CαRMSD cutoff. The best that a selection scheme can do is the lowest CαRMSD decoy generated by Rosetta, shown by the thick solid line on the left of the plots. In the ideal case, all 78 predictions would be under 1 Å CαRMSD to the native structure, but the performance of the structure generation algorithm limits the predictions. The worst a scheme can do is random, shown by the thick broken line on the right in Figure 2. In Figure 2(A), we show the performance of various selection schemes. Rosetta's original centroid backbone (CNBB) energy,[7] (filled diamonds), does just better than random most of the time. We get an improvement if we use the all-atom (AA) energy on the starting structures as shown by the filled circle line. If we cluster the top 33% structures by the all-atom energy using either starting or refined structures, we can do as well if we choose the top 33% by CαRMSD. This set is somewhat biased since it is augmented with extra low CαRMSD structures. For a more objective test of the performance, we generated 100,000 decoy structures for each protein. The lowest scoring 1,000 structures were selected and subjected to the standard clustering procedure used to select Rosetta models (see Methods). As is evident in Figure 2(B), the selection with the full atom scoring function performs better than clustering 1,000 random structures, but not as well as clustering of 1,000 selected by CαRMSD.

As an even more stringent test of our procedure, we used it for predictions in the fourth Critical Assessment of Structure Prediction (CASP4).[4,27] Figure 3 shows our results for two proteins. In Figure 3(A), we show the distribution of target 102's (1e68, bacteriocin AS-48[28]) all-atom energies for all the decoys generated in the top plot. In Figure 3(A) (bottom) is the selected and relaxed top 33%. The bottom scatter is not a direct subset of the top one, since the refinement changes the structures and their energies. The open circles are the clusters and the closest prediction is shown next to the native structure on the bottom. The all-atom energy for the cluster centers is correlated with their CαRMSD. Also, our best guess was just under 4 Å CαRMSD to the native structure, while if we

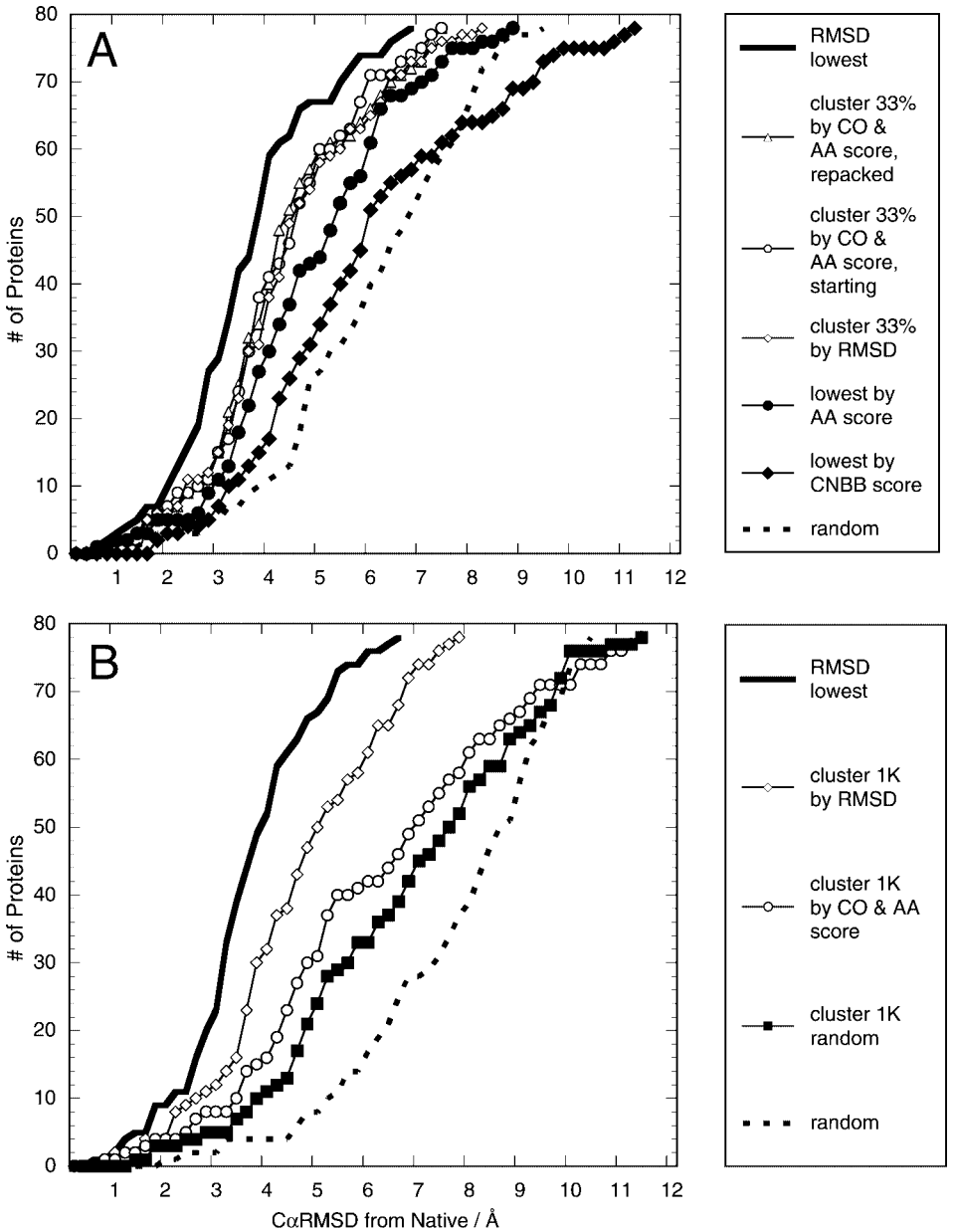

Fig. 2. Performance of decoy selection methods. The y axis is the number of proteins for which the lowest CαRMSD of five selected decoys is at or below the CαRMSD on the x axis. For both A and B, the bold black line labeled RMSD is the best-case scenario of selecting the lowest CαRMSD decoy for each protein. If our decoy set were near perfect, then the bold line would be close to vertical along the y-axis, indicating all our protein decoys sets contained near native structures with CαRMSDs close to 0. The remaining lines plot the performance of different selection procedures in a way that imitates the conditions used in CASP.[27] We selected the top five decoys from a decoy set based on the post-filtering selection protocol (see Methods for more details) for each of the 78 target proteins in Table I. For energy functions, we used either the all-atom energy (AA) or the centroid backbone energy (CNBB). As described in the post-filtering section of the Methods, Cluster indicates that we clustered that percent or number of decoys, and CO indicates we used contact order along with the energy function to normalize our contact order distribution before clustering. To compare against the worst-case scenario, we plotted the average of the lowest RMSD s of 100 random selections of 5 structures at a time. **A:** The performance on the Rosetta all-atom decoy set with ~1,400 structures per native structure. Starting indicates that the energies were calculated before refinement and Repacked indicates energies after refinement. **B:** The performance on the 100,000 decoys with the AA energy without refinement. In this part, we also show results from clustering the top 1,000 closest decoys to the native structure (cluster 1K by RMSD), and the results of an average over 100 sets when we clustered 1,000 decoys chosen at random (cluster 1K by random).

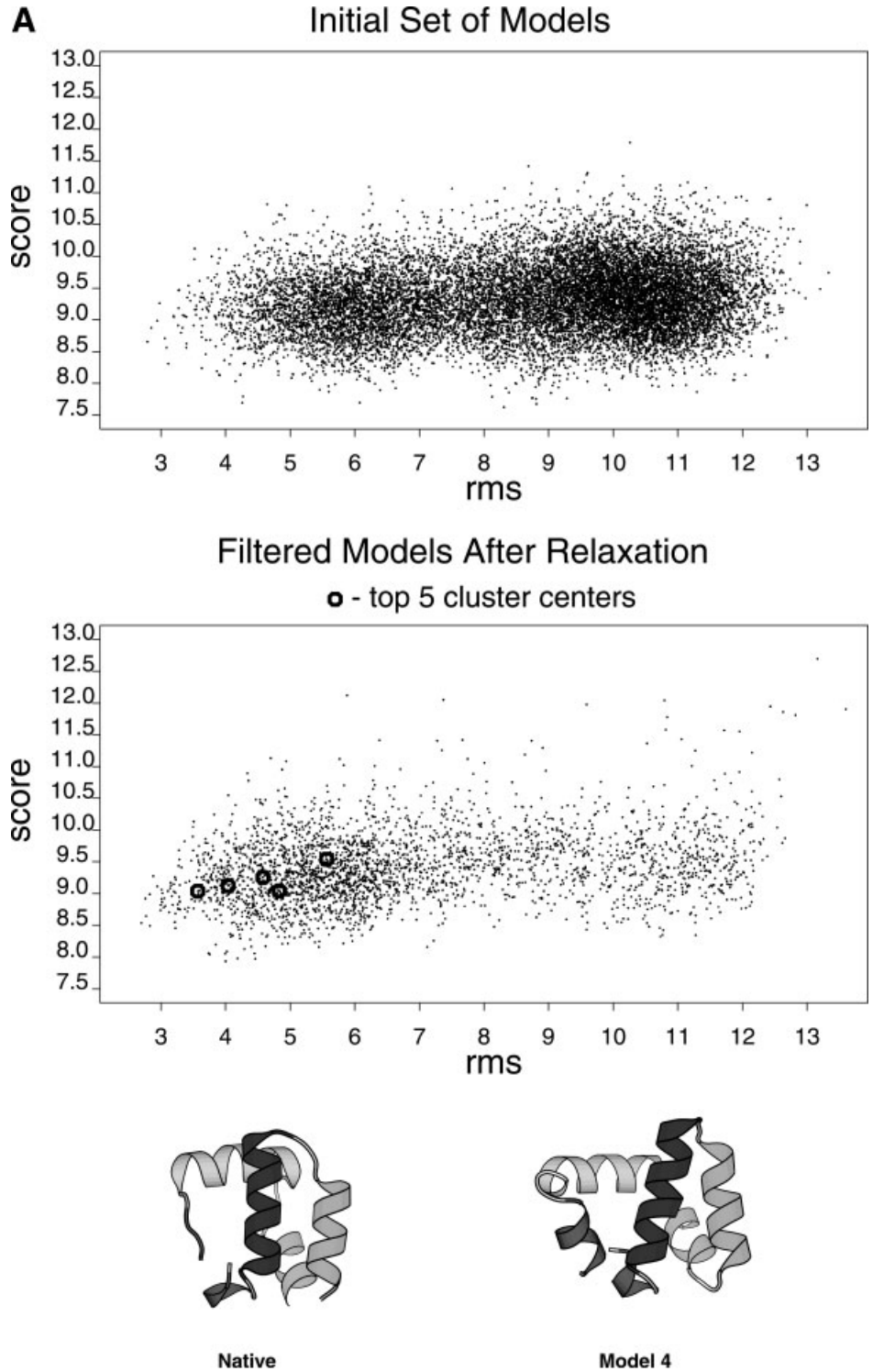

Fig. 3. CASP 4 Targets. **A:** Results from target 102 (1e68, bacteriocin AS-48[28]). **Top:** The all-atom energy for all the decoys generated. **Bottom:** The top 33% selected and refined by the all-atom energy with clusters represented by open circles. At the bottom, the structure of the closest match is shown next to the native structure.

had chosen based on energy alone, we would have picked the structure at just over 4 Å CαRMSD. In Figure 3(B), we show the same plots for target 106 (1ijx, secreted frizzled protein 3[29]). In this case, the cluster centers' energies do not correlate with CαRMSD, but the lowest energy structures were some of the closest to the native structure.

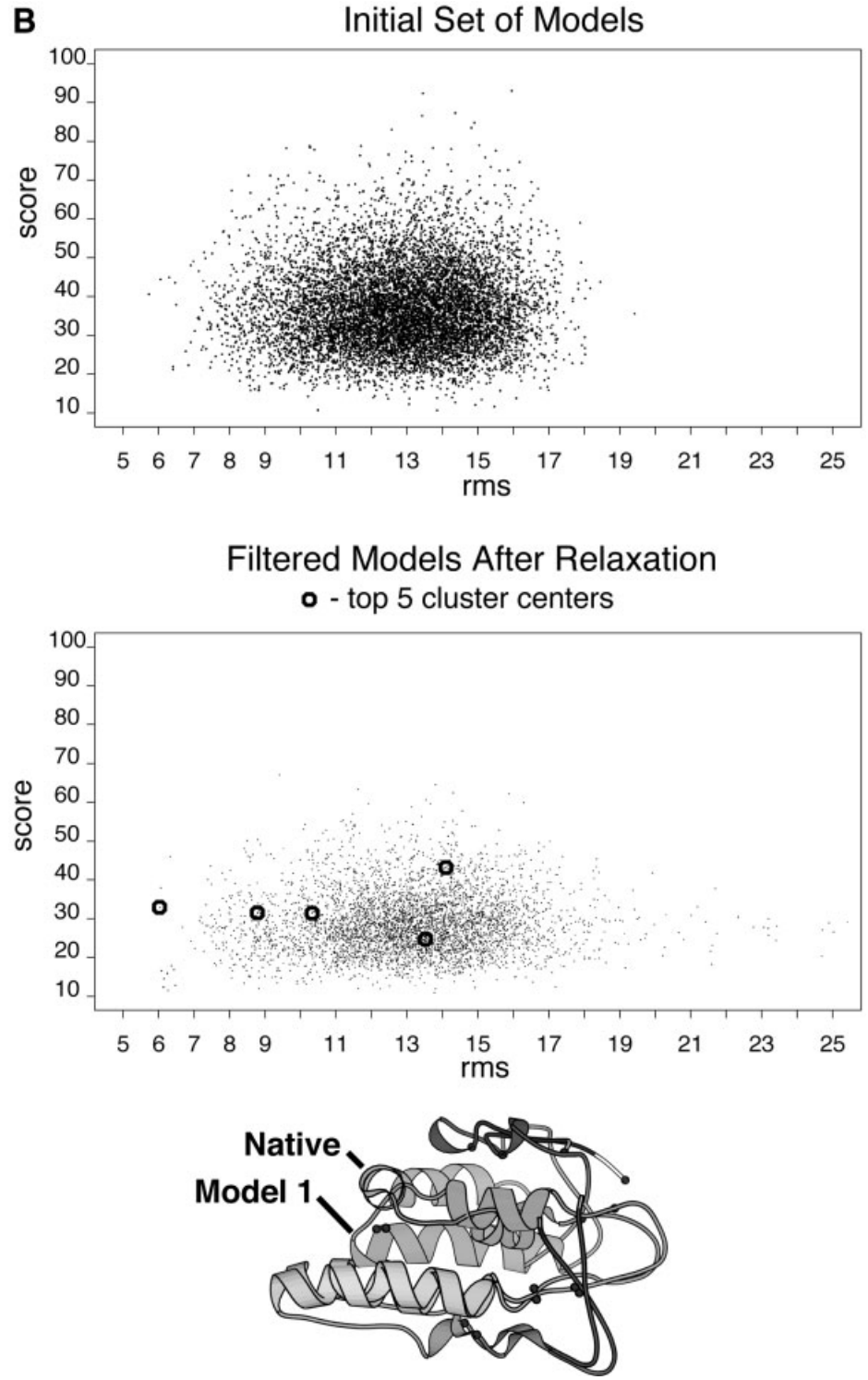

Figure 3. **B:** Similar to A for target 106 (1ijx, secreted frizzled protein 3[29]). The native structure and the closest decoy are shown superimposed. Circles on structures indicate placement of cysteine groups for disulfide bonds.

DISCUSSION

We have developed a large and diverse decoy set that provides a stringent test for evaluating energy functions. The set consists of 78 single domain proteins with varying degrees of secondary structure and length from 25 to 87 residues (Table I). By augmenting the set with decoys close to the native structures, the set also provides a good challenge for scoring functions and selection schemes to

test themselves against the local minima around the native state. As has been previously shown, many decoy sets suffer some weakness that can be exploited to find a good correlation between an energy and CαRMSD.[3] Such weaknesses result from sampling only structures within the native state well or sampling of structures perturbed from the native state. Since this decoy set was not generated with a bias for the native state, it samples many states outside of the native well and in conformational areas that are likely false minima for scoring functions. Table II and Figure 2 illustrates the value of this decoy set in effectively evaluating energy functions. The results indicate that scoring functions are discriminatory for the native structure, but are not good at finding the absolute closest decoy structure. We anticipate this Rosetta All-atom decoy set should be broadly useful to developers of potential functions in assessing the abilities of scoring functions and selection schemes. Using the all-atom energy function, we have been able to improve the discriminatory ability of the Rosetta algorithm for de novo, protein structure prediction. There is room for improvement, however, because this procedure cannot consistently identify nearest native structures (Fig. 2). Future work will encompass further testing of energies as well as developing better methods for increased refinement of decoys towards the native structure.

METHODS

Rosetta Method and Decoys

As described in Simons et al.,[6] Rosetta is a fragment-based, *de novo*-structure generation method. Two fragment libraries (3mer and 9mer) are built based on the secondary structure prediction of the target sequence.[4] Starting from an extended amino acid chain, the method inserts a fragment and then evaluates a centroid/backbone-based potential function[7] consisting of knowledge-based terms, secondary structure terms, and a check for overlapping residues. New configurations are accepted based on Monte Carlo Metropolis criterion. Ten thousand 3mer insertions follow 10,000 9mer insertions. The resulting structures have been previously described[7] and provide a point of departure for this work.

Enriching Set for Low CαRMSD Decoys

To enrich for decoys that were closer to the native structure but were not biased by the native structure, we used the Rosetta method described above, but only output structures within 15% CαRMSD of the previously lowest CαRMSD structure. Additional decoys (~400) were produced in this way for each set.

Scoring Functions

Lennard-Jones (LJ)

We used two functions for the LJ repulsive. The one referred to as "capped" in Table II was the standard $1/r^{12}$ repulsive component of a 6–12 potential with a cutoff at 100 kcal/mol. The other called "linear" switched to a linear function from 0 to 10 kcal/mol for all repulsive values. The "linear" function was used in the final all-atom scoring function described in Table III and in Figures 2 and 3.

Solvent accessible surface area (SASA)

The SASA was computed using a fast, approximate method,[30] where the surface of an atom was represented by grid points and stored in binary. Overlap (buried) grid points were pre-computed based on distance and angle between atom centers and were switched off using binary operators. The grid points that remained "on" represented exposed surface. The surface area was calculated by summing the areas of the exposed grid points. For the scoring, we multiplied the SASA by an atomic solvation parameter (ASP).[20,23]

Addition of Sidechains and Refinement

Side chains were added to the centroid/backbone decoy structures using a simulated annealing method described previously[31] and a backbone-dependent library of rotamers.[32] The move set for relaxing the backbone consisted of two types: small random changes in phi, psi torsion angels of a single residue (small move), and a three-residue fragment insertion followed by conjugate gradient minimization of the perturbation on the structure by varying the backbone torsion angels of the flanking residues (wobble move). The number of moves for a particular structure was set to four times the number of residues. The potential used during refinement included the original Rosetta centroid/backbone-based terms supplemented with hydrogen bonding and the LJ function with the linear repulsive term. The procedure began with an initial minimization over the entire structure. This stage was followed by a set of small moves using a reduced rotamer set comprised of the top three most prevalent rotamers for a buried residue and two for an exposed one. Next, the resulting structure was minimized using a slowly increasing weight for the LJ repulsive term in three steps. This phase was followed by a set of small moves followed by wobble moves, again using the reduced rotamer set for packing. After another set of minimizations, a full set of rotamers was used. As before, the order was a set of small moves, minimization, a set of small moves with minimization and wobble moves, and final minimization.

Scoring Potential Optimization/Fitting

Logistic regression aimed at optimizing the recognition of the lowest CαRMSD 5% of the decoys for each protein was used to weight the components of the energy function. Instead of treating the protein set as whole, we split the set of structures into three groups based on secondary structure (α, αβ, and β; see Table I) as a logical way to separate protein environments and improve the discrimination of our function. All fitting was done using the program SPLUS © Mathsoft. Weights with a negative correlation have been set to zero in the total energy function. Table III shows the results of the fitting, where those with negative correlations are given a dash.

Calculation of Enrichment

Enrichment was calculated based on the union of the top 15% of decoys by energy and top 15% by CαRMSD to the native structure. Dividing this number by what would be

expected for a uniform distribution (15% · 15% · number in set) yields the enrichment. Values greater than one indicate an enrichment over a uniform distribution.

Post Filtering

We experimented with several different procedures to select "good" decoys from the large decoy sets using an energy function. The simplest method, and certainly the best given a perfect energy function, is simply to take the lowest energy decoys. However, given the imperfections in current functions, it is useful to take into account previous observations of the power of clustering methods to identify near native structures,[26] and of the tendency of Rosetta to generate an excess of low contact order structures.[25] If it is assumed that noise in the energy function prohibits accurate ranking of the quality of the decoys, but does allow the exclusion of physically implausible structures, a reasonable protocol is to select the lowest energy 1–10% of decoys, and then cluster this subset to identify the broadest minima in the energy landscape. Furthermore, to compensate for the uneven contact order distribution sampled by Rosetta, this procedure can be further elaborated by taking not the lowest energy 1–10% of decoys in the overall population, but a fixed number of low-energy structures in each of a number of independently considered contact order ranges. For example, if 90% of decoys for a given protein fall in a contact order range only populated by 10% of native proteins in the same length range, simply selecting by energy could produce a considerable excess of low-contact order conformations. This can be remedied by selecting an equal number of low-energy structures from across the contact order range, which results in a population of low-energy structures evenly distributed with respect to contact order. This low-energy, contact order normalized population can then be clustered and the five largest centers selected as above.

The contact order of native proteins increases with increasing length, and this must be taken into account in defining contact order ranges in which equal numbers of native protein structures are expected to fall. We separated native proteins ranging from 50 to 160 residues into all α, all β, and αβ, and computed the mean contact order (or 50th percentile), the 5th percentile value, and the 95th percentile for each 10-residue, protein length interval. Each of these three sets of points was then fit to simple linear functions of protein length. The slopes (m) and y-intercepts (b) of the lines thus obtained are given in Table IV, where x is the number of residues and y is either the 5th, 50th, or 95th percentile contact order value. For the contact order normalization described in the previous paragraph, 5% of structures were taken from below the 5th percentile value for the length of the protein, 45% from between this value and the 50th percentile value, and so on.

For example, to reduce a population of 100,000 decoy structures to 1,000 prior to clustering, from decoy structures in the <5% bin, we would choose the lowest energy 5% (50 structures) based on the energy function, from the 5 to 50% bin, we would choose the best 45% (450 structures) based on the energy function, the same number for the 50 to 95% bin, and finally the lowest energy 50 from the >95% bin for a total of 5 + 450 + 450 + 5 = 1,000 structures to be clustered. Once clustered, the top 5 centers from the largest clusters are selected.

TABLE IV. Parameters for Contact Order Cutoff Lines: Slope m and y-intercept b

	α		αβ		β	
	m	b	m	b	m	b
Upper 95th percentile	0.22	5.50	0.25	8.75	0.34	2.00
Middle 50th percentile	0.18	2.07	0.19	7.36	0.24	6.00
Lower 5th percentile	0.15	3.00	0.14	3.25	0.15	8.00

ACKNOWLEDGMENTS

The authors thank Ingo Ruczinski for SPLUS expertise, Eric Alm for the SASA code, and Keith Laidig for tireless technical support and system administration. We thank Charlie Strauss, Dylan Chivian, and especially Kira Misura for discussion and careful reading of the manuscript. J.T. thanks the National Science Foundation (Biological Informatics Fellowship). B.K. is a fellow of the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation. R.B., A.M., C.A.R., and D.B. were supported by the Howard Hughes Medical Institute. Thanks also to J. Brad Holmes and Jordan Tayce for preparing the decoy set for public use.

REFERENCES

1. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. Curr Opin Struct Biol 2002;12:176–181.
2. Felts AK, Gallicchio E, Wallqvist A, Levy RM. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. Proteins 2002;48:404–422.
3. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;2664:831–846.
4. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;45(Suppl 5):119–126.
5. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;(Suppl 3):171–176.
6. Simons KT, 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 1997;268:209–225.
7. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
8. Eyrich,VA, Standley DM, Friesner RA. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. J Mol Biol 1999;288:725–742.
9. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
10. Huang ES, Subbiah S, Tsai J, Levitt M. Using a hydrophobic contact potential to evaluate native and near-native folds gener-

ated by molecular dynamics simulations. J Mol Biol 1996;257:716–725.
11. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258: 367–392.
12. Chung SY, Subbiah S. How similar must a template protein be for homology modeling by side-chain packing methods? Pac Symp Biocomput 1996;126–141.
13. Liang J, Dill KA. Are proteins well-packed? Biophys J 2001;81:751–766.
14. Chothia C. Principles that determine the structure of proteins. Annu Rev Biochem 1984;53:537–572.
15. Still WC, Tempczyk A, Hawley RC, Hendrickson TJ. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
16. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117: 5179–5197.
17. Hawkins GD, Cramer CJ, Truhlar DG. Pairwise solute descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.
18. Hawkins GD, Cramer CJ, Truhlar DH. Parameterized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. J Phys Chem 1996;100:19824–19839.
19. Jayaram B, Sprous D, Beveridge DL. Solvation free energy of biomacromolecules: Parameters for a modified Generalized Born Model consistent with the AMBER force field. J Chem Phys 1998;102:9571–9576.
20. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
21. Lazaridis T, Karplus M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. Science 1997;278:1928–1931.
22. Qiu D, Sherkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation, A fast analytical method for the calculation of approximate Born radii. J Phys Chem A 1997;1997:3005–3014.
23. Koehl P, Delarue M. Polar and nonpolar atomic environments in the protein core: implications for folding and binding. Proteins 1994;20:264–278.
24. Gordon DB, Mayo SL. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. J Comp Chem 1998;19:1505–1514.
25. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002:11:1937–1944.
26. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
27. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): Round IV. Proteins 2001;45(Suppl 5):2–7.
28. Gonzalez C, Langdon GM, Bruix M, Galvez A, Valdivia E, Maqueda M, Rico M. Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. Proc Natl Acad Sci USA 2000;97:11221–11226.
29. Dann CE, Hsieh JC, Rattner A, Sharma D, Nathans J, Leahy DJ. Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. Nature 2001;412:86–90.
30. LeGrand S, Merz KM. Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. J Comput Chem 1993;14:349–352.
31. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97: 10383–10388.
32. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230:543–574.
33. Morozov AV, Kortemme T, Baker D. Evaluation of models of electrostatic potentials. J Phys Chem B 2003;107:2075–2090.
34. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediciton of specificity and structure for proteins and protein-protein complexes. J Mol Bio 2003;326:1239–1259.