R. Kolodny[1]
M. Levitt[2]
[1] *Department of Computer Science, Stanford University, Stanford, CA 94305-5126*

[2] *Department of Structural Biology, Stanford University, Stanford, CA 94305-5126*

Received 18 July 2002; accepted 23 July 2002

Protein Decoy Assembly Using Short Fragments Under Geometric Constraints

Abstract: *A small set of protein fragments can represent adequately all known local protein structure. This set of fragments, along with a construction scheme that assembles these fragments into structures, defines a discrete (relatively small) conformation space, which approximates protein structures accurately. We generate protein decoys by sampling geometrically valid structures from this conformation space, biased by the secondary structure prediction for the protein. Unlike other methods, secondary structure prediction is the only protein-specific information used for generating the decoys. Nevertheless, these decoys are qualitatively similar to those found by others. The method works well for all-α proteins, and shows promising results for α and β proteins.* © 2003 Wiley Periodicals, Inc. Biopolymers 68: 278–285, 2003

Keywords: *protein decoy assembly; discrete comformation space; geometric constraints*

INTRODUCTION

The structural biology community has long focused on the very hard task of developing algorithms for solving the ab initio protein folding problem—namely, predicting protein structure from sequence. It seems appropriate in a volume dedicated to the memory of Shneior Lifson to recollect Levitt's first memories of discussing this problem with Lifson. This occurred at the seventh Ciba Foundation Symposium on Polymerization in Biological Systems, which was organized by Ephraim Katzir and held in London in the autumn of 1971. Levitt was in the third year of his Ph.D. and was only there as Aaron Klug had asked him to attend in his place. This meeting was attended by all the big shots of the time and Levitt was completely out of his depth. His paper was on the folding of nucleic acids[1] and in the extensive discussion period Lifson asked in his provocative way, "Who cares about how a protein folds. It just happens like a leaf falling from a tree." At the time Levitt was lost for words and had no real answer other than: "Because the problem is there we must tackle it. . .."

Reflecting back, we now understand more about what Lifson meant. He realized earlier than many that protein folding was a very complicated many-body problem dependent on the details of the underlying

Correspondence to: M. Levitt; email: Michael.Levitt@stanford.edu
Biopolymers, Vol. 68, 278–285 (2003)
© 2003 Wiley Periodicals, Inc.

energy function. He felt that one could probably approximate the energy function as he demonstrated so elegantly with his pioneering work on the Consistent Force Field.[2,3] He also appreciated then that, even if the underlying forces were simple, the dynamics would be complicated, just like a leaf falling through the air. Each time a sequence folded it would fold differently, so why attempt the impossible task of simulation? Well, although his argument showed great insight, he may have been wrong. We need to tackle these messy, many-body problems because that is what biology is all about. Although Shneior Lifson laid the foundation for modern computational biology (see Levitt[4]) and applied computers to what were very complicated organic molecules, in 1971 he may not have fully appreciated that the key to dealing with the messy details caused by the intrinsic complexity of biology must be the same computer simulations and computer models that he pioneered.

Ab initio prediction is commonly viewed as composed of two subproblems (1) generating candidate folds—called decoys; and (2) devising a scoring function, or energy potential, that discriminates between near native folds and others non-native folds amongst the decoys. Many studies[5–7] consider the parallel question of searching for conformations that have low energy, and address it using minimization-based methods that search the landscape of the scoring functions. Huang et al.[8] pointed out that a clean separation of the two subproblems is advantageous as it allows an easier comparison between different energy potentials, without penalizing scoring functions with a rugged landscape. This work is focused on the first subproblem—namely, sampling the conformation space for many near native structures.

A common strategy for generating protein fold decoy structures considers discretized versions of the conformation space. Conformation space has been discretized by means of lattice models[9–11] and off-lattice models.[12,13] Typically, there is a trade-off between the complexity, i.e., the size of the conformation space, and the precision with which shapes from the discretized space can approximate real protein structure.[13] Although the size of these discrete conformation spaces can be fairly large, they are finite and thus offer the hope that they could be explored sufficiently well to produce structures that will serve as good decoys for a protein structure predictor. Here, we explore structures in a novel discrete conformation space.

The number of proteins with known structures in the Protein Data Bank (PDB) increases steadily and is currently greater than 15,000.[14] Various studies have extracted knowledge from this database of structures to improve protein structure prediction. For instance, Simons et al.[15] exploited correlations between sequence and local protein structure for decoy generation; Fain and Levitt[16] devised helical protein decoys by constructing structures consistent with observed interhelical characteristics.

Our study relies on the observation that a small database-derived library of short fragments can adequately represent all protein structures, and uses this library to generate sets of protein decoys. We construct self-avoiding and compact protein decoys by repeatedly assembling pieces from a library of common protein fragments. The pieces used for the assembly of the chains are chosen at random, biased by the secondary structure sequence of the protein considered. The restrictions and the bias used in our scheme are based solely on the geometric nature of the protein and completely ignore all specific details of its amino acid sequence. Despite the extreme simplicity of this method, the sets of decoys generated include many structures that have a coordinate root mean square (cRMS) deviation smaller than 6 Å from the native conformation. This method works well for small all-α proteins, and reasonably well for an α and β protein.

METHODS

The representation we use for protein structures is based on a library of 20 fragments of protein backbone. Each fragment is 5 residues long and consists of the three-dimensional coordinates of its 5 Cα atoms. To generate the library we consider 200 protein domains whose three-dimensional structure was accurately determined. Each of these domains was broken into a series of nonoverlapping fragments, thus obtaining a total of 7133 fragments. These fragments were then grouped based on their cRMS deviations from one another, into 20 clusters: each a collection of geometrically similar fragments represented in the library by a single element—the cluster's centroid. Note that the fragments do not include any sequence specifics of the proteins, only structural information. This work has been described in detail elsewhere (Kolodny et al.[17]).

The 20 elements in our library serve as building blocks used in constructing protein fold decoys. To build a decoy, copies of fragments from the library are repeatedly added to extend a chain, until reaching the target length of n residues. Each added fragment is positioned by superimposing[18] the first three residues of the fragment on the last three residues of the growing chain, extending the chain by $5 - 3 = 2$ residues. The positioning of the first three residues in space determines the orientation of the fragment, so the two extending residues have a unique position. Figure 1 illustrates this procedure in two dimensions. To summarize, a string of m fragments encodes a unique structure of $5 + 2(m - 1)$

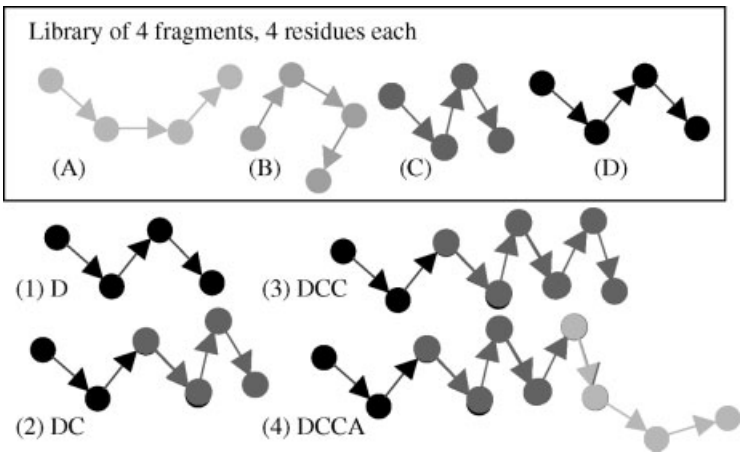

FIGURE 1 A two-dimensional example of constructing a chain from a library of 4 fragments with 4 residues in each. We construct the chain DCCA and illustrate the 4 construction steps. In each extension, the first two residues of the new fragment are superimposed on the last two residues of the chain (only 2 residues are used for positioning because the example is in 2D). Notice that library fragments can be reused and that the chain has a direction.

$= 2m + 3$ residues; equivalently, a structure of n residues that is constructed from library fragments is encoded by a string of $(n - 3)/2$ library elements.

The structures that can be constructed using the above scheme are protein-like in that they consist of fragments whose local chain conformations are like the ones common in real proteins. Structures generated this way also approximate real proteins well: combinations of overlapping fragments fit the 145 proteins of the Park and Levitt set[13] to 1.88 Å cRMS deviation on average.

Decoy generation procedures aim to build structures that are similar to the structure of a target protein. By the above, an obvious decoy generation scheme for a protein of length n is to enumerate all possible strings of length $(n - 3)/2$ over a 20-letter alphabet, and construct the corresponding structure for each one. The number of possible strings, $20^{(n-3)/2} \cong 4.47^n$, grows exponentially with the protein length, making enumeration tractable only for short proteins. This suggests an alternative scheme for generating decoys: sample the space of strings and construct the corresponding structures.

Secondary structure offers a coarse description of the backbone shape and can be predicted fairly well.[19] For a protein p of n residues, the first approximation secondary structure sequence is a string of symbols H, E, and C that indicates, for each residue, whether it is a part of a helix (H for helix), a strand (E for extended), or neither (C for coil). More formally: $s(p) \in \{H,E,C\}^n$. We refine our scheme to include the predicted secondary structure of the target protein. For every possible characterization of a secondary structure pair extending the chain, we consider the distribution of the different library elements that should be used for this extension. Equivalently, we calculate the conditional probability $P(l|\sigma)$ for every library element l, and every two-letter secondary structure sequence [string of two letters or $\sigma \in \{H,E,C\}^2$; see Appendix for details]. In the refined scheme, denoted as *biased sampling*, the ith fragment of the chain is chosen at random, according to a distribution of the library elements conditioned by $s^i(p)$, the secondary structure sequence of the two residues added by this fragment. The secondary structure sequence for a protein is derived from its native structure using STRIDE,[20] reflecting our assumption of a perfect secondary structure predictor. We emphasize that in this study our knowledge of the protein sequence influences the decoy generation process only via the use of the "predicted" secondary structure sequence.

We generate decoy structures that satisfy just two simple geometric properties of proteins: self-avoidance and compactness. Specifically, we enforce self-avoidance by requiring that any two Cα atoms be separated by at least 2.5 Å. Compactness is enforced by allowing only decoys in which all Cα atoms are separated by at most B Å, with B varying from 20 to 60 Å. As could be expected, the majority of the structures are either not compact or self-intersecting. Sampling the space and discarding the structures that fail these geometric tests is increasingly inefficient as the compactness condition is enhanced (smaller B). Furthermore, the construction of a decoy structure of n residues is an expensive computation involving $(n - 3)/2$ superpositioning steps. Consequently, a strategy that repeatedly constructs decoy chain prefixes that are eventually discarded is wasteful. To solve this problem, we use the technique suggested by Rosenbluth and Rosenbluth[21] for sampling only allowed chain paths. This technique assures us that whenever chain extension is possible the chain is extended, while its statistical weight is changed to account for the altered distribution. The details of this sampling method, as well as the modifications implied by it to random variable estimation, are discussed in the Appendix.

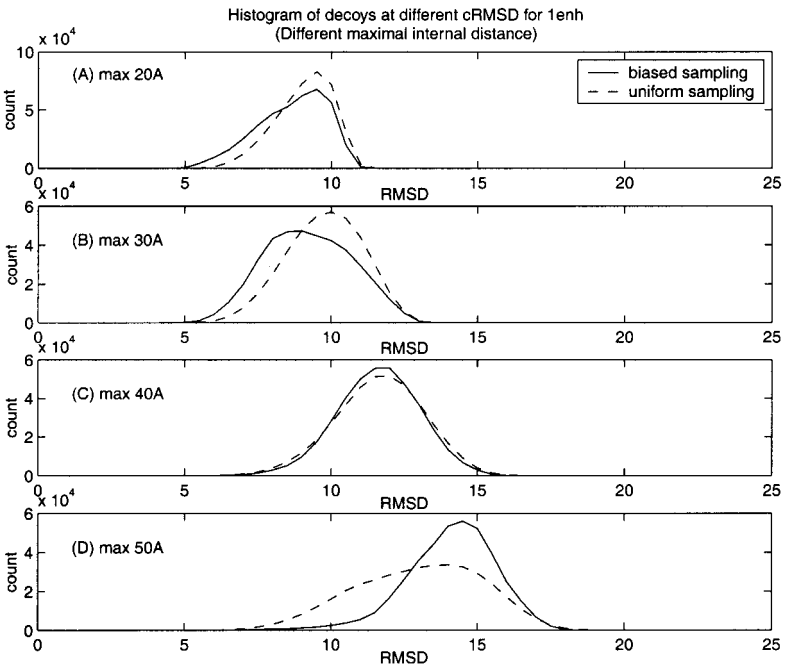

FIGURE 2 Distribution of the cRMS deviations (in Å) of 400,000 decoys for protein 1enh when enforcing maximum distance constraints of (A) 20 Å, (B) 30 Å, (C) 40 Å, and (D) 50 Å, using biased and uniform sampling.

RESULTS

We generate decoys for four proteins with PDB codes 1enh, 4icb, 2cro, and 1ctf, all having relatively short chains (54, 76, 65, and 69 residues, respectively). The first three proteins are all α: 1enh has three helices, 4icb and 2cro each have four helices; 1ctf is an α and β protein with three helices and three strands. We verify that the discretized conformation space that we use has structures that are close enough to the proteins studied, by constructing approximations to these proteins using the known native structures. The best approximations we found have a cRMS deviation of 1.41, 1.60, 1.41, and 1.43 Å from the native structures of 1enh, 4icb, 2cro, and 1ctf, respectively.

We investigate the different decoy structures generated for 1enh using uniform and biased sampling, with biased sampling based on the secondary structure sequence of the protein. We generate 400,000 decoys with maximal distance constraints of 20, 30, 40, and 50 Å (the native state of 1enh satisfies a maximal distance constraint of 26 Å). For every decoy we construct, we measure its cRMS deviation from the native state of the protein. Figure 2 plots the number of decoys found for a range of cRMS deviations from the native state in the various experiments. The plots show that the histograms shift to the left as the maximal distance constraint is tightened, demonstrating that decreasing the maximal distance generates more native-like decoys. The biased sampling technique offers better decoys than the uniform sampling when the maximal distance is 20 and 30 Å, but gives worse decoys when the maximal distance constraint is relaxed to 40 and 50 Å. This phenomenon can be explained by the following observation: decoy structures generated when the maximal distance constraint is greater than 40 Å tend to be relatively open. The decoys generated by the biased sampling are open chains with rigid helical parts embedded in the appropriate places. As all decoys are fairly far from the native state, the more compact ones seem to be best (even though they do not actually resemble the protein's structure). The biased decoys, with rigid parts along the chain, have fewer positions (the nonrigid ones) in which there is a probability of turning back and making a compact structure, and thus appear to be worse. This explanation is supported by the fact that the uniform sampling histogram is broader when the maximal distance constraint is relaxed while the biased sampling histogram is only shifted to the right. This phenomenon is even more pronounced when we considered larger maximal distances (e.g., 80 Å, data not shown). Similar behavior of (1) improved decoys as the distance constraint is enforced more strictly and (2) better decoys in the biased sampling when the maximal distance is small, is observed for all the

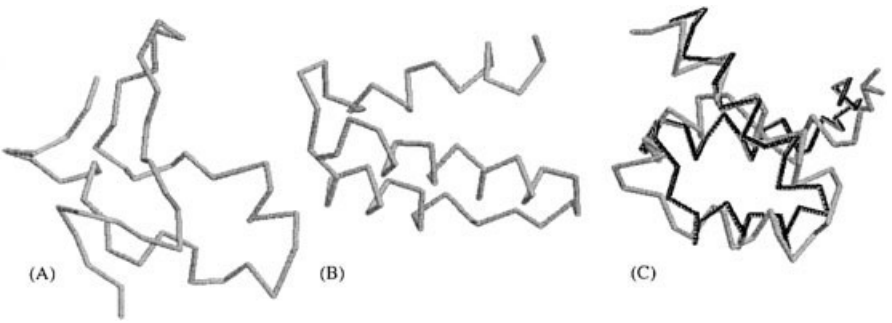

FIGURE 3 Decoy structures generated by our procedure for 1enh: (A) A typical structure generated by uniform sampling (5.94 Å cRMS deviation from the native structure). (B). A typical structure generated by biased sampling (5.72 Å cRMS deviation). (C) The native state superimposed on the best decoy found by biased sampling (3.9 Å cRMS deviation, the decoy is in the darker tone).

proteins we studied; their specific histograms are not shown for brevity.

Decoys generated using the biased sampling are more protein-like as they have secondary structure, while those generated by the uniform sampling have a more "scrambled" look (compare Figures. 3A and 3B). With secondary structure, the best of the decoys does look very-native-like (Figure 3C). We noticed similar behavior for all proteins studied.

We follow Reva et al.[22] and consider a decoy "good" when its cRMS deviation is smaller than 6 Å from the native structure of the protein it targets. For the proteins 1enh, 4icb, 2cro, and 1ctf, we count the number of good decoys found by our sampling procedures and plot the results comparing this measure for uniform and biased sampling (Figure 4). The maximal distance of the native state of the proteins' structures is 26 Å in 1enh, 30 Å in 4icb, 26 Å in 2cro, and 30 Å in 1ctf. In all cases, the biased sampling found more good decoys than the uniform sampling—this effect is very strong for 1enh and more subtle for 1ctf. Good decoys are found only when a sufficiently tight compactness restriction is enforced. The total number of good decoys found is encouragingly high at 1507, 280, 269, and 30 structures out of 400,000 generated for 1enh, 4icb, 2cro, and 1ctf, respectively using a

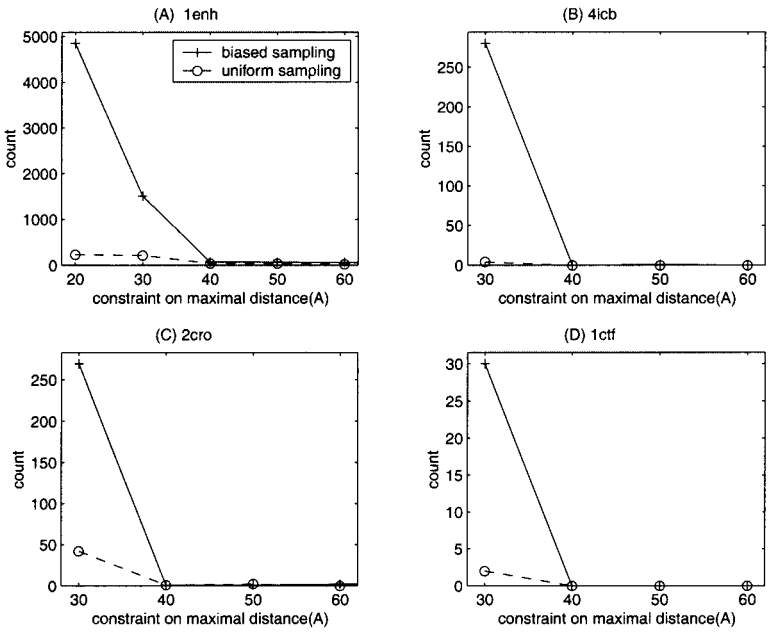

FIGURE 4 The number of good decoys (cRMS deviation <6 Å) as a function of the compactness constraint for 1enh, 4icb, 2cro, and 1ctf. Each of the graphs compares the number found when generating decoys using biased (solid) and uniform sampling (dashed).

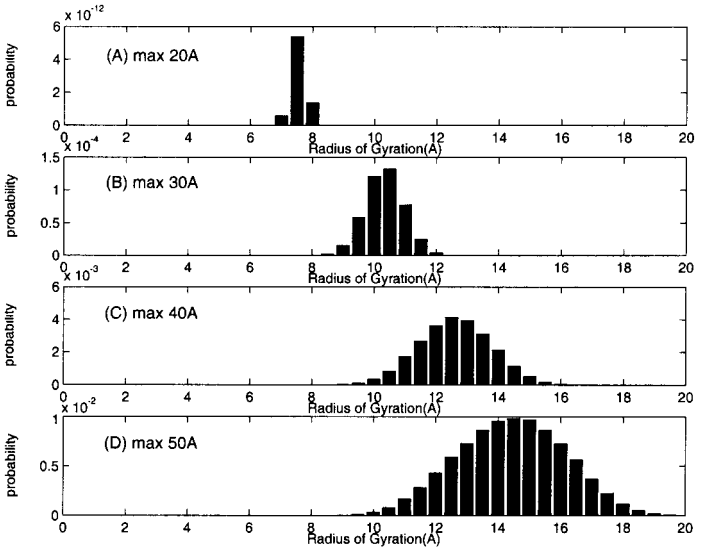

FIGURE 5 The estimation of the distribution of radius of gyration of 1enh decoys with biased sampling. The maximum distance constraint is (A) 20 Å, (B) 30 Å, (C) 40 Å, and (D) 50 Å.

compactness constraint of 30 Å. For 1enh, the results are even better with a compactness constraint of 20 Å with 4849 good decoy structures.

The cRMS deviation of the best decoys found was less than 5.0 Å from the corresponding native states: 3.94 Å for 1enh, 4.23 Å for 4icb, 4.94 Å for 2cro and 4.95 Å for 1ctf. In all cases, the best decoy structure is found with a 30 Å compactness constraint and biased sampling. Figure 3(C) shows the best decoy of 1enh superimposed on the native structure.

We calculate the radius of gyration of the decoy structures generated with biased sampling for 1enh and plot the results in Figure 5. The histogram values are normalized by dividing them by 400,000 so that the integral of the histogram gives the fraction of space occupied by geometrically valid decoys. As expected, decoys generated with a smaller maximal distance constraint have smaller radii of gyration. We can also see that the fraction of space occupied by the valid decoy structures decreases dramatically as the maximal allowed distance is decreased.

DISCUSSION

We have aimed to generate decoys by sampling a discrete and relatively small space of conformations that approximates proteins well. As the space is far larger than our sampling set, we hope to sample the relevant regions, i.e., those compact, self-avoiding folds with the secondary structure of the protein considered. The two prominent characteristics of our decoy generation scheme are as follows: (1) the amino acid sequence of the target proteins is not considered—only its secondary structure is used; (2) the construction is based on the local geometry of other native proteins.

Almost all decoy generation techniques incorporate the amino acid sequence of the protein studied. The information flow from the amino acid sequence into the construction of the decoys varies: some import it via the scoring function,[5] while others search the PDB for structures with similar amino acids and generate decoys with resembling structural pieces. In particular, Simons et al.[15] seek structures with similar consecutive triplets of amino acids while Huang et al.[8] consider nonconsecutive pairs of residues with similar amino acids. Our method explores the boundaries of decoy generation schemes by addressing the question: "How good are decoys when using only secondary structure information?" Surprisingly, although we use significantly less information about the proteins, the decoys we generate are qualitatively very similar to those found by others. For instance, the best decoy for 4icb found by Huang et al.[8] has a cRMS deviation of 5.0 Å from the protein (compare with our value of 4.9 Å). Simons et al.[15,23] found better decoys than ours: the best 2cro decoy at 4.2 Å (compare with 4.9 Å), the best 1ctf at 3.46 Å (compare with 4.9 Å) and the best 5icb decoy at 3.74 Å (compare with 4.2 Å).

Our decoy generation method uses only the local geometric properties of protein backbones as implied by the shape of the fragments in the library. We do not employ any observations regarding common geomet-

ric structures of nonconsecutive residues. The method works best when generating decoys for all-α proteins, and not as well for α and β proteins. Our study suggests that conforming to correct local geometric features along the chain may suffice to imply the structure of an all-α protein. In a way, it is complementary to the work of Fain and Levitt,[16] which showed that conforming to correct nonlocal geometric features suffices to imply the structure of an all-α protein. Indeed, the stabilizing interactions in helices are close along the chain, while those that involve β strands are not, and hence are not captured correctly by our scheme.

These observations suggest several directions for future study: (1) incorporate nonlocal geometric features; (2) incorporate a scoring function into the scheme allowing searching rather than sampling; (3) enrich the library with amino acid sequence information, and use this information to bias the sampling.

APPENDIX

Estimation of the Probabilities of the Library Elements, Conditioned by the Secondary Structure Sequence

We associate with each fragment f a secondary structure sequence $s(f) \in \{\mathrm{H,E,C}\}^2$ of its last two residues (the first three residues are used only for positioning), implied by the secondary structure sequence of its originating protein. These sequences are calculated with STRIDE.[20] Our experiments show similar results regardless of the exact positioning in the fragment of the two residues, for which the secondary structure sequence is considered. It is hard to associate a longer sequence with each fragment (of three residues or more) due to insufficient statistics. We calculate $P(l|\sigma)$ using Bayes law:

$$P(l|\sigma) = \frac{P(\sigma|l)P(l)}{P(\sigma)} \tag{1}$$

where the probabilities $P(\sigma|l)$ and $P(l)$ are estimated by their observed frequencies. Denote by *cluster*(l) the cluster of fragments having l as is its centroid and by $|\cdot|$ the cardinality of a set. $P(\sigma|l)$ is estimated by the frequency of the fragments in *cluster*(l) that have the secondary structure σ,

$$P(\sigma|l) = \frac{|\{f \in cluster(l) | s(f) = \sigma\}|}{|cluster(l)|} \tag{2}$$

As we are estimating probabilities by frequencies, the probability of an event may be too small for our sampling: when we do not observe any instances of a secondary structure sequence σ in the cluster of a library fragment l, we assign a small probability $P(\sigma|l) = \epsilon$ (rather than 0, the observed frequency) to compensate for our finite sampling. We estimate $P(l)$ by the frequency of the fragments represented by l in the total population of clustered fragments:

$$P(l) = \frac{|cluster(l)|}{\sum_{l' \in library} |cluster(l')|} \tag{3}$$

The term $P(\sigma)$ is an easily calculated normalizing factor.

The probability of sampling the sequence of library fragments l^1, l^2,. . .,$l^{(n-3)/2}$, given the secondary structure sequence $s(p)$ is

$$P[l^1, l^2, \ldots, l^{(n-3)/2} | s(p)] = \prod_{i=1}^{(n-3)/2} P[l^i | s^i(p)] \tag{4}$$

Chain Growth Monte Carlo Technique

We sample only decoy structures that satisfy the geometric constraints of compactness and self-avoidance, using the method of Rosenbluth and Rosenbluth.[20] For l = 1,. . ., 20, define $v^i(l)$ to be 1 if l is valid as the ith fragment in the chain and 0 otherwise. For every extension of the chain, $v^i(l)$ is computed for all l values, increasing the computational cost of every chain extension by 20 superpositioning computations. We sample the extending fragment, either uniformly or with a bias, from the subset of valid fragments for the ith position, namely from $\{l \mid v^i(l) = 1\}$. If no library fragment is valid at the ith position, we discard the chain and restart the decoy generation process. This procedure assures that unless a "dead end" is encountered, the already constructed chain is maintained. In cases where some of the library fragments are nonvalid, we renormalize the probabilities to sum to 1 by dividing the probability by $w_i = \Sigma_{i=1}^{20} v^i(l) P[l|s^i(p)]$, the total weight of all valid fragments for that position. Notice that when sampling uniformly, the secondary structure sequence does not influence the probability value $[P(l|s^i(p)] = 1/20$ for all $s^i(p)$.

Enforcing the sampled structures to be valid implies that sampling is done from a different distribution than the one intended. Consequently, when estimating the value of a random variable, one needs to

account for the restricted sampling (only from a subset of all fragments) in different positions along the chain. The random variable of a sampled chain should be weighted by the portions of space we restricted ourselves to during the construction process. The weight of a decoy structure is $w = \Pi_{i=1}^{(n-3)/2} w_i$, the product of all the constraints enforced along the way. The estimate for a random variable r from N samples should therefore be $\frac{1}{N} \sum_{k=0}^{N-1} r_k w_k$, where r_k is the random variable value for the kth constructed chain and w_k is its weight.

We thank Yu Xia for useful discussions. We thank Patrice Koehl, Tanya Raschke, and Daniel Russel for useful discussions as well as helpful comments regarding the manuscript. Finally, we thank Patrice Koehl for providing access to his superefficient FORTRAN routines for superimposing structures.

REFERENCES

1. Levitt, M. Ciba Foundation Symposium 7; Elsevier: Amsterdam, 1972; pp 146–171.
2. Lifson, S.; Warshel, A. J Chem Phys 1968, 49, 5116.
3. Warshel, A.; Lifson, S. J Chem Phys 1970, 53, 582.
4. Levitt, M. Nature Struct Biol 2001, 8, 392–393.
5. Sun, S. J.; Thomas, P. D.; Dill, K. A. Protein Eng 1995, 8(8), 769–778.
6. Pederson, J. T.; Moult, J. J Mol Biol 1997, 269, 240–259.
7. Foreman, K. W.; Phillips, A. T.; Rosen, J. B.; Dill, K. A. J Comp Chem 1999, 20(14), 1527–1532.
8. Huang, E. S.; Samudrala, R.; Ponder, J. W. J Mol Biol 1999, 290, 267–281.
9. Hinds, D. A.; Levitt, M. Proc Natl Acad Sci USA 1992, 89, 2536–2540.
10. Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. Protein Sci 1995, 4, 561–602.
11. Yue, K.; Fiebig, K. M.; Thomas, P. D.; Chan, H. S.; Shakhnovich, E. I. Proc Natl Acad Sci USA 1995, 92, 325–329.
12. Rooman, M. J.; Kocher, J. I.; Wodak S. J. Biochemistry 1992, 31, 10226–10238.
13. Park, B.; Levitt, M. J Mol Biol 1995, 249, 493–507.
14. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Research 2000, 28, 235–242.
15. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. J Mol Biol 1997, 268, 209–255.
16. Fain, B.; Levitt, M. J Mol Biol 2001, 305, 911–201.
17. Kolodny, R.; Koehl, P.; Guibas, L.; Levitt, M. J Mol Biol 2002, 323, 297–307.
18. Kabsch, W. Acta Crystallog Sect A 1978, 34, 827–828.
19. Jones, D. T. J Mol Biol 1999, 292(2), 195–202.
20. Frishman, D.; Argos, P. Proteins Structure Funct Genet 1995, 23, 566–579.
21. Rosenbluth, M. N; Rosenbluth, A. W. J Chem Phys 1955, 23, 356–359.
22. Reva, B.; Finkelstein, A. V.; Skolnick, J. Folding Design 1998, 3, 141–147.
23. Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. Proteins Struct Funct Genet 1999 S3, 171–176.