Probabilistic Sampling of Protein Conformations: New Hope for Brute Force?

Howard J. Feldman[1,2] **and Christopher W.V. Hogue**[1,2]*
[1]*Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada*
[2]*Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada*

ABSTRACT **Protein structure prediction from sequence alone by "brute force" random methods is a computationally expensive problem. Estimates have suggested that it could take all the computers in the world longer than the age of the universe to compute the structure of a single 200-residue protein. Here we investigate the use of a faster version of our FOLDTRAJ probabilistic all-atom protein-structure-sampling algorithm. We have improved the method so that it is now over twenty times faster than originally reported, and capable of rapidly sampling conformational space without lattices. It uses geometrical constraints and a Leonard-Jones type potential for self-avoidance. We have also implemented a novel method to add secondary structure-prediction information to make protein-like amounts of secondary structure in sampled structures. In a set of 100,000 probabilistic conformers of 1VII, 1ENH, and 1PMC generated, the structures with smallest Cα RMSD from native are 3.95, 5.12, and 5.95Å, respectively. Expanding this test to a set of 17 distinct protein folds, we find that all-helical structures are "hit" by brute force more frequently than β or mixed structures. For small helical proteins or very small non-helical ones, this approach should have a "hit" close enough to detect with a good scoring function in a pool of several million conformers. By fitting the distribution of RMSDs from the native state of each of the 17 sets of conformers to the extreme value distribution, we are able to estimate the size of conformational space for each. With a 0.5Å RMSD cutoff, the number of conformers is roughly 2^N where N is the number of residues in the protein. This is smaller than previous estimates, indicating an average of only two possible conformations per residue when sterics are accounted for. Our method reduces the effective number of conformations available at each residue by probabilistic bias, without requiring any particular discretization of residue conformational space, and is the fastest method of its kind. With computer speeds doubling every 18 months and parallel and distributed computing becoming more practical, the brute force approach to protein structure prediction may yet have some hope in the near future. Proteins 2002;46:8–23.** © 2001 Wiley-Liss, Inc.

DOI 10.1002/prot.1163

Key words: conformational space; protein folding; tertiary structure prediction; extreme-value distribution; Levinthal paradox

INTRODUCTION

One of the most interesting aspects of the protein folding problem is understanding the true size and complexity of protein conformational space. Approximations have been made in an attempt to determine the number of possible conformations of a protein sequence of given length. Perhaps the most famous was given by Levinthal[1,2] who approximated the total number of conformations as 10^N,[3] with N being the number of residues. This so-called paradox has been a subject of great interest for some time.[4–6]

On a lattice, it is possible to enumerate the conformations exactly for small N.[7,8] But in real-space, there is the additional issue of how similar two conformers must be to consider them as the "same" structure. The obvious answer may be that any two structures similar to within the RMSD tolerances of X-ray crystallography or NMR should be considered "identical" for all practical purposes.

Another definition is that two structures are the "same" if all their backbone Φ and Ψ torsions lie in approximately the same regions of Ramachandran space, and all sidechain χ angles fall into the same local energy minima, i.e., they have the same rotamers at each residue.[9,10] At first this sounds like a good way of enumerating conformations but there is a major flaw in doing so. The variability in bond lengths and bond angles, which normally are assumed constant, can build up over the length of a protein so that two structures with identical Ramachandran plots may

Abbreviations: CD, circular dichroism; CDF, cumulative distribution function; CPU, central processing unit; DSSP, dictionary of secondary structures of proteins; EVD, extreme value distribution; MMDB, molecular modelling database; MoBiDiCK, modular big distributed computing kernel; NCBI, National Centre for Biotechnology Information; NMR, nuclear magnetic resonance; PDB, protein data bank; PDF, probability distribution function; PHD, profile network prediction HeiDelberg; RMSD, root mean square deviation; SCOP, structural classification of proteins.

Grant sponsor: Natural Sciences and Engineering Research Council of Canada.

*Correspondence to: Christopher W.V. Hogue, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Ave, Toronto, Ontario, Canada M5G 1X5. E-mail: hogue@mshri.on.ca

Received 20 February 2001; Accepted 2 July 2001

have very different backbones. As an example, a 154 residue myoglobin can be generated with identical Φ, Ψ and ω to the crystal structure and standard bond lengths and angles, with a Cα RMSD of 4.9Å from the native structure.

It is well established that performing an exhaustive search of conformational space for all but the smallest of peptides is not possible with our current level of computational power. However, it is still of great interest to know how much conformational space one can sample in a reasonable amount of time. For this work, we employ a Beowulf cluster computer comprising 216 Pentium III-450 CPUs. We consider a reasonable amount of time to be a run that can be computed in a few hours on this system.

One way to quantify how close brute force can get in this "reasonable time" is to measure the RMSD of each predicted structure from the true structure and to report the smallest RMSD found in the sample. Several groups have studied the relationship between RMSD, distance matrix error, and protein length. They have used simplified models of protein structure such as confined Cα random walks[3] compared to crystal structures, unconfined Cα random walk structures compared to each other,[11] and comparing crystal structures to compact fragments of other, non-homologous crystal structures.[8,12] These latter two groups were able to assign probabilities of observing a given RMSD for a given sequence length by observing the distribution of RMSDs in their respective pseudo-random structures. However, both groups used fragments of crystal structures as their set of random conformers. These may not serve as a fair sampling of conformational space because many sterically plausible conformations may not appear in known crystal structures, and any statistical distributions derived from such approaches must be interpreted with caution. Such probabilities are perhaps best suited for evaluating threading predictions.

By making use of secondary structure prediction, one can, in principle, reduce the conformational space available to a protein. However, integrating secondary structure prediction into a three-dimensional protein structure sampling system is a challenge that first must be addressed. Numerous secondary structure prediction algorithms exist in the literature ranging from ones based on neural networks (PHDsec),[13] or information theory (GOR),[14] to other methods employing sequence similarity (SOPMA).[15] No algorithm has yet exceeded 80% prediction accuracy on average, although any one particular prediction may be nearly 100% correct. Chen et al.[16] have shown that imperfect secondary structure information can still greatly improve structure prediction and computation methods. We are used to secondary structure prediction reporting the most probable secondary structure as a 1-state prediction at each residue, either helix, strand or coil. Directly converting this symbolic prediction of helix, strand or coil into three-dimensional space will cause between 20–40% of the residues to be predicted incorrectly in the 3-D structure!

We want to be sure that the secondary structure prediction method adds information across the ensemble, but does not preclude an observation of the correct secondary structure at any given residue in some members of the protein ensemble. We present a novel method to map one-dimensional 3-state secondary structure information into three-dimensional conformational space. We use this method to generate ensembles of structures sampling a wide range of conformational space. We also demonstrate that the ensembles generated do have significant amounts of secondary structure, like the residual secondary structure observed in denatured or unfolded proteins in buffer, measured for example, by CD spectra[17] or fluorescence resonance energy transfer.[18,19] We apply a 3-state prediction to help guide 3-D structure generation, significantly narrowing down the conformational space but importantly, without excluding the native state. We obtain much more realistic conformers in our ensembles that, as expected, deviate from simple Flory-like[20,21] homopolymer behavior.

We have made use of a modified version of the algorithm used in FOLDTRAJ[22] to perform extensive, though non-exhaustive, probabilistic sampling of protein conformational space for a set of 17 protein chains of different fold classes, which are summarized in Table I. FOLDTRAJ produces probabilistic all-atom off-lattice protein conformers that are geometrically and sterically valid, but unconfined. The resulting polymers thus have a much broader distribution of radii of gyration than has been considered before.[3,8,12] The walk is biased by a three-state secondary structure prediction using PHDsec[13] (though any method providing a three-state prediction may be used) mapped onto probability distribution functions (PDF) in Ramachandran space.

The structures we generate with FOLDTRAJ may also be slightly biased due to the fact that they are built from N- to C-terminus using a kinetic random walk.[23] However, for the less compact structures (which are in the majority) collisions play a relatively minor role in the resultant conformations. We point out here that the bias from the directed walk is a very weak bias compared to the secondary structure PDFs that control the sampling scheme. Hence we are carrying out probabilistic sampling, not random sampling.

These probabilistic conformers serve as the ideal reference population for determining just how "good" a given protein structure prediction is, and we find, as discussed below, that the observed RMSD distributions for all probabilistic protein ensembles tested fit well to an Extreme Value Distribution (EVD) with approximately the same mean and variance as the data and not a Gaussian as previously postulated.[12]

MATERIALS AND METHODS

Choice of Folds

The 17 distinct protein folds used in this study were chosen using the SCOP database[24,25] and were intended to represent a wide range of structural classes. They also were chosen to span a range of chain lengths, and most sequences chosen fold into compact, globular structures. The proteins, their lengths and structural classes are

TABLE I. Protein Chains Used in This Study

PDB code[a]	Protein	Chain length	Structural classification[b]	SCOP fold
1VII	Thermostable subdomain from chicken villin headpiece from Chicken	36	Alpha	Thermostable subdomain from chicken villin headpiece
1ENH	Engrailed homeodomain from *Drosophila melanogaster*	54	Alpha	DNA/RNA-binding 3-helical bundle
4ICB	Calbindin D9K from Bovine	76	Alpha	EF Hand-like
1YCC	Mitochondrial cytochrome c from Baker's yeast	107	Alpha	Cytochrome c
1MBD	Myoglobin from Sperm whale	153	Alpha	Globin-like
1PMC	Proteinase inhibitor PMP-C from Migratory locust	36	Beta	Proteinase inhibitor PMP-C
1SHG	α-Spectrin, SH3 domain from Chicken	62	Beta	SH3-like barrel
3PDZ	Phosphatase hPTP1e from Human	96	Beta	PDZ domain-like
1NEU	Myelin membrane adhesion molecule P0 from Rat	124	Beta	Immunoglobulin-like beta-sandwich
1TNR_A	Tumor necrosis factor (TNF) from Human	144	Beta	TNF-like
1BTB	Barstar (barnase inhibitor) from *Bacillus Amyloliquefaciens*	89	Alpha/Beta	Barstar (barnase inhibitor)
1CDZ	BRCT domain from DNA-repair protein XRCC1 from Human	96	Alpha/Beta	BRCT domain from DNA-repair protein XRCC1
5PNT	Tyrosine phosphatase from Human	157	Alpha/Beta	Phosphotyrosine protein phosphatases I-like
5PTI	Pancreatic trypsin inhibitor, BPTI from Bovine	58	Alpha+Beta	BPTI-like
1KPT_A	Virally encoded KP4 toxin from *Ustilago maydis*, P4 strain	105	Alpha+Beta	Yest killer toxins
135L	Lysozyme from Turkey	129	Alpha+Beta	Lysozyme-like
1DIV[c]	Ribosomal protein L9 from *Bacillus Stearothermophilus*	149	Alpha+Beta	Ribosomal protein L9 N-domain/ C-domain

[a]First four characters are standard PDB code, any following characters indicate the chain.
[b]According to SCOP.
[c]1DIV is actually a two-domain protein and so contains two distinct SCOP folds.

summarized in Table I. Very few α/β proteins under 150 residues could be found so these are less represented than the other structural classes. 1DIV is a two-domain protein with the domains joined by a very long helix (about 35 residues). This was added to the test set to see if non-globular structures had similar RMSD distributions to probabilistic structures as globular ones did. All other protein chains were single domains.

Secondary Structure Prediction Methods

We chose to evaluate the performance of two secondary structure prediction methods that provide 3-state predictions—GOR[14,26] and PHDsec[13]—in order to choose the one that performed best for our purposes. For the GOR method, we used an alternate "training" database, the same non-redundant set of 834 proteins used to generate the trajectory distributions, with secondary structure assignments derived from NCBI's MMDB assignment.[27]

Two post-processing filters within GOR usually correct improbable structures, for example changing helices of length one to sheet or coil for the 1-state prediction. Following these filters, we have adjusted each residue's corresponding probabilities for the 3-state predictions. The resulting N×3 matrix of predicted secondary structure probabilities is then convolved with the trajectory distributions from the protein generator, creating trajectory distributions biased in accordance with these structural predictions. We use three trajectory distributions for each of the 20 amino acids, plus *cis*-proline, for a total of 63 base trajectory distributions.

Secondary Structure Prediction Accuracy

The training database we used for the GOR method was cross-validated or "jackknifed." This gave an average prediction accuracy of 60.9% ± 10.2%, with an approximately Gaussian distribution [Fig. 1(a)], and Q3 = 60.6%, slightly below the 62.7% obtained using this method on the standard GOR database of 267 proteins.[26]

To analyze the quality of 3-state prediction, an additional score was recorded during jackknifing. Instead of just counting the number of correct residues in a given protein, we step through it one residue at a time, and look at the actual secondary structure, recording the percentage predicted probability for the correct secondary structure type at each residue. This is averaged over the entire protein to give an amount of correct "information" within the 3-state prediction scheme. The information score expected by chance is 33% since there are three possible secondary structure assignments.

This information score gave an average of 50.3% ± 6.0% correct information at each residue for GOR [Fig. 1(b)]. That is, we have on average a 50% chance at each residue of choosing the correct secondary structure, whereas with 1-state prediction, we have a 100% chance at 61% of the

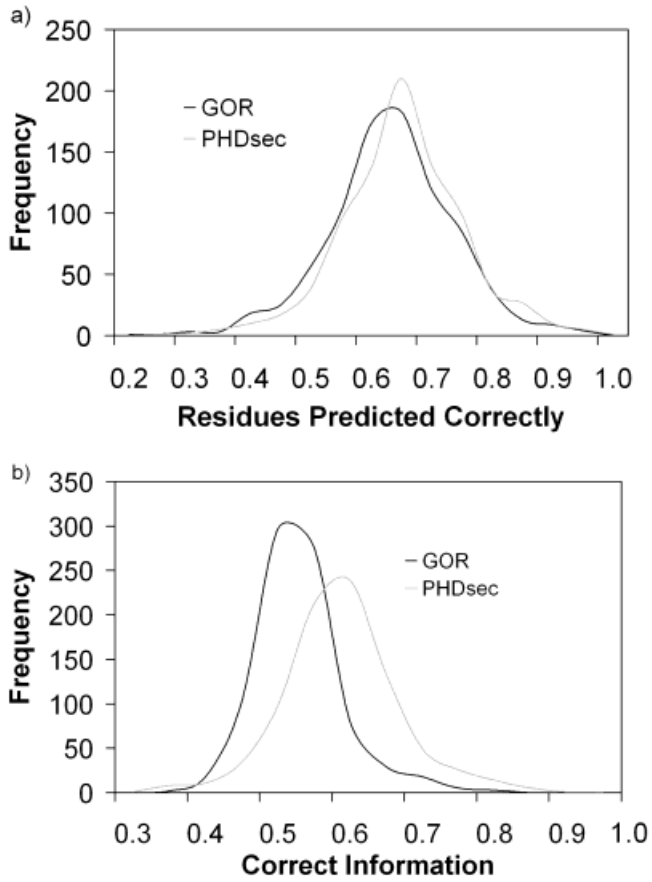

Fig. 1. **a:** Cross-validation results for the GOR 1-state secondary structure prediction, using the non-redundant set of 834 proteins. The mean of the distribution is 60.9%, compared with 33.3% expected by chance. This estimates the probability a given residue is predicted correctly. The 1-state PHDsec predictions were computed for each protein as well (but using its own, smaller training set) having a mean of 62.7%. **b:** The distribution of "correct information" using the 3-state GOR and PHDsec predictions. This is the average value over a protein of the probability assigned to the correct secondary structure at each residue; 33.3% is still expected by chance, and the mean value of the distribution is 50.3% for GOR, and 56.3% for PHDsec. b tells us the chance of choosing the correct region of space at any residue, using 3-state prediction; that is, we always have at least a chance of choosing correctly. Conversely, a tells us the percentage of residues we expect to choose correctly, using 1-state prediction, an all or nothing approach.

residues, and a 0% chance of being right at the remaining 39%!

The greatest benefit of using the 3-state method is that the correct region of space is never forbidden to us, even when a wrong secondary structure prediction is made, as is the case with the 1-state prediction. Also, the minimum value in the distribution of information score from the jackknifing is 32.9%, much higher than the 19.1% correct minimum with 1-state prediction. This assures us that the 3-state method always works at least as well as chance, which cannot be said for the 1-state prediction. The narrower distribution also shows that it is more consistent.

PHDsec v0.1[13] was also run on the set of non-redundant structures and its 1- and 3-state prediction ability compared with GOR. In this case, the neural net provided with the program was used to predict each of the structures, without retraining the net for each prediction. However, the set of test proteins is much larger than the training set so any bias is thought to be negligible. The 1-state prediction accuracy afforded a slight improvement over GOR to an average of 62.7% ± 10.1% on our test set and with similar extreme values to the GOR [Fig. 1(a)]. However the information score was superior [Fig. 1(b)] by 6%, 56.3 ± 8.1% and with a maximum information score of 88.7% compared to only 77.5% for the GOR. This suggested that the PHDsec program provided superior 3-state predictions, and so it was chosen for the remaining experiments.

Markovian Backbone Growth

While secondary structure prediction is able to influence individual residues to take on certain conformations, trajectory distributions are by nature mutually independent of each other so that the choice of space to occupy at residue i does not directly influence the choice of trajectory space for residue i+1. That is, no Markovian information is present in the trajectory distributions. So, for example, although a section several residues long may be predicted as 90% helical, there is still a 10% probability at each residue of some other conformation being chosen, regardless of how far into the helix we are. There is no guarantee to avoid getting a very kinked helix, for example, in this manner.

To determine its effect, Markovian information was added to the trajectory distributions in the following way. For a given residue i in a protein, consider its a-carbon trajectory space (i.e., spherical) co-ordinates (ϕ_i,θ_i). Here ϕ_i is the supplement of the angle between $C\alpha_{i-1}$, $C\alpha_i$, and $C\alpha_{i+1}$, while θ_i is the dihedral angle between $C\alpha_{i-2}$, $C\alpha_{i-1}$, $C\alpha_i$; and $C\alpha_{i+1}$. Comparing these to the co-ordinates of the previous residue in the backbone, we can obtain $\Delta\phi_i = \phi_i - \phi_{i-1}$ and $\Delta\theta_i = \min(|\theta_i - \theta_{i-1}|, 360 - |\theta_i - \theta_{i-1}|)$. Trajectory co-ordinates themselves are relative to the local backbone, so these incremental co-ordinates correspond to deviation of the backbone from its current conformation, with a large value corresponding to transitions in secondary structure, for example, from helix to strand. To calculate the actual relative change in distance on the surface of a sphere between residue i and $i-1$ in trajectory space, it can be shown that:[28]

$$\cos d_i = \cos \Delta\phi_i + \sin \phi_i \sin \phi_{i-1}(\cos \Delta\theta_i - 1) \qquad (1)$$

where d_i is the distance in radians on the unit sphere.

Using the non-redundant set from the PDB, the distribution of d_i was recorded and separated into two distinct regions: $-50° < \theta < 110°$ (helical) and $\theta < -50°$ or $\theta > 110°$ (extended). The two resultant curves were then normalized so that the largest peak had a height of 1 (Fig. 2). Then, when a new α-carbon was chosen during the build-up procedure, d_i was computed and the value of the probability on the corresponding helical or extended distribution was used as an acceptance probability (see Fig. 2), with failure resulting in backtracking and a new choice of Cα.

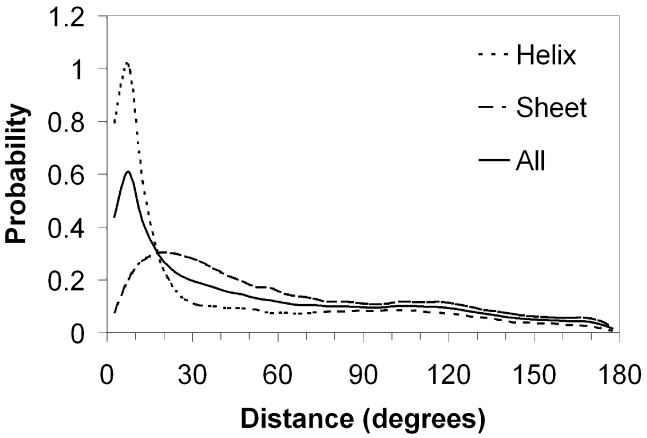

Fig. 2. Distribution of Markovian information based on secondary structure type in the non-redundant set from the PDB. The curves are probability distributions (normalized to make the helix peak at 1) for the magnitude of the change in backbone orientation (as measured by Cα-Cα trajectory vectors) between consecutive residues. To interpret, looking at 10° on the x-axis, suppose a given residue is in the helical conformation, there would be about an 85% chance of accepting a placement with its trajectory space co-ordinates (ϕ_i,θ_i) making an angle of 10° with those of the previous residue $(\phi_{i-1},\theta_{i-1})$ when both vectors are viewed on a common unit sphere, with the chance dropping to about 25% if residue i were in the sheet conformation, and 55% overall if the conformation of residue i is not considered.

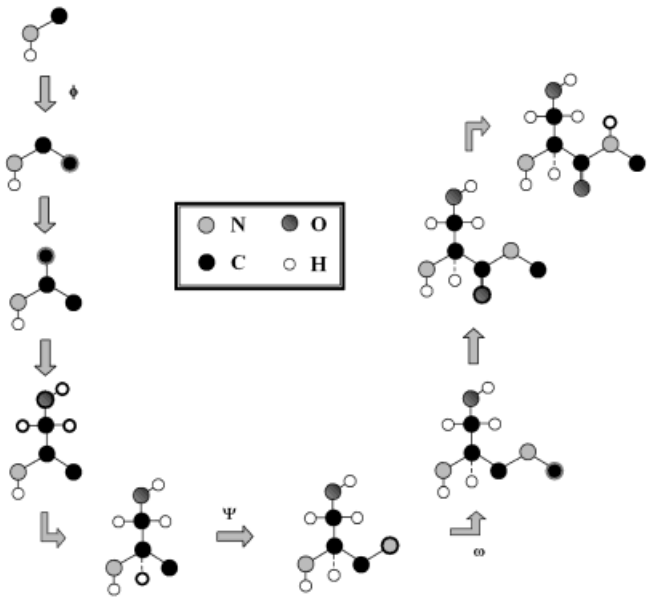

Fig. 3. A snapshot of the protein building pipeline of FOLDTRAJ using Ramachandran trajectory distributions. The protein chain is built up one atom at a time, as detailed in Materials and Methods. This process continues until a van der Waals collision occurs (in which case we backtrack) or the end of the protein is reached. The atom(s) added at each step are outlined.

Probabilistic Structure Sampling

Structures were generated by a kinetic self-avoiding random walk algorithm biased by secondary structure prediction, similar to that previously reported.[22] The major difference between the original algorithm and the present technique is that the walk is done with the trajectory distributions now in Ramachandran space[29] rather than Cα—Cα space. This means that the backbone atoms can now be placed in their bonded order using Φ and Ψ chosen from their observed distributions in the MMDB[30,31] using 3-state secondary structure prediction to determine the exact distribution used at each residue. ω is chosen randomly with mean 179.8° and standard deviation 4.5°.[32,33] Also, the carbonyl oxygen is now placed using the planarity of the peptide bond, and a constant Cα—C—O angle of 120.8° and C═O bond length of 1.231Å.[34] Incidentally, since no energy minimization is required to place the peptide N and C atoms using this newer algorithm, the method is over twenty times faster than the original Cα walk, taking only 3.6 seconds on average to make a 154-residue myoglobin on a Pentium II-350 CPU running Linux, as opposed to 94 seconds using the original approach.

Lastly, the Cβ direction can be computed exactly using residue-dependent angles N—Cα—Cβ and C—Cα—Cβ along with a Cα chirality constraint. The calculation is carried out as follows. Let the N—Cα—Cβ angle be α and the C—Cα—Cβ angle β, both given. Denote the unit vector in the direction from Cα to N by $\mathbf{N} = (\mathbf{N_x},\mathbf{N_y},\mathbf{N_z})$, from Cα to C by $\mathbf{C} = (\mathbf{C_x},\mathbf{C_y},\mathbf{C_z})$, and from Cα to Cβ as $\mathbf{B} = (\mathbf{B_x},\mathbf{B_y},\mathbf{B_z})$. Thus we have the following system of 3 equations:

$$\mathbf{N} \cdot \mathbf{B} = \cos\alpha \tag{2}$$

$$\mathbf{B} \cdot \mathbf{C} = \cos\beta \tag{3}$$

$$\|\mathbf{B}\| = 1 \tag{4}$$

which we must solve to obtain the three unknowns, the components of **B**. With a little algebra we can (aside from a few special cases) arrive at a quadratic equation for $\mathbf{B_y}$, of the form $a\,\mathbf{B}_{y2} + b\,\mathbf{B_y} + c = 0$, with:

$$a = \mathbf{N_z^2} * (1 + (\tau_{xy}/\tau_{zx})^2 + (\tau_{zy}/\tau_{zx})^2) \tag{5}$$

$$b = (2 * \mathbf{N_z^2}/\tau_{zx}^2) * (\phi_x * \tau_{xy} + \phi_z * \tau_{zy}) \tag{6}$$

$$c = (\mathbf{N_z^2}/\tau_{zx}^2) * (\phi_z^2 + \phi_x^2 - \tau_{zx}^2) \tag{7}$$

where

$$\tau_{ij} = \mathbf{C_i N_j} - \mathbf{C_j N_i} \text{ and } \phi_i = \mathbf{N_i}\cos\beta - \mathbf{C_i}\cos\alpha$$

This yields two possible solutions for $\mathbf{B_y}$. We also get:

$$\mathbf{B_x} = -(\phi_z + \tau_{zy}\mathbf{B}_y)/\tau_{zx} \tag{8}$$

and

$$\mathbf{B_z} = (\cos\alpha - \mathbf{N_x B_x} - \mathbf{N_y B_y})/\mathbf{N_z} \tag{9}$$

for each of the two possible values of $\mathbf{B_y}$ giving two potential values for **B**. However only one of these will result in an L- amino acid (while the other gives a D-amino acid). We choose the unique **B**, which satisfies $(\mathbf{N} \times \mathbf{C}) \cdot \mathbf{B} > 0$. If $\tau_{zx} = 0$ or $\mathbf{N_z} = 0$ we must calculate **B** in a slightly different way to avoid dividing by zero but the derivation is nearly identical.

Note that while the sampling of conformational space performed on the structures in Table I used Ramachand-

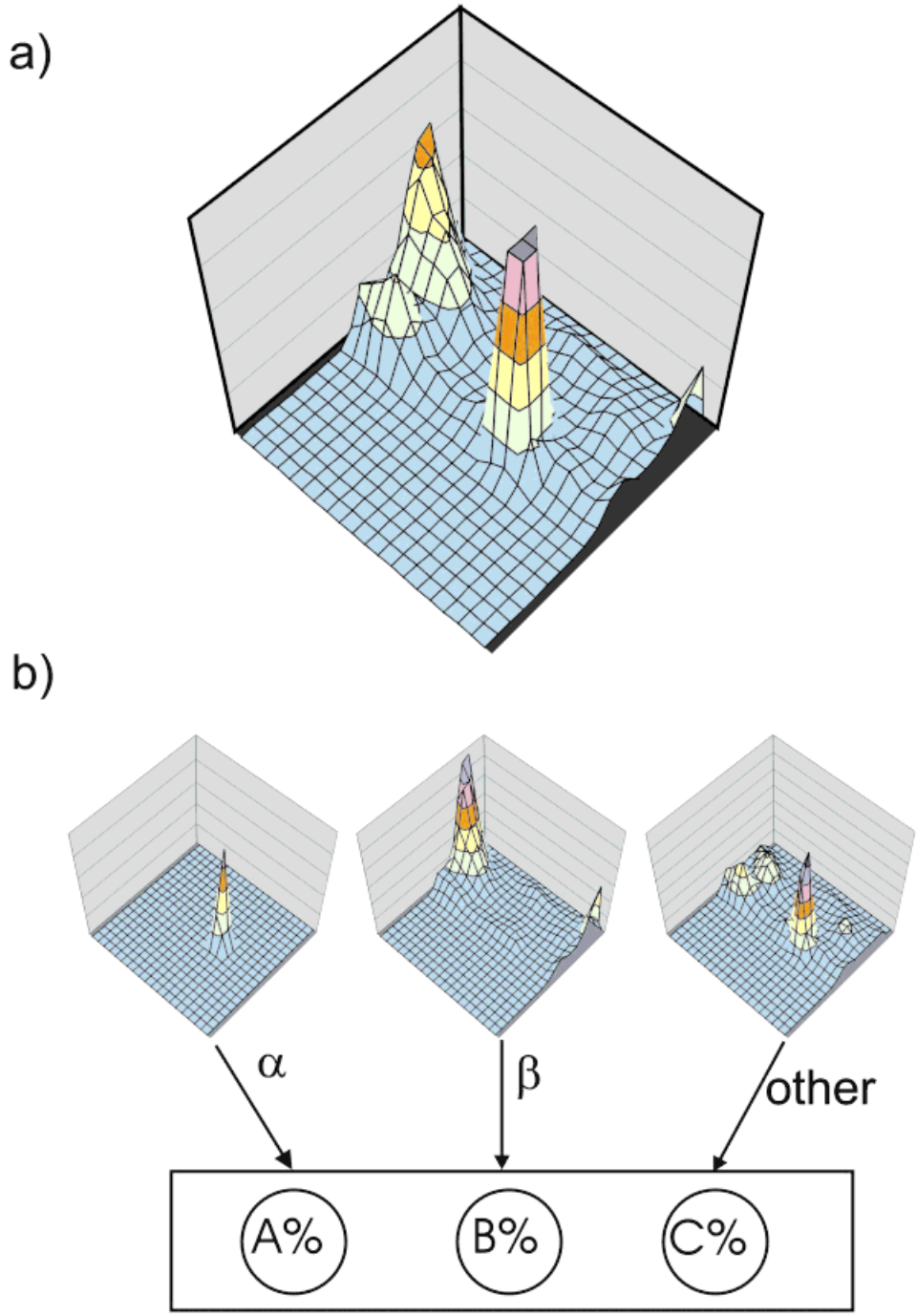

Fig. 4. In this work, initial trajectory distributions were chosen in one of two ways. **a:** Amino-acid based trajectory distributions use the frequency distribution of a non-redundant set of structures from the PDB in trajectory space, sorted by residue type, as the initial trajectory distributions, for a total of twenty possibilities at each residue, plus *cis*-proline. **b:** The 3-state secondary structure method operates in a similar way to a, with the addition that frequency distributions from the PDB are separated both by amino acid type and secondary structure assignment, for a total of 63 distributions. A 3-state secondary structure prediction is made, giving probabilities for helix, sheet, and coil at each residue (rather than just choosing one). Then at a given residue, the probabilities are multiplied by the three trajectory distributions that correspond to its amino acid type, and these are added together to give an initial trajectory distribution for the residue. The number of possible starting trajectory distributions for a residue using this method is unlimited. Trajectory distributions shown here are in alpha carbon space. For Ramachandran space distributions, the method is identical, using 63 distributions in Φ-Ψ space as the starting point.

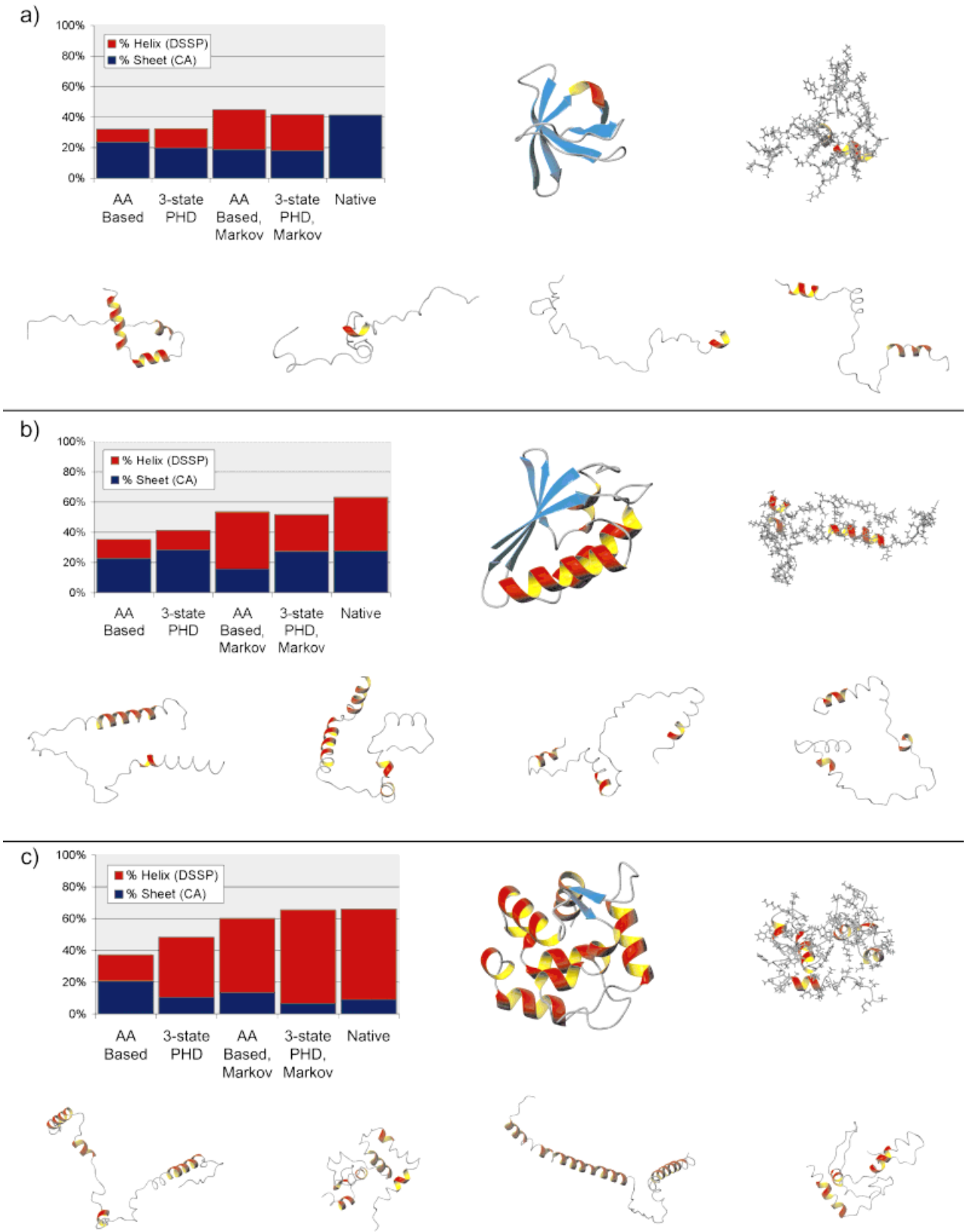
a)
100%
80%
60%
40%
20%
0%
% Helix (DSSP)
% Sheet (CA)
AA Based
3-state PHD
AA Based, Markov
3-state PHD, Markov
Native
b)
100%
80%
60%
40%
20%
0%
% Helix (DSSP)
% Sheet (CA)
AA Based
3-state PHD
AA Based, Markov
3-state PHD, Markov
Native
c)
100%
80%
60%
40%
20%
0%
% Helix (DSSP)
% Sheet (CA)
AA Based
3-state PHD
AA Based, Markov
3-state PHD, Markov
Native

Figure 5.

ran trajectory distributions, the tests involving secondary structure content with and without Markovian information used the older Cα—Cα trajectory distributions so that the Markovian bias could be implemented as described above.

Making Probabilistic Structures

For each protein in the test set, separate runs to create ensembles of 1,000, 10,000, and 100,000 structures of each were performed using our MoBiDiCK distributed computing system[35] to distribute the jobs across our 216 CPU Pentium III-450 cluster computer.[36] Several important quantities including the radius of gyration, RMSD from native structure, exposed surface area, helical content, and so on, were recorded for each structure in the ensemble for later analysis. The RMSD distributions were observed to be skewed, and so were fit to the EVD, which has the cumulative distribution function (CDF):

$$P(X < x) = e^{-e^{(x_c - x)/w}} \tag{10}$$

Other common skewed distributions were found to give slightly poorer fits. Using the fact that for the EVD,

$$\mu = x_c + w\gamma \tag{11}$$

and

$$\sigma^2 = \pi^2 w^2/6 \tag{12}$$

where x_c and w parameterize the distribution (Eq. 10) and γ is the Euler-Mascheroni constant (http://www.mathsoft.com/asolve/constant/euler/euler.html), the mean and standard deviation of the fitted distributions could be easily computed and compared to those of the observed distribution as a measure of goodness of fit.

RESULTS

Creating Ensembles of Structures

FOLDTRAJ uses our previously reported algorithm[22] to make off-lattice all-atom structures of a given protein sequence. Briefly, a map of conformational space available at each residue, called a "trajectory distribution," is generated based on the sequence, and used as a probability distribution to guide a kinetic self-avoiding random walk, filling in the backbone and sidechains as it proceeds and backtracking when collisions occur. Trajectory distributions, the input to FOLDTRAJ, are PDFs determined from database structures that represent alpha carbon (as in our earlier work) or, now optionally, Ramachandran angle trajectory probabilities based on residue type, used in a pipelined N—C directed build up method. The build-up method in Ramachandran space is shown in Figure 3.

Trajectory distributions [Fig. 4(a)] can be further divided into α-helix, β-strand, and "other" components, as shown in Figure 4(b). This allows for the addition of the very useful information available from secondary structure predictions. Many prediction schemes provide both a 1-state prediction (the most probable state at each residue) along with a 3- or more state prediction (the probability of each state at each residue). The 3-state prediction contains much more information than the 1-state model. These probabilities indicate a residue's tendency towards each of the areas of conformational space, based on its context. By incorporating this information into our trajectory distributions, more correct secondary structure on average can be produced and we overcome the problem of 1-state secondary structure prediction that could force the protein build-up into incorrectly predicted regions of space. This is not to say it will not wander into incorrect regions by chance though.

Four different possible starting trajectory distributions could be easily generated. These are (1) Uniform distributions, where the PDF is simply a flat plane at each residue, so any conformation is equally likely; (2) Amino acid-based trajectory distributions, where the initial trajectory distribution at each residue depends only on the residue type. This is the method originally employed in previous work[22]; (3) 1-state secondary structure prediction method, in which the initial trajectory distribution used at each residue depends on amino acid type and the absolute predicted secondary structure (helix, sheet, or coil); (4) 3-state secondary structure prediction method, where initial trajectory distributions depend on amino acid and probabilities of each of helix, sheet, or coil as predicted by some secondary structure prediction method.

As an example of the latter [Fig. 4(b)], if an Ala residue was predicted as 50% helix, 30% sheet, 20% coil, then we could take 0.5*(trajectory distribution for Ala helix)+0.3*(trajectory distribution for Ala sheet)+0.2*(trajectory distribution for Ala coil) as our initial trajectory distribution at that residue. Each of the three types of Ala trajectory distribution would have first been normalized to the same total volume under the curve. A 1-state secondary structure prediction is thus a degenerate case of this where the probabilities are always some permutation of 100, 0, 0%. The 3-state method was used in structure generation unless otherwise stated.

Secondary Structure Content of Ensembles

Protein backbones generated by pure self-avoiding random walk tend not to have a lot of secondary structure elements since these typically require fairly repetitive segments which do not occur by chance frequently enough to correspond to real folded proteins. We are able to increase the chance of getting repetitive segments by introducing a Markovian bias (Fig. 2) to the build-up

Fig. 5. Secondary structure content and sample structures of ensembles of probabilistically generated conformers of (**a**) 1SEM (all β), (**b**) 2HPR (α/β), and (**c**) 1RTP (all α). The first structure in each case is the native, and the remaining five are generated conformers. The conformer next to the native structure is shown with sidechains to illustrate that conformers are indeed all-atom. Each bar in the graphs represents an ensemble average over 10,000 conformers. Helicity was detected with DSSP,[37] consisting of at least four residues with at least one hydrogen bond; however, DSSP detected under 0.1% average hydrogen bonded sheet in all cases (others have experienced similar problems[23,46,47]) so alternate criteria were used to detect extended structure, simply by looking for windows of five consecutive α-carbons spanning 13.25 Å or more, which gave good agreement with secondary structure assignment in the PDB files. Secondary structure compositions of native structures calculated using the same criteria are shown for comparison.

TABLE II. Range of Radius of Gyration and RMSD Observed in Ensembles of 100,000 Probabilistic Conformers

PDB code	Chain length	R_{gyr} min (Å)[a]	R_{gyr} max (Å)[a]	R_{gyr} Native (Å)	R_{gyr} max (Å)[b]	Mean RMSD (Å)[a]	Best RMSD (Å)[a]	Z^c_{best}	Z^c_{worst}
1VII	36	8.3	19.4	8.8	39.6	9.0	4.0	−2.83	4.07
1ENH	54	10.1	28.8	10.1	59.4	14.4	5.1	−3.22	4.40
4ICB	76	11.5	38.1	11.3	83.6	17.0	6.6	−2.93	5.01
1YCC	107	15.3	56.5	12.7	117.7	25.4	11.5	−2.26	5.56
1MBD	153	16.2	57.8	15.2	168.3	26.7	11.2	−2.81	5.56
1PMC	36	8.8	26.6	10.0	39.6	13.7	6.0	−3.28	4.23
1SHG	62	11.5	43.1	10.4	68.2	22.1	9.0	−2.67	4.12
3PDZ	96	15.4	63.4	12.1	105.6	30.2	12.2	−2.57	4.94
1NEU	124	19.3	78.9	14.4	136.4	36.2	14.2	−2.63	5.06
1TNR_A	144	17.6	77.0	16.5	158.4	33.5	15.5	−2.38	5.66
1BTB	89	13.0	40.4	11.4	97.9	19.7	9.1	−2.80	5.38
1CDZ	96	14.0	50.9	12.6	105.6	23.0	10.3	−2.63	5.56
5PNT	157	17.1	70.6	14.5	172.7	31.0	14.5	−2.41	5.67
5PTI	58	11.2	37.7	10.7	63.8	18.2	7.3	−2.98	4.58
1KPT_A	105	14.9	59.8	12.3	115.5	29.2	12.2	−2.51	5.26
135L	129	15.4	65.8	13.7	141.9	28.7	11.1	−2.39	5.85
1DIV	149	17.3	64.8	24.9	163.9	24.3	12.9	−2.87	7.24

[a]From an ensemble of 100,000 conformers.
[b]Theoretical maximum radius of gyration for a completely extended chain.
[c]Z-score for RMSD values calculated as $Z = (x - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the RMSD distribution respectively and x is the RMSD value of interest.

procedure (see Materials and Methods). This bias along with secondary structure prediction help to increase the amount of extended and especially helical structure. To quantitate their effects, 10,000 structures of 1SEM (an all β SH3 domain), 2HPR (an α/β open sandwich phosphocarrier protein) and 1RTP (all α, α-parvalbumin) were generated using trajectory distributions that were amino acid-based or biased by 3-state PHD[13] secondary structure prediction. Each of these was tested with and without the Markov bias. The average helical and extended fractions are plotted in Figure 5. These quantities are ensemble averages, so that for each set of conditions, there may exist a group of structures in the ensemble very close in secondary structure composition to the native composition, regardless of the average value. Some typical structures from each ensemble are shown with the plots.

In all cases, using the Markovian information results in a significant increase in helical content as well as a less pronounced decrease in extended structure. This is desirable for helical proteins, of course, but not for sheet proteins like 1SEM. The 3-state PHD biased Markov structures for 1SEM had 23.6% helix while the amino-acid biased Markov structures had 26.1% helix, and PHD predicted 19% helix for this all β protein. Here the 3-state secondary structure prediction effectively removed some of the helix in the pool of conformers. 2HPR structures without Markov information bias tend to have less than native amounts of helix while the Markov structures have roughly the correct amounts of helix and slightly less strand. Only marginal differences exist between the AA-based and 3-state PHD structures for 2HPR with the latter having close to native strand content and less helix than the former. Lastly, 1RTP structures generated with Markov information have, on average, close to native amounts of both helix and strand, much closer than those without Markov information. The differences between AA-based and 3-state PHD are more subtle and typically concern the placement and length of individual secondary structural elements.

The major effect of using the Markov information to bias the walk is an increase in helical content in those proteins with high propensity to form helices and a small decrease in extended structure. The Markovian bias improves the amount of correct secondary structure in the proteins but also adds some background "helical noise" to all-β proteins due to the strong tendency to stay in the helical region of space once it is entered. This is understandable since once a helix starts forming, there is a very high probability of keeping placements which are in very similar backbone conformations, to continue the helix, and a relatively low probability that the helix will be broken. Thus we elected not to use the Markov method in our evaluation of the conformational space available to proteins, described next.

Sampling

FOLDTRAJ was used to sample the conformational space of the 17 proteins in Table I by generating large ensembles of probabilistic structures as described in Materials and Methods. The probabilistically generated structures possessed a wide range of structural properties indicating that they sample a wide spectrum of the conformational space of each protein, as summarized in Table II Figure 6 shows the distributions of radii of gyration. The maximum theoretical radius of gyration, corresponding to a completely extended chain, was calculated as $L\sqrt{(N^2 - 1)/12}$ where $L = 3.81$Å, the separation between polymer units, and N is the number of residues. In most cases, the largest observed radius of gyration in the pool of probabilistic structures was about half that of the theoretical maximum, so that although many unfolded conforma-

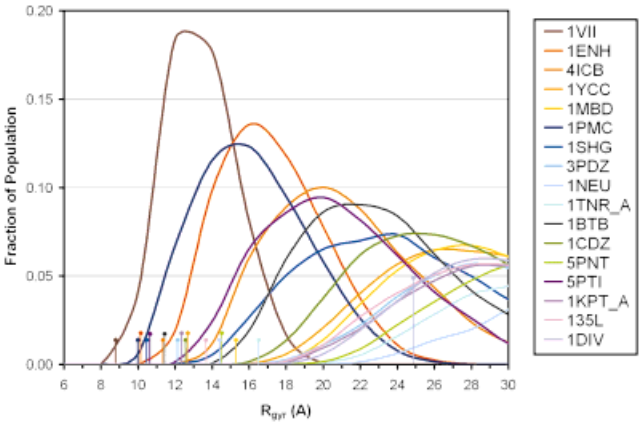

Fig. 6. Distributions of R_{gyr} (in Å) for 17 different proteins of various folds and structural classes. Altogether, 100,000 samples of probabilistic conformational space were used to generate each distribution, which has been normalized to have an integral of one. Vertical lines indicate the R_{gyr} of the native structures, so that any part of the distributions with area to the left of the vertical line of the same color correspond to generated conformations that are more compact than the native state.

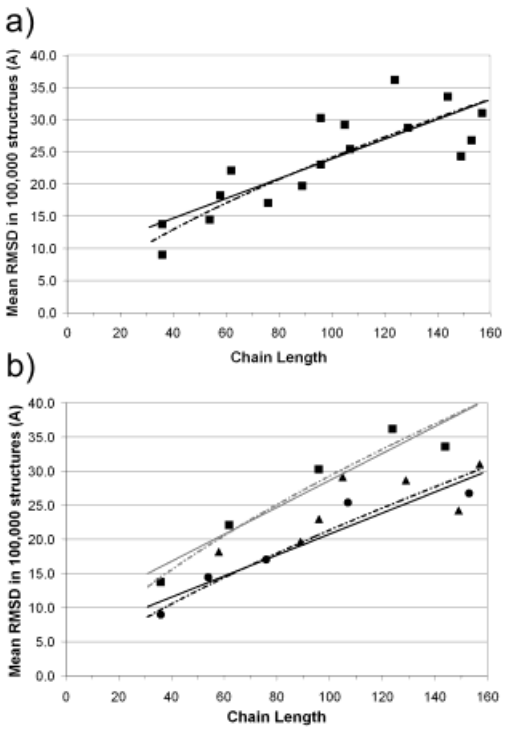

Fig. 7. **a:** Mean RMSD, in Å, between 100,000 probabilistically generated structures of each protein and their native conformations. Each point corresponds to one protein. Solid line: line of best fit. Dashed line: best fit of the form $y = Ax^N$. **b:** The same data as in a but with secondary structure distinguished. Circles are helical proteins, squares are sheet proteins, and triangles are mixed α/β or $\alpha+\beta$. The gray lines represent the best fit for sheet proteins (solid and dashed as in a) and the black lines are the best fit for helical proteins.

tions were sampled, none were completely linear or very nearly so. The smallest radius of gyration was usually on the order of the native value, sometimes slightly higher and sometimes lower, depending how tightly packed the native structure was. The distributions in Figure 6 show

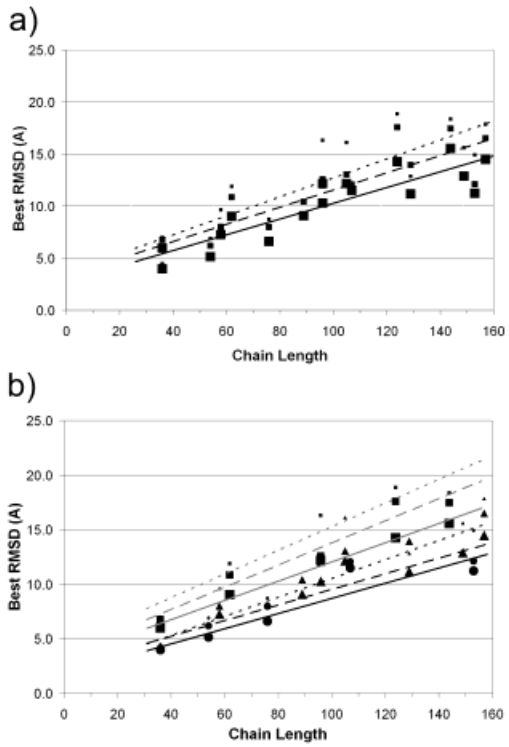

Fig. 8. **a:** Smallest RMSD conformer generated for each protein, in an ensemble of 1,000 (small squares), 10,000 (medium squares), and 100,000 (large squares) structures. Each square corresponds to one ensemble of protein structures. Short dashed line: best fit for ensembles of 1,000 structures. Long dashed line: best fit for ensembles of 10,000 structures. Solid line: best fit for ensembles of 100,000 structures. **b:** The same data as in a but with secondary structure distinguished. Circles are helical proteins, squares are sheet proteins and triangles are mixed α/β or $\alpha+\beta$ (small, medium, and large correspond to ensemble size as in a) The gray lines represent the best fit for sheet proteins (solid and dashed as in a) and the black lines are the best fit for helical proteins.

that we have obtained a good sampling of both compact and unfolded structures for each sequence. On average, the structures were less compact than folded proteins, but the speed of generating the structures offsets any benefits we may gain from trying to confine the conformers to tight spheres, for example, which would slow down the method. The surface area was on average about twice that of the compact native structures, with the most compact conformers generally having slightly larger surface areas than the corresponding crystal structures.

The secondary structure measured by DSSP[37] was also recorded for each structure, with the average helical content being less than that in the native structures (for those with helices). Generally, the most helical of the probabilistic samples had similar amounts of helix to their native counterparts. Perfectly hydrogen-bonded sheet rarely is formed during probabilistic build-up methods, so a different method was used to detect extended stretches of backbone (see Fig. 5 legend). Again, the mean was generally less than the amount of extended structure found in the corresponding crystal structure but the samples with the most extended segments had as much or more than the native. Helical proteins had very little extended structure. For example, probabilistic conformers of myoglobin (1MBD)

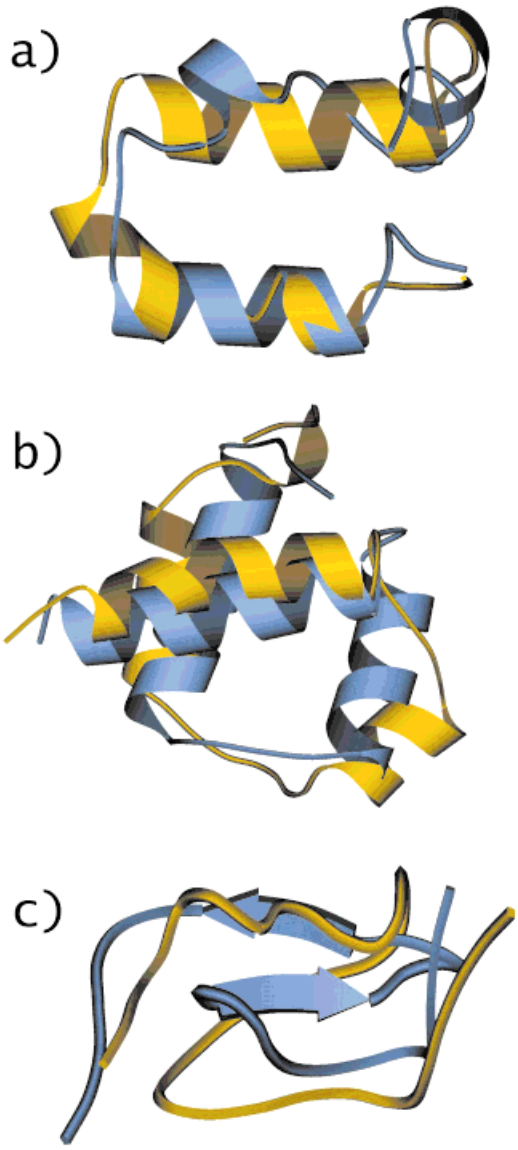

Fig. 9. The structures obtained in a pool of 100,000 probabilistic conformers with smallest Cα RMSD from the native state for (**a**) 1VII (RMSD = 3.95Å), (**b**) 1ENH (RMSD = 5.12Å), and (**c**) 1PMC (RMSD = 5.95Å). In each case, the native structure is blue and probabilistic conformer is yellow. Structures are shown as cartoons for clarity but all atoms are present in each model. Computation times were 70, 123, and 45 CPU-hours, respectively, on a set of Pentium III-450 machines.

had on average only 5 extended residues but 65 helical residues per structure (actual: 5 extended, 113 helix). In contrast, TNF (1TNR) had an average of 8 helical and 41 extended residues (actual: 86 extended, 0 helix).

RMSD vs. Chain Length

The mean RMSD between the ensembles of 100,000 probabilistic structures and their corresponding native states was expected to increase with chain length N,[3,8,12] and indeed was found to in an approximately linear fashion ($R^2 = 0.67$) as seen in Figure 7(a). However, a better fit is given by RMSD $\propto N^{0.68}$ ($R^2 = 0.78$). RMSD can be expressed as[8,11]:

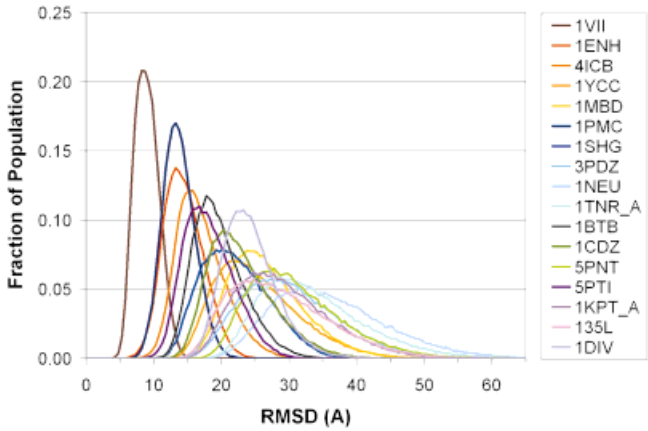

Fig. 10. Distributions of RMSD (in Å) from native state for 17 different proteins of various folds and structural classes. The same coloring scheme is used as in Figure 6. Altogether, 100,000 samples of probabilistic conformational space were used to generate each distribution, which has been normalized to have an integral of one.

$$\mathrm{RMSD}^2 = R_A^2 + R_B^2 - 2v \tag{13}$$

where R_A and R_B are the radii of gyration of the two structures and $2v$ is a term arising from the optimal alignment of the structures. For a given protein, the size of the crystal structure R_B is constant and when R_A is much larger, it will tend to dominate the equation (v is zero on average for random generated structures). The radius of gyration of sterically self-avoiding structures has been shown empirically to vary as $N^{0.5722}$, close to the theoretical values found in the literature.[38–40] The best fit exponent of 0.68 is somewhat swelled due to the added effect of the constant radius of the native structure, which varies as $N^{1/3}$ for globular structures.[11]

Both fits can be improved if we separate proteins into their structural classes, as shown in Figure 7(b). Helical proteins have a lower average RMSD from probabilistic structures than do all-beta proteins, with mixed alpha-beta ones falling somewhere in between. That is to say, helical proteins are easier to "guess" by brute force than sheet proteins. The linear correlation coefficients for both alpha and beta fits is $R^2 = 0.90$, and increases when fit logarithmically to RMSD $\propto N^{0.77}$ for alpha ($R^2 = 0.95$) and RMSD $\propto N^{0.69}$ for beta ($R^2 = 0.96$). It is well known that secondary structure prediction is typically more accurate for all-helical proteins,[26] and predicted helices have much less conformational freedom than beta structure.

We have also examined the correlation between the smallest RMSD structure found for each protein and protein length. Splitting the proteins up by structural class again greatly improves the linear fit with alpha proteins being "predicted" best and sheet proteins worst, with mixed structures lying in between. The linear fit R^2 for N = 100,000 structures improves from 0.79 [Fig. 8(a)] to 0.99 for sheet and 0.85 for helix [Fig. 8(b)]; helical R^2 is 0.99 when the outlier 1YCC is removed. The improvements in the best structure as we move into the tail of the EVD can be seen in Figure 8, going from N = 1,000 to N = 10,000 to N = 100,000. The best fit lines all have approxi-

mately the same slope, but translate towards the origin as N is increased. This is expected because more conformational space is being sampled.

The best structures obtained from the N = 100,000 set for 1VII, 1ENH, and 1PMC, the smallest proteins in the test set, are shown superimposed on their native folds in Figure 9 and have Cα RMSDs of 3.95, 5.12, and 5.95Å, respectively. All have the correct topology and approximately correct secondary structure, differing only in precise placement. Specifically, the 1VII structure forms a single helix where the native structure contains a rather coily turn between two shorter helices at the N-terminus. The 1ENH structure is very close but is missing a turn at the end of the second helix and a turn of the first helix. The 1PMC conformer has the correct topology but doesn't form hydrogen-bonded β-sheet.

It seems remarkable that such good all-atom structures could be obtained by a probabilistic sampling process, using only geometrical constraints, secondary structure prediction and a van der Waals energy potential, in about 50 CPU-hours (Pentium III-450 CPUs) each. These structures could be further refined by molecular dynamics simulation to resolve minor steric clashes and adjust sidechain orientations to nearby local minima.

Fitting the RMSD Distribution

The conformers with the smallest RMSD from the native state, along with the mean RMSD for each structure and the Z-score for the best and worst structures, are summarized in Table II. The Z-score for the best structures is much less in magnitude than the Z-score of the worst structures, indicating that RMSD should fit a skewed distribution.

The RMSD distributions of the structures from their respective native conformations, shown in Figure 10, were found to fit well to the EVD. This is not surprising, since RMSD is a minimum possible RMS difference between two sets of vectors aligned optimally in space. A sample fit is shown in Figure 11 for 100,000 probabilistic conformers of 1NEU. In all cases, the best fits had an R^2 of greater than 0.99 with all except the shortest proteins having $R^2 > 0.995$. The residuals for the fit in each case is shown in Figure 12 and the parameters for the best fits are given in Table III. When the fit was slightly off, it was usually towards the tail for large values of RMSD (see Figs. 11 and 12). We are interested, however, in the tail fit at low RMSDs to determine both the size of conformational space and the probability of observing a given RMSD by chance.

To confirm that the fit to the EVD was accurate, we assumed that our EVD fit was perfect, and compared the expected mean and standard deviation (Eqs. 11 and 12, below) to that observed for a pool of 100,000 probabilistic structures. The difference is given in Table III, showing that the mean is almost exactly as expected except for the very short proteins where the mean is slightly smaller than predicted. The standard deviation for the generated structures is consistently 10–20% smaller than expected due to the fact that the EVD fit is only an approximation. Taken together with the Pearson correlation coefficient R^2,

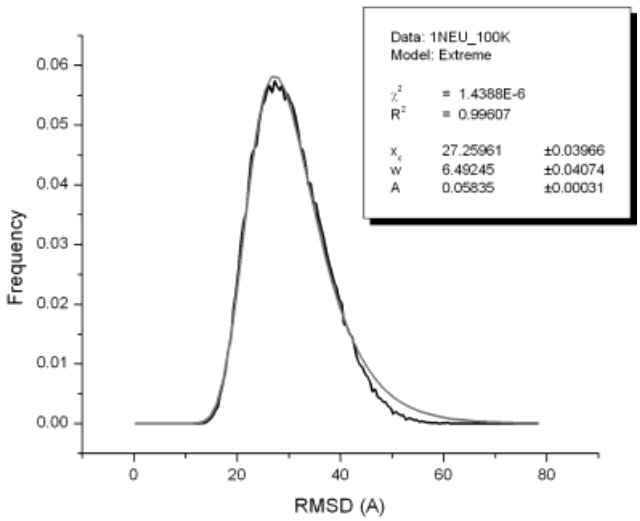

Fig. 11. Fit of the normalized RMSD distribution for 1NEU (black) to the extreme value distribution (gray). The χ^2 and R^2 indicate a very good fit. x_c and w are the parameters of the EVD (see Eq. 10), and A the peak amplitude, which can be shown to equal $(ew)^{-1}$ when normalized.

these give a good idea of which proteins fit best, which were mostly proteins longer than 75 amino acids. For a second test of the fit near the tail, we looked at the smallest RMSD structure computed in 100,000 structures and calculated the number of structures we would expect to need to get one of that accuracy. Using the CDF for the EVD (Eq. 10), we can directly obtain the probability of observing the structure with the smallest RMSD in the ensemble. Taking the reciprocal gives the number of structures needed to get such a hit by chance. Table IV shows that this provides reasonable results in most cases. However, for a few of the proteins: 1ENH, 4ICB, 1PMC, 5PTI, and 1DIV we were either very lucky, or the tail does not fit so well to the EVD for these structures. These, along with 1VII, do indeed have the largest residuals (see Fig. 12). Since the very short helical proteins consist of only three or four mostly rigid secondary structural elements, their conformational space may be more restricted than for the larger proteins as is evident by their poorer fit to the EVD. To further examine this, 1VII conformers were generated without biasing by the PHD prediction, and the fit improved to $R^2 = 0.998$, with the difference in observed and expected standard deviation (Ds on Table III) getting smaller too (data not shown). 1PMC simply does not fit the EVD distribution well at the tail even without the secondary structure bias, having more near-native structures than expected by the best fit, but in this case we are encouraged to find that it works *in favor* of our sampling procedure.

To compare our conformational space estimates to those given by Reva et al.,[12] we computed the number of compact conformations within 6Å RMSD of the native structure of the proteins. Defining compact as $R_{gyr} < 1.25R_{gyr\ native}$, we can use Eq. 13 to obtain an approximate upper bound on RMSD of $RMSD < 1.6R_{gyr\ native}$ for compact structures. Then using Eq. 10 we get the number of structures with this RMSD or less, and the number with an RMSD of 6Å or

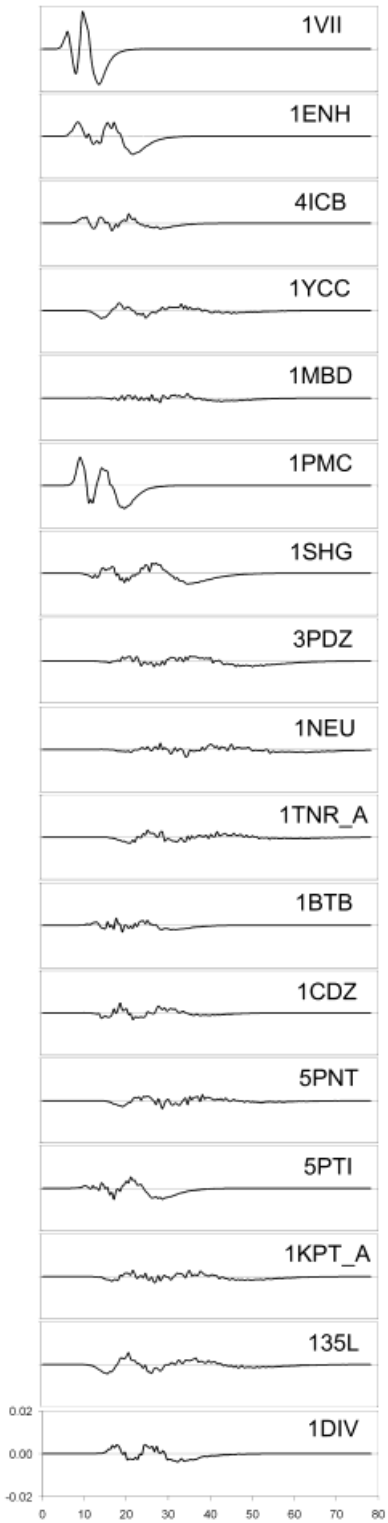

Fig. 12. Residual plots for the fits of distributions in Figure 10 to the EVD. All residuals are plotted on the same scale to allow easy comparison. Difference is plotted as actual distribution minus fitted EVD curve.

better, and divide. The results shown in Table IV are similar to the values given by Reva et al. of 10^4–10^5 for chains of 60–80 residues and 10^{11}–10^{12} for chains of 160–180 residues. Our estimates are a few orders of magnitude larger for several reasons. Our cutoff of protein size based on RMSD does not exclude the less compact structures that happen to have fairly good RMSDs. Also as we have shown, the conformational space available depends on protein structural class, and our probabilistic structures have more conformational flexibility than those sampled by Reva et al.[12] These differences also give us an idea of the size of the possible error in these estimates.

We can estimate the probability of finding a structure within a given RMSD to the native for each protein using the EVD curve fit, and with our assumption that FOLDTRAJ has sampled conformational space fairly. We can thus obtain an estimate of the true size of protein conformational space, where distinct conformations are defined as having a particular minimal RMSD from the native structure. Sample calculations are shown in Table IV for RMSD cutoffs of (1) 0Å, (2) 0.5Å, (3) 3Å and (4) 6Å, which correspond respectively to (1) exact matches; (2) the difference commonly found between two symmetric chains in a crystallographic unit cell; (3) the cutoff often used for correct comparative modeling or threading predictions[41]; (4) the cutoff used by previous workers in a similar study.[12] Plotting log (size of conformational space) against number of residues results in a fairly good correlation at a cutoff of 6Å [Fig. 13(a)]. The slope of this curve indicates the log of the number of conformations added when a protein is extended by one residue. For example, using the 6Å cutoff we find that each residue contributes on average $10^{0.131} \approx 1.35$ conformations within this gross tolerance. Using the 0.5Å cutoff [Fig. 13(b)] we observe a weaker correlation, probably since it is very sensitive here to inaccuracies of the tail approximation and the relative magnitude of the sample size. Also there are notable outliers: 1PMC, 1BTB, 135L, and 1DIV (the latter is not shown on the plot). The slope of best fit is 0.424 giving 2.65 conformations per residue with a 0.5Å cutoff. The effect of the probabilistic bias in the sampling of conformational space is to reduce the effective number of amino acid conformations, but without confining them to any particular discretization of space.

DISCUSSION

The results indicate that the conformational space of a protein is a complex function of both its length and its sequence of amino acids. Local structural preferences will tend to limit the conformations that stretches of amino acids and their side chains can take on. We find that probabilistic self-avoiding walks tend to produce structures that have an average RMSD roughly proportional to the protein length, when compared to the native fold of a globular protein. It is well known that high values of RMSD between a random and a compact, tightly packed structure will on average correlate well with the radius of gyration of the random structure (Eq. 13) and this is found to be the case here as well. However, unlike previous workers[3,8,12] we also find a much better fit when structures are separated into the classes of all α, all β, and mixed. We find that α structures are "hit" by chance much

TABLE III. Parameters for Fit of RMSD Distributions to the EVD

PDB code	$\mathbf{x}_c$ (Å)[a]	$\mathbf{w}$ (Å)[a]	$\mathbf{R}^2$	Observed μ (Å)	Observed σ (Å)	Δμ (%)[b]	Δσ (%)[b]
1VII	8.3	1.7	0.991	9.0	1.8	4	26
1ENH	13.4	2.7	0.994	14.4	2.9	4	20
4ICB	15.5	3.1	0.999	17.0	3.6	1	10
1YCC	22.5	5.3	0.997	25.4	6.1	1	10
1MBD	24.3	4.8	0.999	26.7	5.5	1	11
1PMC	12.9	2.2	0.992	13.7	2.4	4	20
1SHG	20.0	4.7	0.994	22.1	4.9	3	23
3PDZ	27.3	6.5	0.996	30.2	7.0	3	19
1NEU	32.5	7.4	0.997	36.2	8.3	2	14
1TNR_A	30.0	6.4	0.997	33.5	7.6	0	8
1BTB	18.0	3.3	0.999	19.7	3.8	1	10
1CDZ	20.8	4.1	0.998	23.0	4.8	0	8
5PNT	27.8	5.8	0.998	31.0	6.9	0	8
5PTI	16.7	3.4	0.997	18.2	3.7	2	18
1KPT	26.1	6.0	0.998	29.2	6.8	2	14
135L	25.2	6.5	0.994	28.7	7.3	1	14
1DIV	22.7	3.4	0.997	24.3	4.0	2	11

[a]Parameters of fit of the RMSD distribution to the EVD (see Eq. 10).
[b]Difference between observed mean and standard deviation and those calculated using EVD fit and Eqs. 11 and 12.

TABLE IV. Estimated Size of Protein Conformational Space

PDB code	Chain length	No. needed for best observed[a]	No. compact structures to get within 6 Å	$\log_{10}$ (size of conformational space) with given similarity cutoff			
				0 Å	0.5 Å	3 Å	6 Å
1VII	36	190,000	$4.11 * 10^1$	51	38	9	1.6
1ENH	54	$1.5 * 10^9$	$2.84 * 10^6$	61	51	20	7
4ICB	76	90,000,000	$2.94 * 10^9$	69	58	26	10
1YCC	107	3300	$2.36 * 10^9$	31	29	18	10
1MBD	153	5,200,000	$3.51 * 10^{19}$	70	63	37	20
1PMC	36	$1.6 * 10^{10}$	$7.58 * 10^9$	151	120	39	10
1SHG	62	36,000	$5.75 * 10^7$	31	28	16	9
3PDZ	96	28,000	$1.09 * 10^{10}$	29	27	18	11
1NEU	124	120,000	$8.28 * 10^{13}$	35	32	23	15
1TNR_A	144	15,000	$7.65 * 10^{17}$	48	44	30	19
1BTB	89	6,400,000	$1.33 * 10^{17}$	111	95	44	18
1CDZ	96	580,000	$1.00 * 10^{16}$	72	64	35	17
5PNT	157	22,000	$5.77 * 10^{17}$	53	48	31	19
5PTI	58	11,000,000	$6.58 * 10^9$	60	52	25	10
1KPT_A	105	27,000	$1.04 * 10^{11}$	33	31	20	12
135L	129	5700	$3.44 * 10^7$	21	19	13	8
1DIV	149	29,000,000	N/A	>300	273	132	55

[a]For best RMSD observed in 100,000 conformers, estimated using Eq. 10.

better than β structures. The difference is also clear for the lowest RMSD structure achieved in each run, a measure similar to the significance cutoff used by previous authors.[8,12] This can be understood by noting that helical proteins have a lot less conformational flexibility and are quite rigid relative to β-strand structure, which can bend quite easily and has a larger region of Ramachandran space[29] available to it. For a helical protein, we need only get most of the turns correctly oriented and the rest of the structure will fall into place, while with β-sheets we are not guaranteed individual strands will line up to form a hydrogen bonded sheet even with approximately correct turns, due to their flexibility.

This tells us that the brute force probabilistic approach to searching conformational space works well for small (under about 60 residues) helical proteins or very small (about 30 residue) non-helical proteins. At these sizes, the best structure generated in a pool of several million should be close enough to the native to detect with a good scoring function (for example see references[42–45]). The advantage to our approach is that an arbitrary scoring function may be applied since the process is not biased by it, and the one that works best used.

For larger proteins, additional information is needed to arrive at correct folded structures. While pieces of structure on the order of 30 residues in length may be close to their native conformation, no single structure will be very close to the true folded structure. We simply cannot sample enough yet for large proteins. Being able to identify and combine these native-like substructures will be the

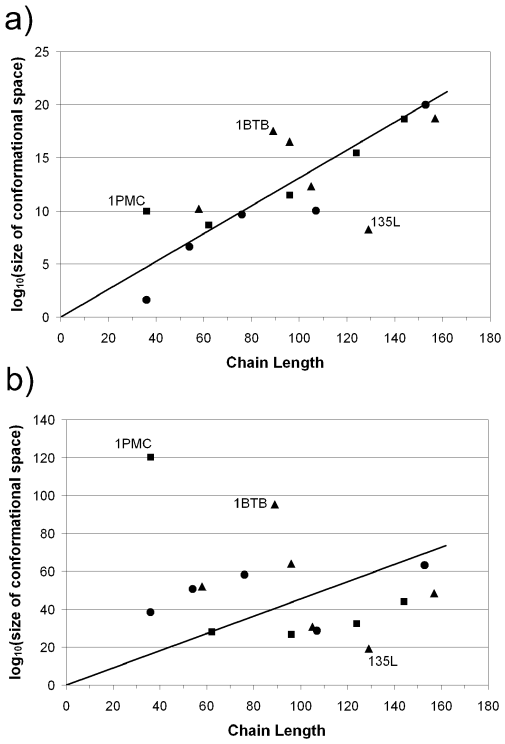

Fig. 13. Estimated size of protein conformational space for the proteins used in this study, plotted logarithmically against chain length. Circles are helical proteins, squares are sheet proteins, and triangles are mixed α/β or $\alpha+\beta$. The solid line indicates the best fit that passes through the origin. All conformations within (a) 6Å RMSD or (b) 0.5Å RMSD from the native state are considered to be identical conformations when enumerating them. These represent different resolutions, or degrees of precision in how exactly we define a conformer. Outliers 1PMC, 1BTB, and 135L are labelled on both plots. 1DIV is a major outlier, most likely due to its non-globular nature, and is omitted for clarity.

subject of future research and may lead to more compact and native-like probabilistic structures for larger proteins.

We have provided what we believe is the best estimate to date of the size of protein conformational space, taking into account steric effects and the allowed regions of Ramachandran space, as well as inherent amino acid propensities and local structural preferences. Our approximation for the size of conformational space with an RMSD cutoff of 0.5Å, of Ω^N where Ω is about 2 for larger proteins (over 100 amino acids) and up to about 10 for very short proteins [see Fig. 13(b)], is a lot smaller than that given, for example, by Cohen and Sternberg[3] who suggested as a crude approximation that $\Omega = 10$. This illustrates how greatly conformational space is reduced by mostly steric effects and probabilistic biases, without having to resort to discretization or lattice models. Unfortunately, a 100-residue protein would still take approximately 10^{22} CPU-years to fold given that a structure was sampled every second. The fact that the structures of sequences of length 30 or less can be reliably "predicted" by brute force to within a given tolerance may be surprising, and this limit will only get larger as computing power continues to increase and techniques improve. Attempts to further prune conformational space could succeed in reducing the protein folding problem to a computable one.

ACKNOWLEDGMENTS

This research was supported by grants to C.W.V. Hogue by the Natural Sciences and Engineering Research Council of Canada. The cluster computer and MoBiDiCK are funded by the Canadian Foundation for Innovation, the Ontario Research and Development Challenge Fund and MDS-SCIEX. H.J. Feldman is supported in part by an Ontario Graduate Scholarship.

REFERENCES

1. Levinthal C. Are there pathways for protein folding? J Chim Phys 1968;65:44–45.
2. Levinthal C. How to fold graciously. In: Debrunner P, Tsibris JCM, Münck E, editors. Mossbauer spectroscopy in biological systems, Proceedings of a Meeting held at Allerton House, Monticello, Illinois. Urbana: University of Illinois Press; 1969. p 22.
3. Cohen FE, Sternberg MJ. On the prediction of protein structure: The significance of the root-mean-square deviation. J Mol Biol 1980;138:321–333.
4. Karplus M. The Levinthal paradox: yesterday and today. Fold Des 1997;2:S69–S75
5. Ngo JT, Marks J. Computational complexity of a problem in molecular structure prediction. Protein Eng 1992;5:313–321.
6. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci USA 1992;89:20–22.
7. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding: a perspective from simple exact models. Protein Sci 1995;4:561–602.
8. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol 1994;235:625–634.
9. Dunbrack RLJ, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1994;1:334–340.
10. Dunbrack RLJ, Karplus M. Backbone-dependent rotamer library for proteins. Application to sidechain prediction. J Mol Biol 1993;230:543–574.
11. McLachlan AD. How alike are the shapes of two random chains? Biopolymers 1984;23:1325–1331.
12. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an RMSD of 6Å? Fold Des 1998;3:141–147.
13. Rost B, Sander C, Schneider R. PHD - an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994; 10:53–60.
14. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.
15. Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 1995;11:681–684.
16. Chen CC, Singh JP, Altman RB. Using imperfect secondary structure predictions to improve molecular structure computations. Bioinformatics 1999;15:53–65.
17. Heyn MP. Circular dichroism for determining secondary structure and state aggregation of membrane proteins. Methods Enzymol 1989;172:575–584.
18. Amir D, Krausz S, Haas E. Detection of local structures in reduced unfolded bovine pancreatic trypsin inhibitor. Proteins 1992;13:162–173.
19. Gottfried DS, Haas E. Nonlocal interactions stabilize compact folding intermediates in reduced unfolded bovine pancreatic trypsin inhibitor. Biochemistry 1992;31:12353–12362.
20. Brant DA, Flory PJ. The configuration of random polypeptide chains. I. Experimental results. J Am Chem Soc 1965;87:2788–2791.
21. Brant DA, Flory PJ. The configuration of random polypeptide chains. II. Theory. J Am Chem Soc 1965;87:2791–2800.
22. Feldman HJ, Hogue CWV. A fast method to sample real protein conformational space. Proteins 2000;39:112–131.
23. Gregoret LM, Cohen FE. Protein folding. Effect of packing density on chain conformation. J Mol Biol 1991;219:109–122.
24. Lo CL, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C.

SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
25. 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.
26. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540–553.
27. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
28. Ryan PJ. Geometry on the sphere. In: Euclidean and non-Euclidean geometry: an analytical approach. Cambridge: Cambridge University Press; 1992. p 84–123.
29. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283–438.
30. Wang Y, Addess KJ, Geer L, Madej T, Marchler-Bauer A, Zimmerman D, Bryant SH. MMDB: 3D structure data in Entrez. Nucleic Acids Res 2000;28:243–245.
31. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN.1 specification for macromolecular structure. Proc Intell Sys Mol Biol 1995;3:259–267.
32. Stewart DE, Sarkar A, Wampler JE. Occurrence and role of cis peptide bonds in protein structures. J Mol Biol 1990;214:253–260.
33. MacArthur MW, Thornton JM. Deviations from planarity of the peptide bond in peptides and proteins. J Mol Biol 1996;264:1180–1195.
34. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst 1991;A47:392–400.
35. Dharsee M, Hogue CWV. MoBiDiCK: A tool for distributed computing on the Internet. In: Raghavendra C, editor. Proceedings, 9th Heterogeneous Computing Workshop. Los Alamitos: IEEE Computer Society; 2000. p 323–335.
36. Michalickova K, Dharsee M, Hogue CWV. Sequence analysis on a 216-processor Beowulf cluster. In: Beckman P, Greenberg D, Hankins G, Ts'o T, editors. Proceedings of the 4th Annual Linux Showcase & Conference. Berkeley: USENIX Association; 2000. p 111–119.
37. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
38. Edwards SF. The statistical mechanics of polymers with excluded volume. Proc Phys Soc Lond 1965;85:613–624.
39. Pietronero L. Survival probability for kinetic self-avoiding walks. Phys Rev Lett 1985;55:2025–2027.
40. Kremer K, Lyklema JW. Kinetic growth models. Phys Rev Lett 1985;55:2091
41. Bryant SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.
42. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
43. Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. Proteins 1999;54–67.
44. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
45. 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.
46. Hunt NG, Gregoret LM, Cohen FE. The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search. J Mol Biol 1994;241:214–225.
47. 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.