Reducing the dimensionality of the protein-folding search problem

George D. Chellapa and George D. Rose*

T.C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland 21218

Received 19 April 2012; Revised 4 June 2012; Accepted 5 June 2012

DOI: 10.1002/pro.2106

Published online 12 June 2012 proteinscience.org

Abstract: How does a folding protein negotiate a vast, featureless conformational landscape and adopt its native structure in biological real time? Motivated by this search problem, we developed a novel algorithm to compare protein structures. Procedures to identify structural analogs are typically conducted in three-dimensional space: the tertiary structure of a target protein is matched against each candidate in a database of structures, and goodness of fit is evaluated by a distance-based measure, such as the root-mean-square distance between target and candidate. This is an expensive approach because three-dimensional space is complex. Here, we transform the problem into a simpler one-dimensional procedure. Specifically, we identify and label the 11 most populated residue basins in a database of high-resolution protein structures. Using this 11-letter alphabet, any protein's three-dimensional structure can be transformed into a one-dimensional string by mapping each residue onto its corresponding basin. Similarity between the resultant basin strings can then be evaluated by conventional sequence-based comparison. The disorder $\rightarrow$ order folding transition is abridged on both sides. At the onset, folding conditions necessitate formation of hydrogen-bonded scaffold elements on which proteins are assembled, severely restricting the magnitude of accessible conformational space. Near the end, chain topology is established prior to emergence of the close-packed native state. At this latter stage of folding, the chain remains molten, and residues populate natural basins that are approximated by the 11 basins derived here. In essence, our algorithm reduces the protein-folding search problem to mapping the amino acid sequence onto a restricted basin string.

Keywords: protein folding; conformational searching; search algorithm; conformational

Introduction

Two major obstacles are thought to impede progress in ab initio protein folding: (i) accurate representation of conformational energy terms and (ii) the search problem. The former typically uses a force-

field such as Amber¹ or CHARMM² from which the energy of a given conformation can be calculated by summing over intramolecular interactions plus intermolecular solvent contributions. The latter seeks to identify relevant conformations within a vast energy landscape, but the apparent magnitude of this problem defeats any possibility of doing so by systematic search.^{3–5} In the sense used here, “ab initio” means an approach that is not based on assembling a structure using segments extracted from the protein database.⁶

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: George D. Rose, Jenkins Department of Biophysics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218. E-mail: grose@jhu.edu

We question both perceived obstacles. The calculation of conformational stability is simplified by the realization that the forces stabilizing native topology are separable from those forces stabilizing native packing. At least for some proteins, native topology is already well developed at the dry molten globule stage, prior to the formation of side chain close-packing.⁷ It follows that the intricate details of side chain close-packing can be neglected when seeking to predict native topology for such proteins.

Here, our main focus is on the search problem. Specifically, we seek to disentangle the folding step at which native topology is established from the step at which side chain close-packing occurs. Toward this goal, we show that topology can be mapped from three-dimensional coordinate space (Euclidian 3-space) into a discrete linear space, where the discrete elements correspond to labeled ϕ, ψ -basins embedded in a Ramachandran plot.⁸ Conformational comparisons can then be performed over an alphabet of basin labels, using conventional sequence-based search tools such as BLAST.⁹

To develop a practical algorithm, we define 11 discrete ϕ, ψ -basins, each $20^\circ \times 20^\circ$; in the aggregate, they cover approximately 3.4% of the total ϕ, ψ -space. The basins are labeled uniquely and constitute an 11-letter alphabet. Any database of protein structures can be transformed into a companion database of basin sequences by mapping each residue of every protein structure onto its nearest basin. In other words, each protein structure in the companion database is represented simply as a string of letters in the 11-letter alphabet. A target protein of interest can then be compared to candidates in the companion database by performing a conventional sequence-based search, as described later. Of course, this algorithm will only be sufficient if the 11 chosen basins provide enough specificity to back-transform a linear basin string into a tight cluster of native-like three-dimensional structures, comparable to reconstructing tertiary structure from a knowledge of secondary structure.^{10,11}

Is the algorithm described here merely a technical artifice, or does it reflect an underlying experimental verity? To probe this question, the algorithmically derived entropy loss (basin string $\rightarrow$ structure) is compared with an experimentally derived estimate of entropy loss (dry molten globule $\rightarrow$ close-packed native structure). The values differ by approximately a factor of 2.7, approximating physiological thermal energy ($RT \ln 2.7 \sim 0.6$ kcal/mol; R = the gas constant, T = 300 K) and compatible with a folding process where trajectories converge on native-like topology prior to final annealing.

Several aspects of our approach are familiar topics in the existing literature. Structural alphabets and simplified representations of protein structure

Figure 1. The eleven $20^\circ \times 20^\circ$ labeled basins abstracted from high resolution X-ray structures.

have a long history,^{12–24} for example, and it is not our primary aim to further refine such methods here. Rather, our specific threefold goal is to show that:

- (1) A simple 11-letter structural alphabet extracted from populated basins in proteins of known structure^{25,26} together with a modified BLAST search⁹ that uses a structure-based substitution matrix is sufficient to anchor a practical search algorithm.
- (2) Information embedded within this alphabet is also sufficient to encode protein topology, which can be successfully recovered from the linear basin-string representation, albeit approximately, without resorting to either detailed energy minimization or extensive conformational searching.
- (3) For proteins with dry molten globule intermediates, overall topology is established prior to ultimate side chain annealing.⁷ Assessed within this framework, our results are compatible with a substantial reduction in the apparent magnitude of accessible conformational space.

Results

Basin classification and mapping

Highly populated ϕ, ψ -regions were identified in a contour plot of high-resolution X-ray crystal structures (resolution ≤ 1.6 Å), taken from a nonredundant (sequence similarity $\leq 25\%$) PISCES list²⁷ of PDB structures⁶ that ranges from 16 to 1015 residues, with an average length of 198 ± 130 residues. Eleven regions were abstracted from the plot and classified as labeled $20^\circ \times 20^\circ$ basins (Fig. 1 and Table I).

Using this classification, three-dimensional X-ray structures were transformed into linear basin strings by mapping each ϕ, ψ -angle onto its nearest

Table I. *Basin Definitions*^a

A: -62; -42	R: -68; -18	V: -93; 2	B: -120; 135	P: -64; 139
G: -93; 95	D: -134; 70	L: 51; 42	U: 82; -3	T: 55; -129
Y (Gly only) 77; -171				

^a Basin labels and ϕ, ψ -centroids in degrees. Basins were defined as ϕ, ψ -regions that extend $\pm 10^\circ$ beyond the centroid in each direction.

Table II. *Ten Representative Proteins*^a

2UBP (100); 2.0 Å	1UBQ (76); 1.8 Å	1YRF (35); 1.07 Å	1LMB (92); 1.8 Å	1C75 (71); 0.97 Å
1PGB (56); 1.92 Å	1X6Z (123); 0.78 Å	2PHY (125); 1.4 Å	1PZ4 (116); 1.35 Å	1HEL (129); 1.7 Å

^a Table entries list PDB identifiers, number of residues (in parentheses), and resolution for each protein.

basin. Residues with *cis*-isomers were identified by lower case labels, and their ω -angles were generated accordingly (see Methods).

When does a protein's topology emerge?

Operationally, proteins from a PISCES list²⁷ were mapped onto their respective basin strings and accumulated in a companion database. A conventional sequence-based algorithm was then used to search this database for candidates matching a given target protein (see Methods). This approach provides a substantial reduction in search time by transforming a three-dimensional search into a corresponding linear search.

However, it is important to emphasize that the primary motivation for this article is to test whether the search problem can be solved by mapping the amino sequence onto its basin string. Little or no conformational indeterminacy (i.e., “frustration”⁴) remains beyond this state, although substantial residual conformational entropy still persists.⁷

To test this hypothesis, it is sufficient to show that a protein can be identified uniquely and reconstructed faithfully from knowledge of its basin string. This section presents data for 10 representative proteins (Table II) that were chosen as search targets and identified successfully, with no false positives. The following section describes the three-dimensional structures that were back-calculated from the basin strings of these 10 proteins.

A search for the 10 target proteins, each in turn, was performed against a companion database derived from a PISCES list²⁷ of 2322 proteins (resolution at least 1.6 Å; sequence identity $\leq 25\%$). The companion database was created by mapping all PDB structures from the PISCES list onto their corresponding basin strings. The target protein itself was appended to the PISCES list if not already included, and a few additional nearly identical structures were also added in order to assess search sensitivity. Similarity scores were calculated using gapped sequence alignment and a “basin substitution matrix” (see Methods). Detailed results are shown for hen egg lysozyme (1HEL) and summarized for the remaining nine proteins.

Results from the database search for 1HEL are shown in Figure 2. In this case, the PISCES list was augmented with eight hen egg lysozymes (listed in Table III), and scores for the target protein, 1HEL, with itself and in alignment with all other proteins were calculated and plotted. From the plot, it is clear that the basin string representation is sensitive enough to differentiate 1HEL from other candidates, even highly similar structural analogs. Two proteins of identical sequence to 1HEL could not have been distinguished in a conventional sequence-based search^{9,28} but can be successfully differentiated here because their X-ray structures differ slightly. Specifically, two or more residues map to different basins at 38 sites across the 129-residue sequence, despite the fact that eight of the nine hen egg lysozymes crystallized in the same space group.

Similar plots for the remaining nine proteins in Table II are provided in the Supporting Information Figure S1. Z-scores (i.e., number of standard deviations above the mean) for this set range between 2.13 and 3.36, with no false positives.

Figure 2. Basin sequence alignment scores. Scores were fitted to an extreme value distribution⁷⁶ (dashed curve) with mean = -73.5 and $\sigma = 59.0$ and an extended tail of decreasing similarity. 1HEL has a score of 135, 3.5σ , above the mean, at the leading edge of a tight cluster of nine structurally analogous proteins with similar scores. In addition to 1HEL, there are seven other added lysozyme analogs, plus a previously undetected analog (2VB1) that was already included in the original PISCES list. This nine-protein cluster is expanded in the inset. Reading right to left, the nine proteins are 1HEL, 1HEN, 1HEO+1HEM, 1HER, 1RFP+1FLQ, 1BWI, and 2VB1.

Table III. *1HEL Analogs Added to the Test Database^a*

1HEL (0.0 Å); 100%	1RFP (0.11 Å); 100%	1BWI (0.12 Å); 100%	1HEO (0.09 Å); 99%	1HER (0.09 Å); 99%
1HEM (0.11 Å); 99%	1HEN (0.12 Å); 98%	1FLQ (0.13 Å); 99%	2VB1 ^b (0.70 Å); 100%	

^a PDB identifiers of hen egg lysozyme structural analogs, RMSD with 1HEL (in parentheses), and percentage aligned sequence identity with 1HEL.

^b A previously undetected protein (2VB1) already included in the original PISCES list.

Table IV. *Best-Fitting Replica for the Ten Representative Proteins^a*

1YRF; 0.55 Å	1PGB; 1.43 Å	1UBQ; 1.78 Å	1C75; 2.30 Å	1LMB; 2.63 Å
2UBP; 2.62 Å	1X6Z; 4.00 Å	1PZ4; 4.18 Å	1HEL; 7.75 Å	2PHY; 7.90 Å

^a Table entries list PDB identifiers and RMSD between the closest basin string replica and its X-ray counterpart.

Recovering the three-dimensional structure from a basin string

The 10 proteins were minimized into their X-ray structures as described in Methods. The objective is to show that a close replica of the X-ray structure can be recovered from each protein's basin string, simply by generating a set of viable structures with ϕ, ψ -angles chosen from their respective basins, as described previously.²⁹ Table IV lists the RMSD of the best-fitting (i.e., least root-mean-square-difference) member of this set for each of the 10 proteins, and Figure 3 illustrates the fit of this replica to the native structure. A more general approach would be to recover the X-ray structure by minimizing an energy function, and, indeed, we have shown previously that knowledge of a protein's secondary structure is sufficient to closely model its tertiary structure.^{10,11} However, analysis of viable replicas with further fine-tuning of basins is left to future work. Here, the immediate goal was to demonstrate that the 11 basins are sufficient to identify a protein structure uniquely (When Does a Protein's Topology Emerge? section) but specific enough to allow back-calculation of a close-fitting three-dimensional model of the original X-ray structure.

The RMSD was calculated using α -carbons. Poor fits in Table IV are caused by only few aberrant residues, often glycines or cysteines. For example, in 1HEL (RMSD = 7.75 Å) G102 maps to the U-basin, but its ϕ -angle is 127.8°, placing it 45° from the basin centroid. The worst fragment in Figure 4(a), residues 83–128, is the result of a perturbing disulfide bridge between C115 and C30; C115 maps to the A basin, but with a ϕ -angle that is 51° from the basin centroid. Splitting this fragment at the perturbing cystine results in two piecewise contiguous segments that match the X-ray structure quite well (residues 83–115; RMSD = 0.75 Å and residues 115–128; RMSD = 0.57 Å). Similarly for 2PHY, the overall high RMSD is the result of a few outlying residues, which coincide in this case with ligand-binding sites. Again, the protein can be split into three piecewise contiguous fragments that match the X-ray structure quite well [Fig. 4(b)].

These are typical examples. A poor overall RMSD is the result of a few outliers, as shown by splitting the protein into a few piecewise contiguous

fragments that match the X-ray structure closely. With the exception of these anomalous residues, the overall three-dimensional structure can be captured successfully using the 11 basins in Table I. Despite these anomalies, the basin string is sensitive enough to identify the protein uniquely, even in the presence of structurally analogous decoys (Fig. 2).

Thermodynamically defined basins in dry molten globule folding intermediates

A dry molten globule (DMG) folding intermediate is an expanded form of the native fold in which most solvent water has been expelled from the protein

Figure 3. Superimposed image of the three-dimensional structure recovered from its basin string (cyan) on the X-ray structure (purple) for each of the 10 proteins in our test set.

Figure 4. Superposition of three piecewise contiguous fragments for (a) 1HEL and (b) 2PHY, the two proteins in our test set with the worst overall RMSDs. It is apparent that the poor overall RMSD fits are the result of a few outlying residues and, with these exceptions, the reconstructed backbone follows main-chain topology quite well.

core but buried side chains lack the close-packed character that is a familiar hallmark of the native state.^{30,7} When an unfolding protein expands to the dry molten globule state, its side chains unlock and gain conformational entropy, side chain interactions become increasingly liquid-like, but nevertheless the backbone structure remains intact or largely so. A DMG state has been persuasively documented in several experimental studies,^{31–35} and additional examples are on the immediate horizon.

Ideally, our basin sizes would be based on the thermodynamics of the folding reaction, not simply the investigator-derived values defined in Figure 1. In principle, the natural basins are those present in the DMG, where chain topology has been established but substantial conformational mobility remains. A back-of-the-envelope basin size estimate can be obtained from the villin headpiece subdomain. Based on data from triplet-triplet energy transfer,³⁵ the transition between the native state and the DMG (N – DMG) in the unfolding direction is associated with a large, unfavorable enthalpy ($\Delta H^0 = 8.37$ kcal/mol) that is compensated by a large favorable entropy ($\Delta S^0 = 0.24$ cal/mol-K). Assuming negligible change in heat capacity, the DMG basin would be 1.47 times the size of the native basin ($\Delta S = 26.77$ kcal/mol-K/35 residues = 0.76 kcal/mol-K/residue. At constant T and ΔC_p , $\Delta S/\text{residue} = R \ln \frac{V_{\text{basin}}}{V_{\text{native}}} = 0.76$ kcal/mol-K, a volume ratio = 1.47). Using this crude estimate, our investigator-derived basin size exceeds that of the DMG by a factor of ~ 2.7 for native state basins of $10^\circ \times 10^\circ$, a difference that approxi-

mates physiological thermal energy ($RT \ln 2.7 \sim 0.6$ kcal/mol; R = the gas constant, $T = 300$ K).

Conformational space is highly degenerate

Following Levinthal,³ it has become customary to assess the magnitude of conformational space by counting distinguishable ϕ, ψ -angles, with constraints imposed by excluded volume assumptions and other plausible restrictions.^{36–38} The resultant picture portrays conformational space as vast and largely featureless. For example, with residues restricted to only two possible conformations (e.g., α or β), there would be $2^{100} \approx 10^{30}$ possible conformers for a 100-residue chain. Thus, even an overly conservative approximation yields an inordinately large number. Such estimates have buttressed the conviction that a protein's native conformation resides within a vast energetic landscape.

However, just seven recurrent structural motifs account for $\sim 90\%$ of protein structure on average,²⁵ and these can be rationalized by simple models that include only sterics and hydrogen-bond satisfaction.²⁶ The predominant motifs include α -helices, β -strands, polyproline II, and β -turns together with three additional minor motifs.²⁵ All seven motifs are familiar structures that range in length from 3 to 5 residues. Consequently, mapping ϕ, ψ -angles onto these iconic structures reveals a dramatic and simplifying degeneracy that is built into the physical chemistry of the protein backbone.

Table V. Basin-String Encodings of Structural Motifs^a

Structural motif	Basin string
α -Helix	AAAAA
β -Strand/PII	Any combination of B and P
β -Turn (residues i+1, i+2)	RV, AV, LU, PU, TV, AA, AR, RR, RA, LL
Inverse γ -turn	G

^a See Table I for basin definitions.

By analogy, the number of codon sequences compatible with even a small protein-like BPTI (57 residues) is three orders of magnitude larger than Avogadro's constant. Accordingly, algorithms to identify protein homologs are based on protein sequences rather than DNA sequences.³⁹ Similarly, the conventional approach of mapping either ϕ, ψ -angles (or basins) into undifferentiated conformational space obscures the underlying degeneracy, much as it would if protein sequences were represented by codons instead of amino acids. Here, each of the seven structural motifs can be encoded by multiple basin strings, so alternative encodings accumulate exponentially. For example, with an average of two possible basins for each residue position, there are $2^{100} \approx 10^{30}$ possibilities for a 100-residue chain, recapitulating the familiar Levinthal conundrum.

Overcoming degeneracy by inverting the problem, we assess the magnitude of conformational space that is accessible to the seven structural motifs²⁵ and then work backward to basin-string representations of these motifs. Specifically, each structural motif is represented by the set of basin strings that can encode it (Table V). Summing the coverage of each protein's complete basin string over all proteins in the PISCES list of 2363 proteins (resolution ≤ 1.6 -Å resolution, sequence similarity $\leq 25\%$), these motif strings account for 89.59% of the entire population of 515,791 residues. Remaining fragments are typically short, as shown in Figure 5.

Notably, the basins used in Perskie *et al.*²⁶ are substantially larger than those used here (Fig. 1 and Table I), and the basin-string representations of these seven structural motifs²⁵ in Table V are overly stringent. Consequently, our finding of $\sim 90\%$ coverage is deliberately conservative. Many of the ostensibly remaining fragments (Fig. 5) are non-ideal α -helices, helix caps,⁴⁰ and distorted β -turns.⁴¹

Our strategy of inverting the conventional approach reduces the protein folding search problem to manageable proportions by collapsing ϕ, ψ -space onto the repertoire of discrete motifs known to populate folded proteins. Larger structures are essentially mix-and-match composites of these primitive components.

Discussion

The supposition that small proteins fold via a two-state reaction, $U(nfolded) \rightleftharpoons N(ative)$, has become

an anchoring premise in protein folding studies.⁴² Typically, the U state is viewed as structurally featureless because its energy landscape is consistent with a vast number of accessible conformers: literature estimates a range from 10^{15} to 10^{80} possibilities for a protein of 100 residues.³⁷ Under unfolding conditions, only minor energy barriers, of order $k_B T$, separate these conformers. Accordingly, any given molecule in the population can adopt any accessible conformation readily. Following a shift from unfolding to folding conditions, each chain in the population must wend its way from this energetically arid landscape to the stable native state and do so in biological real-time. Framed in this way, the search problem is expected to be a formidable obstacle in the design of folding algorithms. Consistent with this expectation, lack of success in protein folding simulations is often attributed to an inadequacy in either search time or search strategy. Yet, proteins do fold successfully on a biologically relevant time scale, prompting suggestions that they solve the search problem by tracing energetically preferred routes.^{3-5,43} Here, we suggest that the entire issue should be re-evaluated.

It is usually assumed that little or no structural discrimination occurs in the unfolded state because U is structurally featureless. Further, residue backbones are identical (with the minor exception of glycine and proline), so discrimination must arise from interactions between and among residue side chains. Such interactions are incomplete until final close packing is achieved, implying that the search proceeds until the native state is attained. This plausible folding paradigm has conditioned our thinking for decades. In its stead, we offer an alternative explanation of the experimental facts.

Completely denatured protein chains have a mean radius of gyration, $\langle R_g \rangle$, identical to that of a random coil polymer in good solvent.^{44,36,45} This experimental finding and related data have prompted the field to conclude that denatured proteins are

Figure 5. Fragments that remain after eliminating all basin-string encoded structural motifs. Our basin-string definitions of these motifs are deliberately stringent (Table V), and many of these remaining short fragments would be subsumed within a more liberal definition. The number of fragments at each length decreases as $\sim 1/\sqrt{n}$.

random coils, although the founders such as Kauzmann⁴⁶ and Tanford⁴⁴ raised prescient cautions about jumping to such conclusions. For example, a contrived model consisting of rigidly structured segments interconnected by flexible joints would also have the $\langle R_g \rangle$ of a random coil.⁴⁷ Further, it has been shown that the unfolded population is preferentially enriched in conformers with backbone dihedral angles in the polyproline II region of the ϕ, ψ -map.^{48–54}

An even more compelling case for re-evaluation comes from the thermodynamics of organic osmolytes, small uncharged molecules, ubiquitous throughout nature, that can modulate solvent quality.⁵⁵ The addition of osmolyte co-solvents shifts the $U \rightleftharpoons N$ equilibrium, protecting osmolytes, such as TMAO and glycine betaine, shift to the right, forcing folding.⁵⁶ Conversely, the familiar denaturing osmolyte urea shifts to the left forcing unfolding.⁵⁷ In contra-distinction to the conventional paradigm, the osmolyte effect is exerted predominantly on the backbone in the unfolded state.⁵⁸ This observation is well supported by experimental measurements at equilibrium. The underlying osmophobic mechanism is controversial,^{59–65} but clearly the addition of an osmolyte co-solvent can dial backbone–solvent interactions up/down by changing solvent quality.^{55,64,66}

In our model, solvent quality controls the balance between intramolecular H-bonding and backbone–solvent H-bonding.^{58,60} In turn, this balance influences folding by either favoring (in poor solvent) or disfavoring (in good solvent) the formation of α -helices and strands of β -sheet, the hydrogen-bonded scaffold elements on which—of necessity—proteins are built. These two secondary structure elements, and only these two, can be extended indefinitely without encountering a steric clash while at the same time providing readymade H-bond partners for every backbone polar group removed from access to solvent water. An unsatisfied H-bond in just one solvent-inaccessible residue would come at an energetic cost of ~ 5 kcal/mol,^{67,41} rivaling the entire free energy of stabilization, $\Delta G_{U \rightarrow N} \approx [-5 \text{ to } -15]$ kcal/mol, for a typical globular protein.⁶⁸

In short, as folding is initiated and backbone polar groups become sequestered, the loss of solvent H-bond partners is compensated by the formation of secondary structure, the fundamental interplay between energy and structure. A typical 100-residue protein would have approximately 10 segments of α -helix and/or β -strand, amounting to $\sim 2^{10} = 1024$ possible scaffolds. The resultant number of energetically accessible conformers is limited by the number of ways these scaffold elements can be interconnected by intervening “coil” segments. In a globular protein, the length of such interconnecting segments is highly restrictive; the distribution of lengths⁶⁹ peaks at 3–4 residues⁷⁰ and then diminishes expo-

nentially. As a consequence, the backbone alone constrains fold space to a few thousand distinct folds for a protein the size of lysozyme or ribonuclease,¹⁴ in agreement with other estimates.^{71,72}

The backbone-based winnowing of fold space is sequence-independent, and it takes effect as proteins fold.⁵⁸ Side chains are then responsible for selecting the native conformation from the available repertoire of a few thousand remaining possibilities. However, it is unlikely that specific side chain–side chain interactions are primarily responsible for conformational selection because the topology is well established before side chains have annealed.⁷ Although the 11 basins defined here cover only 3.4% of the total ϕ, ψ -space, protein topology is determined uniquely when each residue is confined to its nearest available basin. At this point, remaining conformational indeterminacy has been abolished, but substantial side chain conformation entropy remains.

The alternative explanation of the experimental facts described here leads to a revised estimate of the magnitude of thermodynamically accessible conformational space: only $\sim 10^4$ viable backbone scaffolds are possible for a lysozyme-sized protein, regardless of sequence. Further reduction comes from the realization that protein topology is established prior to the endpoint of folding, when side chains are fully annealed. Notably, excluding forces (e.g., steric clash, polar groups lacking H-bond partners) are primarily responsible for this revised estimate, not the usual attractive forces treated in a forcefield. Of particular note, these excluding forces operate predominantly in the unfolded state and evade detection in Gō models⁷³ or knowledge-based potentials.

The preceding demonstrates that prediction of protein topology can be reduced to the problem of predicting the basin string from the amino acid sequence. It remains to be seen whether the wealth of available sequence-structure information in the PDB will enable a ready solution to this problem.

Materials and Methods

Determination of basins

Nearly 90% of all protein structure can be captured by a small number of ϕ, ψ -basins,²⁵ each associated with only one or two structural motifs.²⁶ Extending this earlier work, high-resolution PDB structures⁶ were extracted from a PISCES list,²⁷ mapped onto a ϕ, ψ -plot, and 11 major basins were identified (Fig. 1).

Generation of structures

A three-dimensional protein structure of N residues was transformed to a linear basin string of N letters by mapping each residue onto its nearest ϕ, ψ -basin. This basin string was then subdivided into overlapping six residue fragments, each overlaying the next

Table VI. The Basin Substitution Matrix Used in This Study

	A	B	D	G	L	P	R	T	U	V	Y
A	0										
B	-3	1									
D	-1	0	1								
G	-2	1	1	1							
L	-2	1	1	1	3						
P	-2	1	1	1	1	1					
R	0	0	0	0	0	0	1				
T	-2	1	1	1	1	1	1	4			
U	-2	1	1	1	2	2	0	1	3		
V	-1	0	1	0	1	1	0	1	1	2	
Y	-2	1	1	1	2	1	0	1	2	1	1

by one residue. If necessary, lengths were extended to avoid either glycine or proline at fragment termini. Fragment structures were generated and then merged to form the entire protein structure by overlaying adjacent fragments at their residue in common. Specifically, two adjacent fragments were joined at their residue in common, r and r' , (i.e., last residue of the first fragment, r , and first residue of the second fragment, r') by translating and rotating r' into the reference frame defined by r , which is held fixed. This transformation is then applied to remaining residues in the second fragment, after which r' is removed. All residues except glycine and proline were represented by alanine.

In greater detail, 10^4 attempts were made to generate a structure for every fragment. An attempt involved generating a conformation for each residue in the fragment by selecting a ϕ, ψ angle at random from its appropriate basin. A random ω -angle was selected also from a Gaussian-distributed $\pm 5^\circ$ interval centered at 180° , unless the residue in question was a *cis*-proline, in which case the interval was $0^\circ \pm 20^\circ$. Puckering of the pyrrolidine ring was taken into account,⁷⁴ using statistics extracted from the same PISCES list (*cis*-proline: 85% DOWN, 14% UP; *trans*-proline: 46% DOWN, 54% UP). The resultant trial structure was retained without further action if it was determined to be (i) clash-free and (ii) within 1.5-Å RMSD of its corresponding PDB structure. Otherwise, the trial structure was minimized and then retained if it met the two criteria or rejected if it failed to do so. The fragment with the smallest RMSD was then selected from the list of successful attempts and joined to its adjacent chain neighbors to form a complete protein.

Basin string comparisons

To test whether the one-dimensional basin string can capture the three-dimensional X-ray structure, a conventional gapped BLAST-type search⁹ was performed against our basin sequence database derived from the PISCES list.²⁷ Gaps were penalized by 10 points for gap-opening and 0.5 points for each suc-

cessive extension. Alignment scores were calculated using a basin substitution matrix (Table VI).

The substitution matrix tracks the rate at which one basin changes to another over time. To generate the matrix, a set 1533 high-resolution (≤ 1.6 Å), low-sequence similarity ($\leq 5\%$) structures was extracted from a recent (Aug 2011) PISCES list.²⁷ These structures were mapped onto their corresponding basin sequences and all possible consecutive triplets were extracted. The number of different basin pairs occurring in the central basin of each triplet (11 basins, 66 basin pairs) was counted, converted to a $\log_2$ likelihood, and used to form the matrix. In other words, the substitution matrix gives the $\log_2$ odds of converting basin $i \rightarrow$ basin j as reckoned over the set of all triples. The usual simplifying assumption was made that basin $i \rightarrow$ basin j converts at the same rate as basin $j \rightarrow$ basin i .⁷⁵

Conclusions

We have developed a procedure to transform the three-dimensional description of a protein backbone into a linear string by mapping the ϕ, ψ -angles obtained from X-ray coordinates onto 11 discrete basins. Although these basins cover only 3.4% of ϕ, ψ -space, they are sufficient to recover the original structure to a close approximation, as demonstrated. Furthermore, a small number of familiar structural motifs account for approximately 90% of protein structure, and mapping basin sequences onto these iconic structures reveals a dramatic and simplifying degeneracy that is built into the physical chemistry of the protein backbone. The approach described here demonstrates that, contrary to popular opinion, conformational searching is not the major obstacle in solving the protein folding problem.

Acknowledgments

We are indebted to Lauren Porter and Buzz Baldwin for useful comments and to Brian Matthews for critical insight. Support from the National Science Foundation and the Mathers Charitable Foundation is gratefully acknowledged.

References

1. Cornell WD, Cieplak P, Bayly C, Gould IR, Merz KMJ, Ferguson DM, Sepplmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins and nucleic acids. *J Am Chem Soc* 117:5179–5197.
2. MacKerell AD, Jr, Bashford D, Bellott M, Dunbrack RL, Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Gao H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WEI, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.

3. Levinthal C. How to fold gracefully. In: Debrunner P, Tsibris JCM, Münck E, Eds. (1969) *Mössbauer Spectroscopy in Biological Systems*. Urbana: University of Illinois Press, pp. 22–24.
4. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21:167–195.
5. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10–19.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
7. Baldwin RL, Frieden C, Rose GD (2010) Dry molten globule intermediates and the mechanism of protein unfolding. *Proteins* 78:2725–2737.
8. Ramachandran GN, Sasisekharan V (1968) Conformation of polypeptides and proteins. *Adv Prot Chem* 23:283–438.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
10. Gong H, Fleming PJ, Rose GD (2005) Building native protein conformation from highly approximate backbone torsion angles. *Proc Natl Acad Sci USA* 102:16227–16232.
11. Fleming PJ, Gong H, Rose GD (2006) Secondary structure determines protein topology. *Protein Sci* 15:1829–1834.
12. Vasquez M, Scheraga HA (1988) Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data. *J Biomol Struct Dynam* 5:705–755.
13. Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 5:355–373.
14. Przytycka T, Aurora R, Rose GD (1999) A protein taxonomy based on secondary structure. *Nat Struct Biol* 6:672–682.
15. Tsai CJ, Maizel JV, Jr, Nussinov R (2000) Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc Natl Acad Sci USA* 97:12038–12043.
16. Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* 323:297–307.
17. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A (2004) Geometry and symmetry prescript the free-energy landscape of proteins. *Proc Natl Acad Sci USA* 101:7960–7964.
18. Etchebest C, Benros C, Hazout S, de Brevern AG (2005) A structural alphabet for local protein structures: improved prediction methods. *Proteins* 59:810–827.
19. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc Natl Acad Sci USA* 103:2605–2610.
20. Friedberg I, Harder T, Kolodny R, Sitbon E, Li Z, Godzik A (2007) Using an alignment of fragment strings for comparing protein structures. *Bioinformatics* 23:e219–224.
21. Minary P, Levitt M (2008) Probing protein fold space with a simplified model. *J Mol Biol* 375:920–933.
22. Shu N, Hovmöller S, Zhou T (2008) Describing and comparing protein structures using shape strings. *Curr Prot Prot Sci* 9:310–324.
23. Le Q, Pollastri G, Koehl P (2009) Structural alphabets for protein structure classification: a comparison study. *J Mol Biol* 387:431–450.
24. Budowski-Tal I, Nov Y, Kolodny R (2010) FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc Natl Acad Sci USA* 107:3481–3486.
25. Perskie LL, Street TO, Rose GD (2008) Structures, basins, and energies: a deconstruction of the Protein Coil Library. *Protein Sci* 17:1151–1161.
26. Perskie LL, Rose GD (2010) Physical-chemical determinants of coil conformations in globular proteins. *Protein Sci* 19:1127–1136.
27. Wang G, Dunbrack RL, Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19:1589–1591.
28. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
29. Fitzkee NC, Rose GD (2005) Sterics and solvation winnow accessible conformational space for unfolded proteins. *J Mol Biol* 353:873–887.
30. Shakhnovich EI, Finkelstein AV (1989) Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* 28:1667–1680.
31. Hoeltzli SD, Frieden C (1995) Stopped-flow NMR spectroscopy: real-time unfolding studies of 6–19F-tryptophan-labeled *Escherichia coli* dihydrofolate reductase. *Proc Natl Acad Sci USA* 92:9318–9322.
32. Kiefhaber T, Labhardt AM, Baldwin RL (1995) Direct NMR evidence for an intermediate preceding the rate-limiting step in the unfolding of ribonuclease A. *Nature* 375:513–515.
33. Vidugiris GJ, Markley JL, Royer CA (1995) Evidence for a molten globule-like transition state in protein folding from determination of activation volumes. *Biochemistry* 34:4909–4912.
34. Jha SK, Udgaonkar JB (2009) Direct evidence for a dry molten globule intermediate during the unfolding of a small protein. *Proc Natl Acad Sci USA* 106:12289–12294.
35. Reiner A, Henklein P, Kiefhaber T (2010) An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain. *Proc Natl Acad Sci USA* 107:4955–4960.
36. Flory PJ (1969) *Statistical mechanics of chain molecules*. New York: Wiley.
37. Dill KA (1999) Polymer principles and protein folding. *Protein Sci* 8:1166–1180.
38. Pappu RV, Srinivasan R, Rose GD (2000) The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci USA* 97:12565–12570.
39. Doolittle RF (1986) *Of Urfs and Orfs*. California: University Science Books.
40. Aurora R, Rose GD (1998) Helix capping. *Protein Sci* 7:21–38.
41. Panasik N, Jr, Fleming PJ, Rose GD (2005) Hydrogen-bonded turns in proteins: the case for a recount. *Protein Sci* 14:2910–2914.
42. Ginsburg A, Carroll WR (1965) Some specific ion effects on the conformation and thermal stability of ribonuclease. *Biochemistry* 4:2159–2174.
43. Maity H, Maity M, Krishna MM, Mayne L, Englander SW (2005) Protein folding: the stepwise assembly of foldon units. *Proc Natl Acad Sci USA* 102:4741–4746.
44. Tanford C (1968) Protein denaturation. *Adv Prot Chem* 23:121–282.
45. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, Ruczinski I, Doniach S, Plaxco KW (2004) Random-coil behavior and the

- dimensions of chemically unfolded proteins. *Proc Natl Acad Sci USA* 101:12491–12496.
46. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
 47. Fitzkee NC, Rose GD (2004) Reassessing random-coil statistics in unfolded proteins. *Proc Natl Acad Sci USA* 101:12497–12502.
 48. Tiffany ML, Krimm S (1968) Circular dichroism of poly-L-proline in an unordered conformation. *Biopolymers* 6:1767–1770.
 49. Pappu RV, Rose GD (2002) A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci* 11:2437–2455.
 50. Shi Z, Olson CA, Rose GD, Baldwin RL, Kallenbach NR (2002) Polyproline II structure in a sequence of seven alanine residues. *Proc Natl Acad Sci USA* 99:9190–9195.
 51. Avbelj F, Baldwin RL (2003) Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of ϕ . *Proc Natl Acad Sci USA* 100:5742–5747.
 52. Ferreón JC, Hilser VJ (2003) The effect of the polyproline II (PPII) conformation on the denatured state entropy. *Protein Sci* 12:447–457.
 53. Mezei M, Fleming PJ, Srinivasan R, Rose GD (2004) Polyproline II helix is the preferred conformation for unfolded polyalanine in water. *Proteins* 55:502–507.
 54. Grdadolnik J, Mohacek-Grosov V, Baldwin RL, Avbelj F (2011) Populations of the three major backbone conformations in 19 amino acid dipeptides. *Proc Natl Acad Sci USA* 108:1794–1798.
 55. Bolen DW, Rose GD (2008) Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu Rev Biochem* 77:339–362.
 56. Auton M, Bolen DW (2005) Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proc Natl Acad Sci USA* 102:15065–15068.
 57. Auton M, Holthauzen LMF, Bolen DW (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci USA* 104:15317–15322.
 58. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. *Proc Natl Acad Sci USA* 103:16623–16633.
 59. Rosgen J, Pettitt BM, Bolen DW (2005) Protein folding, stability, and solvation structure in osmolyte solutions. *Biophys J* 89:2988–2997.
 60. Street TO, Bolen DW, Rose GD (2006) A molecular mechanism for osmolyte-induced protein stability. *Proc Natl Acad Sci USA* 103:13997–14002.
 61. Cannon JG, Anderson CF, Record MT, Jr (2007) Urea-amide preferential interactions in water: quantitative comparison of model compound data with biopolymer results using water accessible surface areas. *J Phys Chem B* 111:9675–9685.
 62. Hua L, Zhou R, Thirumalai D, Berne BJ (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc Natl Acad Sci USA* 105:16928–16933.
 63. O'Brien EP, Ziv G, Haran G, Brooks BR, Thirumalai D (2008) Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc Natl Acad Sci USA* 105:13403–13408.
 64. Hu CY, Lynch GC, Kokubo H, Pettitt BM (2010) Trimethylamine N-oxide influence on the backbone of proteins: an oligoglycine model. *Proteins* 78:695–704.
 65. Knowles DB, LaCroix AS, Deines NF, Shkel I, Record MT, Jr (2011) Separation of preferential interaction and excluded volume effects on DNA duplex and hairpin stability. *Proc Natl Acad Sci USA* 108:12699–12704.
 66. Teufel DP, Johnson CM, Lum JK, Neuweiler H (2011) Backbone-driven collapse in unfolded protein chains. *J Mol Biol* 409:250–262.
 67. Fleming PJ, Rose GD (2005) Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci* 14:1911–1917.
 68. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res* 34:D204–D206.
 69. Fitzkee NC, Fleming PJ, Rose GD (2005) The Protein Coil Library: a structural database of nonhelix, non-strand fragments derived from the PDB. *Proteins* 58:852–854.
 70. Street TO, Fitzkee NC, Perskie LL, Rose GD (2007) Physical-chemical determinants of turn conformations in globular proteins. *Protein Sci* 16:1720–1727.
 71. Chothia C (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357:543–544.
 72. Govindarajan S, Recabarren R, Goldstein RA (1999) Estimating the total number of protein folds. *Proteins* 35:408–414.
 73. Go N (1983) Theoretical studies of protein folding. *Annu Rev Biophys Bioeng* 12:183–210.
 74. Milner-White EJ, Bell LH, Maccallum PH (1992) Pyrrolidine ring puckering in cis- and trans-proline residues in proteins and polypeptides. Different puckers are favoured in certain situations. *J Mol Biol* 228:725–734.
 75. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.
 76. Forbes C, Evans M, Hastings N, Peacock B (2011) Statistical distributions. Hoboken, NJ: John Wiley & Sons.