# (Probably) All Possible Protein Folds at Low Resolution

Gordon M. Crippen and Vladimir N. Maiorov

College of Pharmacy, University of
Michigan, Ann Arbor, MI 48109, U.S.A.

## Abstract

For decades, a large number of investigators have been sifting the database of experimentally determined three-dimensional protein structures to discover recurring patterns of all types. Now that there are over a thousand such structures available, the natural question is whether we have seen all substantially different protein folds, and if not, how many have yet to be discovered? Answering the question can be broken down into three steps: (1) choose the range and domain for a similarity function, then (2) choose a particular similarity function, and (3) construct a corresponding protein model space that can be searched for dissimilar structures. In our analysis of the problem, we first chose to examine different conformations of the same protein, taking into account only $C^{\alpha}$ atomic coordinates. In particular, we do not compare proteins of different chain lengths on the basis of some kind of gapped alignment. Secondly, we use a measure of conformational similarity based on rigid body superposition that emphasizes overall geometric resemblance, rather than agreement in secondary structure, for example. Third, we employed the discrete cosine transform to construct exhaustive sets of globular self-avoiding $C^{\alpha}$ traces that were all different from each other by a given level. These sets of artificial structures were not too large to explicitly enumerate as long as the level of dissimilarity was high, and the chain flexibility was low. For chains flexible enough to match all experimental structures of 170 residue or less that are not $\beta$-barrels, we find 128 artificial structures, of which 28 resemble nothing in the Protein Data Bank.

# 1: Introduction

The hardest, and indeed the most contentious part about estimating the number of different protein folds is to decide what is to be compared on what basis. There are many reasonable choices here, depending on one's leanings toward molecular biology, polymers, biochemistry, or geometry. Clearly the comparison of protein sequences calls for gapped alignments, because there are many functional and evolutionary similarities that would be obscured without realizing proteins are often mutated by adding or deleting a few residues in the middle of a chain. Matching this is the standard observation that segments of polypeptide chain having conserved sequence over a family of related proteins generally correlate well with conserved relative three-dimensional position. These conserved segments also tend to have conserved, well-defined, secondary structure, such as $\alpha$-helix or $\beta$-strand, and lie in the interior of globular proteins. Alternatively, functionally important residues, for example the active site residues of enzymes, tend to be conserved in sequence and three-dimensional position. These observations have led many workers in this field to assess the similarity of different proteins on the basis of various combinations of gapped sequence alignment, matching of secondary structural elements, matching overall visual similarity, rigid body coordinate superposition with gaps, and agreement in biological role [2, 3, 17, 18, 19, 21]. Their underlying philosophy is that "a protein" is really a whole family of proteins from many different organisms whose properties cluster together more tightly than between families. Their corresponding answer to the number of folds really is an assessment of the number of recognizable functional families.

A more geometric view of the problem is to avoid the whole gapped comparison issue by concentrating on comparing alternative conformations of the same polypeptide chain. The difficulty here is that nature only furnishes us with at most one stable conformation for any particular sequence, so conformational variety can be seen only by comparing pieces of experimental structures or by artificially constructing structures. Thus, Cohen and Sternberg [4] examined the root-mean-square

deviation in $C^{\alpha}$ coordinates after optimal superposition ($D$) for pairs of random, compact, self-avoiding chain conformations and estimated $5 \times 10^{13}$ statistically significantly different conformations for BPTI. See also [12, 10, 11].

Our approach has been to compare different artificial conformations of the same protein on an absolute geometric basis that does not depend on the statistical distributions of such comparisons. This avoids the gapped alignment question. Secondly, not being able to rigorously define "folding topology" or "motif", we measure similarity by rigid body superposition of $C^{\alpha}$ coordinates, which is of course sensitive to overall shape, and will even compare helices with extended strands. Finally, the number of different conformations for a given protein chain length we take to be the number of artificial structures we can generate that all differ by at least some specified similarity cutoff. These sets can also be compared with the experimentally determined protein structures seen in the Protein Data Bank (PDB) [1].

## 2: Methods

### 2.1: Conformational Similarity

The usual measure of difference between two conformations, $A$ and $B$, of the same protein having $n_r$ residues is just $D(A, B)$, the root-mean-square distance between corresponding $C^{\alpha}$ atoms after optimal rigid body superposition, where obviously atom $i$ in $A$ corresponds to atom $i$ in $B$. A better, related measure [15] is

$$\rho(A, B) = \frac{2D(A, B)}{(2R^2(A) + 2R^2(B) - D^2(A, B))^{1/2}} \tag{1}$$

where $R$ is the radius of gyration. An intuitive interpretation of this formula is as follows. After optimal superposition, the $C^{\alpha}$ atoms of structures $A$ and $B$ have coordinates $\mathbf{a}_i, \mathbf{b}_i$, $i = 1, \ldots, n_r$, respectively. Imagine the difference structure, having coordinates $(\mathbf{a}_i - \mathbf{b}_i)$, and the mean structure, with coordinates $(\mathbf{a}_i + \mathbf{b}_i)/2$. Then $\rho(A, B)$ is just the ratio of the radius of gyration of the difference structure to that of the mean structure. When $A$ and $B$ are very similar, the difference structure is small compared to the mean structure, and $\rho$ is nearly zero.

To summarize ref. [15], there are key values of $\rho$ that are independent of any statistical interpretation: $\rho(A,B) = 0$ if and only if they are identical up to a rigid body translation and rotation, $\rho \lessapprox 0.3$ to 0.5 for obvious visual similarity, $\rho = \sqrt{4/5} = 0.894...$ is the smallest value for which the mirror image of $B$ could be more similar to $A$ than $B$ is, $\rho$=1 when the mean and difference structures have equal radii of gyration, $\rho = \sqrt{2} = 1.414...$ is the similarity of any structure to its own mirror image, and $\rho$=2 is the maximal dissimilarity possible. Technically, the 0.894 and 1.414 values require the two structures to be scaled so that all three principal moments of inertia are equal. However, typical compact protein structures are spherical enough (axial ratios of about 2) that the scaled and unscaled $\rho$ differ by only about 5%. Otherwise, these values are independent of overall size of $A$ and $B$, and of their relative sizes.

At least $\rho$ provides a quantitative scale of similarity so that we can set a cutoff value for dissimilarity, $\rho_c$, somewhere between 0 and 2, and enumerate a set of dissimilar structures. Naturally, the set size increases with decreasing $\rho_c$, so we will work around $\rho_c$=1 (i.e., enormously dissimilar) and extrapolate down to levels corresponding to visually recognizable similarity. What we really want to know is how big the space of globular protein structures is, but what we will actually do is produce as large a set of artificial structures as we can such that all differ by at least $\rho_c$. This is roughly analogous to measuring the size of a table by randomly placing as many dinner plates on it as we can, subject to the constraint that no plate covers the center of another. Here, the radius of a plate corresponds to $\rho_c$. An initial strategy of random placement enjoys a high success rate at first, but eventually the remaining bare spots are so small that it is more productive to place the new plate over an old one and then randomly move it away so that it covers the center of no old plate. This search strategy is illustrated in Fig. 1. In order to see how this carries over to proteins, we must first explain how we generate artificial chain conformations.

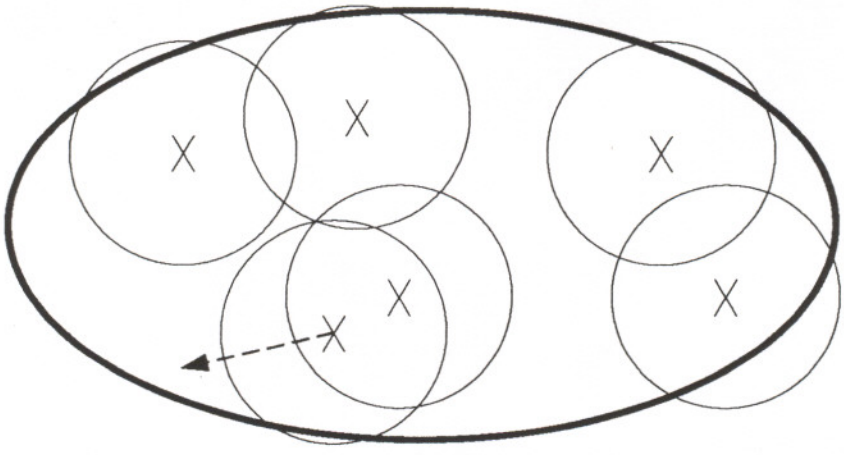

**Figure** 1. Measuring the size of a table by randomly placing plates on it such that the plate centers are not covered.

## 2.2: Discrete Cosine Transform

Suppose we sample a signal $N$ times at uniform intervals, resulting in a sequence of values $x_0, ..., x_{N-1}$. Then the cosine transform coefficients are calculated [20] by

$$\hat{x}_k = \frac{2c_k}{N} \sum_{j=0}^{N-1} x_j \cos\left[\frac{(2j+1)k\pi}{2N}\right] \quad \text{for } k = 0, \ldots, N-1 \tag{2}$$

from which one can return to the precise original signal sample points via the inverse transform

$$x_j = \sum_{k=0}^{N-1} c_k \hat{x}_k \cos\left[\frac{(2j+1)k\pi}{2N}\right] \text{ for } j = 0, \ldots, N-1 \tag{3}$$

where for both transforms

$$c_k = \begin{cases} 1/\sqrt{2}, & k = 0 \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

Note that the $x_j$ and the $\hat{x}_k$ are all real numbers. We represent a path in space by three signals representing the x, y, and z coordinates of a chain of $m$ points, where we choose some $m < n_r$. All such combinations of three signals produce all possible three-dimensional paths, just as all combinations of three real numbers correspond to all possible points in three dimensions. Then to get $n_r$ interpolated

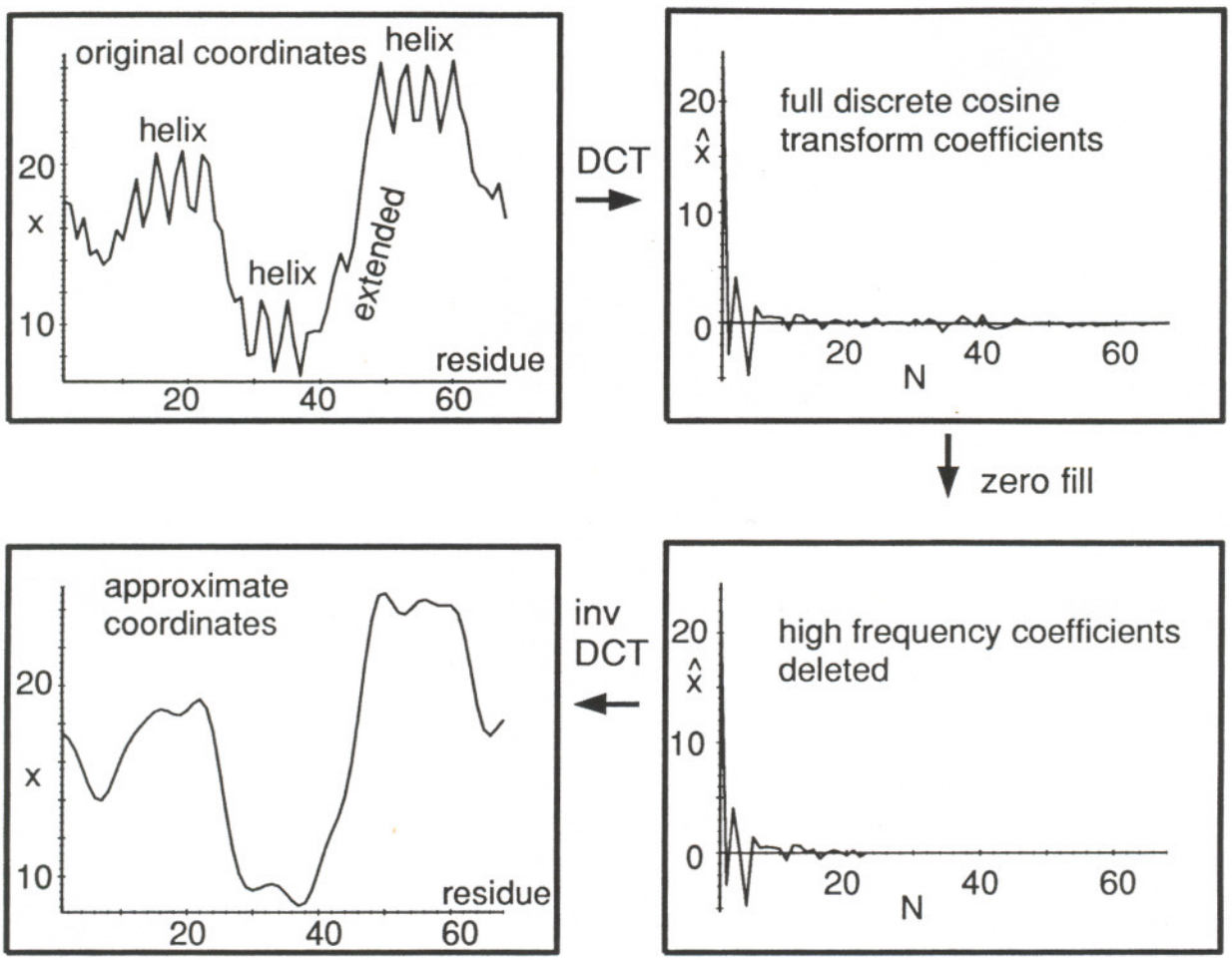


**Figure** 2. Discrete cosine transform filtering of high frequency components of the $C^{\alpha}$ atom $x$ coordinates of protein 1ctf (L7/L12 50S ribosomal protein, 68 residues).

points along the smooth curve, we first determine the DCT coefficients $\hat{x}_k$, $\hat{y}_k$, and $\hat{z}_k$ from equation (2) with $N = m$. Then we backtransform by equation (3) with "zero-filling", that is, $\hat{x}_k = 0$ for $k = m, ..., n_r - 1$ and $N = n_r$. The effect is a high frequency filter, as illustrated in Fig. 2 on a real protein structure. Since we are considering extremely different structures according to a metric that focuses on overall shape, it is not important that our chain representations blur helices into rods.

It turns out that a number of useful features of these artificial structures can be controlled by their DCT coefficients, so we generate them in the transform space

for some small number of terms $N$ and then backtransform them with zero-filling to Cartesian coordinates of the $n_r$ points representing the chain. For example, to get structures having their centroids at the origin, one must set

$$\hat{x}_o = \hat{y}_o = \hat{z}_o = 0 \ . \tag{5}$$

To get a particular radius of gyration $R$ for the final coordinates, requires

$$R^2 = \frac{1}{2} \sum_i \left( \hat{x}_i^2 + \hat{y}_i^2 + \hat{z}_i^2 \right) \ . \tag{6}$$

The RMS deviation in coordinates after optimal rotation is simply proportional to the RMS deviation in transform coefficients after optimal rotation (for $n_r = N$):

$$D(A, B) = (N/2)^{1/2} \hat{D}\left(\hat{A}, \hat{B}\right) \ . \tag{7}$$

Unfortunately, there is not such a simple relation on the coefficients to ensure the structures are self-avoiding.

To generate an exhaustive set of representative artificial structures for comparison with $n_r$–residue proteins, we choose independently with uniform distribution random $\hat{x}_k, \hat{y}_k, \hat{z}_k \in [-1, +1]$ for $k = 1, ..., N-1$, and then scale the coefficients to give the desired radius of gyration. Based on our empirical observation [13] that the most compact globular proteins have radii of gyration

$$R_{min}(n_r) = -1.26 + 2.79 n_r^{1/3} \tag{8}$$

in Å, we use this $R_{min}$. The first random structure is the first representative, and a subsequent random structure is added to the growing set of representatives only if it is sufficiently different ($\rho > \rho_c$) from all the representatives found so far. The typical progress of such a search is shown in Fig. 3, where the total number of structures found increases roughly linearly with $\log_{10} t$, where $t$ is the number of random tries. Curiously, this rate of accumulation is slower than what one would expect if there were simply $n_c$ total representatives that were being chosen at random with equal probability and subsequent replacement.

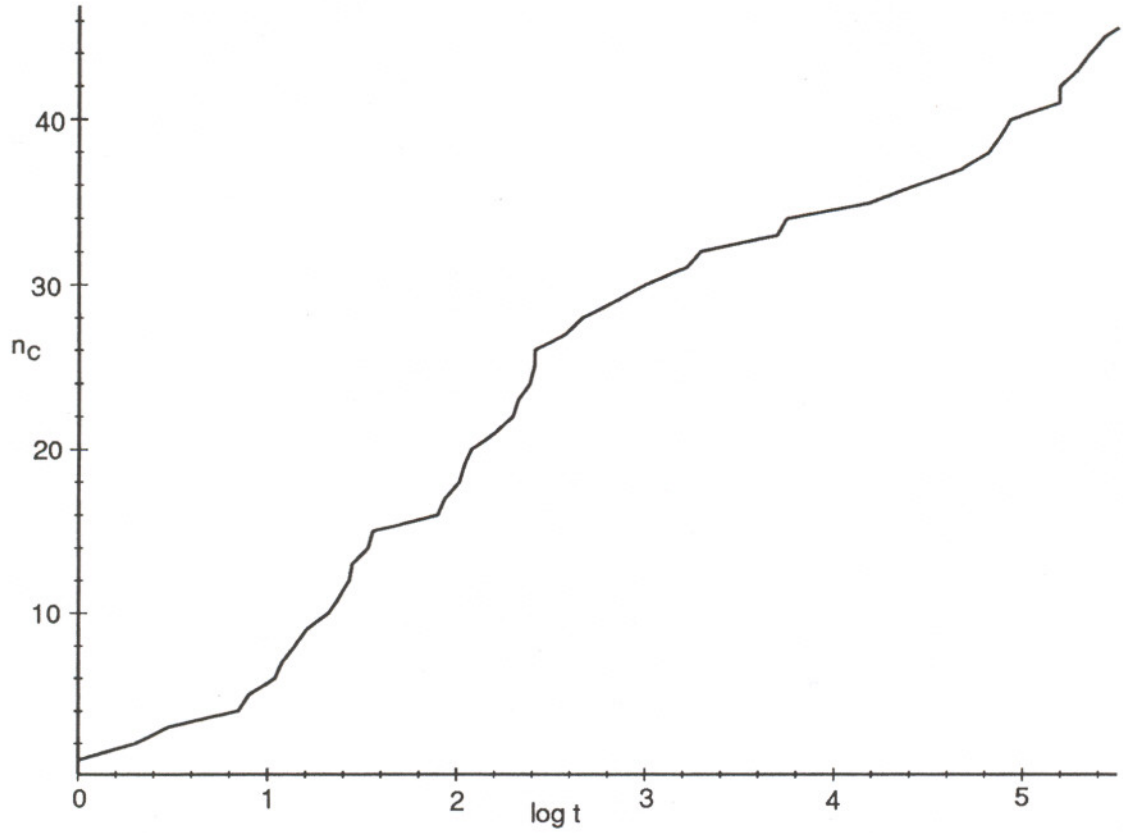


**Figure** 3. Semi-log plot of the number of representatives found, $n_c$, vs. the number of tries, $t$, in the initial random search for $N = 4$ and $\rho_c = 0.5$.

In our experience, the random search for representatives rapidly finds about 70% of them, even when there are many variables, until it becomes unproductive around $t = 10^6$ to $10^7$ iterations.

The second stage of the search is a systematic perturbation of the representatives found by the random search. For each current representative, make a variant by repeatedly incrementing and decrementing each of the $3N - 3$ nonzero coefficients by 0.1, retaining any changes that increase $\rho$ between the variant and the nearest representative. Add the variant to the set of representatives once this $\rho > \rho_c$. Otherwise, discard the variant when no perturbation improves the minimal $\rho$. Quit when no successful variant can be produced from any representative.

The third stage converts the relatively exhaustive set of representatives to self-avoiding representatives. We take as our definition of self-avoiding that the long-range distances between $C^\alpha$ points $d_{ij} > 4.0$ Å for $|i - j| \geq 8$ for the radius

of gyration scaling corresponding to $n_r = 100$. Starting with zero self-avoiding representatives, perturb the coefficients of each non-self-avoiding DCT, keeping any change that improves the score of the corresponding structure. The score is the sum of the worst long-range contact violation plus the $\rho < \rho_c$ violations for previously determined self-avoiding representatives. Perturbation stops when no score reduction can be achieved and the structure is rejected, or when the score reaches zero and the structure is added to the list of self-avoiding representatives.

Note that the final number of self-avoiding representatives, $n_a$, is a function of *two* parameters: $\rho_c$ and $N$. In effect, $N$ controls the flexibility of the artificial chain paths, since $N = 2$ permits only the straight line segment, $N = 3$ allows at most one bend, etc. For great similarity to the natural structures, they should be flexible enough to form $\alpha$-helices, for example, but not so flexible that the interpolated points are scattered at random about the origin. Hao et al. [8] have observed that the polypeptide chain of helical proteins often reverses direction after 2 residues, but $\beta$-sheet proteins tend to reverse direction in about 10 residues. When $\rho_c \approx 1$, the conformational similarity measure essentially views helices and extended strands as vaguely straight rods, so that $N$=10 is appropriate for $n_r = 100$.

## 3: Results

To summarize our recent studies [16], we have enumerated sets of artificial representative structures for several choices of $N$ and $\rho_c$. Consider first some extreme cases. Obviously when $\rho_c$=2, all structures are similar to one another, and the number of conformers, $n_a$, in the set of representatives is just 1. Which conformer it is, is completely irrelevant. At the other extreme, $n_a \to \infty$ as $\rho_c \to 0$, so long as there can be any conformational variation ($N > 2$). When $N = 1$, all conformers are a point at the origin, a case too trivial to consider further. $N = 2$ allows straight line segments running through the origin in various orientations, their lengths set by the desired $R$. For all pairs of these structures, $\rho = 0$ so that $n_a = 1$, regardless of $\rho_c$.

**Table** 1 The numbers of representative self-avoiding conformers, $n_a$, found[a] for $N$ points in the DCT and differing by at least $\rho > \rho_c$.

| | $\rho_c$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $N$ | 1.2 | 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | |
| 5 | 3 | 5 | 6 | 12 | 25 | 41 | | |
| 6 | 3 | 8 | | | | | | |
| 7 | 6 | 17 | | | | | | |
| 8 | 8 | 32 | | | | | | |
| 9 | 10 | 57 | | | | | | |
| 10 | 11 | 84 | | | | | | |
| 11 | 13 | 128 | | | | | | |
| 12 | 17 | | | | | | | |

[a] The number of random trials in the search was $10^7$ for $(N, \rho_c) = (10, 1.0)$, $2 \times 10^6$ for (8, 1.0), and $10^6$ for all the rest.

For $N > 2$ and $0 < \rho_c < 2$, we have to rely on our Monte Carlo estimations. Clearly $n_a$ increases as $\rho_c$ decreases, and increases as $N$ increases, just as smoothing the $C^\alpha$ trace over a narrower window increases the perceived differences between structures. Table 1 shows the number of self-avoiding conformers, $n_a$, as a function of $N$ and $\rho_c$. For $N > 3$, a least-squares fit consistently gives

$$n_a \approx \exp\left[\frac{0.564(N-3)^{1.085}(2-\rho_c)^{1.651}}{\rho_c^{0.969}}\right] \tag{9}$$

with standard deviation of 32.7, independent of the starting values of the four parameters in the curve fitting procedure. The functional form was chosen to enforce the boundary conditions discussed above. If we assume the Monte Carlo estimates are all low by 10%, the revised fit is $n_a \approx \exp\left[\frac{0.628(N-3)^{1.040}(2-\rho_c)^{1.646}}{\rho_c^{0.865}}\right]$.

If $\beta$ proteins tend to have a chain reversal about every 10 residues [8], then even a low-resolution approximation to the backbone of such a 100–residue protein would require $N \approx 10$. We have observed that $\rho_c$=0.4 corresponds to obvious

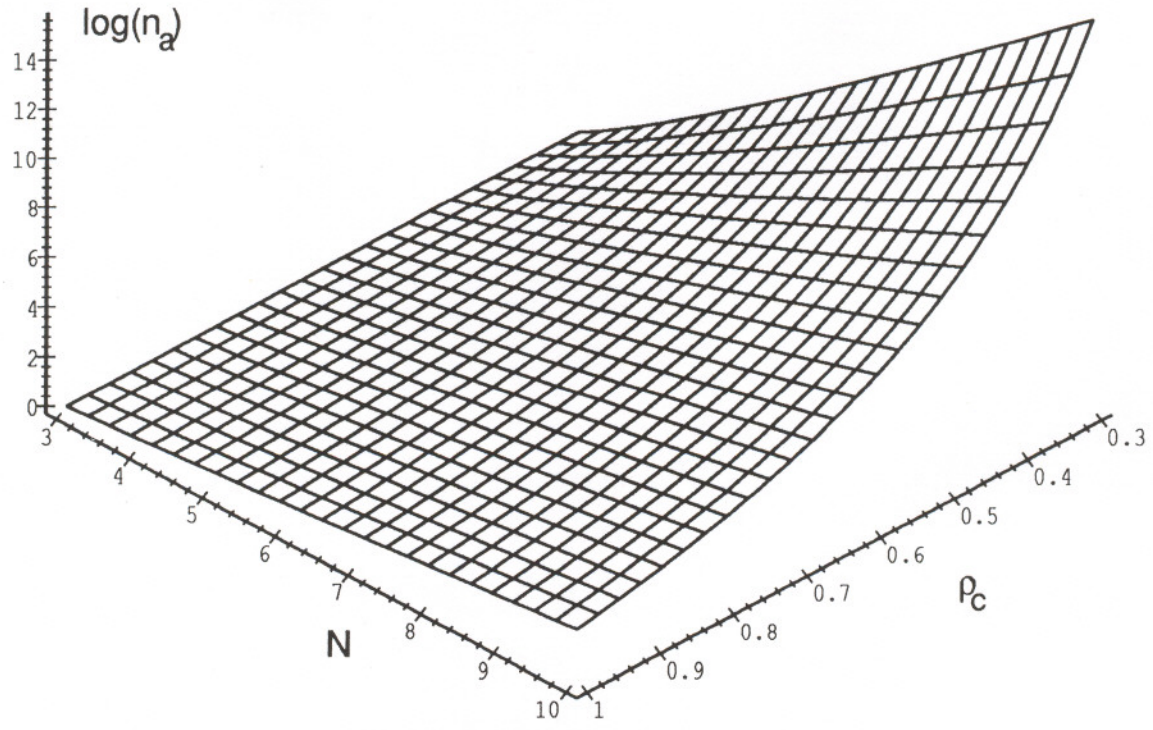


**Figure** 4. The number of significantly different artificial conformations as a function of the $\rho$ cutoff, $\rho_c$, and the number of cosine transform terms used, $N$.

**Table** 2 The estimated number of significantly different protein folds, $n_a$, for single chain globular proteins having $n_r$ residues, as calculated from equation (9), assuming $\rho_c = 0.4$ and $N = n_r/10$.

| $n_r$ | $n_a$ |
|---|---|
| 50 | 600 |
| 100 | $10^{11}$ |
| 200 | $10^{28}$ |
| 300 | $10^{46}$ |

visually recognizable similarity, independent of chain length. Then in order to include all-$\beta$ proteins, we must have $N = n_r/10$, so that equation (9) gives us the estimates shown in Table 2 for the number of protein folds as a function of chain length.

We tested the degree of overlap between real protein structures and our sets of artificial representatives by comparing the 128 self-avoiding artificials for $\rho_c = 1.0$ and $N = 11$ to 1993 PDB entries. These are all the files of the April 1994

release of PDB that contain protein structures having backbone and at least $C^{\beta}$ atomic coordinates, no major breaks in the middle of chains, and were otherwise machine readable. These entries contain a total of 1473 polypeptide chains that have $11 < n_r \leq 170$ and were compact enough that their radius of gyration $R < 1.3R_{min}$ from equation (8). Of course such a set of chains does not evenly sample all experimentally observed structures due to the numerous sets of close homologs. However, our goal here is not to discuss the statistical distribution of structures, but rather to find the full range of our experimental knowledge and compare it to the range of artificial structures. As explained in the Methods section, each one of the artificials is actually a template DCT from which we can produce chain paths having any number of residues and any radius of gyration. In order to compare an artificial representative with an experimental protein structure having an $n_r$–residue chain, we backtransform the DCT with zero-filling to produce a path of $n_r$ points, and then scale the artificial path to match the protein's radius of gyration. That way, $\rho$ is always calculated between two sets of $n_r$ points.

Altogether, 1329 out of 1473 polypeptide chains matched one or more of the artificial representatives to $\rho < 1.0$. (Since there are on the order of $10^{11}$ possible structures with $\rho_c$=0.4, it is not surprising that none of the artificials bore an obvious visual resemblance to any of the natural proteins.) Of the 144 failures, only 69 had optimal matches with $\rho > 1.05$, and of these, only 32 chains had optimal match at a level of $1.1 < \rho < 1.16$: 1spd.A-B, 1sdy.A-D, 1sda.O,Y,B,G, 1cob.A,B, 2sod.O,Y,B,G, 3sod, 1srd.A-D (superoxide dismutases, ca. 150 residues per chain, 8 stranded $\beta$-barrels), 1hlc.A-B, 1slt.A,B (lectins, ca. 130 residues per chain, 11 stranded $\beta$-barrels), 2bfh (human growth factor, 12 strands, all $\beta$), 1opa.A,B, and 1opb.A-D (retinol binding protein II, 135 residues per chain, 10 stranded $\beta$-barrels). As explained above, the many long strands and sharp turns of $\beta$ proteins are relatively hard to fit by our sets of artificial structures, even for $n_r < 171$, and would require artificials with $N > 11$. Otherwise, these 128 artificial representatives can be said to cover the conformation space spanned by nearly all known small and medium proteins.

**Figure** 5. The three most commonly matched self-avoiding artificial structures from the set of representatives having $N = 11$ and $\rho_c = 1.0$. Computer graphics by UCSF MidasPlus [7].

On the other hand, only 100 of the 128 self-avoiding artificials came within $\rho < 1.0$ of one or more of the 1473 proteins. A few of the artificials matched more than 100 proteins, and the most popular one, shown in the upper right of Fig. 5, matched 136 proteins, primarily the numerous T4 lysozyme mutants. The other two matched 123 and 121 proteins, respectively. Even at such a low level of conformational similarity, they are recognizable as cartoons of real proteins with packed helices. That leaves 28 artificials that match nothing in PDB, perhaps because they violate

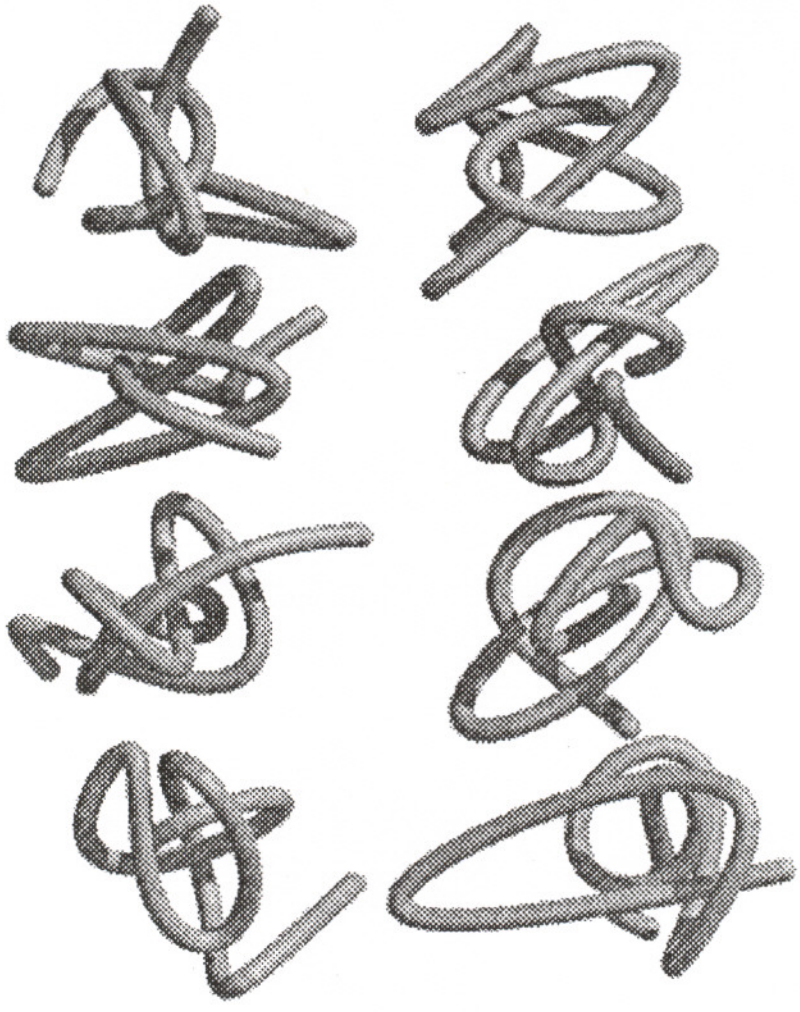

**Figure** 6. Eight self-avoiding artificial structures matching no protein. The threaded loops and complicated entanglement are features unobserved in PDB.

some so far unstated principle of protein folding, or because they correspond to novel protein folds waiting to be discovered. Figs. 6 and 7 show 15 of these 28 that exhibit distinctly alien threading of the ends through broad loops, even allowing for some simplification of the chain crossings, given the low resolution implied by $\rho_c = 1.0$. Some time ago [5] the occurrence of such features was examined in 20 proteins from the PDB. While there were many examples of a part of the chain penetrating a loop formed by another part, these were seen only in polypeptide chains over 200 residues. It would require a fresh survey of the much larger current PDB in order to quantitatively establish whether the entanglements

**Figure** 7. Seven more self-avoiding artificial structures that match no protein and show unnatural entanglements.

seen in these 15 structures are distinctly beyond that observed for chains of 170 residues or fewer. In contrast, the 13 artificial structures shown in Figs. 8 and 9 appear no more snarled than those artificials matching many proteins, yet these 13 match none. It is not unreasonable to hope that some day a novel protein fold will be discovered that matches some of them.

## 4: Conclusions

This work is based on a simple geometric view of the counting all the different protein folds that emphasizes overall spatial similarity and disregards biological

**Figure** 8. Seven self-avoiding artificial structures that are matched by no protein, but appear plausible.

and evolutionary aspects. We are able to use the discrete cosine transform to conveniently generate sets of artificial chain conformations that span the full range of possibilities for self-avoiding, globular structures. How many representative conformers are produced depends on the choice of cutoff between spatial similarity and dissimilarity, and also on the degree of flexibility allowed to the chain. Extrapolating from our explicitly enumerated sets of representative for different levels of flexibility and dissimilarity, we estimate there are $10^{11}$ visually distinct folds for a 100–residue protein (Table 2).

However, the number of different conformers drops to manageable levels when

**Figure** 9. Six more self-avoiding artificial structures that are matched by no protein, but appear plausible.

great spatial dissimilarity is required, leading to a set of 128 "prototype" structures that cover every compact, globular protein of 170 residues or fewer, except for a few $\beta$-barrels. While 100 artificial structures matched one or more proteins, 28 matched none. Some 15 of these can be regarded as suspect because of unusual entanglement, but the other 13 may represent novel protein folds waiting to be discovered. In other words, at this very low level of resolution, 90% of the possible conformational variation has already been seen in PDB, but there are still surprises waiting to be discovered.

## Acknowledgments

This work was supported by a grant from the Ella and Hans Vahlteich Research Endowment Fund, College of Pharmacy, the University of Michigan. We are indebted to all those who deposited their structural data into the Protein Data Bank.

## References

1. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, J. Weng, in *Crystallographic Databases — Information Content, Software Systems, Scientific Applications* F. H. Allen, G. Bergerhoff, R. Sievers, Eds. (Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester, 1987) pp. 107–132.
2. N. I. Aleksandrov, N. Go, *Protein Science* **3**, 866–875 (1994).
3. C. Chothia, *Nature* **357**, 543–544 (1992).
4. F. E. Cohen, M. J. E. Sternberg, *J. Mol. Biol.* **138**, 321–333 (1980).
5. M. L. Connolly, I. D. Kuntz, G. M. Crippen, *Biopolymers* **19**, 1167–1182 (1980).
6. G. M. Crippen, T. F. Havel, *Distance Geometry and Molecular Conformation.* Research Studies Press, Ltd. (Wiley), New York (1988).
7. T. E. Ferrin, C. C. Huang, L. E. Jarvis, R. Langridge, *J. Mol. Graphics* **6**, 13–27 (1988).
8. M.-H. Hao, S. Rackovsky, A. Liwo, M. R. Pincus, H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **89**, 6614–6618 (1992).
9. U. Hobohm, M. Scharf, R. Schneider, C. Sander, *Protein Science* **1**, 409–417 (1992).
10. L. Holm, C. Sander, *J. Mol. Biol.* **233**, 123-138 (1993).
11. L. Holm, C. Sander, *Proteins* **19**, 165-173 (1994).
12. M. Levitt, *J. Mol. Biol.* **170**, 723–764 (1983).
13. V. N. Maiorov, G. M. Crippen, *J. Mol. Biol.* **227**, 876–888 (1992).
14. V. N. Maiorov, G. M. Crippen, *J. Mol. Biol.* **235**, 625–634 (1994).
15. V. N. Maiorov, G. M. Crippen, *Proteins*, in press (1995).

16. V. N. Maiorov, G. M. Crippen, *J. Mol. Biol.*, in press (1995).
17. C. A. Orengo, T. P. Flores, W. R. Taylor, J. M. Thornton, *Prot. Eng.* **6**, 485–500 (1993).
18. C. A. Orengo, D. T. Jones, J. M. Thornton, *Nature* **372**, 631–634 (1994).
19. S. Pascarella, P. Argos, *Protein Eng.* **5**, 121-137 (1992).
20. K. R. Rao, *Discrete Cosine Transform: Algorithms, Advantages, and Applications*, Harcourt Brace Jovanovich, Boston (1990).
21. S. D. Rufino, T. L. Blundell, *J. Comput.-Aided Mol. Design* **8**, 5-27 (1994).
22. D. P. Yee, K. A. Dill, *Protein Science* **2**, 884–899 (1993).