PROTEINS: Structure, Function, and Genetics 35:408–414 (1999)

Estimating the Total Number of Protein Folds

Sridhar Govindarajan,[1] Ruben Recabarren,[1] and Richard A. Goldstein[1,2]*
[1]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*
[2]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **Many seemingly unrelated protein families share common folds. Theoretical models based on structure designability have suggested that a few folds should be very common while many others have low probability. In agreement with the predictions of these models, we show that the distribution of observed protein families over different folds can be modeled with a highly-stretched exponential. Our results suggest that there are approximately 4,000 possible folds, some so unlikely that only approximately 2,000 folds existing among naturally-occurring proteins. Due to the large number of extremely rare folds, constructing a comprehensive database of all existent folds would be difficult. Constructing a database of the most-likely folds representing the vast majority of protein families would be considerably easier. Proteins 1999;35:408–414.** © 1999 Wiley-Liss, Inc.

Key words: protein structures; protein folding; protein families; likelihood estimation; species problem

INTRODUCTION

It has been repeatedly noted that certain folds are greatly over-represented in biological databases. We would expect that proteins of the same protein family that share a clear evolutionary relationship would be structurally similar. But there are numerous examples of seemingly unrelated protein families also sharing the same fold.

Two classes of explanations have arisen that can explain this observation. The most obvious explanation is that there are a limited number of possible protein folds.[1,2] Certain folds are thus seen repeatedly because there are few other options. In contrast, a number of investigators have explained this observation by considering "structure designability." According to this second view, it is much easier to find viable sequences that form into some folds than others. The over-representation of these most-designable folds would be expected. For instance, Govindarajan and Goldstein used analytical and computational models to consider which folds could be most optimized for protein folding, and related the optimal foldability with the number of sequences that would be able to generate that fold.[3,4] Similar results were obtained using lattice models by a number of investigators who counted the number of sequences that would form into various non-degenerate ground-states.[5–7] All of these studies found that there was a broad distribution in the number of sequences that could form into different folds.

We can distinguish between these two different classes of models by considering how the protein families of known structure are distributed among the various folds. If the over-representation of certain folds is due to the small number of possible folds, we would expect to be able to model the observed distribution with the assumption that a small number of folds were possible and roughly equally-likely. In contrast, the designability approaches predict that there should be many rare folds and a few extremely common folds. This again should be detectable in the observed distribution.

The distinction between these two classes of models have consequences for both protein design and protein structure prediction. We are gaining growing abilities to fashion amino acid sequences that form into pre-decided protein structures. If certain folds were over-represented because they were more designable, then these would make attractive targets for such protein engineering attempts. If all folds were equally likely, then the choice of an appropriate target structure can be made based on other criterion. While ab initio protein structure predictions are still currently beyond the limit of feasibility, limited success has been obtained with a simpler problem—recognizing when a given target protein will form into a fold that has been observed previously. This method must fail when the target protein has a novel fold. It is obviously of interest to understand the relative probability of a match with a previously observed fold, both now and in the future. Such protein-structure prediction methods would be most widely applicable if we had a database of all possible protein folds. How possible would it be to assemble such a database? Distinguishing between these two models can also help us to refine our questions regarding basic principles of protein structure. If only a few folds are possible, why is this? What properties characterize "impossible" protein structures? Alternatively, if many protein folds are possible but with varying degrees of likelihood, what determines why nature chooses some folds more often than others? Does this reflect structural or functional constraints, or does it have to do with the nature of biomolecular evolution?

Grant sponsor: National Institutes of Health; Grant number: LM0577; Grant sponsor: National Science Foundation; Grant number: BIR9512955.
Sridhar Govindarajan's present address is Leigh Hall, PO Box 117200, University of Florida, Gainesville, FL 32611-7200.
*Correspondence to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@umich.edu.
Received 29 October; Accepted 4 February 1999

Two additional questions arise: How many different folds are possible? How many should we expect to find in nature? A number of investigators have tried to estimate the total number of existent folds, both observed and currently unobserved.[2,8–13] The estimated number, based on different databases and different sets of assumptions, range from 400 to 8,000. As discussed below, many of the assumptions used in these analyses can be rejected based on the observed data.

The problem of estimating an underlying statistical distribution from a random sample drawn from that distribution has a long history in statistics, and is often referred to as the "species problem."[14–18] Given a random sampling of a number of organisms of different types, what is the best estimate of the total number and distribution of organisms, both observed and unobserved? Two standard approaches have been developed. In the parametric approach, a functional form is assumed for the underlying distribution, and the form is adjusted to match the observed random sample. In the non-parametric approach, statistical properties of the observed samples are used to estimate various characteristics of the underlying distribution directly. While the non-parametric approach involves fewer assumptions, lack of numerical stability often makes the results of less use.

In this paper, we follow the parametric approach and develop a model for the distribution of protein families among the various folds based on a stretched exponential. This allows us to model the relative abundance of rare and common fold types with only two adjustable parameters—the total number of folds, and the "stretchiness" of the exponential. We use a maximum-likelihood analysis to model the overall distribution of sampled structures given the underlying distribution, and compare with the observed distribution of folds as classified by Murzin and co-workers.[19] The fit to the observed data is approximately equal to what we would expect if the model were correct. In contrast, we find that the models used by other investigators are in variance with the observed distribution and can be statistically rejected. We find that the optimal distribution is extremely stretched, as predicted by the computational and analytical models, with a total number of possible folds of approximately 4,000. According to this model, some folds will be common and thus over-represented among observed protein structures. Conversely, many folds will be highly unlikely, so that the total number of folds found in nature would be about 2,000. These results give strong support to the models that focus on protein designability, and suggest that forming a comprehensive database of all existent folds will be difficult. Conversely, because of the extreme range of representation of different folds, a more limited set of folds will be adequate for describing the vast majority of protein families.

METHODS

The Model

In this paper, we use the standard parametric approach to estimate the underlying distribution of protein families among the various folds. (See, for example, Efron and Thisted[16]). We imagine that there is an ensemble of N possible folds $\{S_i\}$, with $1 \geq i \geq N$. The probability that any particular protein family will form into fold S_i is given by λ_i. If all folds are equally likely, then λ_i would be equal to 1/N for all i. In general, we can consider that some folds are more common than others, so that there is a distribution of $\{\lambda_i\}$ values for all of the different folds. Rather than to try to model the discrete values of $\{\lambda_i\}$, we convert this set of values to a continuous distribution characterized by the function $\rho(\lambda_i)$ $(0 \leq \lambda_i \leq 1)$ which describes the probability of any fold having a particular value of λ_i. The distribution is normalized, so that $\rho(\lambda_i)$ satisfies

$$\int_0^1 \rho(\lambda_i) d\lambda_i = 1. \tag{1}$$

Similarly, since all proteins form one of the N possible folds, the average value of λ_i must be equal to $1/N$, or

$$\int_0^1 \lambda_i \rho(\lambda_i) d\lambda_i = 1/N. \tag{2}$$

We are interested in finding an acceptable form for $\rho(\lambda_i)$ that matches the distribution of observed protein structures among the various possible folds. Specifically, there are various protein folds that are observed different numbers of times. The observed experimental data is in the form of a set of values $\{\mu_l\}$, where μ_l is the total number of folds that are observed among exactly l different protein families. μ_0 represents the number of protein folds that have not yet been observed. $L = \Sigma_{l>0}\mu_l = N - \mu_0$ is the total number of folds that have been observed at least once. $M = \Sigma_l l\mu_l$ is the total number of protein families of known structure.

For any given functional form for $\rho(\lambda_i)$, we are interested in varying the adjustable parameters to best fit the observed values of $\{\mu_l\}$. As an example, for the stretched exponential represented in Eq. (6), we would like to find the optimal values of N and β. (Once these values are determined, C and α can be calculated using the normalization conditions expressed in Eq. (1) and (2).) We use the maximum-likelihood approach, and seek to maximize Λ, the log of the probability that a set of $\{\mu_l\}$ values would result with M observations given a particular $\rho(\lambda_i)$ distribution.

For our observed sample of M protein families of known structure, the probability $P_{\lambda_i}(l)$ that exactly l out of the M protein families will form a specific fold with a given value of λ_i (so that $M - l$ proteins do not form into this fold) is given by

$$P_{\lambda_i}(l) = \binom{M}{l} \lambda_i^l (1 - \lambda_i)^{M-l}. \tag{3}$$

As we do not know the value of λ_i for any particular fold, we integrate over the distribution of λ_i values to calculate the total probability $P(l)$ of any particular fold representing l

out of the M native folds

$$P(l) = \binom{M}{l} \int_0^1 \rho(\lambda_i)\lambda_i^l(1-\lambda_i)^{M-l}d\lambda_i. \quad (4)$$

Consider an ordered list of the N possible folds, with the first μ_0 folds each unobserved, the next μ_1 folds observed once, etc. Approximating these as independent events, the probability of such a situation can be written as the product of all of the corresponding $P(l)$ for each of the folds, equal to $\Pi_{l=0}^{\infty} P(l)^{\mu_l}$. We then need to multiply this product by the number of the ways of ordering the N folds consistent with the observed set of $\{\mu_l\}$ values. The probability $P(\{\mu_l\})$ of observing a set of structures described by $\{\mu_l\}$ is then

$$P(\{\mu(l)\}) = \frac{N!}{\mu_0!\mu_1!\ldots}\prod_{l=0}^{\infty} P(l)^{\mu_l} \quad (5)$$

$$= N!\prod_{l=0}^{\infty}\frac{P(l)^{\mu_l}}{\mu_l!}.$$

Note that the products over values of l includes $l = 0$, the folds that are *not* observed. The log-likelihood function Λ is then equal to the log of $P(\{\mu(l)\})$.

We attempt to model the data with expressions for $\rho(\lambda_i)$ that have been preposed and used previously, including a delta-function representing the assumption that all folds are equally-likely,[2] a normal distribution,[10] and a simple exponential.[12] Based on the previously-mentioned designability models, we also use a truncated stretched exponential of the form

$$\rho(\lambda_i) = \begin{cases} C\exp(-\alpha\lambda_i^{\beta}) & | \quad 0 < \lambda_i < 1 \\ 0 & | \quad \text{otherwise} \end{cases} \quad (6)$$

(when $\beta = 1$, $\rho(\lambda_i)$ is a common exponential). For each value of β and N, we use an iterative numerical algorithm to calculate the values of C and α so that equations 1 and 2 are satisfied. We then calculate the log-likelihood by taking the log of $P(\{\mu_l\})$, as computed in Eq. (5). The values of β and N that maximize Λ are found through a standard quadratic interpolation scheme, using a Numerical Algorithms Group (NAG) algorithm.[20]

We must next examine how well the various models fit the data. What log-likelihood value would we expect if the models were correct? Which models can be rejected based on their log-likelihood values? One approach for addressing these types of questions is with parametric bootstrap sampling.[21] In brief, each proposed model is optimized to fit the observed data. For each of these optimized model, 1,000 synthetic datasets are constructed. The corresponding model is then re-optimized for each of these synthetic datasets, and the maximum log-likelihood values computed. We can then compare the log-likelihood value obtained with the observed data with the distribution of log-likelihood values obtained with the synthetic data where we know (by construction) that the model is perfect. We can also use the bootstrap procedure to estimate confidence intervals: the relative range of the parameters obtained through analysis of the synthetic data sets provides an estimate of the relative range of those parameters given the observed data. Furthermore, the ability of the likelihood-maximization method to extract the correct values of the parameters used to generate the synthetic datasets provides independent confirmation of the assumptions used in the analysis, such as the assumption of the independence of the values of l for the different structures.

Once a model is created with a optimized form of $\rho(\lambda_i)$, $\langle\mu_l\rangle$, the expected number of folds that would be found in l different protein families out of a database of M protein families, is given by

$$\langle\mu_l\rangle = NP(l) \quad (7)$$

where $P(l)$ is given by Eq (4). The expected number of observed folds, $\langle L\rangle$ is equal to

$$\langle L\rangle = N - \langle\mu_0\rangle = \sum_l l\langle\mu_l\rangle. \quad (8)$$

The probability that a fold with a given value of λ_i is not represented among the M protein families of known structure is $(1-\lambda_i)^M$. This fold would be found, on average, in λ_i of all protein families. We would expect to find $N\rho(\lambda_i)$ folds with this value of λ_i. Integrating the product of these terms over all values of λ_i gives us the fraction of all protein families with folds that have not yet been observed. f, the fraction of all protein families with folds that have been previously observed, is then one minus this quantity:

$$f = 1 - N\int_0^1 \lambda_i(1-\lambda_i)^M\rho(\lambda_i)d\lambda_i. \quad (9)$$

The Data

A number of investigators have compiled classification systems of protein structures, with somewhat different criterion for deciding whether two proteins have the same fold.[11,19,22] For such pattern-recognition problems it is difficult to find computational approaches that can compare with human judgement. For this reason, we favor the methods with the maximum of expert human intervention and use the most recent SCOP classification (release 1.37) of Murzin and co-workers.[19] According to the SCOP classification, proteins are considered in the same family if they have evident homology detectable based on comparison of their sequences. As described above, proteins from the same family would be expected to share the same fold. We therefore gather statistics of how many protein families form each particular fold. We define two protein families to have the same fold if and only if they are assigned the same fold in the SCOP database. For this study, we omit the proteins classified as multi-domain and membrane proteins. (As these categories contain relatively few examples, comprising approximately 8% of the protein folds, their inclusion or exclusion would not significantly change

TABLE I. Log-Likelihood Λ of the Observed Data for the Optimized Stretched Exponential Model and for Other Models That Have Been Proposed Previously†

Model	Optimal N	Log-likelihood: Observed data	Log-likelihood: Simulated data
Stretched-exponential	3756	−40.7	−41.1 ± 4.2
Exponential	710	−101.8	−26.0 ± 2.2
Gaussian	522	−158.0	−24.5 ± 2.3
Equi-likely	448	−306.9	−19.1 ± 1.7

†Also shown are the log-likelihood values obtained for simulated datasets where the corresponding model was known to be correct by construction. Log-likelihood values substantially lower than those obtained with the corresponding simulated data indicates that the model does not adequately represent the observed data. The only model investigated that cannot be rejected by the data is the stretched exponential. The log-likelihood values for the simulated data are substantially lower for the stretched exponential compared to the other functional forms due to the random sampling of the few examples of highly-likely folds.

our conclusions.) There are $M = 808$ protein families distributed over a total of $L = 375$ observed folds, with 242 folds observed once ($\mu_1 = 242$), fifty-seven folds observed twice ($\mu_2 = 57$, etc.), twenty-seven observed three times, sixteen observed four times, six observed five times, eight observed six times, five observed seven times, five observed eight times, one observed ten times, two observed eleven times, one observed twelve times, one observed fourteen times, one observed nineteen times, one observed twenty times, one observed twenty-six times, and one observed thirty-one times. We of course do not know μ_0, the number of folds that have not yet been observed.

RESULTS

The optimal parameters for the stretched exponential are $N = 3{,}756$ and $\beta = 0.150$, with a log-likelihood $\Lambda = -40.7$. The optimal values of N as well as the corresponding maximum Λ values for the other models are summarized in Table I.

The distribution of values of Λ obtained in the bootstrap analysis with synthetic datasets for each of the proposed models are also summarized in Table I. As is shown, the stretched exponential model is the only model that cannot be rejected based on the observed log-likelihood values. For the synthetic stretched exponential datasets, the median of the distribution of optimal values of N is $= 3577$. The median of the distribution of β values is 0.157. Ninety percent of the fits to the synthetic data have values of N between 2,105 and 8,069, indicating that this is the 90% confidence interval for this parameter. The 68% confidence interval (corresponding to one standard deviation in the case of a normal distribution) is from $N = 2{,}530$ to 5,575. A similar analysis yields a 90% confidence interval for β between 0.106 and 0.227, with a 68% confidence interval of 0.124 and 0.195. The similarity of the parameter values derived from the analysis of the synthetic datasets and the values used to construct the datasets provide support for the approximations used in the analysis, especially the independence of the values of $P(l)$ for each of the different structures.

$\langle \mu_l \rangle$, the expected average number of folds observed l times as calculated in Eq. (7), is plotted in Figure 1 for the various models as well as for various values of N for the stretched exponential. The observed values of $\{\mu_l\}$ in the SCOP database are plotted for comparison. Even with optimal parameters, the stretched exponential model is the only model that provides appreciable probability of observing highly common structures. Although the uncertainty in N is quite large, this causes only modest changes in the values of $\langle \mu_l \rangle$ as well as other quantities calculated below. As the value of N is increased in the range from 2,530 to 5,575, the corresponding value of β that maximizes Λ decreases from 0.187 to 0.124, resulting in an even more highly-stretched exponential. As a result, the additional possible folds tend to be extremely unlikely, with small values of λ_i. These highly unlikely structures do not have a strong impact on measurable quantities.

According to the equally-likely hypothesis, we are observing certain folds regularly because we have already observed most possible folds. We would correspondingly expect the rate of observation of new folds to rapidly decline. In contrast, if the number of repeatedly-observed folds is due to the fact that these folds are more common, that means that less-common folds will continue to be observed. This is shown in Figure 2 which portrays $\langle L \rangle$, the average total number of folds observed, calculated with Eq. (8) as a function of M, the number of protein families of known structure, for the stretched-exponential models with a range of N values as well as the other alternative models.

N represents the number of possible folds, including unlikely ones that have never arisen. Zhang estimates that there are approximately 17,175 protein families for humans.[12] The expected number of folds according to the stretched-exponential model is approximately 1,600 for a sample of this size. Thornton and co-workers estimate that there are a total of 23,100 protein families.[11] The expected number of folds for a sample of this size is approximately 1,740. This number is again relatively insensitive to the uncertainties in N; as N changes from 2,530 to 5,575, the predicted value of L for $M = 23{,}100$ varies from 1,510 to 1,970.

Although according to the stretched-exponential model the vast majority of folds have not been observed, the folds that have been observed represent the folds most common among protein families. As a result f, the proportion of all protein families whose structure is that of a currently-known fold as computed with Eq. (9), is approximately 0.70, significantly higher than the proportion of all folds that have been observed. This quantity as a function of the number of protein families of known structure is shown in Figure 3 for the optimized stretched-exponential and alternative models. f would also represent the probability that the next protein family, if randomly chosen, would have a previously-observed fold. Also shown is the comparison with the proportion of novel folds that have been observed as the database of protein structures has expanded.

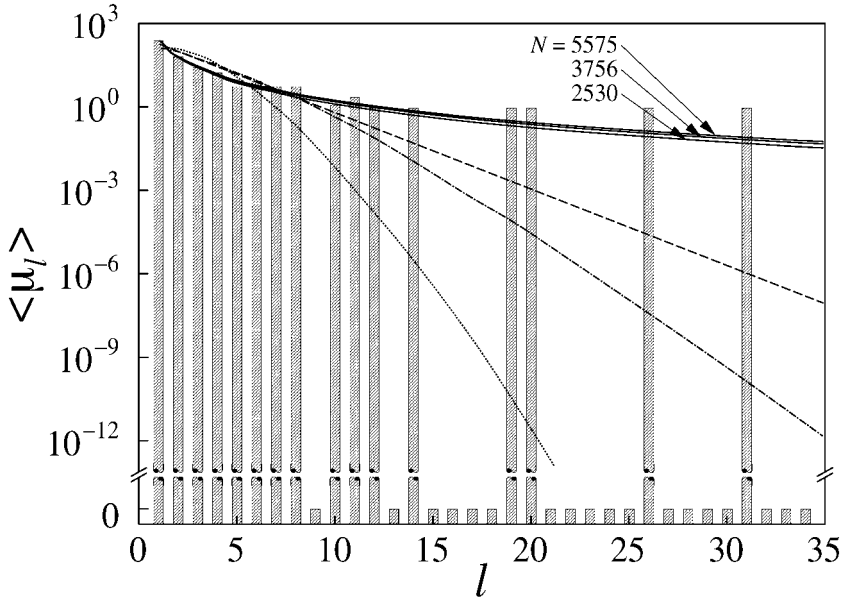

Fig. 1. Comparison of $\langle \mu_l \rangle$, the expected values of μ_l as computed using Eq. (7), for the optimized stretched exponential for three different values of N (———), the exponential model (– – –), the Gaussian model (- · - ·), and the equi-likely model (- - -), compared with the observed distribution of μ_l as catalogued by Murzin et al.[19] For the stretched exponential model, the values of β were optimized for each value of N.

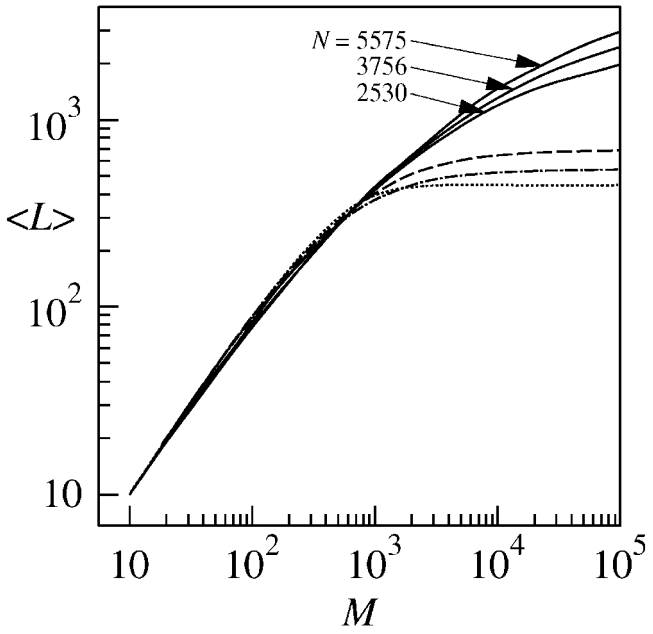

Fig. 2. The expected total number of observed folds, $\langle L \rangle$, computed using Eq. (8) as a function of the number of protein families of known structure, M, for the optimized stretched exponential model for three different values of N (———), compared with the corresponding curves for the exponential model (– – –), the Gaussian model (- · - ·), and the equi-likely model (- - -).

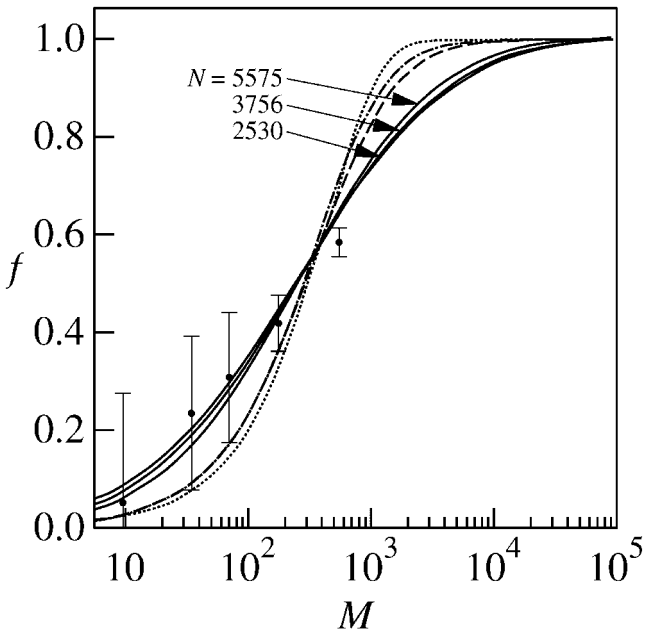

Fig. 3. The expected fraction of all protein families that form a previously-observed fold, f, as calculated using Eq. (9), is plotted as a function of the number of protein families of known structure. The results of the optimized stretched exponential for three different values of N (———) are compared with the results for the exponential model (– – –), the Gaussian model (- · - ·), and the equi-likely model (- - -). Also shown is the observed values of f computed with the SCOP database: measurements were totalled for five blocks of five years. The error bars are computed based on the expectations for random sampling, and represent the 68% confidence limits (thick lines) and the 95% confidence limits (thin lines).

DISCUSSION

Dataset Dependence

A number of different databases exist that categorize proteins of known structure into different folds. We have chosen to use the SCOP database of Murzin and co-workers.[19] Different results are obtained with other databases. For instance, the 3Dee database of Siddiqui and Barton provides a SCOP-like hierarchy that focuses on individual protein domains rather than entire proteins

and protein chains (Siddiqui and Barton, unpublished data). Omitting the peptides and membrane and cell surface proteins, considering the grossest classifications into folds for the proteins whose structures are designated as "defined" and segregating the "undefined" proteins into clusters with a minimum similarity of 1.0 results in a total of 1,171 protein families divided into 484 folds, with 281 folds observed once and one fold observed 42 times. The optimal stretched exponential for this data is with $N = 2828$ folds and a value of β of 0.185, similar to that obtained with the SCOP dataset. The value of Λ is -50.2, while parametric bootstrapping yield a range of Λ of -49.5 ± 4.2, demonstrating the ability of the stretched exponential to represent the observed 3Dee distribution. The CATH database of Thornton and co-workers use a more restrictive criterion for classifying proteins into similar folds. As a result, many sets of proteins that would be considered to have the same fold in the SCOP and 3Dee databases are classified as different folds in the CATH database; the 826 different protein families classified cluster into 590 folds, resulting in an average of only 1.4 families per fold, in contrast to the average of 2.15 in the SCOP database and 2.42 for the 3Dee database. In the CATH database, there are 532 folds found in only single protein families. The large number of such single examples results in an extremely stretched exponential with almost a singularity in $\rho(\lambda_i)$ at small λ_i. Not surprisingly, with distinctions made between highly similar folds, the estimated number of possible folds becomes extremely large and uncertain.

Comparison With Previous Models

A number of previous investigators have tried to estimate the total number of possible folds using different sets of assumptions. One assumption made in some previous estimates of N is to assume that the ratio of observed folds to the number of protein families of known structure will remain constant, so that the relationship between M and L is linear. For instance, Orengo et al.[11] observed a data-set of 130 seemingly non-homologous protein families folding into a total of 80 different folds, with 71 folds defined by only one protein family and the remaining 59 families divided up among the other 9 folds.[11] They estimate that this represents only one-third of the non-homologous "super families" among the 3% of the current sequence databases. Assuming that new folds are obtained at the same rate as the remaining database of sequences is analyzed, this suggests that the total number of folds is $80 \times 33 \times 3 = 7{,}920$. This assumption is identical to assuming that f, the probability that a new protein family will represent a novel fold, is a constant over time. The correctness of this assumption is highly unlikely given the indications that certain folds are much more likely than others; given the preferential early observation of the more frequent folds, the percentage of observed folds that are novel should continuously decline. This assumption also seems to be contradicted by the measured values of f presented in Figure 3.

Zhang recently made an estimate that the total number of folds is less than 5,200 based on a consideration of the "degeneracy" $\overline{d}(t)$, defined as the total number of observed protein families divided by the total number of observed folds, at any particular time t.[12] For the earlier version of the SCOP dataset available at the time, $\overline{d}(t) = 1.955$. Zhang also considered the quantity $a(t)$, defined by the rate of observance of new structures divided by the rate of observance of new folds. This quantity is also an explicit function of time. Zhang estimated the value of $a(t)$ to be approximately 0.3. As new structures are solved, the degeneracy will continue to increase until it is equal to $1/a(t)$. $a(t)$ will tend to decrease as more and more of the common folds are observed. This means that the long-time asymptotic value of $\overline{d}(t)$, notated by $\overline{d}(t^*)$, has to be greater than or equal to the current value of $1/a(t) = 3.3$. Assuming that the total number of protein families in the human data-set is equal to 17,175, Zhang computed that the maximum number of folds is equal to the total number of folds divided by the asymptotic value of $\overline{d}(t)$, or $17{,}175/3.3 = 5{,}200$. In actuality, $\overline{d}(t)$ may not reach its asymptotic value by the time the whole human data-set is available, so a more appropriate limit for $\overline{d}(t)$ is its current value. With the database used in this paper, $\overline{d}(t)$ is currently $808/375 = 2.15$. Using this value forces a revised estimate of the maximum number of folds to 7,988.

Alexandrov and Gō estimated the total number of protein folds at around 6,700 by assuming that the underlying distribution of $\rho(\lambda_i)$ was a normal distribution.[10] Conversely, Wang obtained an estimate of only about 400 folds by considering that all possible folds are equally likely.[2] Both of these models can be rejected based on their inadequate representation of the observed data, as shown in Table I. Wang also considered the possibility that these superfolds were drawn from a different distribution, an ad hoc assumption that violates the spirit of Occum's razor.

In a more recent paper, Zhang and DeLisi use a different model to estimate the number of protein folds.[13] In this model, Γ_x, the probability that a fold is formed by x families is given by

$$\Gamma_x = \left(1 - \frac{N}{M}\right)^{x-1} \frac{N}{M}. \qquad (10)$$

While they show that their model is not obviously inadequate to explain the distribution of relatively rare structures (μ_l for $1 \le l \le 8$), they do not examine how well their model explains the number of more likely structures. We note that according to their formulation and their estimate of the total number of possible folds, the probabilities of having one fold observed twenty-six and thirty-one times in a sample of 808 protein families is approximately 2.9×10^{-5} and 1.3×10^{-6}, respectively, making it highly unlikely that the proposed model is consistent with the observed data.

This approach assumes that proteins of known structure represent a random sampling of real proteins found in nature. This assumption is unlikely to be true. It may be

that certain folds are more likely to solve by crystallography or NMR spectroscopy. It also may be that certain folds are more likely to be found in the organisms under study. Finally, certain families may be represented by multiple examples that share an undetectable evolutionary relationship. One indirect method to identify such biases is through looking at the time evolution of *f*, the number of new protein families that have pre-observed folds. If the sampling biases were constant, we would expect the values of *f* at different times to follow the theoretical curves shown in Figure 3. With technological innovations such as the rise in the use of NMR spectroscopy, we would expect that sampling biases would change, resulting in deviations in the value of *f*. In general, as shown in Figure 3, the observed values of *f* track the theoretical predictions based on the stretched-exponential model, with a possibly-significant deviation at the most recent datapoint. Such a deviation hints that there may be time-dependent biases, so that there are more folds observed multiple times than would be expected with a random sampling. This bias would cause us to underestimate the true value of *N*.

CONCLUSION

According to some of the designability models, there is a wide range in probabilities of different folds among biological proteins, with many rare folds and a few extremely abundant folds.[3,4,7] These models are supported by the success of the stretched exponential model in modeling the observed distribution of protein folds. Conversely, the models that try to explain this distribution based on a relatively small number of equally-likely folds can be statistically rejected.

There has been interest in generating a comprehensive set of all protein folds. The results presented here indicate that this will be a significant challenge, given the large number of highly-unlikely folds. Expansion of the database to include more protein families from, for example, different organisms, will keep increasing the number of relevant folds, as shown in Figure 2. On the other hand, it will be significantly easier to generate a less-than-comprehensive list that has the dominant protein folds. For instance, according to the most optimized model, although we have only observed 375 out of the 3,756 possible folds, this set still includes the structures of 70% of all protein families, even if there were an infinite number of such families. A catalog of only 930 folds would encompass approximately 90% such families.

ACKNOWLEDGMENTS

We would like to thank Edward Rothman, Charles Lawrence, and Brett Larget for helpful discussions.

REFERENCES

1. Crippen GM, Maiorov VN. How many protein folding motifs are there? J. Mol. Biol. 1995;252:144–151.
2. Wang ZX. How many fold types of protein are there in nature? Proteins 1996;26:186–191.
3. Govindarajan S, Goldstein RA. Searching for foldable protein structures using optimized energy functions. Biopolymers 1995;36: 43–51.
4. Govindarajan S, Goldstein RA. Why are some protein structures so common? Proc Natl Acad Sci USA 1996;93:3341–3345.
5. Lipman DJ, Wilbur WJ. Modelling neutral and selective evolution of protein folding. Proc R Soc Lond [Biol] 1991;245:7–11.
6. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–669.
7. Bornberg-Bauer E. How are model protein structures distributed in sequence space? Biophys J 1997;73:2393–2403.
8. Chothia C. One thousand protein families for the molecular biologist. Nature 1992;357:543–544.
9. Blundell T, Johnson MS. Catching a common fold. Protein Sci 1993;2:877–883.
10. Alexandrov NN, Gō N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. Protein Sci 1994;3:866–875.
11. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
12. Zhang CT. Relations of the number of protein sequences, families, and folds. Protein Eng 1997;10:757–761.
13. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284:1301–1305.
14. Fisher RA. A theoretical distribution for the apparent abundance of different species. J Anim Ecology 1943;12:54–58.
15. Good IJ. The population frequencies of species and the estimation of population parameters. Biometrika 1953;40:237–264.
16. Efron B, Thisted R. Estimating the number of unseen species: how many words did Shakespeare know? Biometrika 1976;63:435–447.
17. Hill BM. Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. J Am Stat Assoc 1979;74:668–673.
18. Keener R, Rothman E, Starr N. Distribution on partitions. Ann Stat 1987;15:1466–1481.
19. Murzin AG, Brenner SE, Hubbard TJP, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
20. Numerical Algorithms Group Ltd., University of Oxford.
21. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall; 1993. 436 p.
22. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure families with common folding motifs. Protein Sci 1993;1:1691–1698.