PROTEINS: Structure, Function, and Genetics 46:61–71 (2002)

A Unifold, Mesofold, and Superfold Model of Protein Fold Use

Andrew F. W. Coulson[1*] and John Moult[2]
[1] *Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh, Scotland*
[2] *Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*

ABSTRACT **As more and more protein structures are determined, there is increasing interest in the question of how many different folds have been used in biology. The history of the rate of discovery of new folds and the distribution of sequence families among known folds provide a means of estimating the underlying distribution of fold use. Previous models exploiting these data have led to rather different conclusions on the total number of folds. We present a new model, based on the notion that the folds used in biology fall naturally into three classes: unifolds, that is, folds found only in a single narrow sequence family; mesofolds, found in an intermediate number of families; and the previously noted superfolds, found in many protein families. We show that this model fits the available data well and has predicted the development of SCOP over the past 2 years. The principle implications of the model are as follows: (1) The vast majority of folds will be found in only a single sequence family; (2) the total number of folds is at least 10,000; and (3) 80% of sequence families have one of about 400 folds, most of which are already known. Proteins 2002;46:61–71.**

Key words: protein folds; sequence families; protein evolution; unifolds; mesofolds; superfolds

INTRODUCTION

Knowledge of the complete genome sequences of a number of organisms has led naturally to the idea of completion in other areas of molecular biology. In particular, the notion of completeness of the set of protein structures has arisen. The experimental pursuit of that goal is often termed "structural genomics," and a central question in planning its execution is how many different structures or "folds" have been used in biology. We investigate one method of estimating that quantity, based on the experimental sampling of structure space that has occurred so far. There have been several previous estimates, and we build on that work. In addition to estimating the number of different structures, we also consider the distribution of structure use in sequence space, and the insight this provides into the nature of the evolutionary processes that gave rise to the set of proteins seen today.

The first estimates of the number of folds were based on the apparent number of evolutionarily independent sequence families. Zuckerkandl[1] suggested that the number of protein classes was "…perhaps considerably less… than 1000," and Barker and Dayhoff[2] also estimated that there are approximately 1000 such sequence "superfamilies." Because overall structure is believed to be conserved in evolution, at least out to the limits of detectable sequence relationships, it follows that there would be not more than a thousand independent different folds. As more folds were determined experimentally, it became possible to use structure rather than sequence as the basis for estimating the total number. Chothia[3] observed that the fraction of sequences in genomes that were clearly related to some sequence already in the Swissprot databank was approximately independent of the organism considered, at about one third, and that in turn, about one fourth of the Swissprot sequences were clearly related to one of the 83 then-known folds. From these relationships, he estimated that there are approximately 1000 evolutionarily independent families.

Chothia's model established the principle of using the record of experimentally determined structures to estimate the number of folds, and more sophisticated treatments have followed. All of the methods rely on the association of each fold with one or more sequence families and use statistical models to derive the expected current distribution of fold use.

A convenient and popular catalog of sequence families, superfamilies, and folds is provided by the SCOP database.[4] The Chothia argument implicitly assumes that all folds are approximately equally used in sequence space. Analysis[5] of the protein databank showed that in fact some folds are found in many sequence families and others, so far, in only one. Zhang and DeLisi[6] pointed out that a more reasonable assumption is that all folds were equally likely to be adopted by newly evolving sequence families. Such a process would produce a nonuniform distribution of fold use, with the exact form dependent on the total number of sequence families and the total number of folds. They explored this model, using a simple random sampling process, with no parameters other than the total number of sequence families, to estimate the

Grant sponsor: National Institutes of Health; Grant number: P01 GM5790.

*Correspondence to: Andrew F. W. Coulson, Institute of Cell and Molecular Biology, University of Edinburgh, Swann Building, King's Buildings, Edinburgh EH9 3JR, Scotland. E-mail: a.coulson@ed.ac.uk

15 December 2001; 2 August 2001

DOI 10.1002/prot.10011

distribution of fold use expected and the total number of underlying different folds. An impressive verification of the model was its broad compatibility with the then-current SCOP fold/"sequence family" distribution. Surprisingly, the most likely total number of folds was found to be only about 700. More recently, Govindarajan et al.[7]noted that the Zhang and DeLisi model does not account for the existence of superfolds,[5] that is, the set of approximately nine folds that have been seen to represent an anomalously large number of sequence families. These authors found that studies of simple lattice models suggest some folds are able to accept many more different sets of amino acid sequences than others, such that a stretched exponential would be appropriate for describing the distribution of fold use in biology.[8] The two parameters of this distribution were adjusted to fit the current observed distribution of fold use. In contrast to Zhang and DeLisi,[6] they conclude that many folds are rare in biology, producing an estimate of at least 4000 different possible folds. There have been a number of other estimates of the total number of folds, most recently by Wolf et al.[9]

Figure 1 shows two comparisons of predictions from the Zhang and DeLisi and Govindarajan et al. models with the data in SCOP. Figure 1(A) shows the number of folds that were observed in different numbers of sequence families, according to SCOP, release 1.37. Most folds are only found in a single sequence family (257 from a total of 394), and there is a rapid fall off in the number of observations with increasing sequence family count. The distribution has a tail (not shown) for the superfolds, with the largest number of families for a fold at 31 for the TIM barrel fold. The Zhang and DeLisi model underestimates the number of folds that have only been seen associated with a single sequence family and overestimates the numbers that have been seen in two, three, four, and five families. The Govindarajan et al. model overestimates the number of folds seen in one or in two sequence families.

A second comparison with experiment is provided by the history of the appearance of new folds in the PDB as a fraction of new protein families. Figure 1(B) shows these data. As Govindarajan et al. point out, there are reasons to be cautious about attempts to fit this curve too closely. In the early days of experimental structure determination, the choice of proteins was limited to those easily obtained, whereas in recent years, improved molecular biology and structure determination techniques have provided access to a high fraction of soluble proteins. A second consideration is that the definition of a sequence family within SCOP may have changed over time. However, most data have been added since the experimental techniques matured, and Figure 1(B) is based on a single release of data. Thus, although it may be unreasonable to demand an exact fit, it is worthwhile asking whether the observed data are broadly consistent with the proposed models. It is apparent that the Zhang and DeLisi model underestimates the fraction of new folds that should be seen in recent times, whereas Govindarajan et al. consistently overestimate it.

Is there a simple model that fits the complete observed fold-use histogram, follows the history of new fold discovery reasonably, and is based on a simple and reasonable underlying model of fold use in biology? Below we describe one such model and discuss its implications.

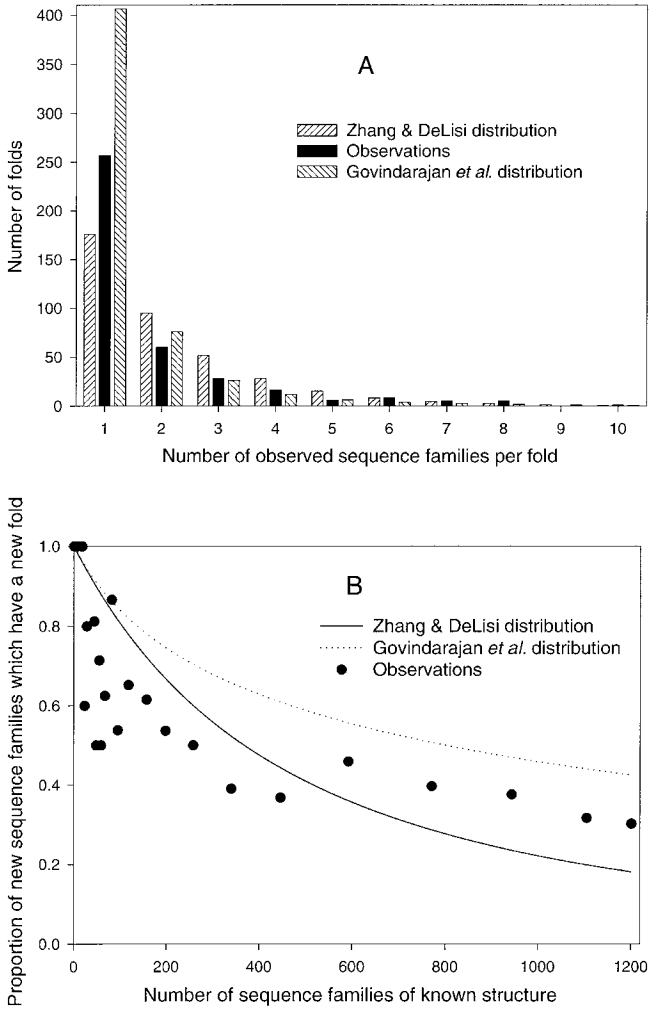

Fig. 1. Fit of earlier models with SCOP 1.37. (**A**) Histogram of the number of folds found in 1,2,3... sequence families in release 1.37 of SCOP,[4] compared with the values expected according to the models of Zhang and DeLisi[6] and of Govindarajan et al.[7] The superfold region is not included. The Zhang and DeLisi distribution underestimates the number of folds so far seen in a single sequence family, whereas that of Govindarajan et al. overestimates this quantity. (**B**) History of accumulation of new folds in the Protein Data Bank. Points (one per year) show the fraction of new sequence families classified as representing new folds (according to the SCOP definitions), as a function of the number of sequence families of known structure. The solid line shows the record expected according to the Zhang and DeLisi model, and the dashed line that according to the model of Govindarajan et al. model. The ratio of new folds to new families has been approximately constant since 1993, a period during which about two thirds of the families currently represented in PDB were added. Both models appear to underestimate the initial rapid drop in the ratio of new folds to new families. For the period since 1993, the Zhang and DeLisi distribution underestimates the ratio (giving fewer new folds than observed), whereas the distribution of Govindarajan et al. overestimates it.

THE MODEL

The key concept behind the model is that fold space divides naturally into three different zones: a zone of

unifolds, which are found in only a single SCOP family; a zone of mesofolds, where fold use follows the Zhang and DeLisi model; and the zone of superfolds,[5] which are seen much more frequently than the Zhang and DeLisi model can support.

Sequence families are therefore divided into three fractions:

$$f_U + f_M + f_S = 1$$

where f_U is the fraction of families represented by unifolds (folds with only one sequence family), f_M is the fraction of sequence families represented by mesofolds (obeying a Zhang and DeLisi-like model of fold use), and f_S is the fraction of sequence families represented by superfolds. The model has four parameters: f_U, f_M , and f_S, and the total number of sequence families, R. R is made up of R_U unifold families, R_M families with mesofold structures, and R_S families belonging to superfolds:

$$R_U + R_M + R_S = R$$

and the total number of folds, N, is the sum of the number of unifolds, N_U, the number of mesofolds, N_M, and the number of superfolds, N_S:

$$N = N_U + N_M + N_S$$

Unifolds

The number of unifold families, R_U is equal to the number of unifolds, N_U, and so

$$f_U = R_U/R = N_U/R$$

Mesofolds

The N_M mesofolds are distributed over the f_M families using the Zhang and DeLisi random fold sampling model. That is, each family is randomly assigned, with equal probability, to one of the N_M folds. Then, following Zhang and DeLisi, the number of mesofolds having i families each, N_{Mi}, is given by

$$N_{Mi} = N_M \cdot (1 - p_b)^{i-1} \cdot p_b,$$

where $p_b = N_M/R_M$, and $R_M = R \cdot (1 - f_U - f_S)$.

Superfolds

Orengo et al.[5] introduced the term superfold to refer to folds that are observed to occur in many apparently evolutionarily independent families. These folds are also associated with an abnormally large number of sequence families, and as a result, clearly do not fit the Zhang and DeLisi model. We therefore treat them separately. The exact number of folds in this category is uncertain, but inclusion of the nine original superfolds produces satisfactory results. So we assume that $N_S = 9$ and that superfolds have been so thoroughly sampled that the currently observed prevalence corresponds to the underlying distribution. Then the number of families, R_{Si}, represented by superfold i is given by

$$R_{Si} = f_{Si} \cdot R \text{ and } f_S = \sum_{i=1,9} f_{Si}$$

where f_{Si} is the fraction of all families in the current distribution that belong to superfold i. Assuming the only superfolds are the nine folds with the largest number of sequence families and that superfold sequence families have been adequately sampled,

$$f_S \cong \frac{R'_S}{R'}$$

where R' is the total number of sequence families whose structure is known, and R'_s is the number of sequence families currently seen in superfolds. For SCOP 1.37, f_S = 0.18.

Number of Sequence Families

As noted by Zhang and DeLisi,[6] the total number of sequence families, R, has little effect on the estimated number of mesofolds. However, in our model, the value of R does affect the estimated number of unifolds. We have used the value of 23100[5] for this parameter, but also explore reasonable limits on its value of 10,000 and 50,000.

Fitted Parameters

The two remaining parameters are f_U and f_M. An analytical solution for values of these quantities may be obtained using a fit to the total number of sequence families and folds so far observed (see Methods).

More refined estimates of the parameters of the model were obtained by adjusting the preliminary values to fit the histogram of the currently observed distribution of fold use. N_M and f_U were systematically adjusted to minimize the sum of the squares of the differences between the observed and predicted histograms (excluding the superfold regions). f_S was then adjusted so that the model predicted the correct number of currently observed superfold families. Finally, it was shown that the values of N_M and f_U still minimized the sum of the squares of the histogram residuals (all combinations of 1% changes in the parameters produced an increase in this measure).

METHODS

Fold, Family, and Superfamily Definitions

The terms "fold" and "family" used in this article refer to the definitions in SCOP.[4] This database provides a hierarchical classification of the protein domains in PDB entries. The analysis used releases 1.37 and 1.48. The numbers and definitions of SCOP Classes changed somewhat between these releases, but our analysis used only Classes 1–5 and 7 (globular nonmembrane proteins including small proteins), for which the definitions were unaltered. SCOP has five levels; within each class there are a number of folds, each containing one or more superfamilies. Superfamilies contain one or more families. Only the fold and family levels are used in the present study.

SCOP data were downloaded in flat file form. The SCOP classification numbers were used to identify members of the same fold and family, respectively. The date of first

deposition on the corresponding PDB files was used to identify the year a structure was deposited. A small fraction of PDB files are replacements of earlier entries and so carry an inappropriate date for our purposes. This introduces small errors into the counts used, too small to affect the conclusions. Histogram data for the numbers of families per fold were generated by a Maple[15] program that counted the number of labels for lower-level divisions within each fold class.

Explicit Forms of the Model Distributions

Suppose that there are N folds in the population from which those currently in SCOP are drawn, and that each fold has a fixed number of sequence families. Suppose n_i is the number of folds which have i sequence families and that w_i are weights such that $n_i = N \cdot w_i$. The total number of families is $\Sigma_i i \cdot n_i = N\Sigma_i i \cdot w_i = N \cdot D$, where D is the mean number of families per fold.

Zhang and DeLisi distribution

Using the symbols defined above, the distribution proposed by Zhang and DeLisi[6] has the form:

$$w_i = \frac{(D-1)^{i-1}}{D^i}$$

The authors show that $D = R \cdot \dfrac{(R' - (1 - R'/R)N')}{R'N'}$, where N' is the number of folds for which there is currently at least one structure known. Assuming $R = 23100$[5], this equation applied to SCOP 1.37 (including 395 folds and 833 families) implies $N = 728$, and from SCOP 1.48 (including 509 folds and 1193 families) $N = 855$. Zhang and DeLisi give $N = 687$, based on the June 1997 release of SCOP (361 folds and 736 sequence families).

Distribution of Govindarajan et al.

The optimum form of the model of Govindarajan et al.[7], derived from SCOP (375 folds and 808 families), was found by these authors to be (using the terms defined here) $M = 3756$ and $w_i = C \cdot \exp[-\alpha \cdot (i/R)^{0.15}]$, where C and α are constants. For $R = 23{,}100$, we estimated the values of C and α as 1966 and 38.15, respectively, by numerical iteration using the constraints $\Sigma_{i=1}^{\infty} w_i = 1$ and $\Sigma_{i=1}^{\infty} i \cdot M \cdot w_i = R$.

Three-zone model proposed here

For unifolds, $w_1 = 1$; $w_i = 0$ for $i \neq 1$. For mesofolds,

$$w_i = \frac{(D'-1)^{i-1}}{D'^i},$$

where D' is the mean number of sequence families per mesofold. For the nine superfolds, the relative weights w_j, $j = 1 \ldots 9$ are taken to be equal to the observed relative proportions of the superfolds in the appropriate release of SCOP. For SCOP 1.37, these values were [.227, .177, .128, .092, .085, .078, .071, .071, .071] For SCOP 1.48 the values were [.194, .175, .152, .114, .085, .081, .071, .066, .062].

Relative Rates of Discovery of Folds and Families

We give the symbol q to the proportion of structures that represent new sequence families, which also represent new structural folds. We first derive a general expression for q and then apply it to each of the three model distributions considered here.

Suppose that at some stage in the progress of structural biology, there is a high-resolution structure for at least one example of each of R' families. This represents a fraction, $R'/N \cdot D$ of all families. If these have been drawn effectively at random from the whole population, the same fraction has been seen of any subset of families. Consider the $i \cdot n_i$ families of the folds that have i families per fold. The probability that any one of these families has not yet been chosen for X-ray crystallographic study is $(1 - R'/N \cdot D)$. So the proportion of these folds for which no families have yet been seen (i.e., the fraction of these folds that is currently unknown) is $(1 - R'/N \cdot D)^i$. Hence, the total number of sequences for folds of which no example is yet known is $\Sigma_i i \cdot N \cdot w_i \cdot (1 - R'/N \cdot D)^i$, and the total number of unknown sequence families is $(N \cdot D - R')$. Therefore, at this stage of discovery, the proportion of new families that represent new folds (i.e., the relative rates of discovery of new families and new folds), which we call q, is given by

$$q = \frac{\sum_i i.N.w_i.(1 - R'/N.D)^i}{(N.D - R')} \qquad (1)$$

This may be simplified to

$$q = \frac{\sum_i i.w_i.(1 - R'/N.D)^{i-1}}{\sum_i i.w_i},$$

and this form makes it clear that q is the weighted average of a power series.

Uniform distribution

for $w_i = 1$, for $i = D$; $w_i = 0$, for $i \neq D$.
$q = (1 - R'/N \cdot D)^{D-1}$.

Zhang and DeLisi distribution

$$q = \frac{\sum_i i.N.\dfrac{(D-1)^{i-1}}{D^i} \cdot \dfrac{(N.D - R')^i}{(N.D)^i}}{N.D - R'}$$

$$= \frac{1}{D^2} \cdot \sum_i i.\left(\frac{D-1}{D} \cdot \frac{N.D - R'}{N.D}\right)^{i-1}$$

$$= \frac{1}{D^2} \cdot \sum_i i.[(1 - 1/D).(1 - R'/N.D)]^{i-1}$$

Carrying the sum over i to infinity,

$$q = \frac{1/D^2}{(r-1)^2}, \quad \text{where} \quad r = \left(1 - \frac{1}{D}\right) \cdot \left(1 - \frac{R'}{N.D}\right)$$

Govindarajan et al. distribution

No analytical solution is available for this case, and q was estimated numerically, by summing the first 1000 terms of the explicit form of the distribution:

$$w_i = 1966 \cdot \exp(-38.15 \cdot (i/R)^{0.15})$$

Three-zone model

The q-curve for the three-zone model was calculated as a weighted average of the curves for each of its components.

Expected Current Distribution of Families per Fold

Zhang and DeLisi (1988) show that (using the symbols defined here) the expected number of folds with m families currently observed is

$$H_m = \sum_{i=m}^{\infty} N w_i \left(1 - \frac{R'}{ND}\right)^{i-m} \left(\frac{R'}{ND}\right)^{m} \binom{i}{m} \qquad (2)$$

and hence, for their distribution, that

$$H_m = \frac{(R' - N')^{m-1}}{(R')^m} \cdot (N')^2.$$

For the distribution of Govindarajan et al., the summation in Equation 2 was carried out numerically for $i = m$ to 1000, using the explicit form of their distribution: $w_i = 1966 \cdot \exp[-38.15(i/R)^{0.15}]$.

For the mesofold part of the three-zone model, the Zhang and DeLisi formula was used. Observed unifolds have a single sequence family belonging to them; the total number of observed sequence families in the superfold zone was distributed according to the proportions observed in the appropriate release of SCOP.

Algebraic Solution of the Three-zone Model

In SCOP release 1.48, 211 sequence families belong to the nine superfolds (there are between 13 and 41 families per superfold), so we estimate the proportion of sequence families belonging to superfolds, f_S, as 211/1193 = 0.177. Assuming 23,100 sequence families in all, the fraction of all sequence families whose structure has been determined, p_s, is 1193/23100 = 0.0516. Assuming no overlap between mesofold and superfold zones, only mesofolds have between 2 and 12, inclusive, sequence families per fold. The number of mesofolds expected to have been seen at least twice can be obtained by summing H_m (Eq. 2) for $m = 2$ to ∞, and the expected number of sequence families by summing $m \cdot H_m$ over the same range. Analytical expressions for these sums were derived using Maple. The ratio of these two quantities was given the symbol r, that is,

$$r = \frac{\sum_{y=2}^{\infty} \sum_{x=y}^{\infty} (1/D' \cdot (1 - 1/D')^{(x-1)}) \cdot p_s^y \cdot (1 - p_s)^{(x-y)} \cdot \binom{x}{y}}{\sum_{y=2}^{\infty} y \cdot \sum_{x=y}^{\infty} (1/D' \cdot (1 - 1/D')^{(x-1)}) \cdot p_s^y \cdot (1 - p_s)^{(x-y)} \cdot \binom{x}{y}}$$

and Maple then provided the solution that

$$1/D' = \frac{r \cdot p_s}{-2r + r \cdot p_s + 1}.$$

The numerator of the expression for r is the ratio of the number of mesofolds already seen at least twice to the total number of mesofolds; Maple evaluates this term as

$$\frac{(1 - 1/D') \cdot R'^2}{(R/D' + R' - R'/D')^2}.$$

Substitution of $1/D'$ and simplification gives

$$N_M = \frac{(r-1)^2 \cdot R \cdot C}{(2r-1)(-rR' + 2rR - R)},$$

where C is the number of mesofolds already seen at least twice. $f_M = N_M \cdot D'/R$, and the final parameter can be obtained from $f_U = -f_S - f_M$.

Simulation of Fold Distributions

For a given set of parameter values of f_U, f_M, f_S, and R, the underlying fold use distribution is constructed numerically by assigning each of the R sequence families to a fold, as follows:

(A) The $f_U \cdot R$ unifold families are each assigned a separate fold.

(B) The $f_M \cdot R$ families are assigned to one of the N_M mesofolds using the Zhang and DeLisi procedure: The families are represented by the numbers 1 to R_M. Consecutive family numbers are assigned to individual mesofolds, with family number 1 assigned to mesofold 1, and family number R_M assigned to mesofold N_M. The $N_M - 1$ boundaries between intermediate families along the line 1 to R_M are selected by drawing $N_M - 1$ unique random numbers in the interval 1 to $R_M - 1$ families. Each selected family number then represents the termination of the consecutive set of families belong to a particular mesofold. For example, if the lowest number family selected is number 5, then families 1 through 5 are assigned to the first mesofold. If the next lowest family selected is number 8, families 6 through 8 are assigned to the second mesofold, and so on.

(C) Superfolds are assigned to the $f_S \cdot R$ families according to the relative prevalence of the nine superfolds in the current observed fold use distribution: $f_{S1} \cdot R_S$ families to superfold 1, $f_{S2} \cdot R_S$ to superfold 2, and so on.

Given an underlying fold use distribution constructed as above, a simulated current fold set is generated by randomly selecting families from the full set of R families. The selected set of families may then be analyzed to determine how many times each fold is represented, generating a simulated current fold use histogram.

Repeated simulations produce somewhat different fold use histograms. The average histogram from a number of simulations with the same parameters was found to agree with that obtained by the analytical procedure.

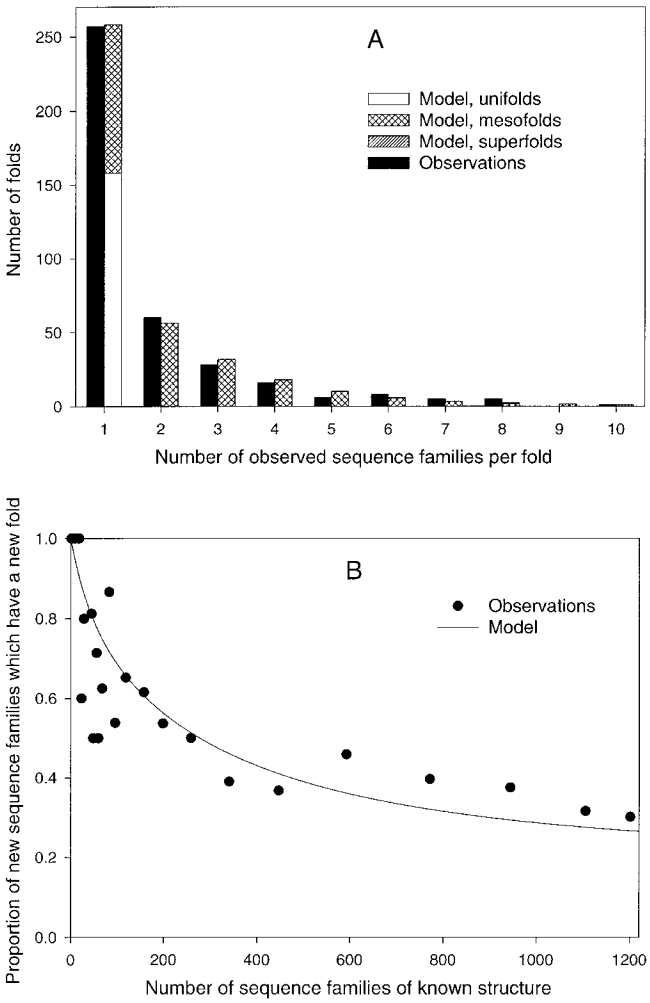

Fig. 2. Fit of the three-zone model with SCOP 1.37. (**A**) Histogram of the number of folds found in 1,2,3… sequence families in release 1.37 of SCOP[4] compared with the values expected according to the model presented here. (**B**) History of accumulation of new folds in the Protein Data Bank. Points show the fraction of new sequence families classified as representing new folds (according to the SCOP definitions), as a function of the number of sequence families of known structure, and the line shows the record expected according to the model presented here.

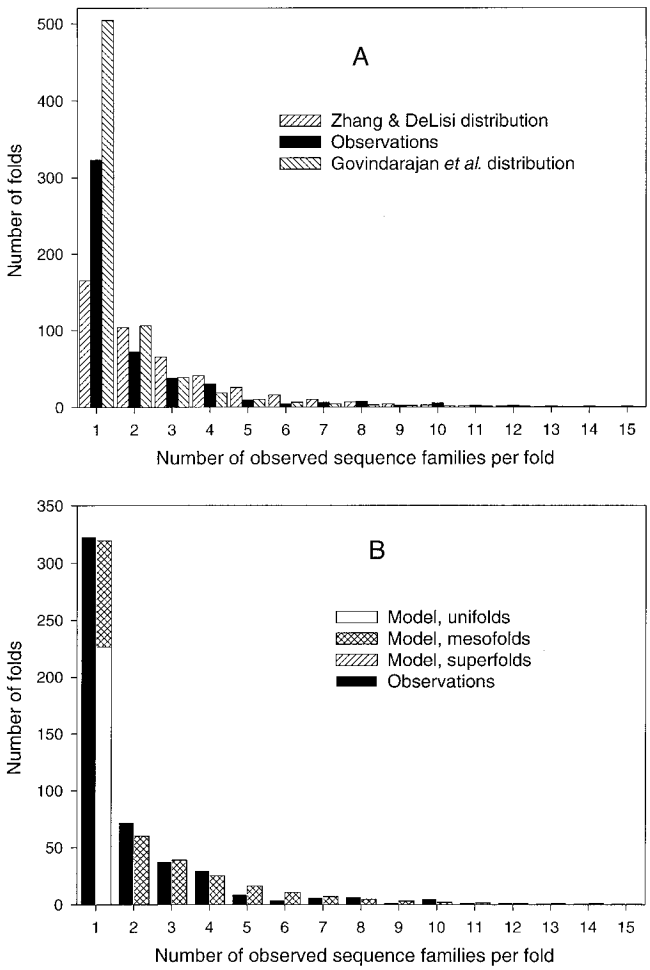

Fig. 3. Comparison of the models with SCOP 1.48. (**A**) Histogram of the number of folds found in 1,2,3… sequence families in release 1.48 of SCOP[4] compared with the values expected according to the models of Zhang and DeLisi[6] and of Govindarajan et al.[7] The superfold region is not included. (**B**) Histogram of the number of folds found in 1,2,3… sequence families in release 1.48 of SCOP[4] compared with the values expected according to the new model, with the parameters derived by fitting the model to data from SCOP Release 1.37. The model continues to match the data well, with approximately 50% more families included.

RESULTS

Fit of the Model to SCOP 1.37

Analytical solution for SCOP 1.37, assuming $R = 23{,}100$ and using $f_S = 0.18$, gave estimates of $N_M = 390$; $f_U = 0.19$. Refinement gave $N_M = 395$, $f_U = 0.190$, and $f_S = 0.182$. Figure 2 compares the observed histogram of the SCOP 1.37 distribution of sequence families over folds and the discovery curve for new folds, with those generated by this set of parameters. The model provides a close fit both to the histogram and to the curve representing the history of discovery of new folds. The latter fit provides an initial test of the model because these observations were not used to obtain the parameters.

Extrapolation of Models to SCOP 1.48

The data and models presented so far refer to SCOP release 1.37, which was current 2 years ago. There has been a substantial increase in the number of sequence families of known structure classified in SCOP since that time (~50%). We have used SCOP 1.48 to test the predictive power of all three models. Figure 3(A) compares the distribution of sequence families over folds seen in SCOP 1.48 with the corresponding distributions predicted by the models of Zhang and Delisi and of Govindarajan et al., using the SCOP 1.37 parameters, and Figure 3(B) shows the same comparison with the model presented here. As before, the model of Govindarajan et al. overpredicts the number of folds so far seen only once and that of Zhang and DeLisi underpredicts the same quantity. In the latter case, the divergence from observation has markedly increased. By contrast, the model presented here has closely predicted the changes in observed fold use, even though the number of sequence families included in the data has increased by 50%. This prediction provides evidence that

the superfold/mesofold/unifold distribution is a significantly more realistic representation of the underlying distribution of sequence families over structural folds than any previous model.

Best-Fit Model to SCOP 1.48.

We derived a new set of model parameters by fitting the model to the enlarged data set represented by SCOP Release 1.48. With the assumption that there are 23100 sequence families in all, the analytical estimates for the other parameters were $f_U = 0.18$, $f_S = 0.18$ and $N_M = 429$. Fitting the parameters to the full histogram of fold use refined these estimates to $f_U = 0.179$, $f_S = 0.175$ and $N_M = 452$. Figure 4 shows the comparison of the best-fit model to the current fold use histogram [Fig.4(A)] and to the history of the ratio of new folds to sequence families [Fig.4(B)].

Sensitivity to Parameters

The sensitivity of the fit to the precise parameter values was explored by a systematic survey of the effects of changes in the parameters on the predicted histogram and history-of-discovery curve. The survey was carried out by selecting a value for f_U and adjusting N_M to give the best fit to the fold-use histogram (omitting the superfold region). f_S was then adjusted so that the model generated the correct value of the currently observed number of sequence families in superfolds, and a final check was made that the sum of the squares of the residuals of the histogram was at a minimum with respect to changes in N_M.

Figure 4 also includes histograms and curves for "flanking" sets of parameters, chosen to give values of the root mean square deviation twice and three times that given by the best-fit set. Values of f_U below 0.179 ($N_M > 452$) systematically underpredict the number of folds for which only one sequence family is now known and overpredict those seen more than once. The first column of the histogram indicates the proportion of unifolds among the folds so far seen only once. According to the model, this proportion increases sharply across the range of parameters considered; for the best-fit parameters, about two thirds of the folds for which only one sequence family is now known are true unifolds.

Sensitivity Analysis of the Model

The SCOP data on current fold use are the result of experimental sampling of a small fraction (~5%) of all sequence families. With such a small fraction, a different set of samples would produce a somewhat different current observed fold distribution, and we therefore ask how likely it is that alternative sampling would have led to a model with significantly different values of the parameters. We addressed this question with a simulation procedure. For a given set of parameters, f_U, f_S, N_M, and R, a complete underlying fold-use distribution was constructed. R' families were selected at random from this distribution and a hypothetical current-fold use histogram constructed. This sampling was repeated 1000 times for each set of parameters. Further details of the procedure are given in Methods.

Figure 5 summarizes the estimates of the model parameters derived from each of 1000 simulations for three sets of parameter values—the best-fitting set to SCOP 1.48, and the two extreme flanking sets from Figure 4. Flanking values have corresponding fits to the experimental histogram that are clearly worse than those obtained with the central value (the root mean square deviation of the fit to the histogram increased by a factor of about 3 in each case). Figure 5(A) shows that about one in a thousand samplings with either of the flanking sets of parameters gives a fitted value of about 0.18 for f_U. That is, the odds are about 1000 to 1 against the true value of f_U being as low as 0.1 or as high as 0.22. Figure 5(B) confirms that the three distributions are qualitatively distinct.

Variation of Sequence Family Numbers

The analysis so far has assumed that there are 23,100 sequence families in all.[5] We repeated the fitting of the three other parameters of the model under the assumptions that there are 10,000 and 50,000 sequence families in all. Table I shows the numbers of folds and sequence families in each class for the best-fitting model in each case. The number of mesofolds and the proportions of sequence families in each type of fold are almost unchanged by variation in the total number of sequence families, whereas the number of unifolds increases sharply as the assumed total number of sequence families rises.

DISCUSSION

Table I shows a markedly different picture, depending on whether the results are viewed from a sequence-space or a fold-space perspective. Although the majority of sequence families are represented by mesofolds, so that the Zhang and Delisi model describes most of sequence family space, an astonishing 90% of folds are excess unifolds. Thus, fold space is dominated by folds representing only one sequence family.

Implications for the Rate of Discovery of New Folds

A reasonable goal for structural genomics is to focus on obtaining at least one representative structure for each sequence family. According to the three-zone model, how rapidly will we complete the set of all folds? Figure 6 shows the expected fraction of sequence families that will have representative structures, up to the stage where 6000 sequence families have been sampled. (SCOP 1.48 has representative structures for 1193 families, indicated on the plot.) The model assumes we have already seen all superfolds. The figure shows that, so far, we have seen about 65% of mesofolds, and this will rise to approximately 90% when there are representative structures for 6000 families. However, at this point we will only have seen approximately 25% of unifolds, with the result that some 70% of folds will still be unknown. The picture is sharply different and more positive when looked at from the point of view of the proportion of sequence families for which the fold has already been observed [Fig. 6(B)]: We already have representative structures for over 70% of families, but this will now rise very slowly to about 82% when 6000

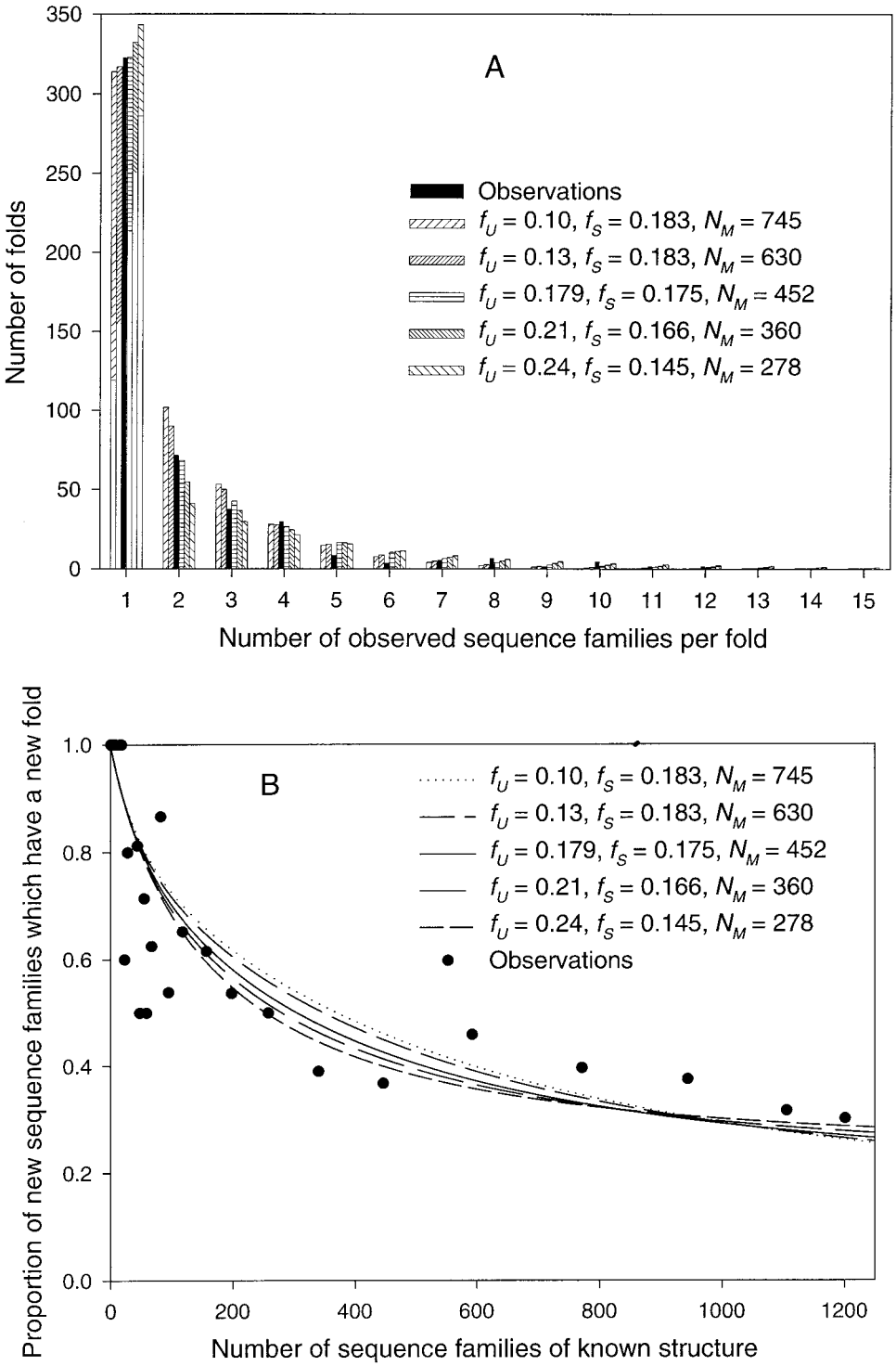

Fig. 4. Parameter sensitivity of the three-zone model. (**A**) Histogram of the number of folds found in 1,2,3… sequence families in release 1.48 of SCOP,[4] compared with the values expected according to the model presented here. Histograms have been calculated with the best-fitting set of parameters, and with four flanking parameter sets, selected to give a two- or threefold increase in the variance of the fit. The parameter values used are shown in the legend. The unpatterened regions of the first set of bars represent the number of folds seen once so far that according to the model are true unifolds. (**B**) History of accumulation of new folds in the Protein Data Bank. Points show the fraction of new sequence families classified as representing new folds (according to the SCOP definitions), as a function of the number of sequence families of known structure, and the lines show the record expected according to the model presented here. The values of the parameters used to calculate the model curves are shown in the legend.

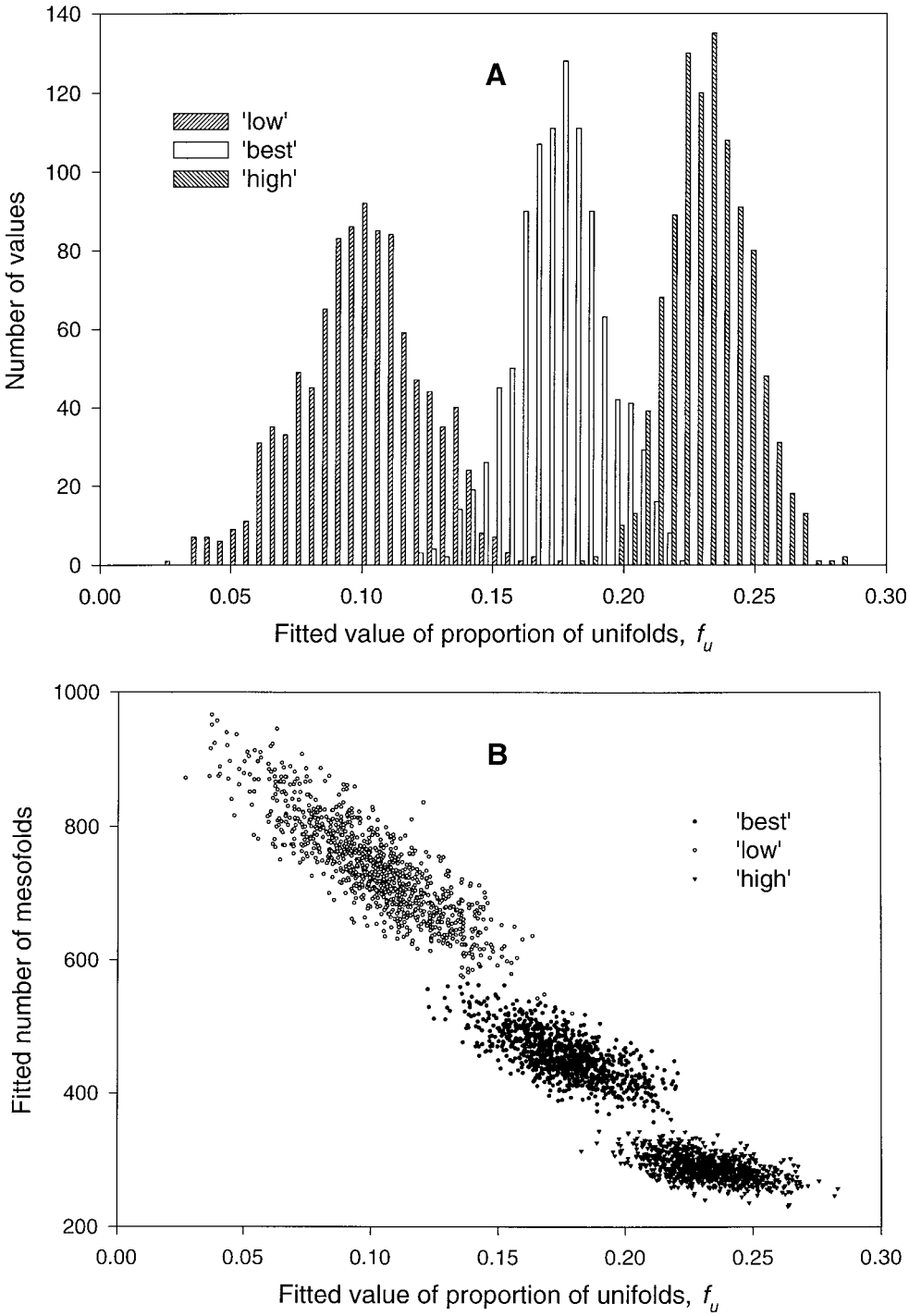

Figure 5. Effect of sample size on uncertainty in the estimates of parameters. Simulated underlying distributions of sequence families over folds were constructed with three versions of the model, using the "best-fit" parameter set (f_U,N_M, f_S) = (0.179, 452, 0.175); a "low f_u" set , (0.10, 745, 0.183); and a "high f_u" set, (0.24, 278, 0.145). For each distribution, 1000 random drawings were made, each of 1193 sequence families (the number of sequence families in SCOP 1.48). Estimates of the parameters were derived for each of these simulated "currently observed" fold-use distributions, using the method described in Methods. (**A**) Distribution of estimates of values of f_U. Approximately 1 in 1000 simulations with either the "low" or the "high" parameter sets give a fitted value of about 0.18 for f_u . (**B**) Estimate of f_U plotted against the corresponding estimate of N_M for each simulated distribution. There is a strong correlation between the estimated values of the two parameters, so that the outliers from the flanking distributions which give estimates of f_u close to 0.18 also give estimates of N_M that are significantly larger or smaller than 450. This implies that the odds are about 1000 to 1 against a distribution with a true value of f_u as low as 0.1 or as high as 0.22, giving rise to the combined pair of estimates (f_u = 0.18, N_M = 450).

TABLE I. Effect of Changing Assumed Total Number of Sequence Families on Best-Fitting Parameters of the Model

Total sequence families	23,100		50,000		10,000	
	No. of folds	No. of families	No. of folds	No. of families	No. of folds	No. of families
Unifolds	4135	4135	9250	9250	1850	1850
Mesofolds	452	14923	444	32000	420	6400
Superfolds	9	4042	9	8750	9	1750
Totals	4596	23100	9703	50000	2279	10000

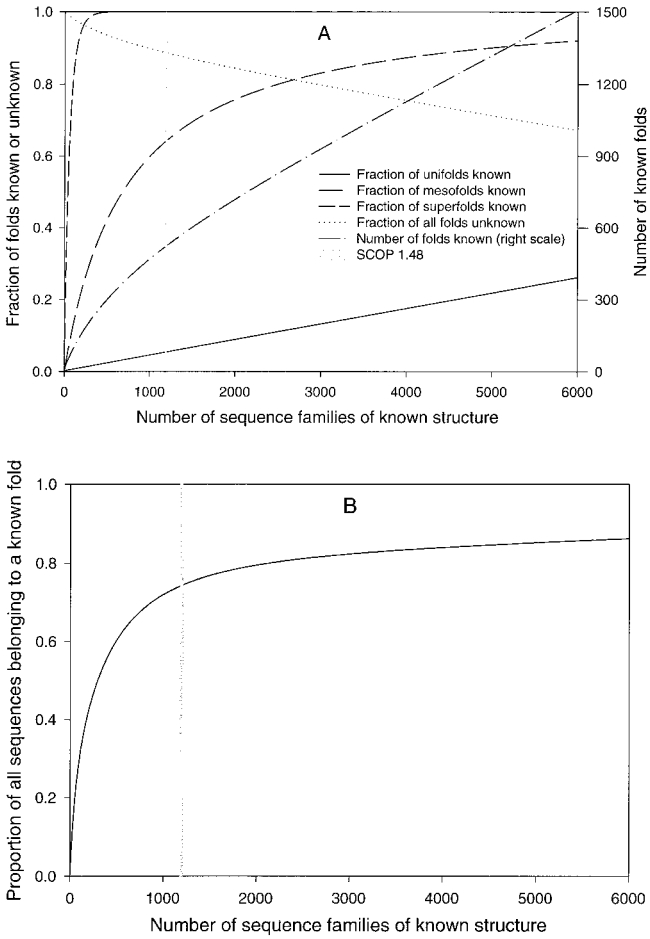

Fig. 6. Predicted Rate of Discovery of all Folds. (**A**) Proportion of each class of fold for which at least one structure will have been determined, as a function of the total number of sequence families for which a representative structure is known, according to the model. The graph also shows the total number of known folds and the proportion of all folds that are still unknown. The vertical grey bar represents the distribution of the 1193 families in SCOP Release 1.48. The extrapolation assumes that there is a total of 23,100 sequence families. When representative structures for 6000 families are known, we will have seen about 90% of all mesofolds, but only about 25% of unifolds. (**B**) Proportion of all sequence families belonging to known folds, as a function of the total number of sequence families for which a representative structure is known, according to the model. The vertical grey bar represents the distribution of the 1193 families in SCOP Release 1.48. When representative structures for 6000 sequence families are known, we will have structures for approximately 80% of all families.

sequence families have been sampled. That is, if it was always possible to correctly assign a fold to a sequence family (providing the fold has already been seen at least once), we would know the folds for 82% of sequence families.

Robustness of the Conclusions

The principal new feature of fold use in the three-zone model is the high fraction of unifolds. That is, most folds represent a rather small number of sequences, all clearly related to each other. What factors in the model could cause this to be an erroneous conclusion? Definitions of a sequence family and a fold are taken directly from SCOP. If SCOP tended to classify folds as different when they should more appropriately be considered the same, folds might appear to represent too few sequence families. A comparison of the SCOP, CATH, and FSSP classifications[10] suggests that if any thing the opposite tendency applies: CATH has some 50% more folds for the set of PDB entries considered than does SCOP. Conversely, excess merging of remote sequences into the same sequence family would reduce the number of new sequence families belonging to already known folds. Although there are some instances of inclusion of remoter sequences in families (e.g., for the globins), these are too rare to affect our conclusions.

The total number of folds depends critically on the total number of SCOP level families. To some extent, the definition of a SCOP family is arbitrary, arising from the state of art in sequence comparison methods a few years ago. The value of 23,100 families suggested by Orengo et al.[5] was also based on data available some years ago. We have made a new estimate, based on analysis of the "pfam A" family collection.[11] Pfam is a hand-curated family set based on a Hidden Markov Model method[12] for detecting evolutionary relations among sequences and so typically produces larger families than those in SCOP. For other purposes,[13] we have determined the number of structures that would need to be solved in order to model all pfam (release 4.4) family members based on 30% or more sequence identity. Because SCOP families are based on approximately this level of sequence identity, this estimate is also approximately the number of SCOP families contained within pfam A. That number is approximately 16,000. Pfam A only covers approximately half the sequences currently in the NR database, so this is clearly a low-end estimate. A simple estimate of the final number of SCOP families can be derived from extrapolation of current coverage of fully sequenced genomes by pfam. For a representative set of genomes, about one fourth of the amino acid residues fall inside a pfam family. Assuming that as more families are added, the rest of sequence space will cluster as well as that represented by the current

pfam, there will therefore be approximately $4 \times 16{,}000 = 64{,}000$ SCOP level families. Thus, the higher limit of 50,000 families considered in Table I is likely to turn out to be the most relevant. The table shows that the principal consequence of a higher number of sequence families is a substantially higher total number of folds, nearly all of them unifolds.

Implications for Evolution of Proteins.

The most striking implication of the three-zone model is that nearly all folds will be unifolds; that is, they will turn out to be narrowly distributed in sequence space. What evolutionary mechanisms might account for such a phenomenon? Four possible explanations suggest themselves:

1. Most unifolds have arisen relatively recently and have not yet had time to radiate far in sequence space. In the limit, this explanation implies three generations of folds—a small number of superfolds arose first and have therefore become most widely adopted in biology. Mesofolds are of intermediate age and have radiated in sequence space according to a Zhang and DeLisi model.
2. Unifolds are those that are less able to adapt to changes in sequence and so are restricted to a small region of sequence space. Studies of model systems[8] have suggested that most folds will be restricted in this way. The Govindarajan and Goldstein model results in a stretched exponential form for the underlying fold use histogram.
3. Unifolds may be associated with functions that are "isolated" in function space. That is, there is no biological need for new functions that could be easily derived from the ones they have. In contrast, mesofolds would then typically represent folds that started with a function that could be usefully modified, but only with the accumulation of substantial sequence changes.
4. Most folds arose at approximately the same time, but covered a limited set of functions. New functions arose from existing ones by adapting existing folds. The more ways a fold has already been adapted, the more likely it becomes that one of the existing forms will be most easily adapted to an additional function. This type of radiation leads to a power law dependence on fold use,[14] approximately like the observed distribution.

Distinguishing among these possibilities requires additional analysis outside the scope of this article.

ACKNOWLEDGEMENTS

The authors thank E. Melamud for assistance with analysis of the SCOP data.

REFERENCES

1. Zuckerkandl E. The appearance of new structures and functions in proteins during evolution. J Mol Evol 1975;7:1–57.
2. Barker WC, Dayhoff MO. Role of gene duplication in the evolution of complex physiological mechanisms: an assessment based on protein sequence data. Stadler Symp. Vol. 11, University of Missouri, 1979, p 125–144.
3. Chothia C. Proteins—1000 families for the molecular biologist. Nature 1992; 357:543–544.
4. Murzin AG, Brenner SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
5. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994; 372: 631–634.
6. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284:1301–1305.
7. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins 1999;35:408–414.
8. Govindarajan S, Goldstein, RA. Why are some protein structures so common? Proc Acad Natl Sci USA 1996;93:3341–3345.
9. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J Mol Biol 2000;299:897–905.
10. Hadley C, Jones DT. A systematic comparison of protein structure classifications. Structure Fold Des 1999;7:1099–1112.
11. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The pfam protein family database. Nucleic Acids Res 2000;28:263–266.
12. Eddy S, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 1995;2:9–23.
13. Vitkup V, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–666.
14. Unger R, Uleil R, Havlin S. Scaling law in sizes of protein families. 2001. Submitted for publication.
15. Maple V Release 4, Waterloo, Maple Inc.