The Number of Protein Folds and Their Distribution Over Families in Nature

Xinsheng Liu, Ke Fan, and Wei Wang*
National Lab of Solid State Microstructure, Department of Physics and Institute of Biophysics, Nanjing University, Nanjing, China

ABSTRACT **Currently, of the 10^6 known protein sequences, only about 10^4 structures have been solved. Based on homologies and similarities, proteins are grouped into different families in which each has a structural prototype, namely, the fold, and some share the same folds. However, the total number of folds and families, and furthermore, the distribution of folds over families in nature, are still an enigma. Here, we report a study on the distribution of folds over families and the total number of folds in nature, using a maximum probability principle and the moment method of estimation. A quadratic relation between the numbers of families and folds is found for the number of families in an interval from 6000 to 30,000. For example, about 2700 folds for 23,100 families are obtained, among them about 33 superfolds, including more than 100 families each, and the largest superfold comprises about 800 families. Our results suggest that although the majority of folds have only a single family per fold, a considerably larger number of folds include many more families each than in the database, and the distribution of folds over families in nature differs markedly from the sampled distribution. The long tail of fold distribution is first estimated in this article. The results fit the data for different versions of the structural classification of proteins (SCOP) excellently, and the goodness-of-fit tests strongly support the results. In addition, the method of directly "enlarging" the sample to the population may be useful in inferring distributions of species in different fields. Proteins 2004;54:491–499.** © **2003 Wiley-Liss, Inc.**

Key words: protein families; protein folds; sampling; distribution; maximum probability principle

INTRODUCTION

Proteins are clustered together into families based on their homologies and similarities. To some extent, the definition of a *protein family* is somewhat arbitrary, arising from the use of different percentages of sequence identity and a variety of sequence- or structure-based methods.[1–6] A residue identity of more than about 30% for clustering protein sequence pairs together into families is widely accepted in the literature.[1–2,7–13] In particular, according to the structural classification of proteins (SCOP) established by Murzin et al.,[2] proteins are clustered together into families based on one of two criteria: (1) proteins that have residue identities of 30% and greater, and (2) proteins with lower sequence identities, but whose functions and structures are very similar (e.g., globins with sequence identities of 15%). *Families* are defined as having a common fold if their proteins have same major secondary structures in the same arrangement, with the same topologic connections. How many folds exist in nature? Furthermore, what is the distribution of the folds over the families (i.e., the breakdown of folds by number of families) in nature? These questions are still enigmas[1,7–22] in biology; however, they are important for studies of structure prediction[23–25] in structural biology, for proteomics in the postgenome era,[25] and also for systematic biology.[26]

Chothia[19] first estimated the total number of families and proposed a concept that the number of families and folds in nature is rather limited. Since then, many estimates have been made, with various estimations of the total number of folds from N = 400–10,000 and the number of the families from M = 1000–30,000.[7–19] [In this work, the family number is not estimated and is considered a variable.] By assuming a geometric distribution for the breakdown of folds by the number of families in nature, Zhang and DeLisi[12] estimated the number of folds as N = 700,[9] and later as about 1300. Wolf et al.[11] proposed a logarithmic distribution and argued that the fold number is about 1000. Govindarajan et al.,[10] who employed a continuous distribution to approximate the number of families in a fold, estimated the number of folds as 4000. By using a Poisson approximation, Wang[7] estimated the total number of folds to be about 650. The number of folds for human proteins is estimated to be ≤5200 by Zhang[8] from the degenerative degree. Recently, Coulson and Moult[13] indicated the number of folds to be about 4600 for M = 23,100, the number of families estimated by Orengo et al.[1] We note that in the study by Orengo et al., protein sequence pairs with more than 30% residue identity are clustered together into *superfamilies* (or 30 SEQ families),

Grant sponsor: NFS; Grant numbers: 90103031, 10074030, and 10021001.

*Correspondence to: Wei Wang, National Laboratory of Solid State Microstructure, Department of Physics and Institute of Biophysics, Nanjing University, Nanjing 210093, China.
E-mail: wangwei@nju.edu.cn

Received 9 January 2003; Accepted 20 May 2003

which is most consistent with the definition of *family* used by SCOP, as Zhang and DeLisi[9] pointed out.

However, the previous models do not follow well the development of data from SCOP (releases 1.50 to 1.59),[2] especially for the number of superfolds (a superfold including a large number of families). For example, Zhang and DeLisi[9,12] concluded that 33 folds having more than 6 families each exist in nature, whereas that number is already 42 in SCOP release 1.55. Coulson and Moult[13] assume that the number of the folds having more than 12 families each is 9, whereas that number is already 19 in SCOP release 1.57, and so on. On the other hand, different models use different hypotheses for the distribution of folds over families. These hypotheses are critical to the estimation of the total number of folds. For example, the geometric distribution[9,12–13] $\Gamma_x = (1 - q)^{x-1}q$, with $q = N/M$, which is uniquely defined by the ratio of the total numbers of folds and families, is used to describe the probability that a fold is composed of exactly x families. However, it gives the same distribution as long as the total number of families, M, and the total number of folds, N, are the same, respectively, for any two different populations. In fact, the geometric distribution with parameter N/M in the Zhang and DeLisi model[12] may approximately describe the sampled fold distribution (for folds that include less than 6 families per fold), but not necessarily the fold distribution in nature. Another supposed distribution, namely, the logarithmic distribution,[11] seems to have better results than the geometric distribution for fitting the sampled distribution of folds in the database (except for the superfolds that are treated separately by the authors). It is not clear whether it could be used to approximate the distribution in nature, or whether the fitting value of the parameter is appropriate to describe such a distribution. Govindarajan et al.[10] employed a continuous distribution to approximate the distribution of the number of families in a fold. It seems that this distribution should be discrete, as pointed out by Wolf et al.[11]

Why do these models not follow well the development of the database, then provide a reasonable estimate for the total number of folds, N? The main reason is the lack of whole knowledge of the distribution of folds over families in nature, and also the estimation methods. What is the fold distribution in nature? Can a single-parameter distribution, as used in most previous models, also describe approximately the fold distribution in nature? Recently, Qian et al.[27] observed that the occurrence of a protein family and fold in genomes follows a power law, which might imply that the distribution of folds over families in nature is also approximated by a power law, or a single-parameter distribution, as mentioned above. Because any single-parameter distribution ultimately depends on one parameter only, and the distribution in nature and the sampled distribution should be different when $M \gg M_0$ (M_0 represents the total number of families in database), the specific parameters and form of the distribution in nature should differ from those of the sampled distribution in the database. The parameters and form of the distribution of folds in nature are unknown, and it is not easy to derive them by fitting data, because of the complications of the sampling.

In this work, we report the results of a study on the distributions of folds over families in nature and estimates of the total number of folds for different values of the total number of families, M. We first work out these distributions directly from the database (not assuming a special kind of distribution) using a maximum probability principle and the moment method of estimation.[28] From these results, we can determine the long tail of the distribution, the number of families in the largest superfold, and the total number of folds in nature for each value of M, and so on; for example, we find that the number of folds is about 2700 for the total number of families, $M = 23{,}100$.

MODEL AND METHODS

Database

We use the SCOP database for the classification of folds and families, and for the statistics for various releases (1.35–1.59) (http://www.scop.mrc-lmb.cam.ac.uk/scop). The data are derived from the parseable files. For each release, we obtain the number of folds, superfamilies, families for different classes, and the distribution of folds over families. We favor this database, because the methods for recognizing homologs are given with maximum expert human intervention (for pattern-recognition problems, it is difficult to find computational approaches that can compare with human judgment).[10] The SCOP database has been used for estimation of the fold number in almost all the recent articles.[7–13]

Grouping of Folds in the Database

Because the experimental observations of proteins may be a random sampling from a pool of proteins in nature,[9,11–13] we can assume that the observed families with total number M_0, belonging to N_0 folds, are sampled from the pool with total number M of families belonging to N folds in nature (Fig.1). Then, the probability, $P(n,k)$, for a fold in nature having n families, of which k families are sampled by experiments, is given by a hypergeometric distribution,

$$P(n,k) = \binom{M_0}{k}\binom{M - M_0}{n-k}/(Mn)\ ,$$

for $k = 1, \ldots, n$ (i.e., the probability distribution of a random sampling without replacement from a limited population). Because, obviously, $n \le 0.1\,M$, this probability can be well approximated by a binomial distribution:

$$P(n,k) = \binom{n}{k} p^k\,(1-p)^{n-k},$$

with $p = M_0/M$ being the probability of selecting any family (we use this approximation in our method). [A fold having n families in nature is denoted as S_n, and in the database as f_n.] Because the distribution of known folds over families is not uniform, and the number of folds f_n for $n \ge 4$ is rather rare, decreasing rapidly as n increases, as seen in the database, to obtain robust estimates, these

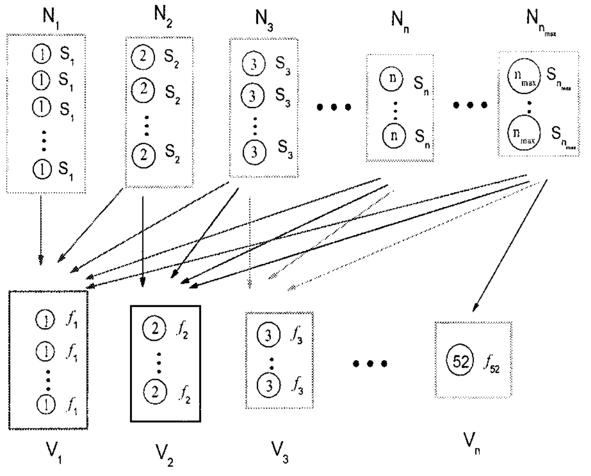

Fig. 1. Relationship between the folds and families in both nature and the observed database. In nature, *M* families are distributed into *N* folds (e.g., N_1 folds of S_1, and N_2 folds of S_2, and so on). Then, the total number of folds is $N = N_1 + N_2 + \ldots + N_n + \ldots + N_{nmax}$, and the total number of families is $M = 1 \times N_1 + 2 \times N_2 + \ldots + n \times N_n + \ldots + nmax \times N_{nmax}$. The observed data may follow the arrows, and the distribution is not uniform. There are V_1 folds of f_1, V_2 folds of f_2, V_n folds of f_n, and so on. For the database SCOP, release 1.55, there are $V_1 = 349$, $V_2 = 85$, ..., and $V_{52} = 1$, with f_{52} the largest fold with 52 families. The total number of folds is $N_0 = V_1 + V_2 + \ldots + V_{52}$. Note that there are some empty terms V_i as a result of the nonuniformity of the observed data.

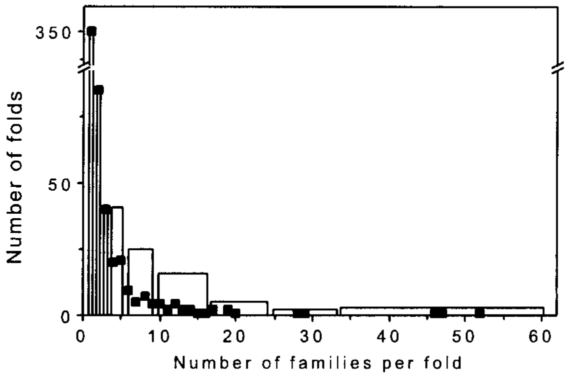

Fig. 2. The distribution of observed data and our grouping. Filled squares represent the distribution of $M_0 = 1494$ protein families among $N_0 = 566$ folds from SCOP release 1.55. The histogram shows the groups of the folds with family interval $[m_i,n_i]$ with [1,1], [2,2], [3,3], [4,5], [6,9], [10,16], [17,24], [25,33] and [34,60], and related numbers of folds O_i = 349, 85, 40, 41, 25, 16, 5, 2, and 3 for $i = 1, \ldots, 9$, respectively. The height of each bar along the ordinate is the sum of the heights of filled squares under the bars, which indicates that the total number of folds is the sum of the numbers of various kinds of folds in this group.

folds should be coarse-grained into some groups. For example, the *i*th group, defined as g_i, includes O_i folds having between m_i and n_i ($m_i \leq n_i$), inclusive, families per fold [for convenience, the interval (m_i, n_i) is hereafter called the *family interval* of the *i*th group in this article]. Such a division should make the number of folds with a different number of families in the same group as close as possible. Thus, for the observed SCOP data (release 1.55), we can divide the folds into 9 groups, each with a specific interval $[m_i,n_i]$ and O_i folds for $i = 1, \ldots, 9$ ($n_i + 1 = m_{i+1}$ for $i = 1, \ldots, 8$). Figure 2 shows the distribution of $M_0 = 1494$ protein families among $N_0 = 566$ folds for these data and their related grouping. For example, group g_4 (the fold numbers, with 4 and 5 families in each fold, being 20 and 21, respectively, are very close) is a remarkable one, with a family interval [4,5] and the fold number $O_4 = 41$. This kind of grouping method is often used in the statistical literature.[29,30] Note that, as in the work by Govindarajan et al.,[10] we omit classes 5 (the multidomain proteins) and 6 (the membrane and cell surface proteins), because they contain relatively few samples. In fact, in SCOP release 1.55, class 5 contains only 28 folds, and class 6, only 11 folds, and most of these folds include only a single family per fold.

Maximum Probability Principle

Obviously, the observed fold, f_k, comes from the fold S_n in nature (here, $k \leq n$), with a probability $P(n,k)$ (Fig. 1). Because $P(n,k)$ is a unimodal function of n, a specific fold S_{t_k} exists, having a maximal probability for f_k; that is, f_k comes from S_{t_k}, with a probability larger than that of other folds. For example, for the largest observed fold f_{52} in SCOP release 1.55, there is a positive integer $t_{52} = 804$, so that f_{52} comes from S_{804}, with maximum probability 5.7×10^{-2} for $M = 23{,}100$. Although we use $M = 23{,}100$ as an illustrative number hereafter in the article, similar results can be obtained for other values of M. Easily, the probability of f_{52} from S_{300} and S_{1500} is 4.6×10^{-11} and 4.5×10^{-8}, respectively. Thus, we have no reason to consider this observed fold coming from S_{300} and S_{1500} rather than from S_{804}. More generally, we can say that the largest observed fold f_{52} comes from a fold including x families with larger probability than that from the others, where x belongs to the interval [620,1000] and the probability is between 0.01 and 0.058. Similarly, the largest observed folds f_{31}, f_{33}, f_{41}, f_{44}, f_{47}, f_{55}, and f_{57} in SCOP releases 1.37, 1.41, 1.48, 1.50, 1.53, 1.57, and 1.59 come from folds S_{886}, S_{799}, S_{811}, S_{814}, S_{824}, S_{781}, and S_{752}, respectively, with maximum probability in nature. These results are amazingly consistent and suggest that the largest superfold [i.e., the triosephosphate isomerase (TIM) barrel fold] in nature includes about 800 families, if $M = 23{,}100$.

Grouping of Folds in Nature

We also arrange the folds in nature into 9 groups. Let X_i be the total number of folds and $[k_i, l_i]$ ($i = 1, \ldots, 9$, $k_1 = 1$, and $l_i + 1 = k_{i+1}$ for $i = 1, \ldots, 8$) be the family interval for group G_i in nature. Then, we can work out k_i and l_i through the endpoints m_i and n_i of the interval $[m_i, n_i]$ of the related group in the observed database, using the maximum probability principle (see the following).

The Moment Method of Estimation

The idea of the moment method of estimation is that the expected sampling number of folds, E_j, for the *j*th group, from nature should be equal to that of the observed folds, O_j. In nature, for group G_j, there are $l_j - k_j + 1$ kinds of folds (i.e., the folds having k_j to l_j families per fold). We consider that the numbers of these various kinds of folds are basically equal or uniformly distributed in the group G_j, and each kind of folds appears with an equal probability $p = 1/(l_j - k_j + 1)$. Thus, the number of folds for each kind is $X_j/(l_j - k_j + 1)$. As mentioned before, the probability of sampling v (here, $v \in m_i, \ldots, n_i$) families (to form a fold in the database) from u families in a fold in nature (here, $u \in k_j, \ldots, l_j$) is $P(u,v)$. Thus, the expected number of folds with v families from $X_j/(l_j - k_j + 1)$ folds having u families each in G_j is $[X_j/(l_j - k_j + 1)]P(u,v)$. Clearly, such a sampled fold with v families may come from any fold having u families [$v \leq u$; if $v \geq u$, set $P(u,v) = 0$] of G_j. Therefore, the expected sampling number of folds having v families per fold from the total X_j folds of G_j is $\sum_{u=k_j}^{l_j} P(u,v)[X_j/(l_j - k_j + 1)]$; furthermore, the total expected sampling number of folds for g_i with the family interval $[m_i,n_i]$ sampled from X_j folds of G_j is $\sum_{v=m_j}^{n_i} [\sum_{u=k_j}^{l_j} (l_j - k_j + 1)^{-1} P(u,v) X_j] = [(l_j - k_j + 1)^{-1} \sum_{u=k_j}^{l_j} \sum_{v=m_j}^{n_i} P(u,v)] X_j = C_{ij} X_j$ for $i,j = 1, \ldots, 9$. Because the sampling of g_i could come from any of the G_j, with $j = 1, \ldots, 9$, we have the expected sampling number of folds of g_i, namely, $E_i = \sum_{j=1}^{9} C_{ij} X_j$. Following the moment method of estimation,[28] then we have

$$\sum_{j=1}^{9} C_{ij} X_j = O_i, \quad \text{with } i = 1, \ldots, 9. \tag{1}$$

Solving these linear equations gives the estimated fold distributions, then the total number of folds,

$$N = \sum_{i=1}^{9} X_i.$$

Consequently, based on the estimated distribution, we obtain the estimated number of families,

$$M' = \sum_{i=1}^{9} (k_i + l_i) X_i/2,$$

which should consistently be equal to the supposed value of M.

The Most Likely Fold Distribution

For any given k_i and l_i, $i = 1, \ldots, 9$, we can easily obtain X_i ($i = 1, \ldots, 9$) by solving linear Eq. (1). However, $k_i$and l_i, $i = 1, \ldots, 9$, are also unknown. We should determine the values of X_i, k_i, and l_i ($i = 1, \ldots, 9$) at the same time. Because all the data indicate that the number of folds decreases as the number of families per fold increases, and that most folds have a single family each, the most likely estimate for the distribution of folds should satisfy the condition that the number of folds having a single family each (i.e., X_1, if $l_1 = 1$) is a lot larger than that of the other kinds of folds, and X_i decreases as i increases, although the length of the family interval $[k_i,l_i]$ increases with an increase in i. With this in mind, in the next section, we determine the fold groups.

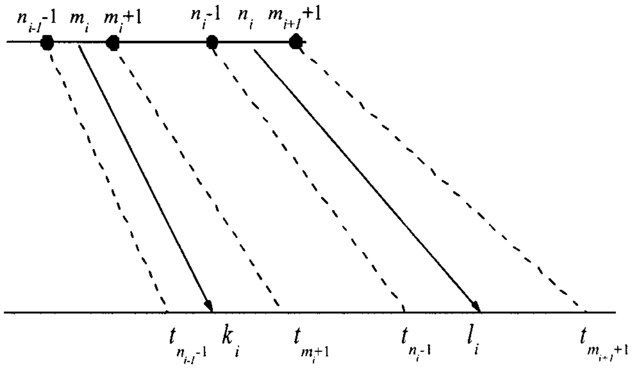

Fig. 3. Determination of the family intervals for the last 5 groups in nature based on the maximum probability principle. Arrows show the mapping of the family interval $[m_i,n_i]$ of the ith group g_i in the observed database to the family interval $[k_i,l_i]$ of the ith group G_i in nature.

Determining Fold Groups in Nature

The last 5 groups in nature can be determined from the related groups in the database by the maximum probability principle. From Figure 2, we know that f_n for $n \geq 5$ is rare, indicating that S_n for $n \geq 5$ in nature is also relatively rare. Thus, the observed folds in the last 5 groups may mainly come from the folds of their related groups in nature, implied by the maximum probability principle. A simple method for determining these fold groups is to choose $k_i = t_a$ (which is like t_{52}), where $a = m_i$ or n_{i-1} ($n_{i-1} + 1 = m_i$), for $i = 5, \ldots, 9$ (then $l_i = k_{i+1} - 1$ for $i = 5, \ldots, 8$), and $l_9 = t_b$, where $b = n_9$. But, in this case, we may obtain estimates X_i, $i = 5, \ldots, 9$, which are not close to integers (especially for X_8 and X_9, because these folds contain many families and too many rounding-off errors may reduce the accuracy of the estimates). In this research, we use the following more elaborate method. From the maximum probability principle, the family interval for G_i with $i \geq 5$ in nature should, at least, contain the interval $[t_u,t_v]$, where $u = m_i + 1$ and $v = n_i - 1$, and at the same time should, at most, contain the interval $[t_q,t_r]$, where $q = n_{i-1} - 1$ and $r = m_{i+1} + 1$ (otherwise, the neighboring groups will not follow the maximum probability principle). Therefore, we obtain the family interval $[k_i,l_i]$ for G_i with $k_i = t_q, \ldots, t_u$, and $l_i = t_v, \ldots, t_r$, for $i = 5, \ldots, 9$ (see Fig. 3). Here, both $k_i$and l_i belong to a set.

Once the last 5 fold groups have settled down, we could divide the interval $[1, k_5 - 1]$ ($k_5 - 1 = l_4$) into 4 arbitrary intervals, in principle, to determine the first 4 fold groups in nature. This should give a consistent estimation for the number of folds in these groups; that is, the longer the interval $[k_i,l_i]$ ($i = 1, \ldots, 4$), the larger the number of related folds). However, to get the most likely estimates mentioned above, it is necessary to search through all the values of $k_i$and l_i. We have a simple method to set ranges for k_i and l_i, $i = 1, \ldots, 4$: Set $k_i = k_{i-1} + 1, \ldots, k_{i+1} - 1$ for $i = 2, 3, 4$ (obviously, $k_1 = 1$, $l_i = k_{i+1} - 1$ for $i = 1, \ldots, 4$).

Now, we search through the above-determined ranges of k_i and l_i to obtain the most likely estimate for the distribution of folds under the following conditions:

TABLE I. Several examples of Groupings of Folds in Nature and Estimated Number of Folds and Distributions Based on the Supposed Number of Families M

M	N	X_1 $[k_1,l_1]$	X_2 $[k_2,l_2]$	X_3 $[k_3,l_3]$	X_4 $[k_4,l_4]$	X_5 $[k_5,l_5]$	X_6 $[k_6,l_6]$	X_7 $[k_7,l_7]$	X_8 $[k_8,l_8]$	$X_9[k_9,l_9]$
6000	1266	714 [1,1]	348 [2,4]	81 [5,10]	80 [11,19]	18 [20,40]	17 [41,64]	3 [65,98]	2 [99,135]	3 [136,240]
9000	1569	874 [1,1]	470 [2,5]	92 [6,14]	90 [15,28]	18 [29,61]	17 [62,96]	3 [97,150]	2 [151,202]	3 [203,360]
12,000	1861	1115 [1,1]	537 [2,7]	85 [8,21]	82 [22,37]	18 [38,83]	17 [84,129]	2 [130,203]	2 [204,271]	3 [272,480]
15,000	2092	1263 [1,1]	608 [2,8]	91 [9,25]	88 [26,46]	18 [47,104]	17 [105,162]	2 [163,259]	2 [260,341]	3 [342,602]
18,000	2360	1515 [1,1]	635 [2,10]	87 [11,33]	80 [34,54]	18 [55,123]	17 [124,192]	3 [193,309]	2 [310,407]	3 [408,722]
21,000	2559	1659 [1,1]	683 [2,11]	91 [12,37]	84 [38,63]	18 [64,144]	17 [145,225]	2 [226,357]	2 [358,478]	3 [479,843]
23,100	2714	1794 [1,1]	703 [2,12]	88 [13,40]	87 [41,69]	18 [70,159]	17 [160,249]	2 [250,396]	2 [397,525]	3 [526,925]
30,000	3081	2051 [1,1]	799 [2,14]	95 [15,49]	93 [50,89]	18 [90,203]	17 [204,317]	3 [318,510]	2 [511,680]	3 [681,1205]

Here, $[k_i;l_i]$ represents the family interval of group G_j, X_i is the estimated fold number for group G_i, and N is the estimated total fold numbers in nature.

1. There is an order in X_j in which $X_1 \geq X_2 \ldots \geq X_8$ [it is not necessary that $X_8 \geq X_9$, because few folds are observed in the last two groups, and $O_8 = 2$, $O_9 = 3$ ($O_9 > O_8$)].
2. Without loss of generality, we choose the solutions of X_j for $j \geq 8$ to be close to an integer, such that we take $1.85 \leq X_8 \leq 2.15$ and $2.85 \leq X_9 \leq 3.15$ (as $O_8 = 2$ and $O_9 = 3$).
3. By choosing as large as possible value of X_1, we obtain the most likely estimate for the distribution of folds, and, simultaneously, the related values of k_i and l_i are settled.

RESULTS AND DISCUSSION

Total Number of Folds and Their Distribution in Nature

For each value of M, we work out the related grouping of folds in nature, namely, the family intervals $[k_i,l_i]$ of group G_i, and the values of X_i for $i = 1, \ldots, 9$, then we obtain the total number of folds N in nature (Table I). Thus, we have directly estimated the distribution of folds over the families and the total number of folds (as a function of the total number of families in nature). For example, we have $N = 2714$, $X_i = 1794, 703, 88, 87, 18, 17, 2, 2, 3$, and $[k_i,l_i] =$ [1,1], [2,12], [13,40], [41,69], [70,159], [160,249], [250,396], [397,525], [526,925], respectively, for G_i for $i = 1, \ldots, 9$ for $M = 23{,}100$ (at the same time the estimated value of families $M' = 23{,}114$. Consistently, the supposed value M and the estimated value M' are quite in agreement with each other. This consistency also exists for the other values of M). Among the 2714 folds there are about 33 superfolds, each including more than 100 families, and the largest superfold composed of about 800 families is obtained from the maximum probability principle. By fitting the numbers for N and M in Table I, a quadratic function,

$$N = -1.217 \times 10^{-6}M^2 + 0.120M + 592.186$$

for $M \in [6000, 30{,}000]$ can be obtained (Fig. 4).

Most recent articles[1,7–13] imply that the most likely number of families in nature may belong to the interval [6000, 30,000]. Also, based on analysis of the "pfam A" family collection,[3] Coulson and Moult[13] suggested that the higher limit of 50,000 families could be reached. In the case in which $M = 50{,}000$, using our methods, we obtain the total number of folds, $N = 4478$, $X_i = 3408, 850, 89, 88, 18, 17, 3, 2, 3$, and $[k_i,l_i] =$ [1,1], [2,24], [25,88], [89,144], [145,339], [340,530], [531,849], [850,1139], [1140,2010], respectively, for G_i with $i = 1, \ldots, 9$. Likewise, for $M = 100{,}000$, the estimated number of folds is about 8000, $X_i = 6609, 910, 94, 86, 18, 18, 2, 2, 3$, and $[k_i,l_i] =$ [1,1], [2,47], [48,180], [181,290], [291,685], [686,1081], [1082,1695], [1696,2279], [2280,4020], respectively. Note that the results for other values of M in interval [30,000, 100,000] are not shown here, and the relationship of N and M in this interval is approximately linear.

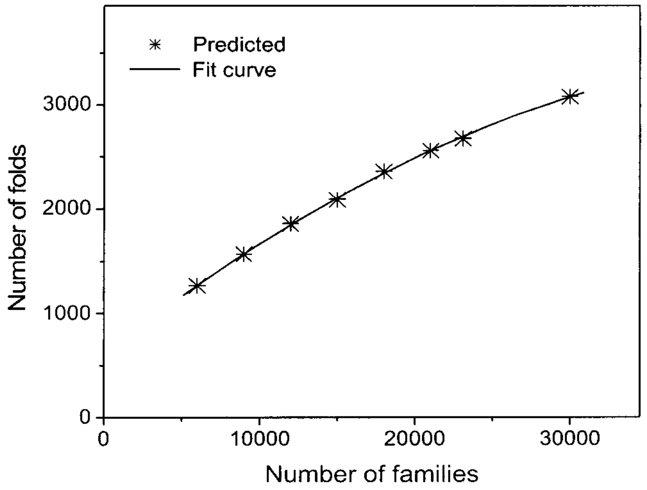

Fig. 4. The estimated number of folds, N, versus the number of families M. Here, N = 1266, 1569, 1861, 2092, 2360, 2559, 2714, and 3081 for M = 6000, 9000, 12,000, 15,000, 18,000, 21,000, 23,100, and 30,000 are used. The solid curve is a fitting with $N = -1.217 \times 10^{-6} M^2 + 0.120\ M + 592.186$ for $M \in [6000, 30{,}000]$.

Robustness of Results Against Changes in Grouping

The estimate of the fold distribution is not sensitive to the change in groupings of the last several fold groups, so long as these formed groups are based on the maximum probability principle. For example, for $M = 23{,}100$, assume that the family intervals of the last 3 fold groups are [250,404], [405,538], and [539,950] instead of [250,396], [397,525], and [526,925] (the corresponding family intervals of the other fold groups are the same; see Table I).

TABLE II. Results for Various Versions of the SCOP Database

Version of SCOP	M_0	N_0	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9
		N_s	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9
			$[m_1,n_1]$	$[m_2,n_2]$	$[m_3,n_3]$	$[m_4,n_4]$	$[m_5,n_5]$	$[m_6,n_6]$	$[m_7,n_7]$	$[m_8,n_8]$	$[m_9,n_9]$
1.37	808	375	242	57	27	16	14	10	5	2	2
		381.93	252.97	58.32	25.56	13.33	13.07	9.32	5.08	2.35	1.93
			[1,1]	[2,2]	[3,3]	[4,4]	[5,6]	[7,9]	[10,14]	[15,21]	[22,32]
1.41	961	426	277	65	29	17	20	11	5	2	2
		428.29	277.50	65.38	29.76	16.13	19.79	10.62	4.95	2.62	1.54
			[1,1]	[2,2]	[3,3]	[4,4]	[5,7]	[8,11]	[12,18]	[19,29]	[30,46]
1.48	1167	489	306	71	37	38	17	11	5	2	2
		484.92	307.26	73.63	34.53	31.73	16.24	10.61	5.98	3.00	1.93
			[1,1]	[2,2]	[3,3]	[4,5]	[6,8]	[9,12]	[13,19]	[20,32]	[33,50]
1.50	1248	515	321	75	39	42	15	13	5	3	2
		505.96	318.16	76.59	36.11	34.20	17.29	11.12	7.03	3.43	2.03
			[1,1]	[2,2]	[3,3]	[4,5]	[6,8]	[9,12]	[13,19]	[20,34]	[35,55]
1.53	1317	528	325	81	39	40	18	13	7	2	3
		523.36	327.15	79.04	37.34	36.24	18.30	11.43	7.96	2.87	3.04
			[1,1]	[2,2]	[3,3]	[4,5]	[6,8]	[9,12]	[13,19]	[20,30]	[31,60]
1.57	1626	605	367	102	39	41	28	17	6	2	3
		595.66	364.52	89.46	41.77	44.17	27.99	16.45	6.16	1.94	3.19
			[1,1]	[2,2]	[3,3]	[4,5]	[6,9]	[10,16]	[17,24]	[25,33]	[34,60]
1.59	1750	640	392	103	43	43	27	18	9	2	3
		622.99	378.37	93.53	43.24	46.63	31.16	16.82	7.53	2.23	3.49
			[1,1]	[2,2]	[3,3]	[4,5]	[6,9]	[10,16]	[17,24]	[25,34]	[35,65]

The observed total numbers of folds N_o and their distributions O_i, and the theoretical sampling total numbers of folds N_s and their distributions E_i are shown for each version of SCOP, with $M = 23{,}100$. Here, $O_{\rm i}$ (or $E_{\rm i}$) represents the observed number of folds (or the expected number of folds) in group g_i, whose family interval is $[m_{\rm i},n_{\rm i}]$.

Then, we have $X_7 = 2.2$, $X_8 = 2.2$, and $X_9 = 2.9$, instead of $X_7 = 2.1$, $X_8 = 2.1$, and $X_9 = 3.0$ (the other corresponding components of these two estimated fold distributions are the same). This gives us the same results for the estimate on both the total number of folds and the distribution.

Though a different classification of the first several groups may give somewhat different estimated numbers of folds for these groups, the estimated distribution of folds is similar, and the estimate for the total number of folds is about the same; that is, the estimated distribution of folds and the estimate for the total number of folds are also robust against change in groupings of the first several fold groups. For example, in the case of $M = 23{,}100$, let the family intervals for the first four groups be respectively [1,1], [2,7], [8,13], and [14,69] instead of [1,1], [2,12], [13,40], and [41,69] (the remaining corresponding family intervals are the same; see Table I). Then we have $X = (1728, 452, 291, 169, 18, 17, 2, 2, 3)$ instead of $X = (1794, 703, 88, 87, 18, 17, 2, 2, 3)$. However, the estimated numbers of folds for both groupings are approximately equal (2682 and 2714, with a relative error within 2%). We can determine that the number of folds of the fourth group for the first estimated distribution is approximately equal to the sum of the number of folds of the third and fourth groups for the second estimated distribution, and the number of folds of the second group for the second estimated distribution is approximately equal to the sum of the number of folds of the second and third groups for the first estimated distribution. It follows that these two estimated distributions are consistent.

Consistency With Other Versions of SCOP

We used SCOP release 1.55 for our estimations. To check whether our results are consistent with the observations of previous or following versions of SCOP, consider the sample of M_0 families (e.g., $M_0 = 808$ for SCOP release 1.37) drawn from the universe of M families distributed among folds in accordance with our estimated distribution. We compare the theoretical sampling numbers of folds N_S with the observed numbers N_0 (here, $M = 23{,}100$ for all comparisons). Our results are in good agreement with the observations, and the relative errors are found to be only within 3% (Table II). It is also noted that we have similar results on the comparison with different values of M.

Next, we further show how well the theoretical sampling distributions of folds compared with the observed ones. To illustrate a problem of hypothesis testing, here, our null hypothesis concerns whether the theoretical sampling distributions fit the observations well. If our null hypothesis is true (i.e., our results fit the observations), the Pearson χ^2 statistic, defined as

$$C = \sum_{i=1}^{9} \frac{(O_i - E_i)^2}{E_i},$$

should satisfy the so-called χ^2 distribution. Here, the degree of freedom is $9 - 1 = 8$, because the expected frequencies E_i (i.e., the theoretical sampling number of folds) are calculated without using any estimate of unknown parameters.[31] For any of the 7 SCOP releases (Table II) in which we obtain a value c of the variable C, then the goodness of fit is given by the probability $g = P(C \geq c)$. Generally, the comparison between the observed and the expected distribution is considered good if the goodness $g > 0.3$. For these 7 releases of SCOP data, we have the values of $g = 0.99$, 0,97, 0.98, 0.91, 0.99, 0.97, and 0.95, respectively, providing strong evidence that our estimated distributions fit the observed data very well.

Several Remarks

Our estimates in this article relate to aspects of the database and the definition of folds in the families. It is assumed in our model (also in previous models[9,11–13]) that the structural database is random samples drawn from the universal populations of protein families. However, the process of selecting protein families for structural determination may not be ideally random. It may be that certain folds are more likely to be solved by crystallography or NMR spectroscopy, or are found easily in the organisms under study, then over-represented. For example, the superfolds are among the most extensively studied taxa in the protein world, and many distantly related families have already been identified with various sequence-matching techniques.[2–6,27] Then the superfolds, such as TIM-barrel, P-loop, and so on, may be over-represented in the database; that is, each contains more families than those found by a random sampling. Thus, the estimated number of families included in each of these folds may be a little larger with use of this database. However, we would argue that the effect of the over-representation of the observed superfolds in the database is small and basically does not affect our estimates, because the increase in the number of families included in each superfold is basically smooth, with the increase in the family count (from the old to the new version of SCOP). Furthermore, as mentioned above, our estimates based on SCOP release 1.55 are consistent with the observations of previous or following releases of the database, and especially, the largest folds in different SCOP releases suggest an amazingly consistent largest superfold in nature (i.e., the TIM barrel fold), based on the same principle.

In contrast, some folds may be inaccessible to the sampling process. It is well known that some protein families are very difficult to crystallize, and their structures will be under-represented in the current structure database. For example, because the crystallization of membrane proteins have presented many technical difficulties in the past, and the classification of transmembrane proteins in SCOP might be oversimplified (but, according to Murzin,will be improved in the future; personal communication), the samples in class 6 (the membrane and cell surface proteins) are few and have been omitted in our study. Therefore, the estimates based on the current structural database may reflect solely the results for nontransmembrane protein folds,[7–13] or more generally, results for the "accessible" folds. However, as Zhang and DeLisi[9] pointed out, though the number of natural transmembrane folds is unknown, it is expected to be small compared with the nontransmembrane folds, because of the severe sequence constraints imposed by the hydrophobic nature of the membrane.

In another example, class 5 (the multidomain proteins) of SCOP also contains few sample folds, whereas there may be many multidomain proteins in SWISS-PROT. This may not present a serious bias for SCOP. In fact, because the unit of classification is usually the protein domain in the SCOP database, there are many multidomain proteins in the whole SCOP database,only a small part of which are classified as class 5. Classes 1 to 4 contain multidomain proteins composed only of domains of the respective classes. Class 5 contains proteins composed of domains of different classes (e.g., α and α/β). A multidomain protein will be split into domains and classified accordingly in SCOP, if there is a good evidence of independent evolutionary origins of the domains. Such evidence might mean the presence of domains of similar sequence or structure in proteins of different domain organization (Murzin, personal communication). There are plenty of multidomain proteins in the Protein Data Bank (PDB). Some are split into domains in SCOP, whereas others remain where they are presently. If there is a bias in the PDB, it is historical in origin. Earlier structures were of smaller proteins or protein domains, because of the limitations of methods for structure determination at the time.

In principle, our estimates of the total folds may be affected by the fluctuation in V_i, namely, the number of folds including exactly i families per fold in the database, due to the randomness of sampling, errors in the methods, and so on, but we expect these to be small, because the data are sufficient in a statistical sense (also see above discussion). In a word, the SCOP database reflects well the "accessible" inter-relationship of the folds and the families.

In addition, the definitions of folds and families can affect the results of this research. For example, CATH,[4] another widely used database, has more folds for the set of PDB entries considered than does SCOP.[6] Based on CATH classification, the estimated number of folds of each group considered in this article may be a little larger than that based on SCOP. However, the physical implications of the model are similar. In this article, we use the SCOP classification directly, and it has been systematically used to estimate the total number of folds.[7–13] As Wolf et al.[11] have pointed out, the SCOP classification is generally compatible with the other, fully automatic classifications, such as CATH, and provides a reasonably robust partitioning of the protein universe.

Comparison with Previous Model and Results

In this article, we have estimated directly the distribution of folds over families in nature by using the observed database most sufficiently. Our results are consistent and depend on the total number of families in nature. However, under the geometric distribution with parameter N/M in the Zhang and DeLisi[12] model, the estimated number of folds is insensitive to the total number of families, M, and the results are not consistent. From their results, the 1250 folds having less than 7 families per fold contain at most $M = 1250 \times 3.5 = 4375$ families, and the 33 folds having more than 6 families per fold contain 423 families. Thus, the total number of families is at most 4789, which may not be in accord with the assumption that $M >> M_0$. In fact, the geometric distribution $\Gamma_x = (1 - q)^{x-1} q$, with $q = N/M$, does not necessarily describe the probability that a fold is composed of exactly x families. Regard M families (belonging to N folds) in nature as sequentially numbered sites along a straight line, with the families belonging to an individual fold represented by consecutive sites (see Fig. 1 in Zhang and DeLisi[9]). Thus, each fold is related to a boundary site and a separate section (there are N folds in nature; therefore, N boundary sites and N sections). Then, the probability Γ_x is accurately the probability of a random sampling of x sites, with replacement in which the first $x - 1$ sites are not boundary sites and the last is a boundary site. In this sampling, the first $x - 1$ sites that are not boundary sites may come from *different* sections (a section means an individual fold), and the last that is a boundary site may come from *any* section. However, when a fold is composed of exactly x families, the x sites come from *the same section*, and this section has only x sites. In the example illustrated in Figure 1 in Zhang and DeLisi,[9] the 3rd, 11th, 14th, 19th, and ($M - 3$)th families are selected (suppose that no other families are selected). Γ_5 does describe the probability of this sampling; however, it does not necessarily describe the probability that a fold is composed of exactly 5 families, because, obviously, these 5 families belong to 4 different folds (folds 1–3 and fold N).

In a recent article,[13] Coulson and Moult divided the universe of folds into 3 zones, that is, the zones of unifolds, mesofolds, and superfolds. This classification may be better than that in previous models. However, first of all, the division of folds is still rough, and the boundary between the zone of mesofolds and the zone of superfolds should depend on the total family number M. If the total number of families in nature is less than 6000, the definition of mesofolds could be suitable (a mesofold includes 2–12, inclusive, families). If M is more than 23,100, the zone of mesofolds should include more kinds of folds; say, for example, that a mesofold has 2–40, inclusive, families (according to our estimation). Second, the geometric distribution according to Zhang and DeLisi model[12] was used to estimate the number of the mesofolds, which results in the inconsistency of their results. For example, 452 mesofolds contain 14,923 families (Table I in Coulson and Moult[13]). Then, a mesofold includes an average number of 33 (14923/452) families for $M = 23{,}100$, which contradicts the definition that a mesofolds has 2–12, inclusive, families. Third, the distribution of the superfolds has not been estimated reasonably. The authors assumed that the number of the superfolds (having more than 12 families each) is only 9, whereas it is already 19 in SCOP release 1.57. Also, based on this assumption, the 9 superfolds contain 4042 families for $M = 23{,}100$ (Coulson and Moult's Table I[13]). Thus, a superfold includes an average number of about 450 families. This number is far larger than the family number in a mesofold; therefore, the zone of mesofolds is far from the zone of superfolds, which is unlikely to occur (no further information about the superfolds). Finally, the estimation of the total number of unifolds relates to the estimated numbers of mesofolds and superfolds in the study by Coulson and Moult.[13] Because the estimated number of mesofolds is insensitive to the total number of families M under the geometric distribution, and the estimated number of superfolds is smaller than that in nature, the estimated number of unifolds is considerably larger when M is large. For example, the estimated number of unifolds is 4135 for $M = 23{,}100$ (Coulson and Moult's Table I[13]). Apart from the above points we have raised, the concept for classifying the folds into three zones may be reliable. Alternatively, in our study, the estimated number of families from the predicted distribution of folds over families is consistent with the supposed number of families showing a good self-consistency, and our results are also consistent with the observations of all previous or following versions of SCOP. Our results indicate that the number of folds having more than 12 families each is about 200 for $M = 23{,}100$, and we work out a definite long tail of fold distribution that shows a smooth fall-off. In addition, the number of superfolds is only about 1800.

CONCLUSIONS AND IMPLICATIONS

The main results of this article are as follows:

1. The distribution of folds over families in nature, in particular, the long tail of the distribution, and the number of the families in the largest superfold in nature, and so on, are first estimated in this article. Our results indicate that although the majority of folds have only a single family per fold, there are a considerably larger number of folds including many families each than in the database. Also, the number of families contained in each of these folds, which depends on the total number of families in nature, is far larger than the number of observations in the database. Therefore, the distribution of folds over families in nature differs markedly from the sampled distribution. The results have been checked by goodness-of-fit tests.
2. We found a quadratic relation between the number of families and folds. From this relation we can estimate the number of folds for any possible number of families belonging to a given interval.
3. We introduce a very useful method that makes a direct "enlargement" of the sample to the population.

We have employed the maximum probability principle by which a fold in nature can be found, such that a certain observed fold in the database derives from it (e.g., according to this principle, all versions of observed data indicate an amazingly consistent largest fold in nature). This principle enables our model to match the real cases. Therefore, our results may provide the first insight into the whole distribution of folds over families quantitatively, and will be very useful for studies on the structure prediction of proteins, proteomics, and systematic biology. Our method has general significance for inferring distribution of species in different fields, such as zoology, botany, and so on.

ACKNOWLEDGMENTS

Our thanks to P.G. Wolynes for discussions of the manuscript and to A.G. Murzin for discussion about SCOP.

REFERENCES

1. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372: 631–634.
2. 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.
3. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The pfam protein family database. Nucleic Acids Res 2000;28:263–266.
4. Pearl FMG, Martin N, Bray JE, Buchan DWA, Harrison AP, Lee D, Reeves GA, Shepherd AJ, Sillitoe I, Todd AE, Thornton JM, Orengo CA. A rapid classification protocol for the CATH domain database to support structure genomics. Nucleic Acids Res 2001;29: 223–227.
5. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res 1999;27:215–219.
6. Hadley C, Jones DT. A systematic comparison of protein structure classifications. Struct Fold Des 1999;7:1099–1112.
7. Wang ZX. How many fold types of protein are there in nature? Proteins 1996;26:186–191.
8. Zhang CT. Relations of the numbers of protein sequences, families and folds. Protein Eng 1997;10:757–761.
9. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284:1301–1305.
10. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins 1999;35:408–414.
11. 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.
12. Zhang C, DeLisi C. Protein folds: Molecular systematics in three dimensions. Cell Mol Life Sci 2001;58:72–79.
13. Coulson AW, Moult J. A unifold, mesofold, and superfold model of protein fold use. Proteins 2002;46:61–71.
14. Berman HM, et al. The Protein Data Bank. Acta Cryst D 2002;58:899–907.
15. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
16. Ptitsyn OB, Finkelstein AV. Similarities of protein topologies: Evolutionary divergence, functional convergence or principles of folding? Q Rev Biophys 1980;13:339–386.
17. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167–339.
18. Chothia C, Finkelstein AV. The classification and origins of protein folding patterns. Annu Rev Biochem 1990;59:1007–1039.
19. Chothia C. One thousand protein families for the molecular biologist. Nature 1992;357:543–544.
20. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
21. Blundell TL, Johnson MS. Catching a common fold. Protein Sci 1993;2:877–883.
22. Alexandrov NN, Go N. Biological meaning, statistical significance, and classification of local spatial similarities in non-homologous proteins. Protein Sci 1994;3:866–875.
23. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
24. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature1992;358:86–89.
25. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
26. Kitano H. Systems biology: A brief overview. Science 2002;295: 1662–1664.
27. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. J Mol Biol 2001;313:673–681.
28. Cramer H. Mathematical methods of statistics. Princeton, NJ: Princeton University Press; 1946.
29. Dykstra RL, Kochar S, Robertson T. Statistical inference for uniform stochastic ordering in several populations. Ann Stat 1991;19:870–888.
30. Wang Y. A likelihood ratio test against stochastic ordering in several populations. J Am Stat Assoc 1996;91:1676–1683.
31. Lehmann EL. Testing statistical hypotheses. New York: John Wiley; 1986.