Monte Carlo Estimation of the Number of Possible Protein Folds: Effects of Sampling Bias and Folds Distributions

Hadas Leonov,[1] **Joseph S.B. Mitchell,**[2] **and Isaiah T. Arkin**[3*]
[1]*School of Computer Science and Engineering. The Hebrew University, Givat-Ram, Jerusalem, Israel*
[2]*Department of Applied Mathematics and Statistics. Stony Brook University, Stony Brook, New York*
[3]*The Alexander Silberman Institute of Life Sciences. Department of Biological Chemistry. The Hebrew University, Givat-Ram, Jerusalem, Israel*

ABSTRACT **The estimation of the number of protein folds in nature is a matter of considerable interest. In this study, a Monte Carlo method employing the broken stick model is used to assign a given number of proteins into a given number of folds. Subsequently, random, integer, non-repeating numbers are generated in order to simulate the process of fold discovery. With this conceptual framework at hand, the effects of two factors upon the fold identification process were investigated: (1) the nature of folds distributions and (2) preferential sampling bias of previously identified folds. Depending on the type of distribution, dividing 100,000 proteins into 1,000 folds resulted in 10–30% of the folds having 10 proteins or less per fold, approximately 10% of the folds having 10–20 proteins per fold, 31–45% having 20–100 proteins per fold, and >30% of the folds having more than 100 proteins per fold. After randomly sampling one tenth of the proteins, 68–96% of the folds were identified. These percentages depend both on folds distribution and biased/non-biased sampling. Only upon increasing the sampling bias for previously identified folds to 1,000, did the model result in a reduction of the number of proteins identified by an order of magnitude (approximately 9%). Thus, assuming the structures of one tenth of the population of proteins in nature have been solved, the results of the Monte Carlo simulation are more consistent with recent lower estimates of the number of folds, ≤1,000. Any deviation from this estimate would reflect significant bias in the experimental sampling of protein structure, and/or substantially nonuniform folds distribution, manifested in a large number of single-fold proteins. Proteins 2003;51:352–359.** **© 2003 Wiley-Liss, Inc.**

Key words: protein folds; proteomics; Monte Carlo

INTRODUCTION

The human genome sequencing project is one of the most ambitious undertakings in the history of science.[1–4] However, in terms of difficulty it may be overshadowed by the second phase of the project: the experimental determination of the structure of each of the gene products, coined "structural genomics."[5,6] Clearly, any theoretical modeling approach that would alleviate the need for the tedious experimental structure determination would prove to be exceptionally advantageous.

One such method is knowledge-based modeling, whereby one models a protein with an unknown structure based on the experimentally determined structure of a close homolog (for reviews see Johnson et al.[7] and Moult[8]). As the body of solved protein structures increases, the chances of finding a protein whose structure has been solved increase, bearing close sequence homology to the protein of unknown structure. However, what happens when no close homolog of the protein in question had its structure solved?

One possibility is to assume, with a certain positive probability, that the structure of the protein in question will in fact be similar to one of the solved proteins' structures, or, in other words, that the fold of the protein in question has been discovered before. If this is the case, then one can use any one of the unique folds in the protein database as a template on which to build the structure of the protein in question; this process is known as threading (for reviews see Jones and Thornton[9] and Bryant and Altschul[10]).

The reliability of this procedure clearly depends on the completeness of the current repository of known folds. If only a fraction of the folds found in nature have been identified, the subsequent ability to predict the structure of an unknown protein will be vanishingly small. However, if in studies so far, the vast majority of folds have already been identified, one can safely assume that threading algorithms will have a good chance of success.

The estimation of the overall number of folds in nature has been reported several times, using varying procedures, leading to a wide range of estimates: 400–8,000.[11–18] As an example, some procedures have plotted the number of

Grant Sponsor: Wellcome Trust (UK); Grant sponsor: UK Biotechnology and Biological Sciences Research Council; Grant sponsor: Israel Science Foundation; Grant Sponsor: National Science Foundation; Grant number: CCR-0098172; Grant sponsor: US-Israel Binational Science Foundation; Grant sponsor: NASA Ames Research Center; Grant number: NAG2-1325.

*Correspondence to: Isaiah T. Arkin, The Alexander Silberman Institute of Life Sciences, Dept. of Biological Chemistry, The Hebrew University, Guat-Ram, Jerusalem, 91904, Israel. E-mail: arkin@cc.huji.ac.il

Received 13 March 2002; Accepted 7 October 2002

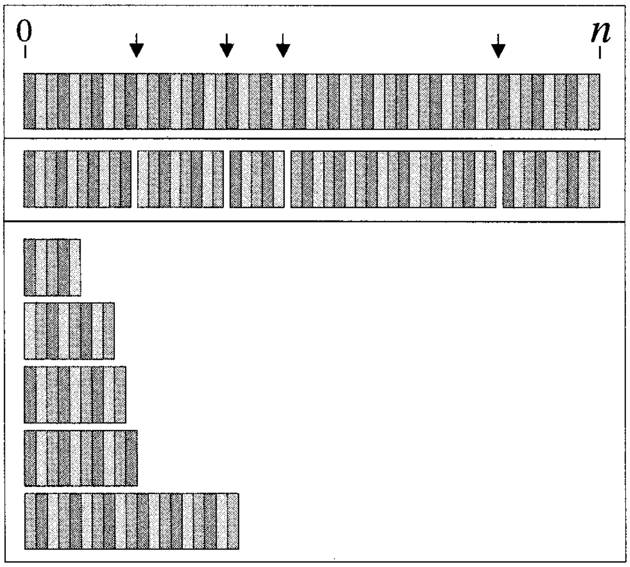

Fig. 1. The delineation of fold sizes. Top panel: a stick of length n is dissected four times in order to result in five folds. Middle panel: the resulting five folds. Bottom panel: the generated folds are arranged according to their sizes (i.e. the number of proteins).

unique folds as a function of the total number of protein structures solved.[11] The diminishing number of new folds identified can then be used to estimate the total number of folds. Other, more sophisticated statistical approaches have attempted to model the division of proteins into families and folds in the existing database and to extrapolate this model to the entire proteome.[15,16,18] Each of the above procedures has used a particular function to describe the distribution of proteins into folds, rather than relying on a systematic variation.

Herein, a different approach to the problem of estimating the total number of folds is taken. Using a Monte-Carlo simulation, the "curve of diminishing returns" is studied for a set number of proteins and any given number of folds. More importantly, the effects of two factors upon the "curve of diminishing returns" are investigated: (1) the folds distribution and (2) the sampling bias.

METHODS

In order to simulate the process of fold discovery, n proteins were divided into f folds. Subsequently, m proteins are randomly sampled, and the fold "affiliation" of the protein is registered. If this fold was not previously "identified" by another sampled protein, the number of known folds is incremented by one; otherwise, the number of previously identified folds remains unchanged. At the end of the simulation, a certain percentage of the folds will have been discovered.

Folds Distribution

The "broken stick model" was used in order to parse a given number, n, of proteins into f folds, as shown schematically in Figure 1. A stick of length n is cut at $f - 1$ breakpoints to yield f smaller sticks, with each stick representing one fold and the stick size representing the number of proteins in the fold. The breakpoints are determined by $f - 1$ random, distinct integers, $X_1, X_2, \ldots, X_{f-1}$, with $0 < X_i < n$. The $f + 1$ numbers $\{0, X_1, X_2, \ldots, X_{f-1}, n\}$ were sorted into an ascending array (of "order statistics" $0 = X^{(0)} < X^{(1)} < X^{(2)} < \ldots < X^{(f-1)} < X^{(f)} = n$), such that the ith fold corresponds to integers j between $X^{(i)}$ and $X^{(i+1)}$: $X^{(i)} < j \le X^{(i+1)}$.

As mentioned, the $f - 1$ breakpoints X_i are selected at random from the set $\{1, 2, \ldots, n - 1\}$ of integers, without replacement. This implies a discrete distribution analysis, thus the discrete random variable $Y = \min_{1 \le i \le f-1} X_i$ has the probability distribution function given by

$$\begin{aligned} P(Y \le x) &= 1 - P(Y > x) \\ &= 1 - P(\text{none of } 1, \ldots, x \text{ are selected}) \\ &= 1 - \frac{\binom{x}{0}\binom{n-1-x}{f-1}}{\binom{n-1}{f-1}} \\ &= 1 - \frac{n-1-x}{n-1} \cdot \frac{n-2-x}{n-2} \cdots \frac{n-x-f+1}{n-f+1}, \end{aligned} \tag{1}$$

for any $x \in \{1, 2, \ldots, n - x\}$. This probability distribution function describes the distribution for any one stick length, for the assumption of a uniform fold distribution.

If the random variables $X_1, X_2, \ldots, X_{f-1}$ are continuous uniform random variables on the interval (0, 1), then the probability density function, $f_y(x)$, for $Y = \min_{1 \le i \le f-1} X_i$ is readily computed:

$$P(Y > x) = \left(\frac{n-x}{n}\right)^{f-1}, \tag{2}$$

$$f_Y(x) = \frac{f-1}{n} \cdot \left(1 - \frac{x}{n}\right)^{f-2}. \tag{3}$$

Since Y is the length of the first stick, and each stick has the same distribution of length, f_Y also gives the probability density function for any one stick length. This can be used as a good approximation to the discrete case.

Non-uniform folds distributions

In order to simulate the effects of a nonuniform distribution, the random breakpoints were generated by nU^y, where U is a uniform random variable in the interval (0, 1). The case $y = 1$ gives the uniform distribution; if $y > 1$, though the breakpoints are more likely to occur close to the left end of the stick.

Sampling

In order to simulate the process of experimental structure determination, m random distinct integers, Z_i, were generated, with $0 < Z_i \le n$, where n is the total number of proteins (see Fig. 2). If the new protein is matched to a fold that was previously unidentified, the number of new folds

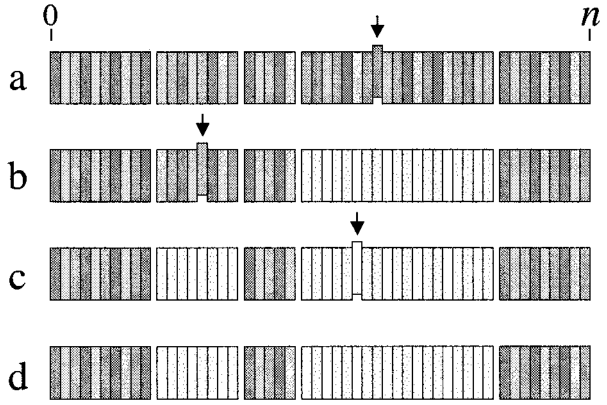

Fig. 2. The sampling of folds. *n* proteins are divided into 5 folds (see Figure 1). (a). Protein number 31 is selected at random, resulting in the identification of the fourth fold. The number of identified folds is now one. (b). A random protein (number 15) identifies a previously unidentified fold: fold number two. The number of identified folds is now two. (c). A third random protein (number 29) is selected, but this time it belongs to a previously identified fold (fold number four). Thus, the number of identified folds remains at two, as in (d).

identified is incremented; otherwise, the number of folds identified remains unchanged.

Biased sampling

Biased sampling was introduced in order to simulate the process of preferential selection (or avoidance) of previously identified folds. In other words, in the case of preferential sampling, new proteins that were selected at random are more likely to belong to previously identified folds, taking into account the compounded number of identified and unidentified folds so far. When sampling with a bias of $b > 1$ towards previously identified folds, the stick lengths belonging to previously identified folds are multiplied by b, thereby increasing the probability to sample the next protein from a previously discovered fold. The decision whether a new fold is discovered or not, is made by selecting a uniform random variable X, $0 \leq X < 1$. Then, if the following expression is true, the sampled protein falls into a previously discovered fold; otherwise, the sampled protein falls into a new fold.

$$X \leq \frac{d \cdot b}{d \cdot b + u}, \tag{4}$$

where d is the total size of all previously discovered folds (i.e., the total number of proteins in those folds), u is the total size of all unidentified folds, and b is the sampling bias (e.g., $b = 0.5$ for a twofold preference towards unidentified folds, $b = 3$ for a threefold preference towards previously identified folds).

RESULTS

Folds Distribution

As a representative test case, 100,000 proteins were divided into 1,000 folds using a Monte Carlo simulation of the broken stick model. Clearly, the aforementioned numbers of proteins or folds are not meant to symbolize the exact number found in nature, but rather are used solely for the purpose of the following analyses.

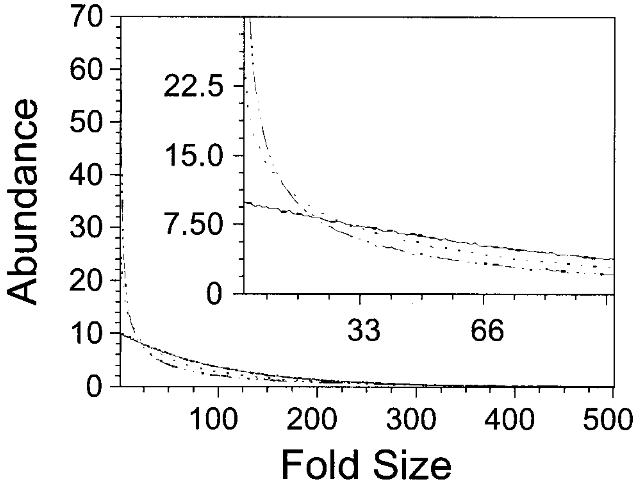

Fig. 3. Distribution of fold sizes. Solid line, uniform distribution, $y = 1$. Dotted line, nonuniform distribution, $y = 2$. Semi-dotted line, nonuniform distribution, $y = 3$. **Inset:** Magnification of the plots for fold sizes of up to 100 proteins per fold.

TABLE I. Total Percentage of Folds of Various Size Ranges, for the Different Distribution Types Used: $y = 1, 2, 3,$ and 20†

	Fold Size			
Folds Distribution	1–10	11–20	21–100	>100
$y = 1$	8.6	8.1	47	37
$y = 2$	16	11	41	33
$y = 3$	28	10	33	29
$y = 20$	79	3	7.1	12

†See Figure 3 and Methods. Shaded cells represent the results for a uniform distribution ($y = 1$).

The distribution of fold sizes is depicted in Figure 3. Table I lists the percentages of folds of various size ranges. Four distribution modes were employed: $y = 1$ (uniform distribution), $y = 2$, $y = 3$, and $y = 20$ (nonuniform distributions, see Methods). As expected, nonuniformity results in a higher percentage of folds having fewer proteins. Specifically, in the uniform distribution, slightly less than 10% of all folds contain 10 or fewer proteins. Most other fold size ranges have more proteins: 47% of the folds contain 21–100 proteins and 37% of the folds contain over 100 proteins per fold. However, in a nonuniform distribution with $y = 3$, the percentage of folds having 10 or less members rises to approximately 30% while the number of larger folds (i.e., folds with more than 20 members) decreases. In the extreme folds distribution in which $y = 20$, one can see that nearly 80% of the folds contain 10 or fewer proteins per fold.

Table II shows the percentage of proteins that belong to folds of a particular size range. As one can see, increasing the nonuniform of the distribution results in:

- An increase in the percentage of proteins in folds that contain a large number of proteins

TABLE II. Total Percentage of Proteins in Folds of Various Size Ranges, for the Different Distribution Types Used: $y = 1, 2, 3$, and 20†

	Fold size (%)			
Folds distribution	1–10	11–20	21–100	>100
$y = 1$	0.5	1.3	25	73
$y = 2$	0.8	1.6	21	76
$y = 3$	1.0	1.6	17	81
$y = 20$	1.0	0.4	3.5	96

†(see Figure 3 and Methods section). Shaded cells represent the results for a uniform distribution ($y = 1$).

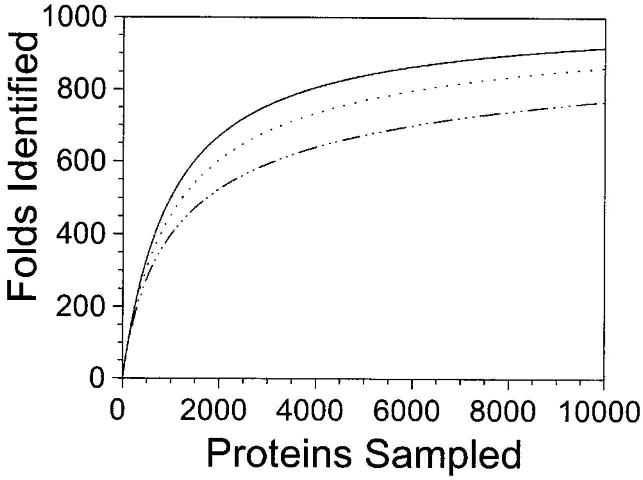

Fig. 4. Results of the unbiased ($b = 1$) sampling at different folds distributions. The solid line corresponds to the uniform distribution ($y = 1$), the dotted line corresponds to a non-uniform distribution with $y = 2$, and the semi-dotted line corresponds to a non-uniform distribution with $y = 3$.

- A mild increase in the percentage of proteins in small folds (1–10 size range).
- A decrease in the percentage of proteins in mid-size folds (i.e., 11–20, 21–100 fold sizes).

However, Table I shows a decrease in the percentage of large folds as the nonuniformity of the distribution increases. Tables I and II imply that there is a large amount of proteins divided into relatively a small number of big folds, and a small amount of proteins divided into a large number of much smaller folds (and mostly to single-protein folds).

Sampling

The results of the Monte Carlo simulation of the folds discovery process are shown in Figure 4 and listed in Table III. After randomly sampling 10,000 proteins (one tenth of the entire population), which were distributed uniformly, 917 folds were identified (92%) out of a total of 1,000 folds. When the folds distribution was not uniform (e.g., $y = 2$ or $y = 3$) a different result was obtained. When $y = 2$, 861 folds were identified, and when $y = 3$, only 768 were identified, a difference of about 15% from the uniform distribution case. In the extreme distribution case of $y = 20$, only 31% of the folds were identified after sampling one tenth of the protein population.

TABLE III. Percentages of All Folds Identified After Sampling 10,000 Proteins at Random, Using Different Values of the Sampling Bias Parameter b†

	Sampling bias b			
Fold Distributions	0.5	1	2	1,000
$y = 1$	96	92	85	9.3
$y = 2$	92	86	78	8.2
$y = 3$	77	84	68	6.8
$y = 20$	38	31	25	2.7

†The results are plotted in Figures 4–7. The shaded cell represents the results for the uniform distribution ($y = 1$) with unbiased ($b = 1$) sampling.

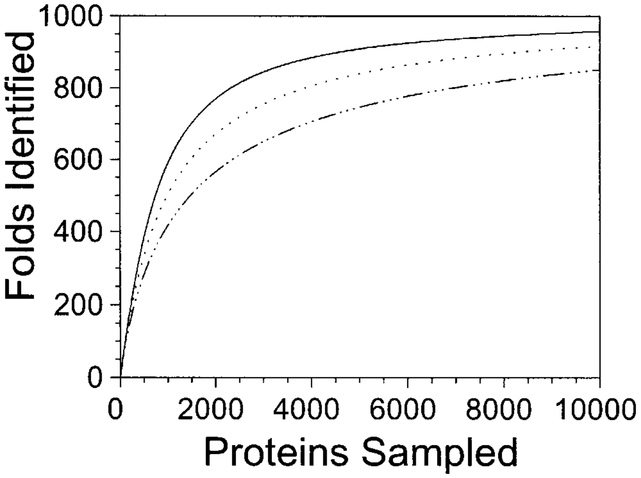

Fig. 5. Results of the biased sampling for a uniform folds distribution ($y = 1$). The dotted line corresponds to non-biased ($b = 1$) sampling, the semi-dotted line corresponds to sampling with a two-fold preference for previously sampled folds ($b = 2$), and the solid line corresponds to sampling with a two-fold preference for unidentified folds ($b = 0.5$).

Biased Sampling

A similar influence was obtained when the sampling was biased (see Methods), as shown in Figure 5 and listed in Table III. When the folds distribution was uniform, unbiased random sampling of 10,000 proteins resulted in the identification of 92% of the folds, while a twofold preference for sampling previously identified folds resulted in a reduction of identified folds to 85%. A twofold preference towards sampling unidentified folds ($b = 0.5$) resulted in an increase in the percentage of identified folds to 96%. A reduction of an order of magnitude in the fold identification (to 9.3%) was observed upon increasing the sampling bias toward previously identified folds to 1,000.

Biased Sampling and Non-Uniform Distributions

The results of the effects of combining a nonuniform folds distributions and biased sampling are shown in Figure 6 (constant distribution parameters $y = 2$ and $y = 3$) and in Figure 7 (constant sampling biases $b = 2$ and $b = 0.5$). Numerical values are listed in Table III. As expected,

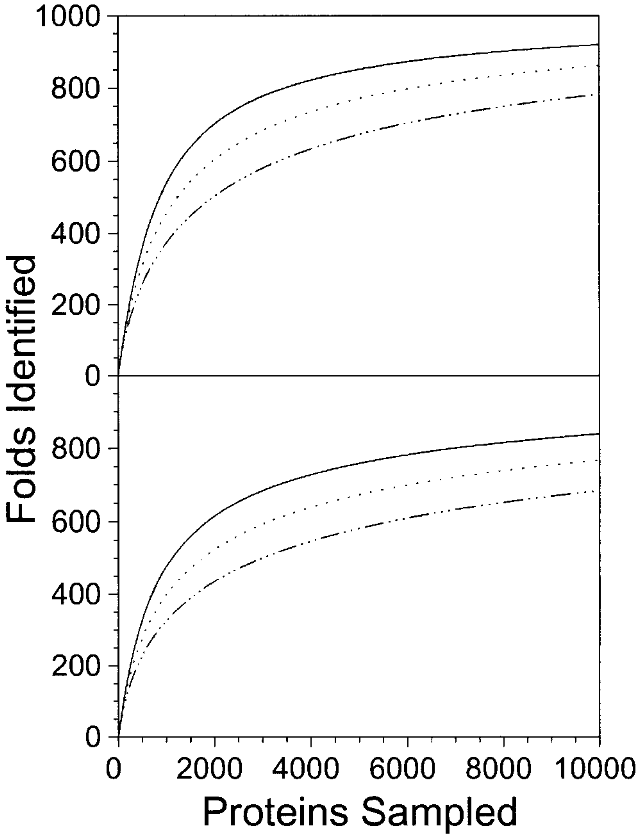

Fig. 6. Results of the biased sampling for non-uniform folds distributions: $y = 2$ (top panel) and $y = 3$ (bottom panel). The dotted line corresponds to non-biased sampling ($b = 1$). The semi-dotted line corresponds to sampling with a two-fold preference for previously sampled folds ($b = 2$). The solid line corresponds to sampling with a two-fold preference for unidentified folds ($b = 0.5$).

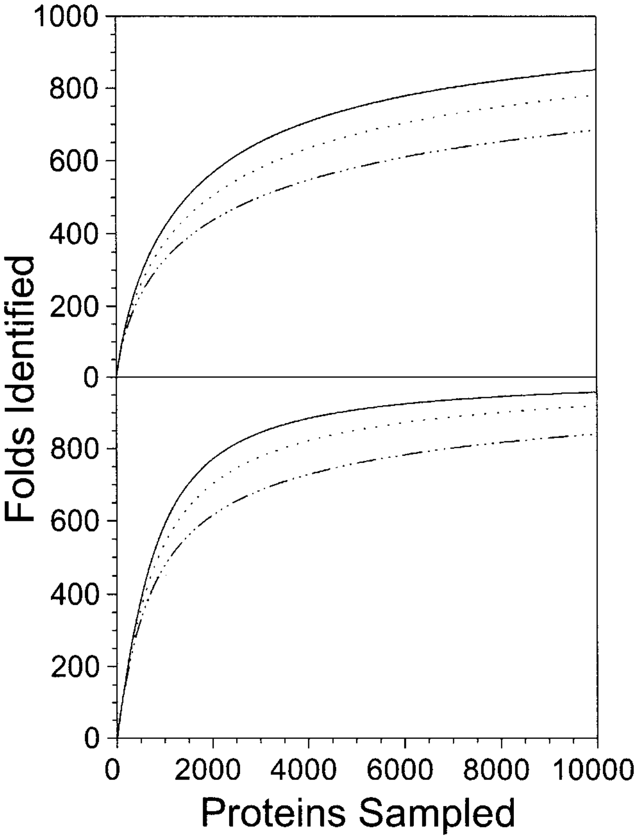

Fig. 7. Results of the biased sampling of uniform and nonuniform folds distributions. The semi-dotted line corresponds to uniform folds distributions, the dotted line corresponds to non-uniform folds distribution with $y = 2$, and the solid line corresponds to non-uniform folds distribution with $y = 3$. Top panel: Sampling is with a two-fold preference towards previously identified folds ($b = 2$). Bottom panel: Sampling is with a two-fold preference towards unidentified folds ($b = 0.5$).

the combined effects of preferential sampling of previously identified folds, along with nonuniform folds distributions, leads to a further reduction in the total percentage of folds identified.

For example, in the nonuniform folds distribution with $y = 2$ (Fig. 6, top), a twofold preference towards previously identified folds resulted in an 8% decrease in fold discovery, while an opposite twofold preference towards previously unidentified folds resulted in a 6% increase in fold discovery. Similar effects were obtained in the nonuniform distribution with $y = 3$ (Fig. 6, bottom).

DISCUSSION

The purpose of this study was not so much to estimate the number of possible folds, but rather to determine the effects of various factors upon the accuracy in estimating the number of folds. Towards that end, a theoretical model was constructed in which 100,000 different proteins were divided into 1,000 folds.[†] With this model at hand, it is possible to simulate the process of fold discovery: i.e., experimental structure determination, followed by counting the number of new folds identified. Furthermore, the simulation can be used to estimate the influence of two factors upon the fold identification process.

Folds Distributions

The folds distribution choice affects the way in which proteins are grouped into folds. For n proteins, the number of folds can be anywhere from 1 to n. Even for a given number of folds, f, the distribution with which proteins are grouped into these folds can vary significantly. Two extreme examples are: (1) each fold contains n/f proteins,[**] or (2) $f - 1$ folds each contain just a single protein and one fold contains $n - f + 1$ proteins (i.e., all the rest). In the first case, each fold may be discovered with an equal probability of (approximately) $1/f$, maximizing the total number of folds that may be discovered with m protein samples. In the latter case, there is a very low probability

[†] Both choices (i.e., the numbers of proteins and folds) are by no means stated to indicate the precise numbers found in nature. They only serve to construct our theoretical framework.

[**] More precisely, each fold contains either the rounded-down integer, $\lfloor n/f \rfloor$, or the rounded-up integer, $\lceil n/f \rceil$, so that the total number of proteins is exactly n.

$(1/n)$ to discover each of the small folds, and there is a high probability $((n - f + 1)/n)$ to discover the large fold; thus, it is very likely to discover the large fold, along with a few of the small folds. Such a folds distribution minimizes the total number of folds discovered. The real folds distribution is likely somewhere in between these two extreme examples.

Sampling Bias

The sampling bias affects the probability of preferentially sampling or avoiding a previously identified fold. There is one major reason for the inadvertent, preferential sampling of previously identified folds: The limitation of the experimental methods that are used to elucidate protein structures.

Both X-ray crystallography and solution NMR spectroscopy place considerable constraints upon the experimental conditions in which the proteins are studied. X-ray crystallography by its very nature requires crystallization of the target protein. Solution NMR spectroscopy is currently capable of studying proteins of limited size and requires that the protein remain soluble at millimolar concentrations. The ease with which these constraints are met will determine if the structure of the protein (and hence the experimental identification of its fold) will be rapidly discovered. If the right experimental conditions are not found, the structure of the protein cannot be solved and its fold remains unknown.

Why would experimental difficulties cause one to preferentially solve previously identified folds? A possible answer is that the scientific community solves the structures that it can solve, not necessarily the structures that it wishes to solve. It is, therefore, very likely that current experimental techniques are better at solving the structures of some folds and not others. This will result in the fact that some folds are preferentially sampled, while others are neglected/avoided due to experimental difficulties. Perhaps the best example of this phenomenon is the difficulty with which the structures of membrane proteins are solved.

One can also think of a reason for which previously identified folds may be preferentially avoided: The desire to investigate dramatically different proteins, assuming that they, indeed, do possess a new fold. While this may be a cause of negative preferential sampling, it may be balanced by the desire to solve the structure of similar proteins with distinct function, thereby further understanding the mechanisms of the protein.

Which of the above factors dominates in determining the sampling bias is unknown; however, it is the opinion of these authors that experimental difficulties that result in preferential sampling of previously identified folds prevail.

Folds Distribution

The Monte Carlo simulation employed herein used the broken stick model. Given a fixed number n of proteins and a fixed number f of folds, it generates a decomposition of a "stick" that represents the set of proteins into f pieces, using a set of $f - 1$ randomly generated breakpoints. Four different folds distributions are obtained by generating these $f - 1$ breakpoints according to four different distributions: a uniform distribution, and three nonuniform distributions corresponding to parameter values $y = 2$, $y = 3$, and $y = 20$ (see Methods). While none of these choices of folds distributions may correspond to the way in which proteins group into folds classes in nature, they do provide a simple conceptual framework to illustrate the effects of nonuniform folds distributions and nonrandom sampling effects upon our ability to estimate reliably the number of folds in nature.

As explained previously, as y increases, so does the number of small folds. Specifically, in the case of uniformly distributed breakpoints used to decompose 100,000 proteins into 1,000 folds, roughly 10% of all proteins belong to folds consisting of 10 proteins or less, while in the nonuniform distribution with $y = 3$, approximately 30% of all proteins belong to folds of size 10 or less. Therefore, the probability to discover such smaller folds decreases, and the total amount of folds identified after sampling 10,000 proteins decreases. This is shown in Figures 4–7. Using the folds distribution model described above, setting the parameter $y = 20$ resulted in a distribution similar to the extreme example (2), in that all of the breakpoints were nearly juxtaposed to one another (see Table II).

Sampling Bias

The effect of preferential sampling (and preferential avoidance) was measured by giving a twofold preference towards previously identified (towards previously unidentified) folds: Given the folds distribution, and the number of identified and unidentified folds, then the probability that the next sampled protein will be from a new fold is lower than the probability described in equation 4 (see Methods section), and therefore results in a reduction in the total number of folds discovered by the end of the simulation. This means that there is a lower probability to discover new folds, and thus a reduction in the total number of discovered folds by the end of the simulation.

Specifically, when 100,000 proteins were uniformly decomposed into 1,000 folds, a biased sampling of twofold preference towards previously identified folds caused a reduction of 7% in folds discovery. However, a twofold preference towards unidentified folds, under the same assumptions, gave only a 4% increase in folds discovery. This can be explained by the fact that the probability of discovering a new fold when 96% of them were already discovered is extremely low, so a greater number of samples (much greater than 10,000) is needed to sample a protein from a new fold. (This phenomenon occurs also in the related simplified probability model known as the "Coupon Collector's Problem"; see, e.g., Ross[11].) Increasing the sampling bias for previously identified folds to 1,000 resulted in lowering the number of identified folds by an order of magnitude to 9.3% (see Table III).

Combining the effects of nonuniform folds distribution and biased sampling leads to a further reduction in the number of discovered folds after 10,000 protein samples. The right side of equation 4 grows rapidly during the first quarter of the simulation (approximately 3,000 samples in a uniform distribution with nonbiased sampling) since the larger folds are usually identified faster. Its rapid growth marks new folds discoveries, so when its growth slows down, it indicates that fewer and fewer new folds are discovered, thereby marking a plateau in the fold discovery graph. Under the assumption of at least a twofold preference towards previously identified folds, the right side of equation 4 grows even faster than in the uniform, nonbiased case; thus a plateau is reached faster (after about 2,000–2,500 samples).

Estimation of Total Number of Folds

As already stated, the purpose of this study was not so much to estimate the total number of possible folds, but rather to understand the effects of various factors upon the estimation potential. Current estimates for the number of folds span a range of more than an order of magnitude, from 400 to 8,000 folds.[11–18] The results obtained in this study in the case of uniformly sampled proteins selected from a protein population uniformly distributed into folds indicate that sampling one tenth of the population results in identifying approximately 90% of the folds. The results obtained are not dramatically different if the model used (100,000 proteins divided into 1,000 folds) is altered. Furthermore, the results obtained regarding twofold preferential sampling (or avoidance) of previously identified folds does not alter the results substantially. Likewise, changing the nature of the distribution of proteins into folds did not affect the result substantially.

Thus, it may be possible to conclude that more recent estimates that point to a total number of folds of 1,000 or less[15–18] are more consistent with our Monte Carlo simulations. An increase from the aforementioned estimate for the total number of folds would reflect the possibility of the presence of at least one of the following two causes:

- The preferential sampling of previously identified folds is substantially greater than that considered in this study (i.e., $\geqslant 2$); or,
- The distribution of proteins into folds is dramatically different from a uniform distribution, in that many proteins belong to sparsely populated (i.e., unique) folds.

Which of the above factors might result in lowering the number of experimentally observed folds can be ascertained from studying what happens to the sampling efficiency under (1) extreme sampling biases or (2) extreme preferential sampling.

(1) Upon skewing the distribution, such that 96% of the folds contains ten or fewer proteins ($y = 100$), sampling 10% of the proteins, results in identifying only 15% of the folds. (2) Upon increasing the preferential sampling of previously identified folds to 1,000, a reduction in the number of identified folds by an order of magnitude (9%) was observed.

While it is difficult to estimate if such extreme distributions are indeed present in nature, it should be kept in mind that a preference for sampling previously identified folds of 1,000 should not be considered extraordinary. As a case in point, the rate at which the structures of membrane proteins are solved is roughly 500 times less than the rate at which water-soluble proteins are solved.[20]

CONCLUSIONS

In summary, using the Monte Carlo simulation and the broken stick model, a hypothetical model was built to examine the effects of two factors upon the process of fold discovery. Given an arbitrary number of proteins and folds, and using different arbitrary parameters for the folds distribution bias and sampling bias, different cases were simulated in which (1) folds are distributed uniformly and there is no preferential sampling; (2) folds are nonuniformly distributed with no preferential sampling; (3) folds are uniformly distributed but there is preferential sampling or avoidance, which attempts to simulate the process of experimental determination of structures; (4) folds are nonuniformly distributed in addition to preferential sampling or avoidance. Using the proposed model with these choices of parameters, one may conclude that a significant fraction of the total number of folds has been discovered. However, as this is only a theoretical model, one can make different assumptions to measure the impact of these factors, and, perhaps, if these differ substantially from those examined here, reach different conclusions.

ACKNOWLEDGMENTS

This work was supported by grants from the Wellcome Trust (UK), the UK Biotechnology and Biological Sciences Research Council, and the Israel Science Foundation to I.T.A. J. Mitchell is partially supported by grants from the National Science Foundation (CCR-0098172), the US-Israel Binational Science Foundation, and NASA Ames Research Center (NAG2-1325).

REFERENCES

1. Venter C, Adams MD, Myers EW, Li, PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji, RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li, Z, Li, J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang J, Wei MH, Wides R, Xiao C, et al. The sequence of the human genome. Science 2001;291:1304–1351.
2. Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, Brady SD, Bushard JM, Bustos VI, Chu A, Chung TR, De Witte A, Denys ME, Dominguez R, Fang NY, Foster BD, Freudenberg RW, Hadley D, Hamilton LR, Jeffrey TJ, Kelly L, Lazzeroni L, Levy MR, Lewis SC, Liu X, Lopez FJ, Louie B, Marquis JP, Martinez RA, Matsuura MK, Misherghi NS, Norton JA, Olshen A, Perkins SM, Perou AJ, Piercy C, Piercy M, Qin F,

Reif T, Sheppard K, Shokoohi V, Smick GA, Sun WL, Stewart EA, Fernando Tejeda J, Tran NM, Trejo T, Vo NT, Yan SCM, Zierten DL, Zhao S, Sachidanandam R, Trask BJ, Myers RM, Cox DR. A high-resolution radiation hybrid map of the human genome draft sequence. Science 2001;291:1298–1302.

3. Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olvier M, Conroy J, Kasprzyk A, Massa H, Yonescu R, Sait S, Thoreen C, Snijders A, Lemyre E, Bailey JA, Bruzel A, Burrill WD, Clegg SM, Collins S, Dhami P, Friedman C, Han CS, Herrick S, Lee J, Ligon AH, Lowry S, Moriey M, Narasimhan S, Osoegawa K, Peng Z, Plajzer-Frick I, Quade BJ, Scott D, Sirotkin K, Thorpe AA, Gray JW, Hudson J, Pinkel D, Ried T, Rowen L, Shen-Ong GL, Strausberg RL, Birney E, Callen DF, Cheng JF, Cox DR, Doggett NA, Carter NP, Eichler EE, Haussler D, Korenberg JR, Morton CC, Albertson D, Schuler G, De Jong PJ, Trask BJ. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature 2001;409:953–958.
4. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, Fulton R, Kucaba TA, Wagner-Mcpherson C, Barbazuk WB, Gregory SG, Humphray SJ, French L, Evans RS, Bethel G, Whittaker A, Holden JL, McCann OT, Dunham A, Soderlund C, Scott CE, Bentley DR, Schuler G, Chen HC, Jang W, Green ED, Idol JR, Maduro VVB, Montgomery KT, Lee E, Miller A, Emerling S, Kucherlapati R, Gibbs R, Scherer S, Gorrell JH, Sodergren E, Clerc-Blankenburg K, Tabor P, Naylor S, Garcia D, De Jong PJ, Catanese JJ, Nowak N, Osoegawa K, Qin S, Rowen L, Madan A, Dors M, Hood L, Trask B, Friedman C, Massa H, Cheung VG, Kirsch IR, Reid T, Yonescu R, Weissenbach J, Bruls T, Heilig R, Branscomb E, Olsen A, Doggett N, Cheng JF, Hawkins T, Myers RM, Shang J, Ramirez L, Schmutz J, Velasquez O, Dixon K, Stone NE, Cox DR, Haussler D, Kent WJ, Furey T, Rogic S, Kennedy S, Jones S, Rosenthal A, Wen G, Schilhabel M, Gloeckner G, Nyakatura G, Siebert R, Schlegelberger B, Korenberg J, Chen XN, Fujiyama A, Hattori M, Toyoda A, Yada T, Park HS, Sakaki Y, Shimizu N, et al. A physical map of the human genome. Nature 2001;409:934–941.
5. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
6. Stevens RC, Yokoyama S, Wilson IA. Global efforts in structural genomics. Science 2001;294:89–92.
7. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. Knowledge-based protein modeling. Crit Rev Bioc Mol Biol 1994;29:1–68.
8. Moult J. The current state of the art in protein structure prediction. Curr Opin Biotech 1996;7:422–427.
9. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
10. Bryant SH, Altschul SF. Statistics of sequence-structure threading. Curr Opin Struct Biol 1995;5:236–244.
11. Chothia C. One thousand families for the molecular biologist. Nature 1992;357:543–544.
12. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
13. Alexandrov NN, Go N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. Protein Sci 1994;3:866–875.
14. Wang ZX. How many fold types of protein are there in nature? Proteins 1996;26:186–191.
15. Wang ZX. A re-estimation for the total numbers of protein folds and superfamilies. Protein Eng 1998;11:621–626.
16. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284:1301–1305.
17. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins 1999;35:408–414.
18. 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.
19. Ross S. A first course in probability, 6th ed. Upper Saddle River, NJ: Prentice Hall; 2002.
20. Bowie JU. Stabilizing membrane proteins. Curr Opin Struct Biol 2001;11:397–402.