Scaling Law in Sizes of Protein Sequence Families: From Super-Families to Orphan Genes

Ron Unger,[1*] Shai Uliel,[1] and Shlomo Havlin[2]
[1]*Faculty of Life Science, Bar-Ilan University, Ramat-Gan 52900, Israel*
[2]*Department of Physics, Bar-Ilan University, Ramat-Gan, Israel*

ABSTRACT **It has been observed that the size of protein sequence families is unevenly distributed, with few super families with a large number of members and many "orphan" proteins that do not belong to any family. Here it is shown that the distribution of sizes of protein families in different databases and classifications (Protomap, Prodom, Cog) follows a power-law behavior with similar scaling exponents, which is characteristic of self-organizing systems. Since large databases are used in this study, a more detailed analysis of the data than in previous studies was possible. Hence, it is shown that the size distribution is governed by two exponents, different for the super families and the orphan proteins. A simple model of protein evolution is proposed, in which proteins are dynamically generated and clustered into families. The model yields a scaling behavior very similar to the distribution observed in the actual sequence databases, including the two distinct regimes for the large and small families, and thus suggests that the existence of "super families" of proteins and "orphan" proteins are two manifestations of the same evolutionary process. Proteins 2003;51:569–576.**
© 2003 Wiley-Liss, Inc.

Key words: protein families; size distribution; scaling; power-law; evolution

INTRODUCTION

It has been noted[1,2] that the universe of proteins is unevenly distributed. On the one hand, there are several "super-families"[3] that include many proteins and, on the other hand, there is a large number of "orphan" proteins[4] that do not show sequence similarity to other proteins. Super families and "orphan" proteins constitute the two extremes of the sizes of protein families. The wealth of genomic sequence data currently available makes it now possible to analyze, quantitatively, the entire distribution of sizes of protein sequence families.

The interest in studying this size distribution originated in the analysis of complex systems. It was found that a common behavior of certain complex systems is the emergence of a scale-free power-law distribution for representative quantities such as size, length, time, etc. Unlike normal or exponential distributions that decay very fast and thus have a characteristic scale, in power-law systems the representative quantities do not have a characteristic scale, hence the term scale-free. This behavior is often associated with self-organizing systems including, for example, the population and size of cities,[5,6] the size of traffic jams,[7] and the degrees of vertex connectivities of computer networks such as the World Wide Web,[8] and very recently, the connectivity pattern of metabolic networks.[9] In such systems, the distribution decays slowly with respect to the representative quantity (e.g., size) N with a power-law as

$$P(N) \propto N^{-\lambda}$$

where λ is the scaling exponent.[10]

The idea that growth proportional to size leads to power-law distribution is not new (see for example reference[11]), but it has gained a lot of interest recently when a mechanism based on two dynamic processes was suggested for the origin of such power-law distributions.[12] These processes are: (1) system expansion and (2) a requirement that new elements are added preferentially to sites that already contain more elements. In the following, we show that similar processes may be operative in protein evolution, i.e., the of sizes of protein families also follow power-law distributions. The observation that the size distribution of protein families in several organisms follows a power law was reported in several recent publications. Previous studies[13,14] analyzed the distribution of the sizes of protein families within single genomes (mostly microbial). In reference [13] the suggestion that sizes of protein families obey a power law was first made. It was argued[14] that due to the small sample size, it is not possible to determine the form of the distribution. A very recent publication[15] contributed significantly to establishing that fold occurrence in genomes follows a power-law distribution for 20 genomes analyzed individually.

Our results validate, in general, these observations but add more relevant details to the picture. We show that the results are independent on the classification scheme used and hold true not only for individual genomes but also when large general (i.e., not organism specific) databases are analyzed. The fact that we pooled together proteins from different organisms enables us to get a large enough

Grant Sponsor: First Foundation of the Israeli Academy of Science.

*Correspondence to: Ron Unger, Faculty of Life Science, Bar-Ilan University, Ramat-Gan, 52900, Israel.
E-mail: ron@biocom1.ls.biu.ac.il

Received 24 April 2002; Accepted 7 November 2002

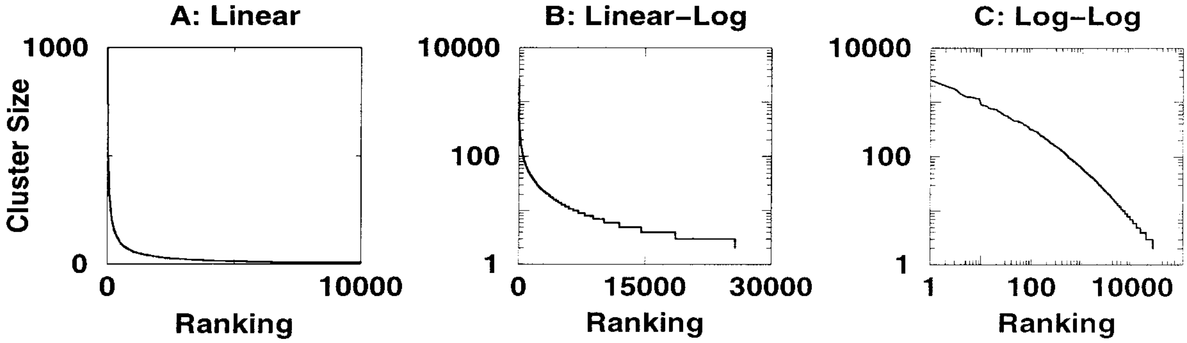

Fig. 1. The distribution of sizes of families in the Prodom database. The distribution of protein family sizes vs. their ranking (1 being the largest) is shown in three scales: (**A**) linear scale, i.e., the original data, (**B**) log-linear, (**C**) log-log. The log-linear plot (**B**) shows that the data do not follow an exponential distribution (otherwise, a straight line is expected). As discussed in the text, the log-log plot (**C**) shows a change in the slopes. When the large and small clusters are analyzed separately, they fit two straight lines with R^2 close to 1.

data set to perform a more detailed analysis and pay attention to small clusters, an aspect that was not studied carefully in previous studies.

We show that a simple model of evolution can generate protein families with size distribution similar to what is observed in the actual databases. We point out that the observed power-law distribution explains both the existence of super families and orphan genes, and thus suggest that they are highly related phenomena.

As an initial demonstration we show (Fig. 1) a plot of the entire classification of protein domains according to the Prodom[16] version 34.2. The plot covers 44,345 clusters, ranging in size from the largest cluster containing 2,574 domains to the smallest clusters: 18,579 clusters including only 2 proteins each. The data are shown as a plot of the ranking of each cluster (1 being the largest) vs. the number of proteins in each cluster. The data are shown in three different scales, the original data (i.e., linear scale) [Fig. 1(A)], log-linear [Fig. 1(B)], and log-log [Fig. 1(C)]. A straight line in a log-log plot is indicative of a power-law distribution. The graph of the observed data is not a perfect straight line, since it seems to be bent in the middle, i.e., there is a change of the slope of the power law between the small and the large clusters. We point out that such changes are typical to finite size power-law systems. A more detailed analysis below shows that this distribution follows, with a very good fit, two scaling exponents, one for the 500 largest clusters, and the other for smaller clusters. In the following, we study this scaling behavior in several classification systems of proteins and in a simple model of protein evolution.

PROTEIN CLASSIFICATION SYSTEMS

Several approaches have been developed in order to classify the ensemble of known proteins into families or clusters, using criteria based mainly on sequence similarities. We mainly study the distribution of sizes of protein clusters in three such systems: ProtoMap[17] (version 2.0), which uses a hierarchical clustering algorithm to classify full proteins found in the Swissprot database; Prodom[16] (version 34.2), which classifies protein domains (defined as segments that appear at least twice) found in Swissprot; and COG,[18] which only includes proteins from organisms whose entire genome has been sequenced (data from October 1999, which includes 21 complete genomes). We also used very recent data about proteins from the Human Genome analyzing the 30 largest human clusters (data from the CluSTr project of the European Bioinformatic Institute: (http://www.ebi.ac.uk/proteome).

TABLE I. Scaling Properties for Classification Systems

Database (size)	No. of clusters	λ_{500}[a]	R^2_{500}[b]	λ_{50}[c]	R^2_{50}[b]
Protomap (81,286)	13,353	−0.65	0.99	−1.83	0.97
Prodom (278,584)	44,345	−0.58	0.99	−1.98	0.99
Cog (28,033)	2,091	−0.44	0.99	−1.91	0.83
Model (80,000)[d]	12,463	−0.63	0.99	−1.90	0.95

[a] The scaling exponent for the 500 largest clusters.
[b] The square of the Pearson correlation coefficient of the corresponding log-log plot to a linear regression line.
[c] The scaling exponent for the frequency of the smallest size clusters. Protomap and the model have clusters of size 1–50, Prodom 2–50, and Cog 3–50.
[d] The model was run with the following three parameters: *M,* mutation rate of 20%; *S,* similarity level of 50%; and *k,* ratio of the survival probabilities P_1 and P_2 of 2.5.

ProtoMap[17] uses a hierarchical clustering algorithm based on sequence comparison of complete proteins. The clusters identified were shown to correspond well with a functional classification of protein families.[17] ProtoMap version 2.0 was used to classify 81,268 proteins from Swissprot, using high level clustering. At that level, ProtoMap produced 13,353 clusters (data shown in Table I). As suggested in Figure 1(C), there seems to be a change in the scaling exponent governing the distribution of sizes of large clusters and small clusters. Hence, we analyze first the largest 500 clusters. For the smaller clusters where there are many clusters of the same small size and ranking is not meaningful, an inverse measure[19] can be used to plot the frequency of clusters of a given size against the cluster size. The exponents of the distributions of the ranking, which we use for the largest 500 clusters, and the distribu-

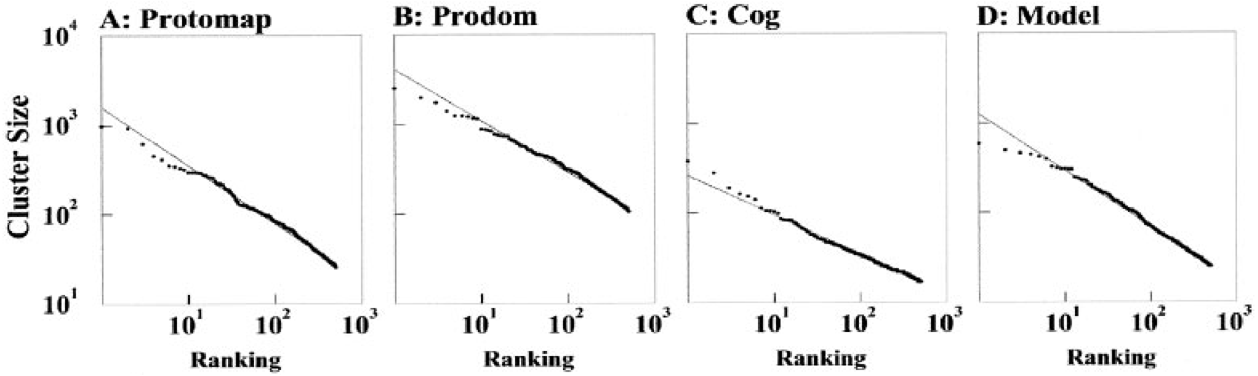

Fig. 2. Size distribution of protein clusters. The size distribution of the largest 500 clusters in four classifications is shown in a log-log plot, together with the regression line. The slopes are (**A**) Protomap: $\lambda_{500} \cong -0.65$; (**B**) Prodom: -0.58; (**C**) Cog: -0.44; (**D**) the model described here: -0.63. While the actual sizes of the clusters depend on the different sizes of the underlying databases, the slopes for all cases are very similar, and the fit to the regression line is high.

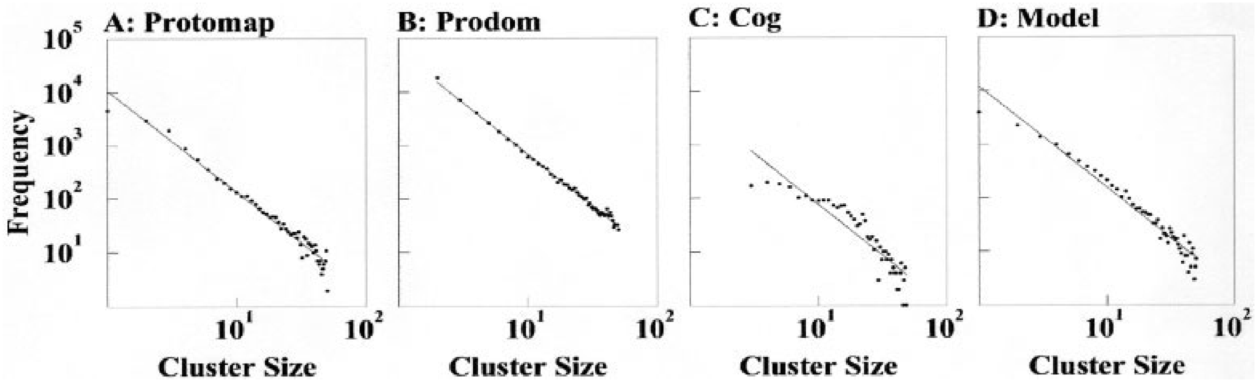

Fig. 3. Distribution of frequencies of small size clusters. In order to focus on the small clusters, the frequencies of clusters of the 50 smallest sizes are shown on a log-log plot, together with their regression line. The slopes are (**A**) Protomap $\lambda_{50} \cong -1.83$; (**B**) Prodom: -1.98; (**C**) Cog: -1.91; (**D**) the model: -1.90. The slopes of the Protomap and Prodom classifications as well as the slope of the model are very similar. The Cog database classifies only proteins that have similarities to other proteins, which may explain the decline in the number of clusters of very small sizes.

tion of size frequency, which we use for the 50 smaller sizes, are related[19] for an ideal power-law distribution by $\lambda_{50} = 1 + 1/\lambda_{500}$. We have computationally validated that, indeed, in a case of a perfect power-law distribution, the exponent for ranking the largest clusters and counting the number of clusters of each of the smaller sizes is related by this relationship. As we discuss below, for finite size systems, a deviation from the relationship of the slopes is expected.

Note that separating the analysis for the large and the small clusters does not lose data points since these two regions overlap. In Protomap, clusters that are ranked as number 216 or lower have less than 50 members and are shown in both representations. Thus, these two complementary representations cover together the entire range, and no data points are lost.

Figure 2(A) shows a log-log plot of cluster size vs. its size ranking (numbered sequentially, starting from the largest cluster) for the 500 largest clusters. The data show the distribution of cluster sizes scales as a power-law with $\lambda_{500} \cong -0.65$. The quality of the fit was evaluated by calculating the R^2 of the regression line to all the data points (500 for the large clusters, and 50 for the small sizes) in the log-log plot. The largest cluster (979 proteins) includes mainly protein kinases, the second cluster (933 proteins) represents mainly G-protein coupled receptors, and the third cluster (621 proteins) includes mainly globins. The size distribution of small clusters ranging in size from 1 to 50 members is shown in Figure 3(A). The log-log plot reveals that this distribution is governed by a power law with $\lambda_{50} \cong -1.83$ (with $R^2 \cong 0.97$).

ProDom is a systematic classification of protein domains[16] based on PSI-BLAST[20] sequence similarity searches. Domains are defined as protein segments that occur (above a threshold of similarity) at least twice. ProDom version 34.2 lists 44,345 such domains. The largest clusters identified by ProDom are: 2,574 domains from serine/threonine kinase proteins, 2,010 coiled coil repeat domains, and 1,765 zinc-finger domains. The scaling exponents for the 500 largest clusters and the 49 small sizes (sizes 2 to 50) are shown as log-log plots in Figures 2B and 3B, respectively. A summary of the analysis is given in Table I.

Finally, a similar analysis was performed using Cog (Clusters of Orthologous Groups of proteins). This is a classification system[18] for protein sequences encoded in 21 completely sequenced genomes (15 bacteria, 5 archaea, and yeast) representing 17 major phylogenetic lineages. As of October 1999, the database includes 28,033 proteins of known or assumed function (out of 43,918 proteins in the 21 organisms) classified into 2,091 clusters. The functions of the largest clusters in this database reflect the fact that mainly bacterial genomes are represented: permeases (384 proteins), methyltransferases (286 proteins), and histidine kinases (195 proteins). The corresponding log-log plots are shown in Figure 2(C) (for the largest 500 clusters) and in Figure 3(C) (frequency of clusters of size 3–50); data are shown in Table I.

It can be argued that the particular distribution of protein cluster sizes in the protein sequence databases is not due to the way proteins have evolved, but is a result of a bias in the choice of proteins for sequencing. Sequencing bias can be based on considerations like availability of genomic data, medical importance, and comparative studies across organisms and it can result in overrepresentation of certain families. For example, hemoglobin appears 562 times in the Swissprot database. This high number does not necessarily mean that hemoglobin form a large family. It happened because the gene is linked to many human diseases, and hence has been studied for a long time by many groups and was sequenced in many organisms. However, the possibility that the scale-free behavior for the large clusters is due to artificial over-representation of important proteins is ruled out by the COG database,[18] which represents only fully sequenced genomes. In such a database, each organism contributes its full genome to the database, and thus there is no artificial overrepresentation of certain protein families in this database. Still, it is likely that sequencing bias does play some role in the observed data, and makes it deviate from the ideal curve.

To further address the possible bias in the data, we used very recent data about the human genome that make it possible to quantify the distribution even within a single organism. Figure 4 shows the size distribution of the 30 largest human clusters (data from the CluSTr project of the European Bioinformatic Institute, URL: www.bioinfo.ebi/clustr). Again, the data show a good fit to a power-law, with a slope of −0.75 and R^2 of 0.97. Thus, we can conclude that power-law scaling is governing protein evolution both within a single organism, and within an ensemble of proteins sampled from various organisms.

The classifications described categorize a large number of different gene products using different clustering algorithms. Nevertheless, they all exhibit a very similar trend: very good fit of the data to a power-law over about two orders of magnitude each, with similar exponents (ranging from −0.44 to −0.65 for the largest clusters and from −1.83 to −1.98 for the small sizes) for the three databases. While the size of the few very large clusters (Fig. 2) or the frequency of the few clusters of very small size (Fig. 3), deviate somewhat from the straight line, the overall fit of the bulk of the data, as measured by R^2, is very good (Table I).

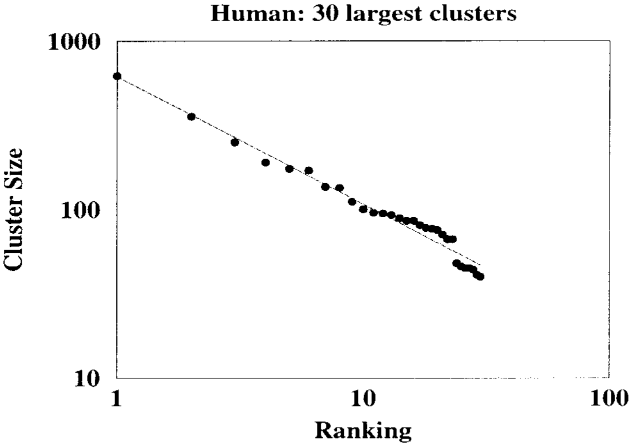

Fig. 4. Size distribution of the 30 largest clusters in human. Log-log plot of the size distribution of the 30 largest clusters in the human genome. The data are taken from analysis of the human genome project by the CluSTr project (www.bioinfo.ebi/clustr). The data show a good fit to a power-law with a slope of −0.75 and R^2 of 0.97.

The one noticeable exception from a power-law distribution is the decline in the frequency of very small size clusters in Cog [Fig. 2(C)]. The decline may be explained by the fact that clusters in Cog are based on seeds that include at least three proteins.[18] Furthermore, the requirement is that the proteins will come from organisms that are very different phylogenetically. As a result of this particular design of the clustering algorithm, Cog excludes rare proteins that appear only once or twice in the database, and also excludes proteins that appear only in a specific branch of the phylogenetic tree. Thus, it is not surprising to see the decline in the number of small clusters in COG, and the deviation of the slope of the distribution for the very small clusters.

To assess the quality of the fit to a power-law distribution, we tried to fit the data to three alternative distributions: log-normal, exponential, and stretched exponential. Both visual inspection [figure is only shown for the exponential case, Fig. 1(B)] and the parameters of the fit revealed that the fit to a power-law was much better than these alternatives: For example, in the case of the 500 largest clusters in Prodom, the R^2 of the quality of the fit to a power-law is 0.99. A fit to an exponential distribution yields a R^2 of 0.85, a fit to stretched exponential distribution yields a R^2 of 0.94, and a fit to a log-normal distribution yields a line with a slope of 1.2, which is different than a slope of 2, which is indicative of a log-normal distribution. We, of course, cannot exclude the possibility of piece-wise fitting to several distributions, but for a single distribution, the fit to power law seems to be the most reasonable.

THE MODEL

In order to explain these findings, we implemented and analyzed a simplified model of protein evolution that is based on biological principles similar to the preferential

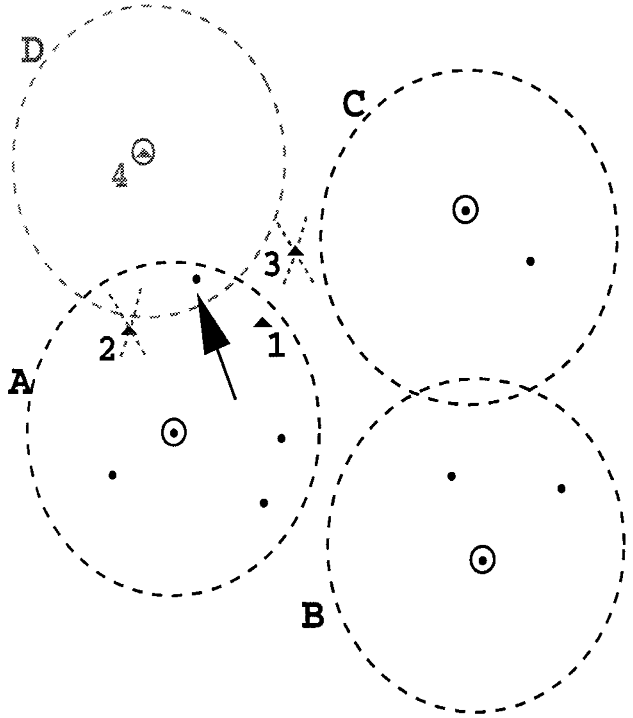

Fig. 5. A schematic view of the model. Each "family of sequences" (**A–C**) has a founder sequence (marked by a double circle) that is the prototype for its biological function. Assume that the sequence marked by the arrow from family A was selected, there are four possibilities for the new replicated sequence (marked as triangles). If it is similar enough to an existing prototype (here of family **A**): (1) with probability P_1, the new sequences is added to that family, and (2) with probability $1 - P_1$, it is discarded. If the new sequence is distinct from any existing prototype: with high probability $1 - P_2$, it is discarded, but with a small probability P_2 ($P_2 < P_1$) it establishes a new family (marked in gray as **D**) and becomes its biological prototype. Then, another sequence is selected to be replicated, and the process is iterated.

expansion processes suggested in Barabasi and Albert[12] for the World Wide Web. Somewhat similar models were used in references[14,15] to simulate single genome evolution. A schematic description of the algorithm is given in Figure 5. The model starts with a single "protein," a random sequence of amino acids. This protein is assigned particular prototypical biological function. For each iteration, one protein is selected at random, with a uniform distribution, to be replicated. The replication process introduces a certain level M of mutations into the new protein, where M actually represents an accumulation of a series of point mutations. Mutations are introduced as random substitutions of amino acids.

The fate of each protein is determined as follows: If its sequence similarity (defined here as the percent of identical residues) to one of the current prototypes (i.e., a founder of a cluster) is greater than a given similarity threshold, S, then with probability P_1 the new protein is considered to be a viable variant within the existing family and it is added to the existing cluster; alternatively, with probability $1 - P_1$, it is discarded. If the new protein differs from any existing prototype by more than S, it is less likely to survive. In this case, with a small probability P_2, ($P_2 < P_1$), the new protein is assumed to develop a novel biological function, and thus a new cluster is established with this protein serving as its prototype founder. Otherwise (probability $1 - P_2$), the newly created protein is discarded. The mutation and selection process is repeated until a predetermined number of proteins are obtained.

The model was run starting with a single protein (a random sequence of 100 amino acids) and stopped when 80,000 proteins (the current Swissprot size) were created; the distribution of cluster sizes was then analyzed. The model is governed by three parameters: *M, S,* and *k,* the ratio between the probability P_1 and P_2. (We find that the important factor is the ratio k between P_1 and P_2, and not their actual values). In the model, one obtains a power-law behavior only for an appropriate selection of these three parameters. For most values of these parameters, power-law behavior does not emerge. Longer runs, up to 300,000 proteins, yielded very similar results.

A systematic study of the parameter space was carried out by changing M and S in increments of 5%, and k in increments of 0.5 from 1 to 10. Out of these 8,000 combinations, only about 10% exhibit a power-law behavior as measured by the fit of the log-log plot to a linear line, R^2, of at least 0.95 for both the large and the small clusters. Only about 1% exhibit exponent values close to those determined for the actual biological databases.

In all of these cases, the scaling exponents for the large and small clusters deviated from the ratio described by the relation $\lambda_{50} = 1 + 1/\lambda_{500}$ expected for an infinite system governed by a single exponent.

Figure 6 shows that the parameter space that leads to power-law distribution tends to be clustered. The mutation rate M and the similarity within a family S form a "line" in the parameter space. When the parameters deviate to one side of the line to values that favor formation of new clusters, the model tends to form a huge number of single member clusters. When the parameters allow clusters to grow, either because the survival rate within clusters is high, or because many mutations are needed to break out from an existing cluster, then only a small number of large clusters are created. The third parameter *k,* which appears in Figure 6 as the radius of the circles around each point, basically determines the slope of the power-law behavior.

Figure 2(D) shows the distribution of cluster sizes obtained using parameters of $M = 20\%$, $S = 50\%$, and $k = 2.5$. These yield the best fit (Table I) to the observed distribution in the Protomap classification. These parameters yield, for the largest 500, a linear log-log plot with $\lambda_{500} \cong - 0.63$ ($R^2 \cong 0.99$) and for the frequency of the smallest size clusters [Fig. 3(D)], $\lambda_{50} \cong -1.90$ ($R^2 \cong 0.95$). The data show that the size distribution of both the small and large clusters is very similar to the observed distributions in the biological classifications.

Note that the fit to the slopes for the large clusters and the small sizes was always found together. In other words, we could not set the parameters to fit only the value for the large clusters and not the value for the small cluster or vice versa. Whenever the parameters yield a distribution that fit the actual value of the slope for the large clusters, the

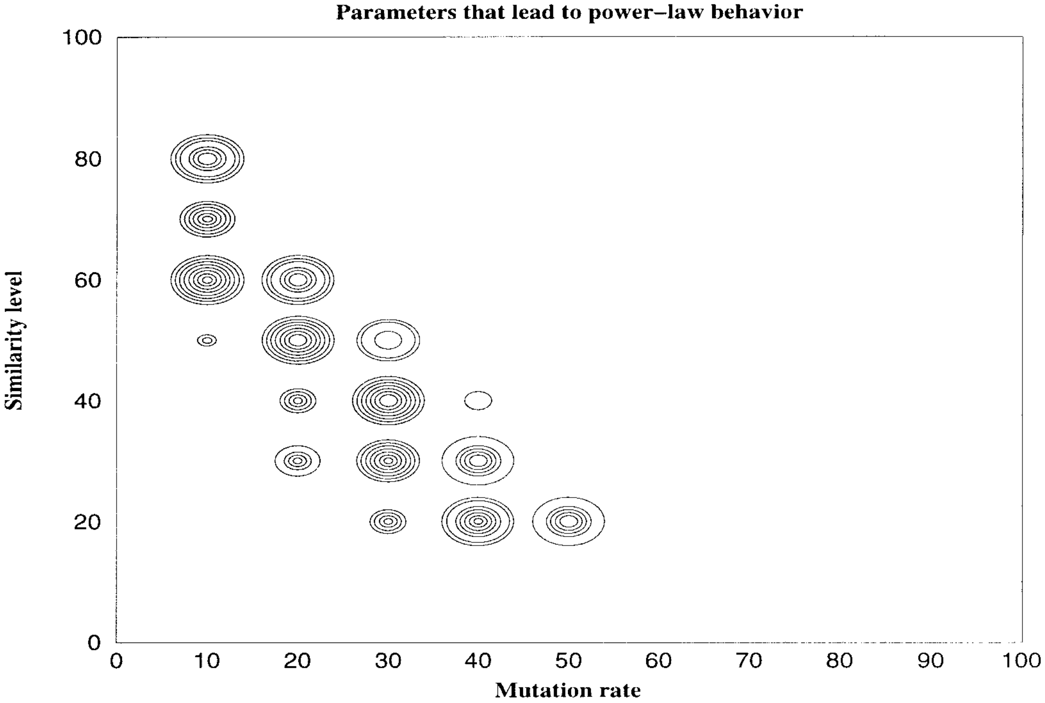

Fig. 6. The parameters that lead to power-law behavior cluster in a region of the parameter space of the model. Shown are the parameters *M*, the mutation rate vs. *S*, the similarity value within a family. The third parameter *k*, the ratio between the survival rate inside and outside a functional family, is indicated by the radius of the circle at each point. It can be seen that only relatively small numbers of combinations of *M* and *S*, which form a kind of line, lead to power-law behavior. For those values of *M* and *S*, most parameters of *k* are allowed. The values of *k* determine the slope of the power-law.

same set of parameters yields values that fit the actual value for the small clusters.

DISCUSSION

In this study, we have examined together, as one pool, the size distribution of protein families. Previous works have concentrated on studying individual organisms. We view these approaches as complementary.

Our approach is to analyze families in the entire universe of identified proteins. This approach has several advantages over a single genome analysis. First, it enabled us to analyze much more data and thus to determine the distribution over a wide range of families, from families with only several members to families with thousands of members.

Second, for each genome, the relative importance and abundance of each protein family is different, which might be the source of the different exponents observed in references.[13,15] Thus, the distribution within individual genomes does not necessarily imply that the combined distribution will follow the same form of distribution. Moreover, studying together the entire available protein database takes into account the interactions between genomes. Organisms do not evolve independently, and the genetic heritage of one organism is the starting point of the evolution of another. Furthermore, transferring genes between organisms, a phenomenon known as lateral gene transfer, is known to play a significant role in the evolution of microbial genomes,[21] and also in higher organisms as revealed by an initial analysis of the human genome.[22]

The similarity level within families and across neighboring families can be estimated from the database. Typically proteins within families have above 30% similarity, and proteins across related but different families have less than 20–25% similarity. These values are within the line of values that lead to power-law behavior. However, we realize that since the model is abstract and generic, there is no direct correlation between the value of parameters in the model and in Nature. This is evident since many combinations of parameters in the model lead to power-law distribution. However, a discussion of the parameter *k,* the survival ratio of mutated proteins within and outside their functional family, is interesting.

Although it might look as if staying within a functional family is much more probable than developing a new function, we think that the main problem of proteins that accumulated many mutations (say about 50%) is to maintain foldability and stability. Clearly, it is extremely rare for a protein to sustain such a high level of mutation and still fold.

However, the difference in survival rate is relevant only for proteins that accumulated many mutations and still fold to a stable conformation. For these highly mutated proteins, changing the function, to the level that will get them from one family to another, may not be so unusual compared to retaining the original function. The problematic issues in relating sequence similarity to functional similarity are described in Devos and Valencia,[23] where examples are shown for proteins that share function with very low sequence similarity in contrast to examples where proteins with high similarity (as high as 41% identity!) have different functions.

In the model, ratios between 1 to 10 in survival rates were scanned and found to be in the right range. Interestingly, the ratio that was found to yield power law slopes similar to the actual databases for most selections of M and S is around 2–3, quite similar to the ratio of sequence similarity (35–40 to 15–20) within and across families.

Clearly, the model is oversimplification of protein evolution. First, the model is assuming that several mutations accumulate before a decision about survival is made, in reality every single mutation is subject to such selection. Second, decisions about survival of new sequences are made in this model at random, while in the course of evolution functional selection plays a major role in determining protein fate. Third, the model operates on the level of individual proteins and does include neither the DNA level in which genetic mutations actually occur, nor the organism level in which mutations are selected. Nevertheless, the model does reproduce the scaling behavior of protein evolution, which suggests that as a conceptual model, it captures the essence of the growth aspect of this dynamic process. A similar model[15] uses three different parameters, the initial numbers of genes, the mutation rate, and the rate of fold acquisition, and yielded similar behavior. Thus, it seems that simple models, using a small number of parameters, can yield a behavior that is similar to the observed distribution of sizes of protein families in the database.

It is also interesting to note that the data consistently show that there is a change in the exponent for the largest and the smallest clusters. The exponents for the largest clusters are around 0.5. Since the exponents are related by $\lambda_{50} = 1 + 1/\lambda_{500}$,[19] this would predict that the exponents for the smaller clusters will be around 3. Nevertheless, in all three classification systems, an exponent of about 2 (see Table I) is observed for the small clusters. Such a change in slopes is actually typical to finite size power-law systems and is not unique to our data or model; for example, see the analysis in Dorogovtsev et al.[24] In an infinite power-law system, there will be one slope governing the behavior of all clusters. In any finite system, there is a change between the slope describing the distribution of the large and small clusters. The reason seems to be that for any system with a finite number of elements, when calculating the expected number of members for clusters of a given size, there is a point where this number starts to be very small (i.e., less than 1), and in this point the statistics are not good enough, clusters will not form for every size, and clusters that do form will include less members than they may have recruited if the system was bigger. In a sense, for a finite system, the large clusters will not have time to gain all the members, and as a result they will be smaller than expected. The emergence, in both the model and the actual datasets, of a similar change of slopes strengthen the claim that the behavior of the model is related to the behavior of the actual data, and that both follow a power-law distribution.

The biological meaning that one can draw from the model is that a family of proteins has to be large enough in order to allow for a rich repertoire of biological functions to further emerge. In our model, the survival of a mutated protein is arbitrarily determined and thus proteins that are created earlier have a much greater probability of forming larger clusters. In evolution, it is not only the age of a family that determines its size. For example, histones, which are considered to be ancient proteins, are not among the 50 largest clusters in Protomap, but immunoglobulins, which are much more recent, constitute the fourth largest family in Protomap.

These examples, of course, do not constitute a rule. In the list of large families, there is a mixture of families that are considered to be ancient like ABC transporters, ATP-related proteins, G-coupled receptors, and families like neurotransmitters, EGF receptors, and tubulins, which are considered to be much more modern. It seems as if biological function rather than age is the dominant factor. We suggest that the survival of a new protein is likely to depend on its biological function in the following way: If the new protein originated from a protein that performs a function for which variability is not required, then there is probably no selective advantage in maintaining several variants. If, on the other hand, the new protein evolved from a large family that already performs an intricate function for which variability is an asset, such as controlling other proteins, and has already developed an enriched repertoire of variants, it is more likely that an additional variant could be utilized, and thus such a protein is more likely to survive and eventually will be part of a large family.

A related idea was discussed recently in Park and Bolser,[25] where it was suggested that critical families of proteins are those families that interact with many types of proteins. Analysis of such interaction networks showed indeed that there are families that evolved to control many types of biological processes and thus become critical in their function and more diverse in their members.

A striking example can be seen in the abundance of kinases (more than 3,000 in Swissprot), proteins that are involved in phosphorylation, which is a key control mechanism in biological processes. We suggest that early in evolution, phosphorylation had only a slight advantage over alternative mechanisms (e.g., sulfation which is currently very rare). This small advantage led to the evolution of more phosphorylation proteins, and co-evolution of supporting machinery (such as ATP processing) created a larger ensemble of functions that further enabled an even larger family. Ultimately, this process led to the current abundance of phosphorylation proteins.

The notion that there are different levels of abundance of protein families is supported by a recent study that discusses the distribution of protein folds.[26] It was demonstrated that there are few super-folds with many members, an intermediate number of folds that appear several times, and a large number of unifolds that are unique. Thus, in addition to its importance to the ongoing debate about the proper counting of the total number of folds in

Nature, this work actually shows that power-law characteristic appear in the distribution of folds in the structural database. This finding is complimentary to the work we describe here that deals with families that are based on sequence similarity. Furthermore, the work of Coulson and Moult[26] might suggest a structural mechanism by which some protein families have grown to be much larger than others. It might be that some folds are more "adaptive" than others, thus these folds will tend to enable more variations, which, in turn, will allow additional modifications.

In particular, we suggest that the model explains the observations[1,2] that the universe of proteins is divided into families with uneven, not normal, size distribution. This distribution, which was shown here to be scale-free, yields clusters ranging from super-families to a large proportion of "orphan" genes with no known similarities to other proteins. It has been estimated[4] that a surprisingly high proportion (about 10–15%) of all genes found in most complete genomes are not homologous to any other known genes. Our model suggests that having a large number of very small clusters (i.e., orphan genes) maybe a natural result of the way the universe of proteins has evolved in a self-organizing manner. In our simulation (with the parameters given in Table I), 7,606 proteins (9.5%) belong to clusters of a size of three or less. These proteins are, almost by definition, "orphan" genes. By analogy, our model suggests that for real proteins a similar phenomenon occurs: the evolutionary process yields a large number of proteins whose sequence is different from any other. The structure and function of some of these proteins may have converged to those of known proteins, although some of these proteins may have unique structural and functional features.

In this regard, our model suggests a testable hypothesis: There are two possible explanations to the observation of the high percentage of orphan genes. One explanation is that the high percentage is mainly an artifact. We simply have not sequenced enough genes from enough organisms and thus there are large gaps in our view of the universe of proteins. The other explanation is that the underlying universe of proteins is fragmented and there are, indeed, isolated, unique sequences. Our model supports the second explanation, namely we predict that the percentage of "orphan" proteins will stay high (say at least 10–15%, as indicated by our results) even when more and more genomes will be sequenced and the size of the databases that are used for similarity searches will increase. This prediction can be already tested, to a certain level, as the percentage of unknown proteins discovered in each genome does not drop, and a preliminary search on the human genome revealed that at least a quarter of the proteins are not similar to other proteins.[27]

Essentially, the model provides a mechanism to explain how a very small advantage, either in function, stability, availability of substrate, etc., can be used as an *amplifier* to produce a huge difference in abundance of protein families. This evolutionary model links together both phenomena of "super families" and "orphan" genes as part of the continuum of size distribution of protein sequence families, and suggests that this distribution is a manifestation of the underlying evolutionary self-organizing process.

ACKNOWLEDGMENTS

We thank John Moult and Elisha Haas for stimulating discussions, and Joel L. Sussman for a critical reading of the manuscript. R.U. is partially supported by the First Foundation of the Israeli Academy of Science.

REFERENCES

1. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
2. Gogarten JP, Olendzenski L. Orthologs, paralogs and genome comparisons. Curr Opin Genet Dev 1999;9:630–636.
3. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
4. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics 1999;15:759–762.
5. Zipf GK. Human behavior and the principle of least efforts. Cambridge MA: Addison-Wesley; 1949.
6. Makse HA, Havlini S, Stanley HE. Modelling urban growth patterns. Nature 1995;377:608–612.
7. Nagel K, Paczuski M. Emergent traffic jams. Physical Rev E 1995;51:2909–2918.
8. Huberman BA, Adamic L. Growth dynamics of the World-Wide Web. Nature 1999;401:131.
9. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature 2000;407: 651–654.
10. Stanley HE, Ostrowskyi N, editors. Correlations and connectivity: geometric aspects of physics, chemistry and biology. Dordrecht: Kluwer; 1990.
11. Hill B, Woodroofe M. Stronger forms of Zipf's law. J Am Stat Assoc 1975;70:349–355.
12. Barabasi AL, Albert R. Emergence of scaling in random networks. Science 1999;286:509–512.
13. Huynen MA, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 1998;15:583–589.
14. Yanai I, Camacho CJ, DeLisi C. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett 2000;85:2641–2644.
15. 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.
16. Corpet F, Gouzy J, Kahn D. The ProDom database of protein domain families. Nucleic Acids Res 1998;26:323–326.
17. Yona G, Linial N, Linial M. ProtoMap: automatic classification of protein sequences, a hier-archy of protein families, and local maps of the protein space. Proteins 1999;37:360–378.
18. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–637.
19. Zipf GK. The psycho-biology of languages, an introduction to dynamic philology. Cambridge MA: MIT Press; 1965.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
21. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000;405:299–304.
22. Lander ES, et. al. Initial sequencing and analysis of the human genome. Nature 2001;495:860–921.
23. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
24. Dorogovtsev SN, Mendes JJF, Samukhin AN. Structure of growing networks with preferential linking. Phys Rev Let 2000;85:4633–4636.
25. Park J, Bolser D. Conservation of protein interaction network in evolution. Genome Inform 2001;12:135–140.
26. Coulson AF, Moult J. A unifold, mesofold, and superfold model of protein fold use. Proteins 2002;46:61–71.
27. Rubin GM. The draft sequences: comparing species. Nature 2001;495:820–821.