Proteins and Genes, Singletons and Species

Branko Kozulić
Gentius Ltd, Petra Kasandrića 6, 23000 Zadar, Croatia

Abstract

Recent experimental data from proteomics and genomics are interpreted here in ways that challenge the predominant viewpoint in biology according to which the four evolutionary processes, including mutation, recombination, natural selection and genetic drift, are sufficient to explain the origination of species. The predominant viewpoint appears incompatible with the finding that the sequenced genome of each species contains hundreds, or even thousands, of unique genes - the genes that are not shared with any other species. These unique genes and proteins, singletons, define the very character of every species. Moreover, the distribution of protein families from the sequenced genomes indicates that the complexity of genomes grows in a manner different from that of self-organizing networks: the dominance of singletons leads to the conclusion that in living organisms a most unlikely phenomenon can be the most common one. In order to provide proper rationale for these conclusions related to the singletons, the paper first treats the frequency of functional proteins among random sequences, followed by a discussion on the protein structure space, and it ends by questioning the idea that protein domains represent conserved units of evolution.

Introduction

One of the first issues encountered in the early studies of proteins was their large size. In 1936, under the assumption that a protein has molecular weight of 20,000, Swiss physicists Charles-Eugène Guye (who experimentally confirmed the prediction of Einstein's special theory of relativity about variation in the mass of electron with its speed) made several calculations with 2,000 atoms arranged in the protein molecule at varying degrees of asymmetry. At his favorite degree (0.9), the calculation showed that the probability of

formation of a particular protein molecule corresponded to one against 10^{321} [1]. Such estimates compelled French biophysicist Pierre Lecomte du Noüy to question any scenario of unguided origination of proteins, for this huge number of different protein molecules, if made, would have a volume many times larger than the volume of the whole universe [2, 3]. In 1953, as part of his Nobel lecture, Hermann Staudinger contrasted the chance of formation of a particular 100,000 molecular weight protein - one in 10^{1270} - to the number of water molecules present in Earth's oceans - a mere 10^{46} [4]. In 1957, Isaac Asimov calculated that if the whole universe were packed with neutrinos, and if each neutrino represented a computer generating per second one billion proteins each of a different sequence over the entire universe's life, the total number of proteins generated would have reached just 10^{179} [5].

Prominent mathematicians and biologists discussed this mathematical challenge to neo-Darwinian evolution at a special meeting in 1966 [6], but, as noted by Salisbury [7], the question is whether the attending biologists understood the nature and magnitude of the challenge. Over subsequent decades, the same challenge has been repeatedly raised by some scientists only to be diffused by others, until its relevance apparently became unclear. Thus physicist Charles Townes could remark: "The biologists may at first seem fortunate because they have not run into brick walls such as physicists hit in finding quantum or relativistic phenomena that are so strange and different. But this may be because biologists have not yet penetrated far enough towards the really difficult problems where radical changes of viewpoints may be essential" [8]. Here I argue that biologists have actually run into brick walls; hence it is time for radical changes of viewpoints.

Size of protein sequence space

One strategy for defusing the problem associated with the finding of functional proteins by random search through the enormous protein sequence space has been to arbitrarily reduce the size of that space. Because the space size is related to protein length (L) as 20^L, where 20 denotes the number of different amino acids of which proteins are made, the number of unique protein sequences will rapidly decrease if one assumes that the number of different amino acids can be less than 20. The same is true if one takes small L values. Dryden et al. used this strategy to illustrate the feasibility of searching through the whole protein sequence

space on Earth, estimating that the maximal number of different proteins that could have been formed on planet Earth in geological time was 4 x 10^{43} [9]. *In laboratory*, researchers have designed functional proteins with fewer than 20 amino acids [10, 11], but *in nature* all living organisms studied thus far, from bacteria to man, use all 20 amino acids to build their proteins. Therefore, the conclusions based on the calculations that rely on fewer than 20 amino acids are irrelevant in biology. Concerning protein length, the reported median lengths of bacterial and eukaryotic proteins are 267 and 361 amino acids, respectively [12]. Furthermore, about 30% of proteins in eukaryotes have more than 500 amino acids, while about 7% of them have more than 1,000 amino acids [13]. The largest known protein, titin, is built of more than 30,000 amino acids [14]. Only such experimentally found values for L are meaningful for calculating the real size of the protein sequence space, which thus corresponds to a median figure of 10^{347} (20^{267}) for bacterial, and 10^{470} (20^{361}) for eukaryotic proteins.

Protein structure space

Even a small protein composed of 100 amino acids comes from a set of 10^{130} different possible sequences. As Lau and Dill stated in 1990 (15), it is essentially impossible for chance to find a particular sequence in a set of such a magnitude, as is for a monkey dancing on a typewriter to produce a Shakespearean play. Because this general argument of low probability gained importance "as support for creationism" [15], Lau and Dill proposed the "structure" hypothesis according to which nature seeks only a compact protein conformation with the proper active site. This is an alternative to the view that nature "seeks" a particular sequence. Since proteins of many different sequences can attain one kind of compact conformation, the structure hypothesis reduced the searchable space, and was thus perceived to increase the likelihood of finding a functional protein by a random process, such as random mutations of neo-Darwinian evolution [15].

In the two decades since the above proposition, scientists have used various criteria to order protein structure space. The primary information about three-dimensional (3D) structure of proteins, obtained mainly using X-ray crystallography and NMR spectrometry, is deposited in Protein Data Bank (PDB). Today there are over 60,000 entries in this databank. The first online database SCOP (structural classification of proteins) was established in 1995, followed by CATH (class, architecture, topology and homology) in 1997 [16, 17]. Both classifications

rely on curators who delineate domains and folds within 3D structure of each individual protein, and both classifications bring in taxonomy to involve evolutionary relationships. Since the two basic entities of classification, domains and folds, are subjectively rather than mathematically derived, the recognition of new folds and the quantification of similarity among folds are difficult [18-23]. Current SCOP recognizes 1,195 and CATH 1,233 different protein folds.

While some argue that a protein fold, and its relationship to other folds, cannot be defined without considering the evolutionary context [23, 24], others define relationships between protein folds purely mathematically in terms of a continuous similarity curve. The number of folds sufficient for describing all protein structures then depends on the chosen similarity cut-off value [18, 25, 26]. Recently, a new classification was described based on supersecondary motifs (Smotifs), which are entities smaller than domains and folds. Smotifs are composed of the two regular secondary structure elements, α-helix and/or β-sheet, linked by a loop. In protein structures Smotifs come in various sequences and orientations with respect to each other. A finite set of 324 such Smotifs appears sufficient for structural classification of each folded protein of any possible sequence: the complete set of Smotifs was identified in the proteins whose structures were known for at least ten years prior to the Smotifs publications [27, 28]. Accordingly, proper description of the whole protein structure space is feasible without any reference to taxonomic relationships that are incorporated in the SCOP database. In another recent paper, the authors studied structural relationships of proteins in selected sets from Protein Data Bank and compared them to artificial, compact, hydrogen-bonded homopolypeptides. They concluded that connectivity features of the structure space stem from intrinsic macromolecular properties of proteins, and that all protein structural relationships can be fully explained without reference to any evolutionary assumptions [29]. Moreover, structure of a protein may be predicable based solely on the data about its amino acid composition [30].

Regardless of whether the 324 Smotifs or 1,233 folds - or a similar number of other basic elements - are sufficient for describing all 3D protein structures, the existence of the enormous number of possible protein sequences necessarily means that a structure defined by any particular fold or combination of Smotifs might be populated by a huge number of unique protein sequences. Instances of proteins having essentially identical structure but different sequences, with sequence similarity even below 10%, are well known [22, 31, 32].

Sequence similarity of 8-9% is characteristic for the proteins of random sequences [33]. Moreover, sequence-similar but structure-dissimilar protein pairs are also numerous in the Protein Data Bank [34]. Figure 1 shows two illustrative pairs. The finding of such orthogonal, independent relationship between structure and sequence is a conundrum for anyone trying to infer evolutionary relatedness (common ancestry) of two proteins from their 3D structural similarity. Exactly at which point would the 3D structural similarity begin to carry more weight than primary sequence similarity for inferring, or not inferring, common ancestry? Evidently, this inference is based on both the degree of shared 3D similarity and someone's sense of how unlikely it is that this similarity could have arisen independently [22]. While the similarity of protein structures can be described in various mathematical terms [35], estimation of the likelihood of independent origination requires evolutionary modeling.

All evolutionary models rely on how certain changes affect fitness. But is changing a protein fold beneficial or detrimental to fitness? Or, is *maintaining* a protein fold beneficial or detrimental? Under physiological conditions, native metamorphic proteins are known to exist in two alternative folds and both of them appear to be beneficial [36-38]. In contrast, when native human prion protein changes its fold, the change causes a deadly disease know as Creutzfeldt-Jakob disease [39]. Thus, fold changing can be beneficial, or it can be detrimental. Furthermore, some native functional proteins lack a defined 3D structure altogether, and thus belong to a group called "intrinsically unstructured proteins" [40]. In view of the above experimental findings, on what basis can one choose the sign and magnitude for the fitness effect due *solely* to, say, a RMSD 1.3 Å difference in the 3D structures of two proteins? And yet fitness estimates are essential for population genetics modeling. The results of such modeling can show whether a particular evolutionary scenario is feasible or not. In this respect, one should bear in mind that in a 300 amino acid protein there are 5,700 (19 x 300) ways for exchanging one amino acid for another, and that each one of these 5,700 possibilities points to a unique direction in the fitness landscape [41]. A single amino acid substitution can trigger a switch from one protein fold to another, but prior to that one, multiple substitutions in the original sequence might be necessary. Thus Alexander et al. [42] described 21 such prior substitutions; each one of them would have represented a crossroad with thousands of directions had these substitutions occurred *in vivo* instead of *in vitro*. Population genetics modeling becomes complicated when dealing with multiple amino acid substitutions in one protein [43-46]. Even more complicated would be the studies involving the fitness effects of multiple amino acid substitutions *and in addition* involving

the fitness effects due to 3D structural changes in a series of proteins undergoing such mutations. But, as a matter of principle, how can one possibly talk about a separate or additional fitness effect due to a 3D structural change if the protein sequence *determines* its structure, and the structure *determines* function and the function *determines* fitness? My literature search for publications describing evolutionary modeling based on fitness effects of protein structures gave no results. And according to a paper published in 2008: "the precise determinants of the evolutionary fitness of protein structures remain unknown" [47] – 18 years since Lau and Dill proposed the „structure hypothesis"[15]. On the other hand, in a number of papers it was shown that all relationships in the protein structure space can be described in purely mathematical terms [18, 25-28], and a most recent study concludes that „these results do not depend on evolution, rather just on the physics of protein structures" [29]. If all relationships in the protein structure space can be described fully without the need to invoke evolutionary explanations, then such explanations should not be invoked at all (Ockham's razor).

Frequency of functional proteins in protein sequence space

A single mutation, an insertion or a deletion, can in theory force a protein to switch its fold and acquire a new function, especially when the number of inserted or deleted nucleotides is not an integer of 3. Such mutations are known as frameshift mutations, as they completely change the amino acid sequence downstream of the mutation point. The probability that the new sequence is functional in combination with the unchanged upstream sequence correlates with the frequency of folds in the protein sequence space. While scientists generally agree that only a minority of all possible protein sequences has the property to fold and create a stable 3D structure, the figure adequate to quantify that minority has been a subject of much debate.

In 1976, Hubert Yockey estimated the probability of about 10^{-65} for finding one cytochrome c sequence among random protein sequences [48]. For bacteriophage λ repressor, Reidhaar-Olson and Sauer estimated that the probability was about 10^{-63} [49]. Based on β-lactamase mutation data, Douglas Axe estimated the prevalence of functional folds to be in the range of 10^{-77} to 10^{-53} [50]. A comparison of these estimates with those concerning the total number of protein molecules synthesized during Earth's history - about 10^{40} [9, 51, 52] - leads to the

conclusion that random assembling of amino acids could not have produced a single enzyme during 4.5 billion years [48, 53]. On the other hand, Taylor et al. estimated that a random protein library of about 10^{24} members would be sufficient for finding one chorismate mutase molecule [54]. Moreover, from an actual library of $6x10^{12}$ proteins each containing 80 contiguous random amino acids, Keefe and Szostak isolated four ATP binding proteins and concluded that the frequency of functional proteins in the sequence space may be as high as 1 in 10^{11}, allowing for their discovery by entirely stochastic means [55]. However, subsequent *in vivo* studies with this man-made ATP binding protein showed that it disrupted the normal energetic balance of the cell, acting essentially as an antibiotic [56]. One can conclude, therefore: had this protein been formed by random mutations, the cell with it would have left no descendants. Furthermore, the probability of its formation in a cell would have been lower than 10^{-11}, because random DNA mutations introduce stop codons and frameshifts whereas Keefe and Szostak avoided stop codons and frameshift mutations by experimental design [55]. The importance of distinguishing the results of *in vitro* from *in vivo* studies is highlighted by the finding that only a tiny fraction, one in about 10^{10}, of the active mutants of triosephosphate isomerase functioned properly *in vivo* [57]. It is also important to note that nucleotide binding protein families are among the most populous of all: the NAD(P)-binding Rossmann-like domains (CATH Code 3.40.50.720) include 70,263 different sequences, while the P-loop containing nucleotide triphosphate hydrolases (CATH Code 3.40.50.300) include 184,999 different sequences [58].

A "macromolecular miracle"

In general, there are two aspects of biological function of every protein, and both depend on correct 3D structure. Each protein specifically recognizes its cellular or extracellular counterpart: for example an enzyme its substrate, hormone its receptor, lectin sugar, repressor DNA, etc. In addition, proteins interact continuously or transiently with other proteins, forming an interactive network. This second aspect is no less important, as illustrated in many studies of protein-protein interactions [59, 60]. Exquisite structural requirements must often be fulfilled for proper functioning of a protein. For example, in enzymes spatial misplacement of catalytic residues by even a few tenths of an angstrom can mean the difference between full activity and none at all [54]. And in the words of Francis Crick, "To produce this miracle of molecular construction all the cell need do is to string together the

amino acids (which make up the polypeptide chain) *in the correct order*" [61, italics in original].

Let us assess *the highest probability* for finding this correct order by random trials and call it, to stay in line with Crick's term, a "macromolecular miracle". The experimental data of Keefe and Szostak indicate - if one disregards the above described reservations - that one from a set of 10^{11} randomly assembled polypeptides can be functional *in vitro*, whereas the data of Silverman et al. [57] show that of the 10^{10} *in vitro* functional proteins just one may function properly *in vivo*. The combination of these two figures then defines a "macromolecular miracle" as a probability of one against 10^{21}. For simplicity, let us round this figure to one against 10^{20}.

It is important to recognize that the one in 10^{20} represents *the upper limit*, and as such this figure is in agreement with all previous lower probability estimates. Moreover, there are two components that contribute to this figure: first, there is a component related to the particular activity of a protein - for example enzymatic activity that can be assayed *in vitro* or *in vivo* - and second, there is a component related to proper functioning of that protein in the cellular context: in a biochemical pathway, cycle or complex. Taking into account both contributions is an essential requirement because a synthetic protein nicely active in the test tube can be lethal in the cellular context, as shown by Stomel et al. for the ATP-binding protein of Keefe and Szostak [55, 56]. Substituting a man-made protein for a natural one might turn out to be easier than reported for triosephosphate isomerase, which is a key enzyme of the glycolysis pathway. One can therefore question the pertinence of combining the estimated contributions from two disparate studies on two unrelated proteins, but presently I am unaware of any other studies more relevant than these two. It is likely that relative contributions of the two components will differ from one future protein to the next, but the upper limit figure of one in 10^{20} is expected to remain valid.

In the context of protein sequences, the figure of one in 10^{20} means that along a polypeptide chain the identity of amino acids at only 15 positions would stay fixed; at each other position there could be any one of the 20 amino acids. With a 50 amino acid peptide, for example, the expectation is then to find 10^{45} functional sequences out of the 10^{65} (20^{50}) possible ones. That expectation seems unrealistic. With the median length in eukaryotes of 361 amino acids, the expectation to find 10^{450} functional proteins and only 10^{20} nonfunctional ones looks utterly

ridiculous. Thus, allowing for the probability of finding one functional protein among 10^{20} random sequences is obviously extremely generous, bordering on unreasonably generous. Nevertheless, for the sake of simplicity let us remain by this figure for "macromolecular miracle" and apply it to all proteins regardless of their length and cellular context.

To put the 10^{20} figure in the context of observable objects, about 10^{20} squares each measuring 1 mm^2 would cover the whole surface of planet Earth (5.1 x 10^{14} m^2). Searching through such squares to find a single one with the correct number, at a rate of 1000 per second, would take 10^{17} seconds, or 3.2 billion years. Yet, based on the above discussed experimental data, one in 10^{20} is *the highest* probability that a blind search has for finding among random sequences an *in vivo* functional protein. This figure denotes the minimal height of the brick wall.

Size of the currently known protein sequence space

One result of rapid advances in DNA sequencing technology is the acquisition of protein sequence data at an exponential rate: a recent extrapolation suggests that the number of known protein sequences will reach one trillion (10^{12}) in 2050 [62]. Currently, several online databases collect protein sequence information and provide various tools for data visualization and analysis. To mention just two of them: present (October 2010) SIMAP database contains over 39 million non-redundant protein sequences, compared to 23 million in September 2009 and 6 million in September 2007. SIMAP stands for the Similarity Matrix of Proteins, and it enables the comparison of a new sequence against all known ones, without biases due to taxonomy [63]. Gene3D provides structural annotation for proteins by assigning them domains from the CATH resource, containing currently (update 9.2.0) over 11 million sequences from 1,867 genomes, in contrast to about 4.5 million sequences from 527 genomes in September 2007 [64].

What have we learned from these tens of millions of protein sequences originating from the genomes of more than one thousand species? When proteins of similar sequences are grouped into families, their distribution follows a power-law [65-72], prompting some authors to suggest that the protein sequence space can be viewed as a network similar to the World Wide Web, electrical power grid or collaboration network of movie actors, due to the similarity of respective distribution graphs. There are thus small numbers of families with

thousands of member proteins having similar sequences, while, at the other extreme, there are thousands of families with just a few members. The most numerous are "families" with only one member; these lone proteins are usually called singletons. This regularity was evident already from the analysis of 20 genomes in 2001 [66], and 83 genomes in 2003 [69]. As more sequences were added to the databases more novel families were discovered, so that according to one estimate about 180,000 families were needed for complete coverage of the sequences in the Pfam database from 2008 [71]. Another study, published in the same year, identified 190,000 protein families with more than 5 members - and additionally about 600,000 singletons - in a set of 1.9 million distinct protein sequences [73].

Novel protein sequences and scaling in self-organizing networks

Systems having many interactive members, where the members are sometimes called nodes or vertices, are often depicted as a network in which connectivity among the members is best described by a scale-free power-law distribution. A power-law, and the related Zipf and Pareto laws [74], generally implies that weak phenomena occur extremely frequently, whereas strong phenomena occur extremely infrequently, so that the number (N) of phenomena with a given occurrence (F) declines according to $N \sim F^{-a}$. Illustrative examples of the phenomena include the Word Wide Web [75], urban growth of various cities [76], citations of scientific papers [77] and collaborations of movie actors [78]. Such a distribution was found to depend on two intrinsic mechanisms: first, networks expand continuously by the addition of new members; and second, the new members attach preferentially to those that are already well connected [78]. The above mentioned self-organizing networks are all associated with human activities, but some natural phenomena, like earthquakes, show the power-law distribution as well [79].

By plotting, on a log-log scale, the number of citations per paper against the total number of citations one obtains the graph shown in Figure 2a, characterized by a disperse tail and a dense head. At the tail, there are groups of small numbers of papers (1, 2, 3 and 4, approximately) achieving citations thousands of times. Only a few individual papers from this dataset approach the 10,000 citations mark. On the other hand, many papers are cited 100 times, even more of them 10 times, while the most numerous are the papers cited just once (apart from those never cited). An analogous plot of earthquake distribution shows many

earthquakes of low magnitudes, and an ever decreasing number of stronger earthquakes (Fig. 2b). Moreover, based on common appearance of actors in the same movie, actors' collaboration network also shows a power-law distribution (Fig 2c). At the tail there are a few superstars who collaborated with thousands of other actors, while newcomers at the head collaborated with just a few.

Distribution of protein families in sequenced genomes is illustrated by a similar graph (Fig. 2d). Comparable distributions have been observed with protein datasets from individual sequenced genomes [65, 80], as well as with the datasets that encompassed all sequenced genomes at various time points [66-72]. Here, at the tail of the distribution there are a few large families each consisting of thousands of proteins having similar sequences, while at the head there are many singletons. The evident similarity of this distribution curve with those of Figure 2a-c has been interpreted as evidence for self-organizing nature of protein networks in living organisms. It was thus inferred that the complexity of genomes grows in the same way as the complexity of WWW, or actors' network. These interpretations, however, are in error because they have failed to take account of a fundamental difference, as described below.

The first condition that the networks of Figure 2 must fulfill is a continuous addition of new members [78]. Thus, continuously new actors appear in movies, new earthquakes happen and new scientific papers get published. Roughly one person in 10^5 acts in a movie, earthquakes make one of less than 10^5 geological phenomena, and the fraction of scientific papers among all publications is higher than one in 10^5. So, to enter the respective network - to become the first point at the head of the distribution - the newcomers must overcome a barrier not higher than one against 10^5. After the entry, to become prominent the newcomers have a chance of about one in 10^5 again. Evidently, the two barriers, of entering and of becoming prominent, are comparable, give or take a few orders of magnitude. What would happen if the entry barrier were one thousand trillion (10^{15}) times higher? Obviously, if just one in 10^{20} persons could become an actor, we would know of no actors: there would be no records of them, and analogously, there would be no records of scientific papers and earthquakes. And without the records, no one could construct distribution graphs.

The frequency of functional proteins among random sequences is at most one in 10^{20} (see above). *The proteins of unrelated sequences are as different as the proteins of random sequences* [22, 81, 82] - *and singletons per definition are exactly such unrelated proteins*.

Thus, to enter the distribution graph as a newcomer (Fig. 2d), each new protein (singleton) must overcome the entry barrier of one against at least 10^{20}. After the entry, singleton's chance of becoming prominent, that is to grow into one of the largest protein families, is about one in 10^5 (Fig. 2d). Thus, it is much more difficult for a protein to become biologically functional than to become, in many variations, widespread: the entry barrier is at least fifteen orders of magnitude higher than the prominence barrier. This huge difference between the entry and prominence barriers is what makes the protein family distribution graph unique. In spite of this high entry barrier, in the sequenced genomes the protein newcomers (singletons) always represent the largest, most common, group: if it were otherwise, the distribution graph would break down. The mathematical models that incorporate data from all sequenced genomes in effect "spy" on nature [21]. With the help of one such model we have just uncovered something remarkable: in living organisms the most unlikely phenomenon can be the most common one. This feature clearly distinguishes the complexity of living organisms from the complexity of self-organizing networks.

Modeling of protein family distributions

Several research groups have attempted to model and explain various aspects of the observed power-law distributions. One key aspect relates to the origin of singletons, while the other concerns the growth of protein families. Huynen and Nimwegen argued that, to obey a power-law distribution, the protein families had to behave in a coherent fashion, that is, the probabilities of gene duplications within a family could not be independent; and likewise, the probabilities of gene deletions could not be independent either [65]. How such coordination might arise and be maintained was not explained. According to Gerstein and coworkers [66, 67], the observed distribution can be replicated only if two conditions are met: first, existing genes must be duplicated for expansion of existing families, and second, novel genes must be introduced by horizontal gene transfer or *ab initio* creation. Koonin and coworkers have developed several versions of their gene birth-death-and-innovation model (BDIM). The power-law distribution, however, could be reproduced only asymptotically, the family evolution time required billions of years when empirical gene duplication rates were brought in, the genes within a family needed to interact, and prodigious gene innovation rate was necessary for maintaining a high influx of singletons [83-87]. Horizontal gene transfer (HGT), rapid sequence divergence and *ab initio* gene creation were mentioned as the possible

sources of singletons. In another attempt, Hughes and Liberles proposed that just gene duplication and different pseudogenisation rates between gene families were sufficient for emergence of the power-law distribution [88]. The authors ruled out horizontal gene transfer and *ab initio* gene creation as the processes that could form new genes, because these processes were rare in eukaryotes but the power-law distribution was observed also with eukaryotic families. The evident problem with this study, however, is in that pseudogenisation per definition leads to a loss of function: the resulting power-law distribution of non-functional protein families is entirely different from the power-law distribution of functional protein families.

Horizontal gene transfer is common in prokaryotes but rare in eukaryotes [89-94], so HGT cannot account for singletons in eukaryotic genomes, including the human genome and the genomes of other mammals. For the origin of unique genes one has to turn to divergence of the existing sequences beyond recognition, or to *ab initio* creation, where the *ab initio* creation can happen either from non-coding DNA sequences present already in the genome or by introduction of novel DNA sequences into the genome. Regardless of which one of these three scenarios, or their combination, we consider, necessarily we come into the wasteland of random sequences or we must start from that wasteland: facing the probability barrier of one against at least 10^{20} cannot be avoided. The formation of each singleton requires surmounting this probability barrier. Without the incorporation of this probability, or perhaps another one that might be better supported by future experimental data, all models aiming to explain the observed protein family distribution will remain unrealistic.

The distribution of protein folds and domains also follows a power-law [21, 66, 67, 70, 72, 80, 83, 87], as predicted by Coulson and Moult [95]. That prediction was considered shocking [13]. Thus, in the sequenced genomes some domains are represented by thousands of different, non-homologous sequences, whereas other domains are represented by a few or by a single, unique sequence [21, 66, 67, 70, 72, 79, 83, 87, 95, 96]. For example, in a set of about 250,000 protein sequences Grant et al. found about 170,000 domains that remained as singletons [96]. These unique domains, called also orphan domains, represent the largest group among all domain groups that make the distributions. This is a feature in common with singletons from the distribution graph of protein sequence families.

Dokholyan et al. have attempted to explain their protein domain universe graph (PDUG) in terms of gene duplication and sequence divergence only [21]. In their explanation, however, implicit was the assumption that in the protein structure space there were just two alternatives: the old domain and a new domain, where each one of the two domains conferred functionality to the protein regardless of the sequence divergence. That assumption is not plausible because a vast majority of proteins would be non-functional after extensive divergence by random mutations. The authors used a cutoff value of 25% sequence identity for differentiating domains, corresponding to the sequence divergence of at least 75%. With the mean domain length of about 160 amino acids [97], the 75% divergence corresponds to 120 substitutions. Experimental data for proteins undergoing 120 substitutions are lacking, so it is currently impossible to provide any figure for the fraction of mutant proteins that might be expected to remain active. On the other hand, experimental data with fewer mutations show that the fraction of proteins retaining function declines exponentially with the increasing numbers of amino acid substitutions [98-101]. The exact percentage of the mutants remaining active is dependent on intrinsic properties of each starting protein; for example, only about 1% of the TEM1 β-lactamase and hen lysozyme mutants remained active after just 5 substitutions [100, 101]. Based on the above, with confidence one can only state that a large fraction of mutant proteins will be inactive following substitution of 75% of the original amino acids. As noted by Drummond et al. [99], exploration of distant regions of sequence space by random mutations alone appears highly inefficient. Mutations are supposed to arise and get fixed in a population sequentially; in order to estimate how probable this is for 120 substitutions, one would need a population genetics model that demonstrates the feasibility of so many substitutions in one single protein - but current models struggle typically with fewer than 10 substitutions [43-46]. In another study that modeled evolutionary dynamics in terms of stability of proteins, the probability of a stable protein native state - equivalent to protein functionality - among random sequences was taken to be 0.23 [102]. This figure is again much too high. In conclusion, all published models seeking to explain the power-law distribution of protein domains, or of protein sequence families, remain deficient unless they incorporate an experimentally supported figure for the probability of finding functional proteins among random sequences.

Singletons, orphans, ORF-ans, TRG-s and POF-s

In addition to the term singleton, other terms, with a similar if not synonymous meaning, have been used to denote proteins and genes having no relatives. Thus, Siew and Fischer define genomic ORFans as orphan open reading frames (ORF) with no significant sequence similarity to other ORFs [103, 104]. Wilson et al. suggest that orphans should be named "taxonomically restricted genes" (TRGs) [105, 106], and state that the abundance of orphan genes is amongst the greatest surprises uncovered by the sequencing of eukaryotic and bacterial genomes [105]. Earlier, Russell Doolittle affirmed that there are large numbers of unidentified genes in a variety of organisms, with the origin and function of these unique sequences remaining "baffling mysteries" [107].

In order to understand why the finding of singletons (ORF-ans, or TRG-s) represented such a great surprise, let us look at the contemporary expectations. They were possibly best outlined by Chothia et al. in 2003 [108]: "all but a small proportion of the protein repertoire is formed by members of families that go back to the origin of eukaryotes or the origin of the different kingdoms." And further: "The earliest evolution of the protein repertoire must have involved the ab initio invention of new proteins. At a very low level, this may still take place. But it is clear that the dominant mechanisms for expansion of the protein repertoire, in biology as we know it, are gene duplication, divergence and recombination." Consequently: "we will be able to trace much of the evolution of complexity by examining the duplication and recombination of these families in different genomes." About 1000 evolutionary independent protein families were expected to encompass all protein diversity [109]. In line with the above, there was an additional expectation of forthcoming grand unification of biology [110]. However, the power-law distribution of protein families and the sheer abundance of singletons have exposed utopian nature of these expectations and, at the same time, opened several important issues.

Siew and Fischer succinctly described the issues at stake: "If proteins in different organisms have descended from common ancestral proteins by duplication and adaptive variation, why is that so many today show no similarity to each other?" And further: "Do these rapidly evolving ORFans correspond to nonessential proteins or to species determinants?" [103].

A recent study, based on 573 sequenced bacterial genomes, has concluded that the entire pool of bacterial genes - the bacterial pan-genome - looks as though of infinite size, because every additional bacterial genome sequenced has added over 200 new singletons [111]. In agreement with this conclusion are the results of the Global Ocean Sampling project reported by Yooseph et al., who found a linear increase in the number of singletons with the number of new protein sequences, even when the number of the new sequences ran into millions [112]. The trend towards higher numbers of singletons per genome seems to coincide with a higher proportion of the eukaryotic genomes sequenced. In other words, eukaryotes generally contain a larger number of singletons than eubacteria and archaea.

When a relative to a singleton is found, together the two proteins create a family. In the absence of biochemical data, nothing can be said about biological function of that protein family as long as no established domain or structural motif is discernable from the amino acid sequences. Such proteins of obscure function, or POFs, make about 25% of the proteins found in each genome [113, 114]. POFs tend to be shorter than the proteins of defined function [114].

Today, almost ten years since the announcement of the first draft of the human genome sequence, no structural assignment is available for about 38% of human proteins [64]: at present we thus lack basic information about a large fraction of the proteins of human proteome [115]. In the initial publications on the sequence of the human genome, functional characterization of all proteins was recognized as one of the research priorities [116, 117], because understanding human biology is impossible without understanding the function of each individual protein. Subsequently, Richard Roberts called for a community-wide action in order to focus research efforts on complete characterization of the proteome of one organism [118]. In contrast, researchers from the Protein Structure Initiative have selected targets for structural characterization with little consideration about the species from which they come: the target had to belong to a large protein family, while proteins from the families with fewer than 10 members were explicitly excluded [58]. If other researchers followed these criteria, structural characterization of all human proteins would never be completed. Functional characterization of all human proteins is important not only for biological but also for commercial reasons, since uncharacterized human proteins represent an unexplored reservoir of drug targets for pharmaceutical and biotech industry [119]. When three-dimensional structures of ORFan proteins are determined, they often resemble previously

observed folds (120, 121). It should be noted that although a solved protein 3D structure represents an important piece of information, alone it is insufficient or even misleading for functional characterization of that protein [122-125]. Classic biochemistry is indispensible.

Cumulative changes in the total number of identified singletons, and their abundance in relation to other protein sequence families, can be followed from the studies that have periodically summarized advances in sequencing of the genomes of various species. Thus, in 2003, based on the data from 83 genomes, Enright et al. [69] identified 41,133 singletons from a total of 449,033 protein sequences. In this dataset the singletons made 9.2% of all proteins. By dividing the number of singletons with the number of genomes (41,133/83), we can see that there were on average 495 singletons in each genome. Interestingly, the same study reported that just 48 protein families were common to the genomes of all species. In this dataset, therefore, on average the unique proteins outnumber the common proteins by an order of magnitude (495 versus 48).

Based on the data from 120 sequenced genomes, in 2004 Grant et al. reported on the presence of 112,000 singletons within 600,000 sequences [96]. This corresponds to 933 singletons per genome. In 2005, Orengo and Thornton reported on the presence of about 150,000 singletons in 150 sequenced genomes [72]. In 2006, within 203 sequenced genomes and 633,546 non-identical sequences Marsden et al. identified 158,798 singletons [97]; thus the singletons made 24% of all sequences and there were on average 782 singletons in each genome. In 2008, Yeats et al. [73] found around 600,000 singletons in 527 species - 50 eukaryotes, 437 eubacteria and 39 archaea - corresponding to 1,139 singletons per species. No information about the number of singletons is available in the most recent summary of the data from over 1100 sequenced genomes encompassing nearly 10 million sequences [64]. In spite of the missing recent data on singletons, the results of the above calculations are sufficient for an unambiguous conclusion: each species possesses hundreds, or even thousands, of unique genes - the genes that are not shared with any other species. This conclusion is in full agreement with the power-law distribution of protein families discussed above.

Singletons as species determinants

A mere idea about the existence of species-specific genes was considered heretic as recently as in 2001 [126, 127]. However, with increasing number of fully sequenced genomes, the recognition and description of species-specific genes has become more and more frequent [113, 114, 128-132]. For example, Gollery et al. estimate that in the sequenced eukaryotic genomes the proteins of obscure function represent about one quarter of all proteins; delineating the origins and function of these species-specific proteins was deemed necessary for understanding an underlying cause of species specificity [113, 114]. In another study, based on the analysis of 1.28 million sequences from 198 genomes, the authors concluded that the majority of sequences were either highly conserved or specific to the species or taxon from which they derive [132].

Figure 3 shows how the number of unique genes (singletons), expressed as an average per each sequenced genome, was changing with the total number of the genomes sequenced. Evidently, the number of singletons tends to increase, from several hundreds to more than one thousand. *The presence of a large number of unique genes in each species represents a new biological reality.* Moreover, the singletons as a group appear to be the most distinctive constituent of all individuals of one species, because that group of singletons is lacking in all individuals of all other species. The conclusion that the singletons are the determinants of biological phenomenon of species then follows logically. In *System of Logic*, John Stuart Mill outlined his Second Canon or Method of Difference [133]: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensible part of the cause, of the phenomenon."

Until recently, most attention has been paid to the genes that are shared among species, instead to those that are different. But when the unique genes are studied, they are found to be the ones that are crucial for the very character of the species, or the whole taxon [134-136]. For example, in *Cnidaria* the proteins encoded by unique genes are essential for construction of stinging cells, the cells that are among the most sophisticated and complex of all cell types in the animal kingdom [134].

Folding of proteins – domains are not basic units of evolution

Structural annotation of proteins from newly sequenced genomes is typically successful for about 50% of all proteins [58, 64, 70, 128]. At first, this result seems surprising in view of the statements about near completeness, or 100% completeness, of the inventory of protein folds [27, 29, 137, 138]. In fact, that success rate is in accordance with the notion that many proteins with unrelated sequences acquire essentially the same 3D structure, as discussed above. The proteins of partially or largely disordered structure, as well as membrane proteins, also contribute to this group of non-annotated proteins [96, 121, 128]. Evidently, structures of just a fraction of novel proteins can be predicted by comparing their sequences against the sequences of those proteins whose 3D structures have already been solved. On the other hand, direct elucidation of 3D structures of new proteins by X-ray crystallography and NMR spectrometry is expensive and slow; structural genomics initiatives are expected to generate just 2,000 to 3,000 new structures in five years [20]. There is also a third way.

The amino acid sequence of a protein determines its structure, which in turn determines its function. In a cell, the structure forms mostly spontaneously by an interplay of attractive and repulsive forces among amino acid side chains, between them and the backbone and among various parts of the backbone, with the participation of hydrophobic interactions, hydrogen bonds, ionic bonds and van der Waals interactions [139-141]. Some proteins complete this folding process and acquire a native conformation in less than a microsecond, while others need seconds: the folding time thus varies over more than eight orders of magnitude [141]. Our understanding of this process has greatly increased during the past 20 years, but accurate prediction of three-dimensional structures of proteins, given just their amino acid sequence, remains a central challenge in computational biology and chemistry [142]. This problem is difficult because a polypeptide chain has many degrees of freedom: many conformational states are possible but the most stable is only a single one, being of the lowest free energy. That native state may be found using computational methods, of which Rosetta is the method most widely used. Rosetta@home operates with 150,000 computers, half of which run at any given time [142]. In general, 3D structure of a protein with up to about 120 amino acids can be solved, requiring sometimes only a few thousand runs. However, even hundreds of millions of computer runs would be insufficient for finding the native folded state of some proteins [142].

As a solution to the problem of limited CPU power for predicting the structure of a protein from its sequence, researchers have developed a scientific discovery game, Foldit. The game integrates human visual problem-solving capacity with computational algorithms. Recently Foldit demonstrated its value with some exciting examples of success, thanks to the efforts of >57,000 volunteers [143, 144]. The combination of human intelligence and computing power drives also the field of computational protein design. The achievements here indicate that scientists are beginning to master practical aspects of protein design [145-150].

The idea that protein domains represent conserved units of evolution [72, 108, 151-155] hinges upon the presumed capability of evolutionary processes - consisting of random mutations, recombination, genetic drift and natural selection [156] - to maintain the 3D structure of a protein while changing its amino acid sequence. These blind processes - which do not know what kind of protein 3D structure they start with, how they change it and in which direction in the structure space they go - thus supposedly possess certain capabilities that are by far superior to those of tens of thousands of computers, or superior to those of tens of thousands people using the computers.

That hypothesis - that evolution strives to preserve a protein domain once it stumbles upon it - contradicts the power-law distribution of domains. The distribution graphs clearly show that unique domains are the most abundant of all domain groups [21, 66, 67, 70, 72, 79, 82, 86, 94, 95], contrary to their expected rarity. Here I predict that the idea of protein domains as the basic units of evolution will be refuted directly by finding in the genome of one species two singletons having identical domain structure. Such a finding will represent the unambiguous and definitive refutation. That finding requires structural characterization of numerous singletons, and it depends on an objective, mathematical rather than a curator's, delineation of the protein structural elements and 3D identity.

Conclusions

The huge amount of DNA sequence data accumulated over the past decade has provided key insights about uniqueness of living organisms. The most important insight is that the genome of each species contains hundreds, or even thousands, of unique genes - the genes that are not

shared with any other species. The origin of species is thus intrinsically related to these unique genes.

Each unique gene, and accordingly each novel functional protein encoded by that gene, however, represents a major problem for evolutionary theory because unique proteins are as unrelated as the proteins of random sequences - and among random sequences functional proteins are exceedingly rare. Experimental data reviewed here suggest that at most one functional protein can be found among 10^{20} proteins of random sequences. Hence every discovery of a novel functional protein (singleton) represents a testimony for successful overcoming of the probability barrier of one against at least 10^{20}, the probability defined here as a "macromolecular miracle". More than one million of such "macromolecular miracles" are present in the genomes of about two thousand species sequenced thus far. Assuming that this correlation will hold with the rest of about 10 million different species that live on Earth [157], the total number of "macromolecular miracles" in all genomes could reach 10 billion. These 10^{10} unique proteins would still represent a tiny fraction of the 10^{470} possible proteins of the median eukaryotic size.

If just 200 unique proteins are present in each species, the probability of their simultaneous appearance is one against at least $10^{4,000}$. Probabilistic resources of our universe are much, much smaller; they allow for a maximum of 10^{149} events [158] and thus could account for a one-time simultaneous appearance of at most 7 unique proteins. The alternative, a sequential appearance of singletons, would require that the descendants of one family live through hundreds of "macromolecular miracles" to become a new species - again a scenario of exceedingly low probability. Therefore, now one can say that each species is a result of a Biological Big Bang; to reserve that term just for the first living organism [21] is not justified anymore. This view about species differs sharply from the predominant one according to which speciation is caused by reproductive isolation of two populations [159, 160] mediated by difficult to find speciation genes [161-163].

Evolutionary biologists of earlier generations have not anticipated [164, 165] the challenge that singletons pose to contemporary biologists. By discovering millions of unique genes biologists have run into brick walls similar to those hit by physicists with the discovery of quantum phenomena. The predominant viewpoint in biology has become untenable: we are witnessing a scientific revolution of unprecedented proportions.

References

1. Guye CE (1942) *L'Evolution Physico-Chimique – Les Frontières de la Physique et de la Biologie.* Herman & Cie (Paris) pp. 203-240.

2. Lecomte du Noüy P (1949) *The Road to Reason*. Longmans, Green and Co. (New York, London, Toronto).

3. Lecomte du Noüy P (1947) *Human Destiny*.: Longmans, Green and Co. (New York, London, Toronto).

4. Staudinger H (1953) Macromolecular chemistry. In: *Nobel Lectures, Chemistry 1942-1962*. Elsevier Publishing Company, 1964 (Amsterdam) pp. 397-419.

5. Asimov A (1957, with new material 1976) *Only a Trillion*. ACE Books (New York) pp 41-57.

6. Moorhead PS, Kaplan MM, eds. (1967) *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*. Wistar Institute Press (Philadelphia).

7. Salisbury FB (1969) Natural selection and the complexity of the gene. Nature 224: 342-343.

8. Townes CH (1998) Logic and uncertainties in science and religion. In: *Science and Religion: The New Consonance.* Peters T, ed. Westview Press, Inc. pp. 296-309.

9. Dryden DTF, Thomson AR, White JH (2008) How much of protein sequence space has been explored by life on Earth? J R Soc Interface 5: 953-956. doi:10.1098/rsif.2008.0085.

10. Walter KU, Vamvaca K, Hilvert D (2005) An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem 280: 37742-37746. doi:10.1074/jbc.M507210200.

11. Tanaka J, Doi N, Takashima H, Yanagawa H (2010) Comparative characterization of random-sequence proteins consisting of 5, 12 and 20 kinds of amino acids. Protein Sci 19: 786-795. doi:10.1002/pro.358.

12. Brocchieri L, Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. 33: 3390-3400. doi:10.1093/nar/gki615.

13. Rost B (2002) Did evolution leap to create the protein universe? Curr Opin Struct Biol 12: 409-416.

14. Bang ML, Centner T, Fornoff T, Geach AJ, Gotthardt M, McNabb M, Witt CC, Labeit D, Gregorio CC, Granzier H, Labeit S (2001) The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. Circ Res 89:1065-1072.

15. Lau KF, Dill KA (1990) Theory for protein mutability and biogenesis. Proc Natl Acad Sci USA 87:638-642.

16. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.

17. Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM (1997) CATH: a hierarchic classification of protein domain structures. Structure 5: 1093-1108.

18. Sippl MJ (2009) Fold space unlimited. Curr Opin Struct Biol 19:312-320. doi:10.1016/j.sbi.2009.03.010.

19. Sadreyev RI, Kim BH, Grishin NV (2009) Discrete-continuous duality of protein structure space. Curr Opin Struct Biol 19: 321-328. doi:10.1016/j.sbi.2009.04.009.

20. Redfern OC, Dessailly B, Orengo CA (2008) Exploring the structure and function paradigm. Curr Opin Struct Biol 18: 394-402. doi:10.1016/j.sbi.2008.05.007.

21. Dokholyan NV, Shakhnovich B, Shakhnovich EI (2002) Expanding protein universe from the biological Big Bang. Proc Natl Acad Sci USA 99: 14132-14136. doi:10.1073/pnas.202497999.

22. Pearson WR, Sierk ML (2005) The limits of protein sequence comparison? Curr Opin Struct Biol 15: 254-260. doi:10.1016/j.sbi.2005.05.005.

23. Taylor WR (2007) Evolutionary transitions in protein fold space. Curr Opin Struct Biol 17:354-361. doi:10-1016/j.sbi.2007.06.002.

24. Valas RE, Yang S, Bourne PE (2009) Nothing about protein structure classification makes sense except in the light of evolution. Curr Opin Struct Biol 19: 329-334. doi:10.1016/j.sbi.2009.03.011.

25. Suhrer SJ, Wiederstein M, Gruber M, Sippl MJ (2009) COPS – a novel workbench for explorations in the fold space. Nucleic Acids Res 37: W539-W544.

26. Sippl MJ, Suhrer SJ, Gruber M, Wiederstein M (2008) A discrete view on fold space. Bioinformatics 24:870-871. doi:10.1093/bioinformatics/btn020.

27. Fernandez-Fuentes N, Dybas JM, Fiser A (2010) Structural characteristics of novel protein folds. PLoS Computational Biology 6:e1000750. doi:10.1371/journal.pcbi.1000750.

28. Fernandez-Fuentes N, Oliva B, Fiser A (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 34:2085-2097.

29. Skolnick J, Arakaki AK, Lee SY, Brylinski M (2009) The continuity of protein structure space is an intrinsic property of proteins. Proc Natl Acad Sci USA 106:15690-15695. doi:10.1073/pnas.0907683106.

30. Rackovsky S (2009) Sequence physical properties encode the global organization of protein structure space. Proc Natl Acad Sci USA 106:14345-14348. doi: 10.1037/pnas.0903433106.

31. Gao J, Li Z (2010) Uncover the conserved property underlying sequence-distant and structure-similar proteins. Biopolymers 93: 340-347. doi:10.1002/bip.21342.

32. Cheng H, Kim BH, Grishin NV (2008) MALISAM: a database of structurally analogous motifs in proteins. Nucleic Acids Res 36: D211-D217. doi:10.1093/nar.gkm698.

33. Rost B (1997) Protein structures sustain evolutionary drift. Fold Des 2: 519-524.

34. Kosloff M, Kolodny R (2007) Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71: 891-902. doi:10.1002/prot.21770.

35. Hasegawa H, Holm L (2009) Advances and pitfalls of protein structural alignment. Curr Opin Struct Biol 19: 341-348. doi:10.1016/j.sbi.2009.04.003.

36. Murzin AG (2008) Metamorhic proteins. Science 320: 1725-1726. doi:10.1126./science.1158868.

37. Bryan PN, Orban J (2010) Proteins that switch folds. Curr Opin Struct Biol 20:482-488. doi: 10.1016/j.sbi.2010.06.002.

38. Gambin Y, Scug A, Lemke EA, Lavinder JL, Ferreon ACM, Magliery TJ, Onuchic JN, Deniz AA (2009) Direct single-molecule observation of a protein living in two opposed native structures. Proc Natl Acad Sci USA 106: 10153-10158. doi:10.1073/pnas.0904461106.

39. Colby DW, Prusiner SB (2011) Prions. Cold Spring Harb Perspect Biol 3:a006833. doi:10.1101/cshperspect.a006833.

40. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nature Reviews: Molecular Cell Biology 6:197-208. doi:10.1038/nrm1589.

41. Bloom JD, Arnold FH (2009) In the light of directed evolution: pathways of adaptive protein evolutions. Proc Natl Acad Sci USA 106: 995-1000. doi:10.1073/pnas.0901522106.

42. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci USA 106: 21149-21154. doi:10.1073/pnas.0906408106.

43. Behe MJ, Snoke DW (2004) Simulating evolution by gene duplication of protein features that require multiple amino acid residues. Protein Sci 13: 2651-2664. doi:10.1110/ps.04802904.

44. Lynch M (2010) Scaling expectations for the time to establishment of complex adaptations. Proc Natl Acad Sci USA 107: 16577-16582. doi:10.1073/pnas.1010836107.

45. Lynch M, Abegg A (2010) The rate of establishment of complex adaptations. Mol Biol Evol 27: 1404-1414. doi:10.1093/molbev/msq020.

46. Axe DD (2010) The limits of complex adaptation: an analysis based on a simple model of structured bacterial populations. Bio-Complexity 4: 1-10. doi:10.5048/BIO-C.2010.4.c.

47. Zeldovich KB, Shakhnovich EI (2008) Understanding protein evolution: from protein physics to Darwinian selection. Annu Rev Phys Chem 59: 105-127.

48. Yockey HP (1977) A calculation of the probability of spontaneous biogenesis by information theory. J Theor Biol 67:377-398.

49. Reidhaar-Olson JF, Sauer RT (1990) Functionally acceptable substitutions in two α-helical regions of λ repressor. Proteins 7:306-316.

50. Axe DD (2004) Estimating the prevalence of protein sequences adopting functional enzyme folds. J Mol Biol 341:1295-1315. doi:10.1016/jmb.2004.06.058.

51. Eden M (1967) Inadequacies of neo-Darwinian evolution as a scientific theory. In: Moorhead PS, Kaplan MM, eds. *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*. Wistar Institute Press. (Philadelphia) pp 109-111.

52. Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal JE (2009) The origin, evolution and structure of the protein world. Biochem J 417: 621-637. doi:10.1042/BJ20082063.

53. Axe DD (2010) The case against a Darwinian origin of protein folds. Bio-Complexity 1: 1-12. doi:10.5048/BIO-C.2010.1.

54. Taylor SV, Walter KU, Kast P, Hilvert D (2001) Searching sequence space for protein catalysts. Proc Natl Acad Sci USA 98:10596-10601. doi:10.1073/pnas.191159298.

55. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410:715-718.

56. Stomel JM, Wilson JW, Leon MA, Stafford P, Chaput JC (2009) A man-made ATP-binding protein evolved independent of nature causes abnormal growth in bacterial cells. PLoS ONE 4:e7385. doi:10.1371/journal.pone.0007385.

57. Silverman JA, Balakrishnan R, Harbury PB (2001) Reverse engineering the $(\beta/\alpha)_8$ barrel fold. Proc Natl Acad Sci USA 98:3092-3097. doi:10.1073/pnas.041613598.

58. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C (2009) PSI-2: Structural genomics to cover protein domain family space. Structure 17: 869-881. doi:10.1016/j.str.2009.03.015.

59. Kelly WP, Stumpf MPH (2008) Protein-protein interactions: from global to local analyses. Curr Opin Biotechnol 19: 396-403. doi:10.1016/j.copbio.2008.06.010.

60. Figeys D. (2008) Mapping the human protein interactome. Cell Research 18: 716-724. doi:10.1038/cr.2008.72.

61. Crick F (1981) *Life itself, Its Origin and Nature.* Simon and Schuster (New York), pp. 51.

62. Levitt M (2009) Nature of the protein universe. Proc Natl Acad Sci USA 106: 11079-11084. doi:10.1073/pnas.0905029106.

63. Rattei T, Tischler P, Götz S, Jehl MA, Hoser J, Arnold R, Conesa A, Mewes HW (2010) SIMAP – a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. Nucleic Acids Res 38: D223-D226. doi:10.1093/nar/gkp949.

64. Lees J, Yeats C, Redfern O, Clegg A, Orengo C (2010) Gene3D: merging structure and function for a thousand genomes. Nucleic Acids Res 38: D296-D300. doi:10.1093/nar/gkp987.

65. Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15: 583-589.

66. Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 313: 673-681. doi:10.1006/jmbi.2001.5079.

67. Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 3:research0040.1-0040.7.

68. Unger R, Uliel S, Havlin S (2003) Scaling law in sizes of protein sequence families: from super-families to orphan genes. Proteins 51: 569-576.

69. Enright AJ, Kunin V, Ouzounis CA (2003) Protein families and TRIBES in genome sequence space. Nucleic Acids Res 31: 4632-4638. doi:10.1093/nar/gkg495.

70. Lee D, Grant A, Marsden RL, Orengo C (2005) Identification and distribution of protein families in 120 completed genomes using Gene3D. Proteins 59: 603-615. doi:10.1002/prot.200409.

71. Sammut SJ, Finn RD, Bateman A (2008) Pfam 10 years on: 10 000 families and still growing. Brief Bioinform 9: 210-219. doi:10.1093/bib/bbn010.

72. Orengo CA, Thornton JM (2005) Protein families and their evolution – a structural perspective. Annu Rev Biochem 74: 867-900. doi:10.1146/annurev.biochem.74.082803.133029.

73. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C (2008) Gene3D: comprehensive structural and functional annotation of genomes. Nuclei Acids Res 36: D414-D418. doi: 10.1093/nar/gkm1019.

74. Adamic LA. Zipf, Power-laws, and Pareto – a ranking tutorial. www.hpl.hp.com/research/idl/papers/ranking/ranking.html.

75. Huberman BA, Adamic LA (1999) Growth dynamics of the World-Wide Web. Nature 401: 131.

76. Makse HA, Havlin S, Stanley HE (1995) Modelling urban growth patterns. Nature 377: 608.

77. Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Eur Phys J B 4: 131-134.

78. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.

79. Gisiger T (2001) Scale invariance in biology: coincidence or footprint of a universal mechanism? Biol Rev 76: 161-209.

80. Wuchty S (2001) Scale-free behavior in protein domain networks. Mol Biol Evol 18: 1694-1702.

81. Lavelle DT, Pearson WR (2010) Globally, unrelated protein sequences appear random. Bioinformatics 26: 310-318. doi:10.1093/bioinformatics/btp660.

82. Weber C, Barton GJ (2001) Estimation of P-values for global alignments of protein sequences. Bioinformatics 17: 1158-1167.

83. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol 2:18.

84. Karev PV, Wolf YI, Koonin EV (2003) Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics 19: 1889-1900. doi:10-1093/bioinformatics/btg351.

85. Karev GP, Wolf YI, Berezovskaya FS, Koonin EV (2004) Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. BMC Evol Biol 4:32. doi:10.1186/1471-2148-4-32.

86. Karev GP, Berezovskaya FS, Koonin EV (2005) Modeling genome evolution with a diffusion approximation of a birth-and-death process. Bioinformatics 21 Suppl. 3: iii12-iii19. doi:10.1093/bioinformatics/bti1202.

87. Novozhilov AS, Karev GP, Koonin VE (2006) Biological applications of the theory of birth-and-death processes. Briefings in Bioinformatics 7:70-85. doi:10.1093/bib/bbk006.

88. Hughes T, Liberles DA (2008) The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families. Gene 414:85-94. doi:10.1016/j.gene.2008.02.014.

89. Boto L. (2010) Horizontal gene transfer in evolution: facts and challenges. Proc R Soc B 277: 819-827. doi:10.1098/rspb.2009.1679.

90. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nature Rev Genet 9: 605-618. doi:10.1038/nrg.2386.

91. Lercher MJ, Pal C (2007) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25: 559-567. doi:10.1093/molbev/msm283.

92. Keeling PJ (2009) Functional and ecological impacts of horizontal gene transfer in eukaryotes. Curr Opin Genet & Develop 19: 613-619. doi: 101.1016/j.gde.2009.10.001.

93. Ragan MA, Beiko RG (2009) Lateral genetic transfer: open issues. Phil Trans R Soc B 364: 2241-2251. doi:10.1098/rstb.2009.0031.

94. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH (2010) High frequency of horizontal gene transfer in the oceans. Science 330: 50. doi:10.1126/science.1192243.

95. Coulson AFW, Moult J (2002) A unifold, mesofold, and superfold model of protein fold use. Proteins 46: 61-71. doi:10.1002/prot.10011.

96. Grant A, Lee D, Orengo C (2004) Progress towards mapping the universe of protein folds. Genome Biol 5:107.

97. Marsden RL, Lee D, Maibaum M, Yeats C, Orengo CA (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. Nucleic Acids Res 34:1066-1080. doi:10.1093/nar/gkj494.

98. Daugherty PS, Chen G, Iversen BL, Georgiou G (2000) Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. Proc Natl Acad Sci 97:2029-2034. doi:10.1073/pnas.030527597.

99. Drummond DA, Iversen BL, Georgiou G, Arnold FH (2005) Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. J. Mol. Biol. 350: 806-816. doi:10.1016/j.jmb.2005.05.023.

100. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH (2005) Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci 102: 606-611. doi:10.1073/pnas.0406744102.

101. Kunichika K, Hashimoto Y, Imoto T (2002) Robustness of hen lysozyme monitored by random mutations. Protein Engineering 15: 805-809.

102. Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI (2007) A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. PloS Comput Biol 3:e139. doi:10.1371/journal.pcbi.0030139.

103. Siew N, Fischer D (2003) Twenty thousand ORFan microbial protein families for the Biologist? Structure 11:7-9.

104. Siew N, Fischer D (2003) Unravelling the ORFan puzzle. Comp Funct Genom 4: 432-441. doi:10.1002/cfg.311.

105. Wilson GA, Bertrand N, Patei Y, Hughes JB, Feil EJ, Field D (2005) Orphans as taxonomically restricted and ecologically important genes. Microbiology 151: 2499-2501. doi.10.1099/mic.0.28146-0.

106. Wilson GA, Feil EJ, Lilley AK, Field D (2007) Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. PLoS ONE 3: e324. doi:10.1371/journal.pone.0000324.

107. Doolittle RF (2002) Microbial genomes multiply. Nature 416: 697-700.

108. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300: 1701-1703.

109. Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357: 543-544.

110. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25:25-29.

111. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25:107-110. doi:10.1016/j.tig.2008.12.004.

112. Yooseph S, Sutton G, Rusch DB et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biology 5:e16. doi: 10.1371/journal.pbio.0050016.

113. Gollery M, Harper J, Cushman J, Mittler T, Mittler R (2007) POFs: what we don't know can hurt us. Trends in Plant Science 12:492-496. doi:10.1016/j.tplants.2007.08.018.

114. Gollery M, Harper J, Cushman J, Mittler T, Girke T, Zhu JK, Bailey-Serres J, Mittler R (2006) What makes species unique? The contribution of proteins with obscure features. Genome Biol 7:R57. doi:10.1186/gb-2006-7-7r57.

115. Hanson AD, Pribat A, Waller JC, de Crecy-Lagard, V (2010) "Unknown" proteins and "orphan" enzymes: the missing half of the engineering parts list – and how to find it. Biochem J 425: 1-11. doi:10.1042/BJ20091328.

116. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

117. Venter JC et al. (2001) The sequence of the human genome. Science 291: 1304-1351. doi:10.1126/science.1058040.

118. Roberts RJ (2004) Identifying protein function: a call for community action. PLoS Biology 2:293-294. doi:10.1371/journal.pbio.0020042.

119. Lespinet O, Labedan B (2006) Orphan enzymes could be an unexplored reservoir of new drug targets. Drug Discovery Today 11: 300-305. doi:10.1016/j.drudis.2006.02.002.

120. Siew N, Fischer D (2004) Structural biology sheds light on the puzzle of genomic ORFans. J Mol Biol 342: 369-373. doi:10.1016/j.jmb.2004.06.073.

121. Jaroszewski L, Li Z, Krishna SS, Bakolista, C et al. (2009) Exploration of uncharted regions of the protein universe. PLoS Biol 7(9): e1000205. doi:10.1371/journal.pbio.10000205.

122. Wong WC, Maurer-Stroh S, Eisenhaber F (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. PLoS Comput Biol 6:e1000867. doi: 10.1371/journal.pcbi.1000867.

123. Gerlt JA (2007) A protein structure (or function?) initiative. Structure 15: 1353-1356. doi:101.1016/j.str/2007.10.003.

124. Omelchenko MV, Galperin MY, Wolf Yi, Koonin EV (2010) Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biology Direct 5:31.

125. Raes J, Harrington ED, Singh AH, Bork P (2007) Protein function space: viewing the limits or limited by our view? Curr Opin Struct Biol 17: 362-369. doi:10.1016/j.sbi.2007.05.010.

126. Long M (2001) Evolution of novel genes. Curr Opin Struct Biol 11: 673-680.

127. Nahon JL (2003) Birth of “human-specific” genes during primate evolution. Genetica 118: 193-208.

128. Marsden RL, Lewis TA, Orengo CA (2007) Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. BMC Bioinformatics 8:86. doi:10.1186/1471-2105-8-86.

129. Schmidt EE, Davies CJ (2007) The origins of polypeptide domains. Bioessays 29: 262-270. doi: 10.1002/bies.20546.

130. Kaessmann H. (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20: 1313-1326. doi:10.1101/gr.101386.109.

131. Marsden RL, Ranea JAG, Sillero A, Redfern O, Yeats C, Maibaum M, Lee D, Addou S, Reeves GA, Dallman TJ, Orengo CA (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. Phil Trans R Soc 361: 425-440. doi:10.1098/rstb.2005.1801

132. Pelegrin-Alvarez JM, Parkinson j (2007) The global landscape of sequence diversity. Genome Biol 8:R238. doi:10.1186/gb-2007.8-11-r238).

133. Mill JS (1882) A System of Logic, Ratiocinative And Inductive, Eighth Edition. Harper & Brothers (New York) [Ebook 27942] pp.483.

134. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG (2009) More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics 25: 404-413. doi:10.1016/j.tig.2009.07.006.

135. Lin H, Moghe G, Ouyang S, Iezzoni A, Shiu SH, Gu X, Buell CR (2010) Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. BMC Evo Biol 10:41.

136. Johnson BR, Tsutsui ND (2011) Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics 12:164.

137. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA (2009) The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res 37:D310-D314. doi:10.1093/nar/gkn877.

138. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. Proc. Natl Acad Sci USA 103: 2605-2610. doi: 10.1073/pnas.0509379103.

139. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289-316. doi:10.1146/annurev.biophys.37.092707.153558.

140. Fersht AR (2008) From the first protein structures to our current knowledge of protein folding: delights and scepticism. Nature Reviews Molecular Cell Biology 9: 650-654. doi:10.1038/nrm.2446.

141. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA (2007) The protein folding problem: when will it be solved? Curr Opin Struct Biol 17: 1-5. doi:10.1016/j.sbi.2007.06.001.

142. Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in de novo protein structure prediction. J Mol Biol 393: 249-260. doi:10.1016/j.jmb.2009.07.063.

143. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popović Z & Foldit players (2010) Predicting protein structures with a multiplayer online game. Nature 466: 756-760. doi:10.1038/nature09304.

144. Cooper S, Treuille A, Barbero J et al. (2010) The challenge of designing scientific discovery games. Proceedings of the 5-th International Conference on the Foundations of Digital Games, Monterey, CA, USA.

145. Koder RL, Dutton PL (2006) Intelligent design: *de novo* engineering of proteins with specified functions. Dalton Trans 3045-3051. doi:10.1039/b514972j.

146. Butterfoss GL, Kuhlman B (2005) Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 35: 49-65. doi:10.1146/annure.biophys.35.040405.102046.

147. Jha RK, Leaver-Fay A, Yin S, Wu Y, Butterfoss GL, Szyperski T, Dokholyan NV, Kuhlman B (2010) Computational design of a PAK1 binding protein. J Mol Biol 400: 257-270. doi:10.1016/j.jmb.2010.05.006.

148. Schmidt am Busch M, Sedano A, Simonson T (2010) Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. PLoS ONE 5:e10410. doi:10.1371/journal.pone.0010410.

149. Leisola M, Turunen O (2007) Protein engineering: opportunities and challenges. Appl Microbiol Biotechnol 75: 1225-1232. doi:10.1007/s00253-007-0964-2.

150. Liu S, Liu S, Zhu X, Liang H, Cao A, Chang Z, Lai L (2007) Nonnatural protein-protein interaction-pair design by key residues grafting. Proc Natl Acad Sci USA 104: 5330-5335. doi:10.1073/pnas.0606198104.

151. Gough J (2005) Convergent evolution of domain architecture (is rare). Bioinformatics 21: 1464-1471. doi:10.1093/bioiformatics/bti204.

152. Vogel C, Teichmann SA, Pereira-Leal J (2005) The relationship between domain duplication and recombination. J Mol biol 346: 355-365. doi:10.1016/j.jmb.2004.11.050.

153. Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. Nature Rev Mol Cell Biol 8:319-330.

154. Bornberg-Bauer E (2010) Signals: Tinkering with Domains. Sci Signal 3: pe31. doi:10.1126/scisignal.3139pe31.

155. Apic G, Russel RB (2010) Domain recombination: a workhorse for evolutionary innovation. Sci Signal 3: pe30. doi:10.1126/scisignal.3139pe30.

156. Lynch M (2007) The frailty of adaptive hypothesis for the origins of organismal complexity. Proc Natl Acad Sci USA 104: 8597-8604. doi:1073/pnas.0702207104.

157. Wilson EO (2003) The encyclopedia of life. Trends Ecol Evol 18: 77-80.

158. Abel DL (2009) The universal plausibility metric (UPM) & principle (UPP). Theor Biol Med Model 8:27. doi:10.1186/1742-4682-6-27.

159. Ayala F, Escalante A, O'Huigin C, Klein J (1994) Molecular genetics of speciation and human origins. Proc Natl Acad Sci 91: 6787-6794.

160. Coyne J, Orr HA (2004) *Speciation*, Sinauer Associates, Sunderland, USA.

161. Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation distortion in Drosphila hybrids. Science 323:376-379. doi: 10.1126/science.1163934.

162. Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J (2008) A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323: 373-375 doi: 10.1126/science.1163601,

163. Nosil P, Schluter D (2011) The genes underlying the process of speciation. Trends Ecol Evol 26: 160-167.

164. Smith JM (1970) Natural selection and the concept of a protein space. Nature 225: 563-564.

165. Jacob F (1977) Evolution and tinkering. Science 186: 1161-1166.

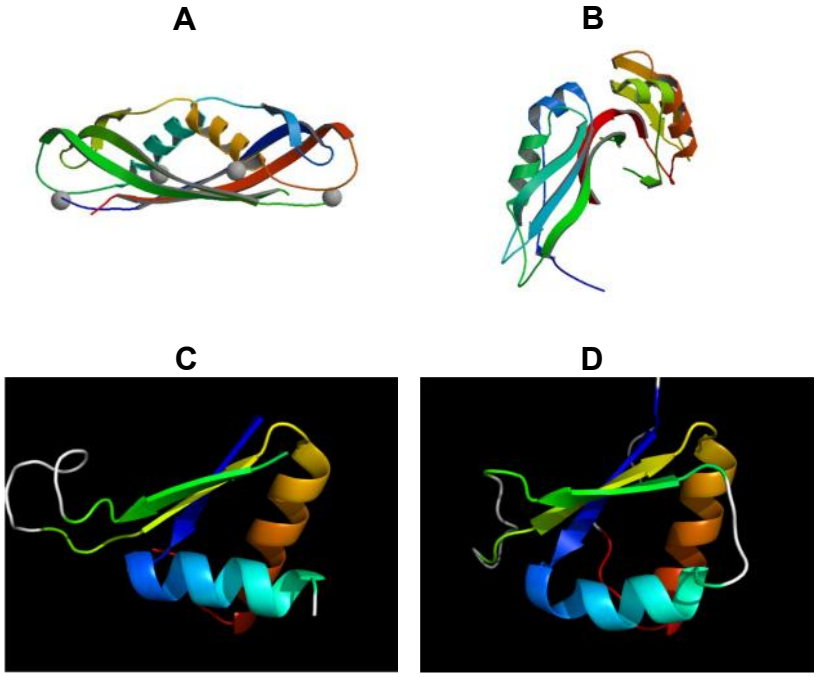

Figure 1.

A pair of proteins of similar sequences but different structures (A-B); and another pair of different sequences but similar structure (C-D). A-TonB protein PDB-ID 1IHR; B-TonB protein PDB-ID 1U07 - sequence similarity is 83%. C–1cs1_Ah, Cystathione gamma-synthetase, CGS; D-1q8i_Ac, exonuclease domain of family B DNA polymerases – sequence similarity is 7%. Source of the second pair: http://prodata.swmed.edu/malisam; ref. 30.

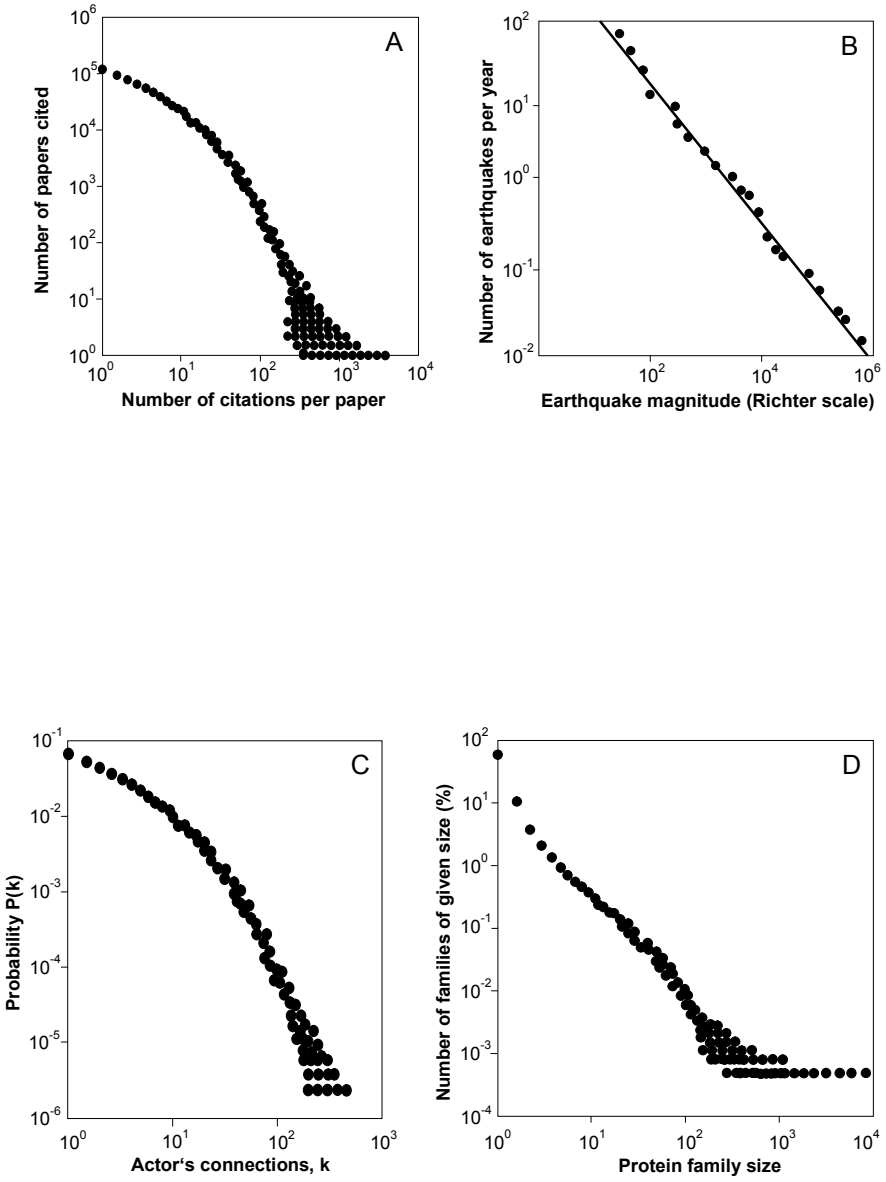

Figure 2.

Power-law distribution graphs of various phenomena. A - Distribution of citations of scientific papers, redrawn from ref. 77. B – Distribution of earthquake strengths, redrawn from ref. 79. C – Distribution of actors' collaboration, redrawn from ref. 78. D – Distribution of protein families, redrawn from ref. 71.

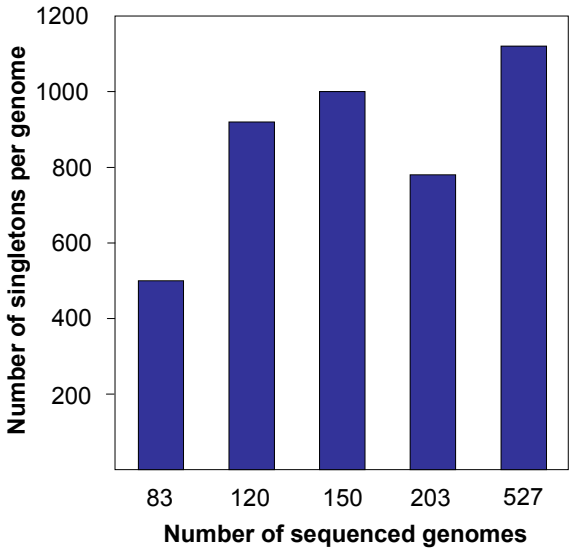

Figure 3.

The average number of singletons present in the genome of one species. The values were obtained by dividing the number of singletons with the number of the sequenced genomes as reported at various time points.