# Evolution of the Protein Universe. Time Scales and Selection

M. Cosentino Lagomarsino,* A.L. Sellerio,* P.D. Heijning,* and B. Bassetti*

(Dated: November 30, 2007)

**The availability of many genome sequences gives us abundant information, which is, however, very difficult to decode. As a consequence, in order to advance our understanding of biological processes at the whole-cell scale, it becomes very important to develop higher-level, synthetic descriptions of the contents of a genome. At the protein level, an effective scale of description is provided by protein domains [1]. Domains are independent unit-shapes (or "folds") forming proteins [2]. They are structurally stable and have thermodynamic origin. A domain determines a set of potential functions and interactions for the protein that carries it, for example DNA- or protein-binding capability or catalytic sites [1, 3]. Therefore, domains underlie many of the known genetic interaction networks. For example, a transcription factor or an interacting pair of proteins need the proper binding domains [4, 5], whose binding sites define transcription networks and protein-protein interaction networks respectively. Protein domains, are found on genomes with notable statistical distributions, which bear a high degree of similarity. A stochastic growth model with two *universal* parameters, related to a minimal number of domains and to the relative time-scale of innovation to duplication reproduces two important features of these distributions: (i) the populations of domain classes (the sets, related to homology classes, containing realizations of the same domain in different proteins) follow common power-laws whose diversity is related to genome size, measured by the total number of proteins or protein domains (ii) the number of domain families is sublinear in genome size. In this evolutionary process, selective pressure can enter both as a global constraint on the innovation time-scale, and as a regulator of the population of specific domain classes, related to their modularity: some shapes are common to all genomes, some are contextual. These two features are sufficient to obtain general quantitative agreement with data from hundreds of genomes, and show that robust self-organizing**

*Università degli Studi di Milano, Dip. Fisica. Via Celoria 16, 20133 Milano, Italy; e-mail address: Marco.Cosentino-Lagomarsino@unimi.it; and I.N.F.N. Milano, Italy. Tel. +39 02 50317477 ; fax +39 02 50317480

**phenomena encase specific selective pressures during evolution.**

Domains are related to sets of sequences of the protein-coding part of genomes. Multiple sequences give rise to the same shape, and the choice of a specific sequence in this set fine-tunes the function, activity and specificity of the inherent physico-chemical properties that characterize a shape. A domain then defines naturally a "domain class", constituted by all its realizations in the genome, or all the proteins using that given shape for some function. Overall, domains can be seen as an "alphabet" of basic elements of the protein universe. Understanding the usage of domains across organisms is as important and challenging as decoding an unknown language. Much as the letters of linguistic alphabets, the domains observable today are few, probably of the order of $10^5$ [3]. This number is surprisingly lower than the number of possible protein sequences (which are in general a hundred orders of magnitude more numerous). In the course of evolution, domains are subject to the dynamics of genome growth (by duplication, mutation, horizontal transfer, gene genesis, etc.) and reshuffling (by recombination etc.), under the constraints of selective pressure [3, 6]. These drives for combinatorial rearrangement, together with the defining modular property of domains, lead to the construction of increasingly richer sets of proteins [7]. In other words, domains are particularly flexible evolutionary building blocks.

In particular, the sequence of two duplicate domains that diverged recently will be very similar, so that one can also give a strictly evolutionary definition of protein domains [3], as regions of protein sequences that are highly conserved. The (interdependent) structural and evolutionary definitions of protein domains given above have been used to produce systematic hierarchical taxonomies of domains that combine information about shapes, functions and sequences [8, 9]. Generally, one considers three layers, each of which is a subclassification of the previous one. The top layer of the hierarchy is occupied by "folds", defined by purely structural means. It is then possible, though it seems quite rare, that a fold is polyphyletic, i.e. found from different paths in evolution. The intermediate "superfamily" class is also mainly defined by spatial shape, with the aid of sequence and functional annotations to guarantee monophyly. Finally, the "family" class is defined by sequence similarity.

The large-scale data stemming from this classification effort enable to tackle the challenge of understanding the alphabet of protein domains [1, 10–12]. In particular, they have been used to evaluate the laws governing the distributions of domains and domain families [6, 13–16]. These laws are notable and have a

high degree of universality. Using the number of domains $n$ to measure the size of a genome, we have the following observations, that confirm (and in part extend) previous ones. (i) The number of domain families (or distinct hits of the same domain) concentrates around a curve $F(n)$ that is markedly sublinear (figure 1A), perhaps saturating. (ii) The number $F(j,n)$ of domain classes having $j$ members (in a genome of size $n$) follows the power-law $\sim 1/j^{1+\alpha}$, where the fitted exponent $1+\alpha$ typically lies between 1 and 2 (figure 2). (iii) The exponent of this power-law appears to decrease with genome size (figure 2A), and there is evidence for a cutoff that increases linearly with $n$ (figure 2C). We tested these observations with data on folds and superfamilies (Supplementary Note S2). Recent modeling efforts focused mainly on (ii), and explored two main directions. First, a "designability" hypothesis [17], which claims that domain occurrence is due to accessibility of shapes in sequence space. While the debate is open, this alone seems to be an insufficient explanation, given for example the monophyly of most folds in the taxonomy [3, 18]. A second, "genome growth" hypothesis ascribing the emergence of power-laws to a generic preferential-attachment principle due to gene duplication seems to be more successful. Growth models were formulated as nonstationary, duplication-innovation models [6, 19, 20] and as stationary birth-death-innovation models [14, 21–23], and were successful in describing to a consistent quantitative extent the observed power laws. However, in both cases, each genome needed a specific set of kinetic coefficients, governing duplication, influx of new domain classes, or death of domains.

Here, we first define and relate to the data a non-stationary duplication-innovation model in the spirit of Gerstein and coworkers [6]. Compared to this work, our main idea is that a newly added domain class is treated as a *dependent* random variable, conditioned by the preexisting genome structure. We will show that this model explains observations (i-iii) with a *unique* underlying stochastic process having only two *universal* parameters of simple biological interpretation, the most important of which is related to the relative time scales of adding a domain belonging to a new family and duplicating an existing one. Subsequently, we argue that the scaling of this parameter can be related to the computational cost for adding a new domain class in a genome, and thus to a global property related to evolutionary selection. Finally, we show how a specific selective property, introduced in the model by inferred data on the usage of domain classes across genomes can predict the saturation of $F(n)$, and recalibrate the two universal parameters to obtain better quantitative agreement with data.

The basic ingredients of the model are $p_O$, the probability to duplicate an old domain (modeling gene duplication), and $p_N$, the probability to add a new domain class with one member (which describes domain innovation, for example by horizontal transfer). Iteratively, either a domain is duplicated with the former probability or a new domain class is added with the latter. A necessary feature for duplication is preferential attachment, stating the fact that duplication is more likely in a larger domain class. In other words, if the duplication probability is split as the sum of *per-class* probabilities $p_O^i$, preferential attachment requires that $p_0^i \propto k_i$, where $k_i$ is the population of class $i$. It is important to notice that in this model, while $n$ can be used as an arbitrary measure of time, the ratio of the time-scales of duplication and innovation is not arbitrary, and is set by the ratio $p_N/p_O$. In the model of Gerstein and coworkers, this is taken as a constant, as the innovation move considered to be statistically independent from the genome content. This choice has two problems. First, it does not give the observed sublinear scaling of $F(n)$. Second, it implies that for larger genomes the influx of new domain families is heavily dominant on the flux of duplicated domains. On the contrary, motivated by the observation (i), we consider dependent moves, or, in the simplest scheme a dependence of $p_N$ by $n$ and $f$ where $f$ is the number of domain classes in the genome. Specifically, we chose $p_N$ to be asymptotically inversely proportional to the mean class population $n/f$. In other words, it is harder to add a new domain class in a larger, or more heavily populated genome. As we will see, this implies $p_N/p_O \to 0$ as $n \to \infty$, and we will show that this choice reproduces properties (i-iii). Precisely, we take $p_O^i = \frac{k_i - \alpha}{n+\theta}$ (hence $p_O = \frac{n - f\alpha}{n+\theta}$), and $p_N = \frac{\theta + f\alpha}{n+\theta}$, where $\theta \geq 0$ and $\alpha \in [0, 1]$ Here, $\alpha$, is the most important parameter, which will set the scaling of the duplication/innovation ratio (table I), and $\theta$ is less important, representing a characteristic number of domain classes needed for the preferential attachment principle to set in, which defines the behavior of $F(n)$ for $n \to 0$.

This kind of model has previously been explored in a different context in the mathematical literature under the name of Pitman-Yor, or Chinese Restaurant Process (CRP) [24–27]. In the Chinese restaurant metaphor, domains are customers and tables are domain classes. In a duplication event, a new customer is seated at a table with a preferential attachment (or packing) principle, and in an innovation event, a new table is added. The natural random variables involved in the process are $f$, the number of tables or domain classes, $k_i$ the population of class $i$, and $n_i$, the size at birth of class $i$. Rigorous results for the probability distribution of the fold usage vector $\{k_1, ..., k_f\}$ confirm the results of our scaling argument. It is important

to note that in this stochastic process, large $n$ limit values of quantities such as $k_i$ and $f$ do not converge to numbers, but rather to random variables [24].

Despite of this property, it is possible to understand the scaling of the averages $K_i$ and $F$ (of $k_i$ and $f$ respectively) at large $n$, writing simple "mean field" equations, for continuous $n$. From the definition of the model, we obtain $\partial_n K_i(n) = \frac{K_i - \alpha}{n+\theta}$, and $\partial_n F(n) = \frac{\alpha F(n) + \theta}{n+\theta}$. These equations have to be solved with initial conditions $K_i(n_i) = 1$, and $F(0) = 1$. Hence, for $\alpha \neq 0$, one has $K_i(n) = (1-\alpha)\frac{n+\theta}{n_i+\theta} + \theta$, and

$$F(n) = \frac{1}{\alpha}\left[(\alpha+\theta)\left(\frac{n+\theta}{\theta}\right)^{\alpha} - \theta\right] \sim n^{\alpha} \ ,$$

while, for $\alpha = 0$

$$F(n) = \theta \log(n+\theta) \sim \log(n) \ .$$

The solution can be used to compute the asymptotics of $P(j,n) = F(j,n)/F(n)$ [28]. This works as follows. From the solution, $j > K_i(n)$ implies $n_i > n^*$, with $n^* = \frac{(1-\alpha)n - \theta(j-1)}{j-\alpha}$, so that the cumulative distribution can be estimated by the ratio of the (average) number of domain classes born before size $n^*$ and the number of classes born before size $n$, $P(K_i(n) > j) = F(n^*)/F(n)$. $P(j,n)$ is then obtained by derivation. For $n, j \to \infty$, and $j/n$ small, we find

$$P(j,n) \sim j^{-(1+\alpha)}$$

for $\alpha \neq 0$, and

$$P(j,n) \sim \frac{\theta}{j}$$

for $\alpha = 0$. The above formulas give the correct average behavior of observation (iii). The trend of the model of Gerstein and coworkers can be found for constant $p_N, p_O$. A comparative scheme of the results is presented in table I. We also verified that these results are stable for introduction of domain loss and global duplications in the model (Supplementary Note S4).

Going beyond scaling, the probability distributions generated by a CRP contain large finite-size effects that are relevant for the experimental genome sizes. In order to evaluate the behavior and estimate parameter values keeping into account stochasticity and the small system sizes, performed numerical simulations of different realizations of the stochastic process (figures 1B and 2B and C). The simulations allow to measure

$f(n)$, and $F(j,n)$, and confirm the asymptotic predictions. Moreover, comparing the histogram of domain occurrence of model and data, it becomes evident that the intrinsic cutoff set by $n$ causes the observed drift in the exponent (figure 2A and B). One can measure the cutoff as the population of the largest domain class, and verify that both model and data follow a linear scaling (figure 2C). This can be expected from the above asymptotic argument, since $K_i(n) \sim n$. While the slope of $F(n)$ is compatible with a model having $\alpha = 0$, the internal distribution of domain families $P(j,n)$ and the behavior of the cutoff resembles more a CRP with $\alpha$ between 0.5 and 0.7 (figure 1B and Supplementary Note S1 and S2). Overall, although one can clearly obtain from the model all the qualitative phenomenology, the quantitative agreement seems to be not completely satisfactory.

We will now show how a simple variant of the model that includes selective pressure based on empirically measured domain family usage can bypass the problem, without upsetting the underlying ideas presented above. Before we do that, we want to argue that the model already contains the signature of selective pressure in the parameter $\alpha$. We suppose that, given a genome with $n$ domains (or for simplicity monodomain genes) and $F$ domain families, the process leading to the acceptance of a new domain family, and thus to a new class of functions, will need a readaptation of the population of all the domain families causing an increase $\delta n$ in the number of genes. This increase is due to an underlying optimization problem that has to adapt the new functions exploited by the acquired family to the existing ones (by rewiring and expanding the interaction networks, etc.) Now, generically, the computational cost for this optimization problem (which, conceptually, may be regarded as a measure of the evolvability of the system) could be constant (and thus $\delta n \sim \delta F$), or else polynomial or exponential in $F$ (i.e. $\delta n \sim F^d \delta F$, where d is some positive exponent, or $\delta n \sim \exp(F)\delta F$ respectively). Integrating and inverting these relations, it is simple to verify that the first choice leads to the scaling of the model of Gerstein and coworkers, while the second two correspond to the CRP, and to a sublinear $F(n)$. In other words, following this argument, the CRP supposes that accepting a new domain family becomes harder with increasing number of already available domain families, which is a (global) effect of selective pressure. On the other hand, there exist also *specific* effects of selection, due to the precise functional significance and interdependence of all domain classes. These give rise to correlations and trends that are clearly visible in the data, which we have analyzed in detail in a parallel study [29]. Here, we will consider simply the empirical probabilities of usage of domain families

for 327 bacterial genomes in the SUPERFAMILY database [30] (figure 1C). The variant of the model can be thought of as a genetic algorithm, where the above duplication-innovation model generates a population of candidate genome domain compositions, subsequently selected using specific criteria that keep into account observed features of the data. We have examined in more detail the analytically approachable case where two virtual moves are considered. The moves are compared on the basis of a fitness function based on the empirical domain usage and only one of the two is chosen (see Supplementary Note S3). With this modification, we introduce a significance to the index of the domain class, or a colored "tablecloth" to the table of the Chinese restaurant. In other words, while the probability distributions in the model are symmetric by switch of labels in domain classes (or *exchangeable* [26]), this cannot be true for the empirical case. The empirical domain family usage can be used to break the symmetry. Figures 1B and 2B show the comparison of simulations with empirical data. The agreement is very good. In particular, the values of $\alpha$ that better agree with the empirical behavior of the number of domain classes as a function of domain size $F(n)$ are also those that generate the best slopes in the internal usage histograms $F(j, n)$. Furthermore, the fitness generates a critical value of $n$, where $F(n)$ becomes flat, as observed empirically. A mean field calculation of the same style as the ones presented above predicts the existence of this plateau (Supplementary Note S3).

In conclusion, model and data together indicate that evolution acts conservatively on domain families, with a preference to exploiting available shapes rather than adding new ones. Specific biological and physical properties, such as function and designability [1, 18, 31] come in at the more detailed level of description of how domains are actually used to form functional proteins. A final point can be made regarding the number of observed domains. The model assumes that the new domain classes are drawn from an infinite family of shapes, which can be even continuous [24], and leads to a discrete and small number of classes at the relevant sizes. Although physical considerations point to the existence of a small "menu" of shapes available to proteins [32], the validity of our model would imply that the empirical observation of a small number of folds in nature does not count as evidence for this thermodynamic property of proteins, but may have been a simple consequence of evolution.

We thank S. Maslov, H. Isambert, F. Bassetti, S. Teichmann, and M. Babu for helpful discussions,

N. Kashtan for suggestion to improve teh text.

---


[1] Orengo, C. A. & Thornton, J. M. Protein families and their evolution-a structural perspective. *Annu Rev Biochem* **74**, 867–900 (2005).

[2] Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland, New York, 1999).

[3] Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420**, 218–23 (2002).

[4] Madan Babu, M. & Teichmann, S. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **31**, 1234–44 (2003).

[5] Nye, T. M., Berzuini, C., Gilks, W. R., Babu, M. M. & Teichmann, S. A. Statistical analysis of domains in interacting protein pairs. *Bioinformatics* **21**, 993–1001 (2005).

[6] Qian, J., Luscombe, N. M. & Gerstein, M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* **313**, 673–81 (2001).

[7] Ranea, J. A., Buchan, D. W., Thornton, J. M. & Orengo, C. A. Evolution of protein superfamilies and bacterial genome size. *J Mol Biol* **336**, 871–87 (2004).

[8] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536–40 (1995).

[9] Orengo, C. A. *et al.* CATH–a hierarchic classification of protein domain structures. *Structure* **5**, 1093–108 (1997).

[10] Ranea, J. A., Sillero, A., Thornton, J. M. & Orengo, C. A. Protein superfamily evolution and the last universal common ancestor (LUCA). *J Mol Evol* **63**, 513–25 (2006).

[11] Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A. & Weiner, J., 3rd. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci* **62**, 435–45 (2005).

[12] Weiner, J., 3rd, Beaussart, F. & Bornberg-Bauer, E. Domain deletions and substitutions in the modular protein evolution. *FEBS J* **273**, 2037–47 (2006).

[13] Huynen, M. A. & van Nimwegen, E. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* **15**, 583–9 (1998).

[14] Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S. & Koonin, E. V. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol* **2**, 18 (2002).

[15] Kuznetsov, V. A. In Zhang, W. & Shmulevich, I. (eds.) *Computational and Statistical Approaches to Genomics*, 125 (Kluwer, Boston, 2002).

[16] Abeln, S. & Deane, C. M. Fold usage on genomes and protein fold evolution. *Proteins* **60**, 690–700 (2005).

[17] Li, H., Tang, C. & Wingreen, N. S. Are protein folds atypical? *Proc Natl Acad Sci U S A* **95**, 4987–90 (1998).

[18] Deeds, E. J. & Shakhnovich, E. I. A structure-centric view of protein evolution, design, and adaptation. *Adv Enzymol Relat Areas Mol Biol* **75**, 133–91, xi–xii (2007).

[19] Kamal, M., Luscombe, N., Qian, J. & Gerstein, M. Analytical Evolutionary Model for Protein Fold Occurrence in Genomes, Accounting for the Effects of Gene Duplication, Deletion, Acquisition and Selective Pressure. In Koonin, E., Wolf, Y. & Karev, G. (eds.) *Power Laws, Scale-Free Networks and Genome Biology*, 165–193 (Spinger, New York, 2006).

[20] Durrett, R. & Schweinsberg, J. Power laws for family sizes in a duplication model. *Ann. Probab.* **33**, 2094–2126 (2005).

[21] Karev, G. P., Wolf, Y. I. & Koonin, E. V. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics* **19**, 1889–900 (2003).

[22] Karev, G. P., Wolf, Y. I., Berezovskaya, F. S. & Koonin, E. V. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol Biol* **4**, 32 (2004).

[23] Karev, G. P., Berezovskaya, F. S. & Koonin, E. V. Modeling genome evolution with a diffusion approximation of a birth-and-death process. *Bioinformatics* **21 Suppl 3**, iii12–9 (2005).

[24] Pitman, J. Combinatorial Stochastic Processes. In *Notes for St. Flour Summer School* (2002).

[25] Pitman, J. & Yor, M. The two-parameter poisson-dirichlet distribution derived from a stable subordinator (1997).

[26] Aldous, D. Exchangeability and related topics. In *Saint-Flour Summer School XIII - 1983* (Springer, Berlin, 1985).

[27] Kingman, J. Random discrete distributions. *J. Roy. Statist. Soc. B* **37**, 1–22 (1975).

[28] Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–12 (1999).

[29] Heijning, P., Sellerio, A., Bassetti, B. & Cosentino Lagomarsino, M. Unpublished.

[30] Wilson, D., Madera, M., Vogel, C., Chothia, C. & Gough, J. The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* **35**, D308–13 (2007).

[31] Zeldovich, K. B., Chen, P., Shakhnovich, B. E. & Shakhnovich, E. I. A First-Principles Model of Early Evolution: Emergence of Gene Families, Species, and Preferred Protein Folds. *PLoS Comput Biol* **3**, e139 (2007).

[32] Banavar, J. R. & Maritan, A. Physics of proteins. *Annu Rev Biophys Biomol Struct* **36**, 261–80 (2007).

| | $K_i$ | $\frac{p_N}{p_O}$ | $\frac{p_N}{p_O^i}$ | $F(n)$ | $F(j,n)/F(n)$ |
|---|---|---|---|---|---|
| CRP $\alpha = 0$ | $\sim n$ | $\sim n^{-1}$ | $\sim n^{-1}$ | $\sim \log(n)$ | $\sim \frac{\theta}{j}$ |
| CRP $\alpha > 0$ | $\sim n$ | $\sim n^{\alpha-1}$ | $\sim n^{\alpha-1}$ | $\sim n^{\alpha}$ | $\sim j^{-(1+\alpha)}$ |
| Qian *et al.* | $\sim n^{p_O}$ | $= R$ | $\sim n^{1-p_O}$ | $\sim n$ | $\sim j^{-(2+R)}$ |

TABLE I: Salient features of the proposed model in terms of scaling of the number of domain classes, compared to the model of Gerstein and coworkers [6, 19]. The first three columns indicate the resulting average population of a class $K_i$, and the ratios of the probability to add a new class $p_N$ to the total and *per-class* probabilities of duplication, as a function of genome size $n$. These latter two quantities are asymptotically zero in the CRP, while they are constant or infinite in the model of Gerstein and coworkers. The last two columns indicate the resulting scaling of number of domain classes $F(n)$ and fraction of classes with $j$ domains $F(j,n)/F(n)$. The results of the CRP agree qualitatively with observations (i-iii) in the text.

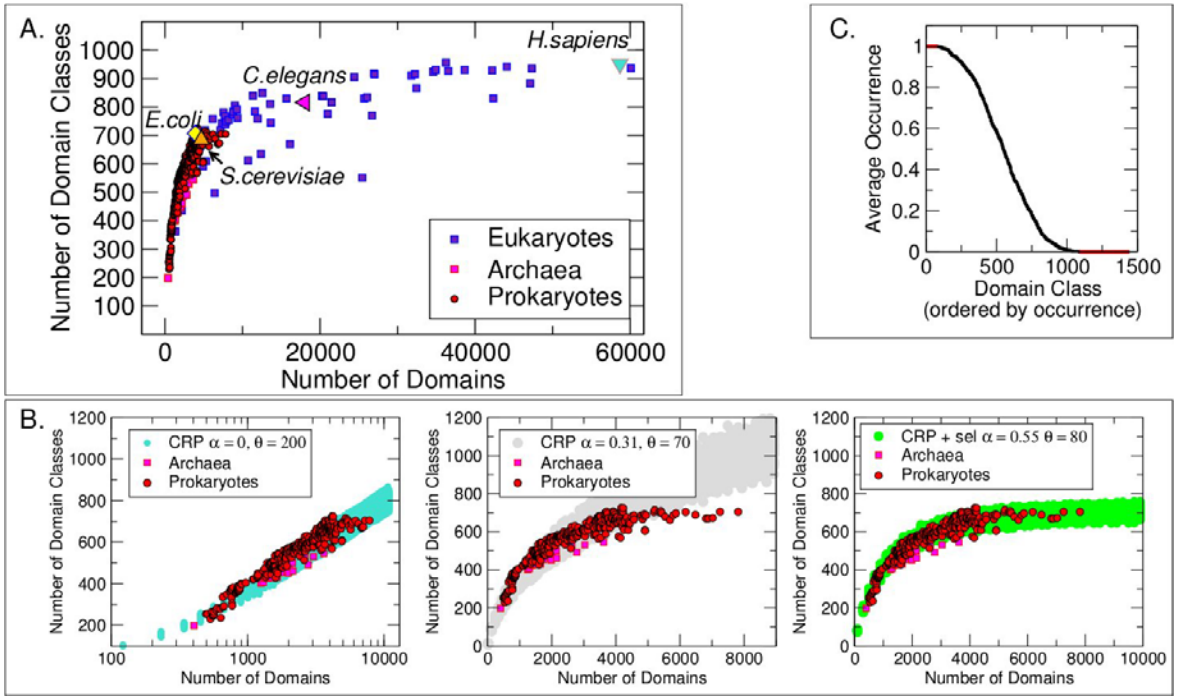



FIG. 1: Number of domain classes versus genome size. A. Plot of empirical data for 327 prokaryotes, 75 eukaryotes, and 27 archaeal genomes. Data refer to superfamily domain classes from the SUPERFAMILY database [30]. Larger data points indicate specific examples. Data on SCOP folds follow the same trend (Supplementary Note S2). B. Comparison of data on prokaryotes (red circles) with simulations of 500 realizations of the model (cyan, grey, and green shade), for fixed parameter values, different in each panel. Data on archaea are also shown (squares). $\alpha = 0$ (left panel, graph in log-linear scale) gives a trend that is more compatible with the observed scaling than $\alpha > 0$ (mid panel). However, the empirical distribution of folds in families is quantitatively more in agreement with $\alpha > 0$ (see table I and figure 2). The model that includes specific selection based on domain family usage (right panel) predicts a saturation of this curve even for high values of $\alpha$, resolving this quantitative conflict. C. Usage profile of SUPERFAMILY domain classes in prokaryotes, used to generate the fitness in the model with selection. In the x-axis, domain families are ordered by the fraction of genomes they occur in. The y-axis reports their occurrence fraction. The red lines indicate occurrence in all or none of the prokarotic genomes of the data set.

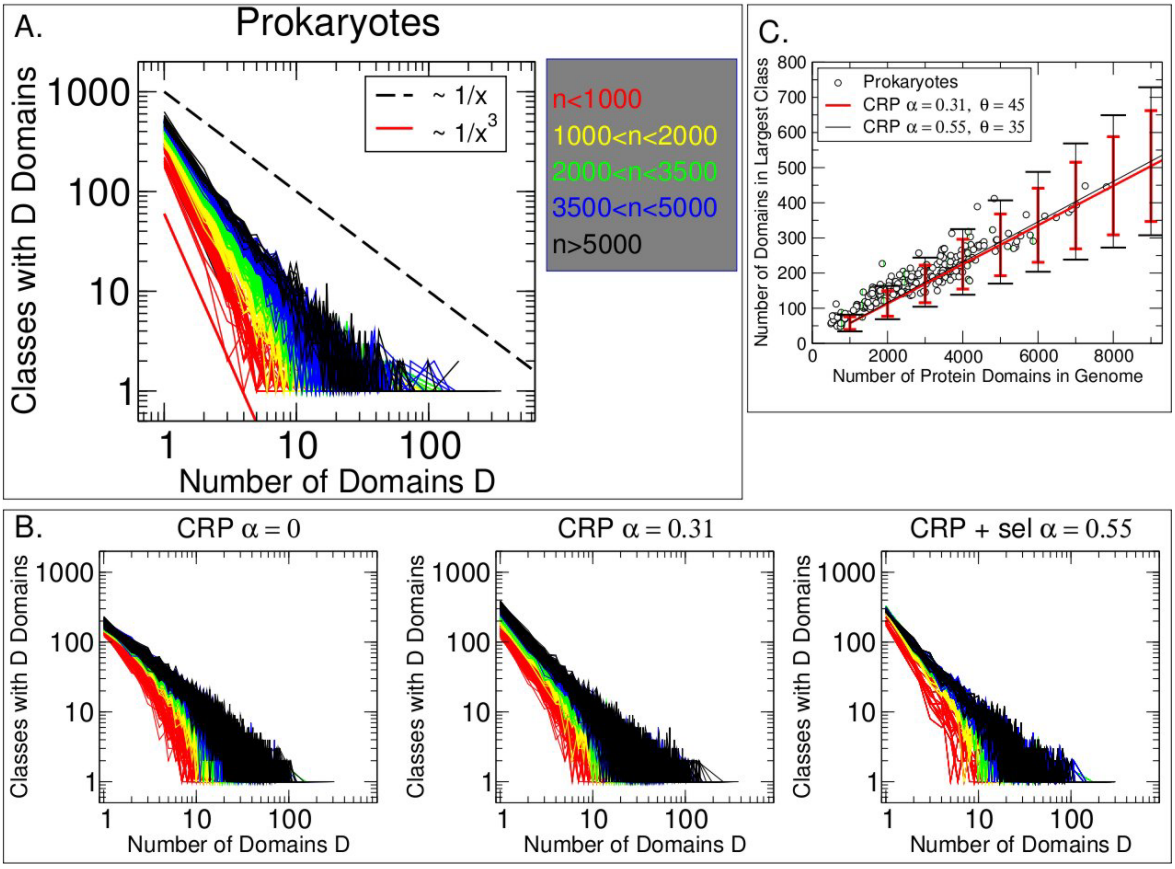
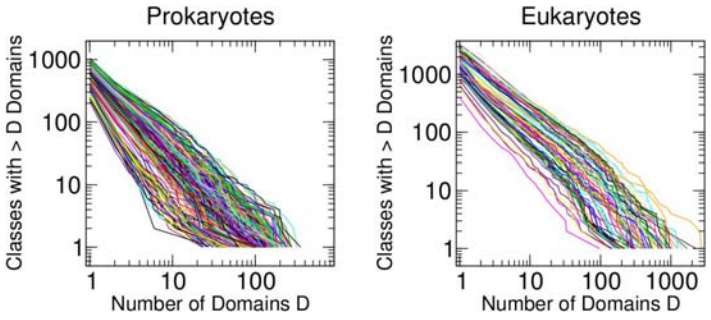



FIG. 2: Internal usage of domains. A. Histograms of domain usage; empirical data for 327 prokaryotes. The x-axis indicates the population of a domain class, and the y-axis reports the number of classes having a given population of domains. Each of the 327 curves is a histogram referring to a different genome. The genome sizes are color-coded as indicated by the legend on the right. Larger genomes (black) tend to have a slower decay, or a larger cutoff, compared to smaller genomes (red). The continuous (red) and dashed (black) lines indicate a decay exponent of 3 and 1 respectively. B. Histograms of domain usage for 50 realizations of the model at genome sizes between 500 and 8000. The color code is the same as in panel A. All data are in qualitative agreement with the empirical ones. However, data at $\alpha = 0$ appear to have a faster decay compared to empirical data. This is also evident looking at the cumulative distributions (Supplementary Note S1). The right panel refers to the model with selection, at parameters values that reproduce well the empirical number of domain classes at a given genome size (figure 1). C. Population of the maximally populated domain class as a function of genome size. Empirical data of prokaryotes (green circles), are compared to realizations of the CRP, for two different values of $\alpha$, the lines indicate averages over 500 realizations, with error bars indicating standard deviation. $\alpha = 0$ can reproduce the empirical trend only qualitatively (not shown). Data from the SUPERFAMILY database[30].

## SUPPLEMENTARY NOTES

### S1. CUMULATIVE DISTRIBUTIONS FOR THE INTERNAL USAGE OF DOMAINS

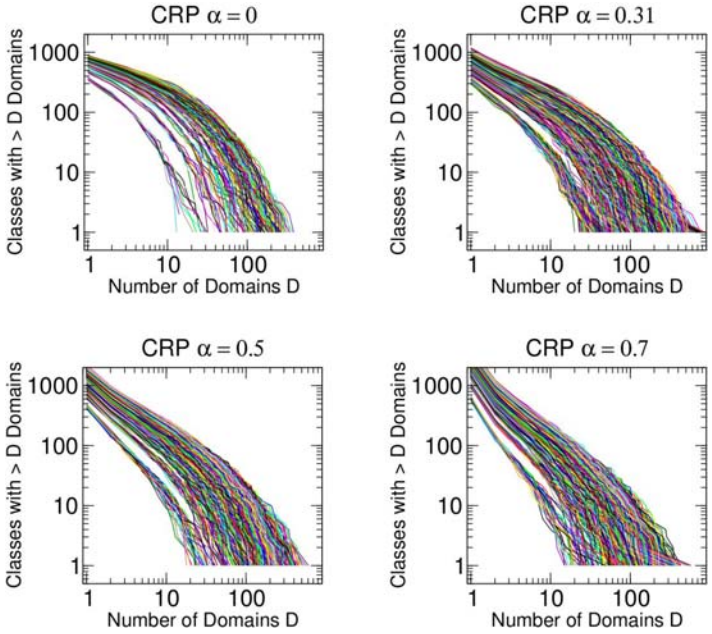
This section briefly discusses the cumulative histograms of domain usage for data and models. Figure S1.1 confirms the markedly power-law behavior observed for the histograms and predicted by the model. Comparison with the predictions of the CRP model (figure S1.2) shows faster decay for $\alpha = 0$. While in good agreement with the observed number of domain classes with increasing size (figure 1B), this parameter choice is unsatisfactory on the quantitative side for the domain distribution in classes. This feature, already visible in figure 2B of the main text, is even more marked from the cumulative histograms. Better-fitting values are in the range $\alpha = 0.5 - 0.7$. The CRP with selection (figure S1.3) has the same qualitative behavior as the standard model for the distributions, while fitting well the scaling of the classes of higher values of $\alpha$ (figure 1B and section S3 below).



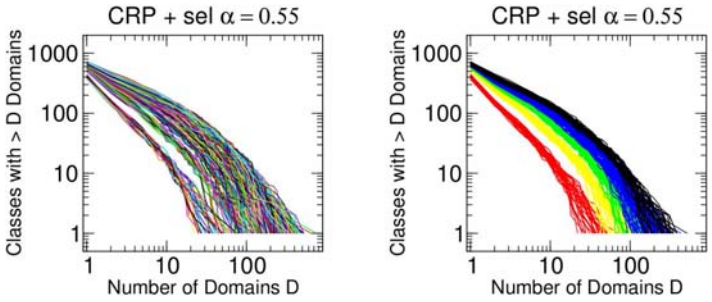

Supporting Figure S1.1: Empirical cumulative distributions of domain usage for domain classes of the SUPERFAMILY database. The x-axis reports domain class sizes in number of domains $D$ while the y-axis refers to the histogram of the number of domain classes containing more than $D$ domains. The left panel is based on the same data on the 327 prokaryotes of figure 2A in the main text. The right panel refers to the 75 eukaryotes in the data set. The genome sizes are not color-coded to show individual plots.

### S2. RESULTS FOR FOLD DOMAIN CLASSES

All data shown in the main text refer to the superfamily taxonomy level, and come from the SUPERFAMILY database. In this section, we report the results of the same analysis in terms of SCOP folds, which show that this category has essentially the same behavior as the previous one (figure S2.4). While

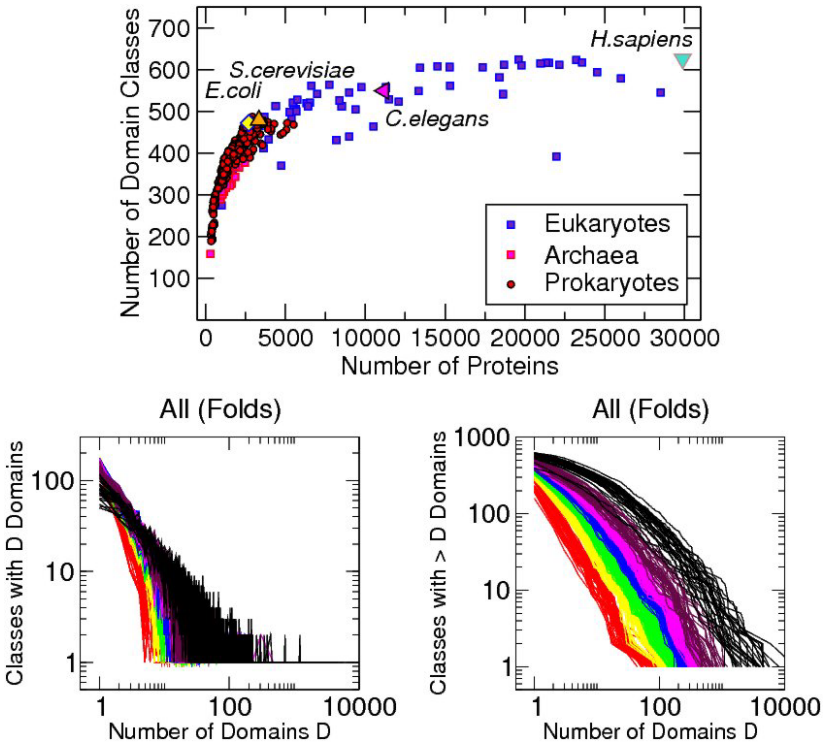



Supporting Figure S1.2: Cumulative histograms of domain usage for 50 realizations of the CRP at genome sizes between 500 and 8000. Increasing values of $\alpha$ are plotted in lexicographic order.

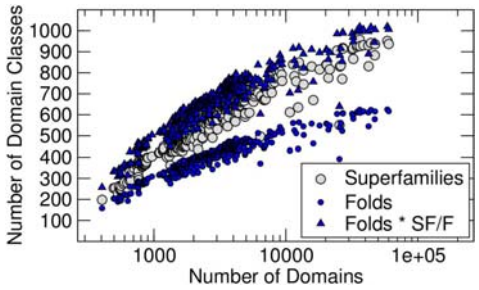



Supporting Figure S1.3: Cumulative histograms of domain usage for 100 realizations of the CRP with selection at genome sizes between 1000 and 8000. In this size range the model variant produces essentially identical distributions to the conventional CRP, with better agreement on the growth in terms of domain classes (see section S3). The left panel is color-coded as figure 2B of the main text.

by definition there are more superfamilies than folds, the number of domain classes versus genome size has very similar scaling in the two cases. The two plots collapse almost exactly, when folds are rescaled by the ratio (1443/884) of superfamilies per folds (S2.5). Furthermore, power-law fits of the experimental data for prokaryotes yield an exponent $\alpha$ between 0.3 and 0.4 for both categories, and logarithmic fits are also in agreement.




Supporting Figure S2.4: Top: Number of fold classes versus genome size, the plot is equivalent to figure 1A, except that the x-axis reports number of proteins scored in the genome, rather than genome size in domains. Since these two quantities are quite markedly linearly related, the two plots are equivalent. Bottom: histogram (left panel) and cumulative histogram (right panel) of domain classes for all genomes in the data set (eukaryotes, prokaryotes and archaea).




Supporting Figure S2.5: Comparison of the scaling of folds and superfamilies plot as a function of genome size. The plots refer to all genomes in the SUPERFAMILY database. The plot for folds (blue small circles) overlaps quite well with the plot for superfamily (large grey circles) when multiplied by the ratio of the total number of domain classes in the two taxonomies (1443/884).

## S3. ANALITYCAL MEAN FIELD EQUATIONS FOR THE CRP MODEL WITH SELECTION

In this section we discuss the analytical treatment of the variant of the CRP model introduced in the main text. We first give some more details on the definition of the model. Generically, we consider the following algorithm. For each genome size $n$, which measures time in some arbitrary units, the configuration is a set or "population" of $M$ genomes $\{g_1(n), ..., g_M(n)\}$, where each genome is a set of $D$ domain classes populated by some domains. An iteration is divided into two steps. A first "proliferation" step generates $qM$ genomes, where $q$ is a positive integer, $\{g_1^{'}(n), ..., g_{qM}^{'}(n)\}$, using the standard CRP move. A second "selection" step discards the $(q-1)M$ individuals with lower fitness. The fitness, for a generic model genome $g$, can be a function $\mathcal{F}(g)$, that takes into account some phenomenological features observed in the data. We choose to include in $\mathcal{F}$ a minimal amount of empirical information on the occurrence of each domain class contained in figure 1C. In other words, we distinguish between "universal" domain classes, used in most of the genomes, and "contextual" ones, occurring only in a few examples. As discussed in the main text , this is sufficient to obtain quantitative agreement with the observed domain distributions (figures 1B and 2B), which are not given to the model as an input. If domain classes are indexed by $i = 1..D$ ($D = 1443$ for Superfamilies), we define the variable $\sigma_i^g$ as follows

$$\sigma_i^l = \begin{cases} 1 & \text{if domain class } i \text{ is present in genome } g \\ -1 & \text{if domain class } i \text{ is absent in genome } g \end{cases} .$$

The fitness of that genome is then defined as

$$\mathcal{F}(g) = \exp\left(\sum_{i=1}^{D} \sigma_i^g \langle \sigma_i^{\mathrm{EMP}} \rangle\right) ,$$

where $\langle \sigma_i^{\mathrm{EMP}} \rangle$ is the empirical average of the same observable:

$$\langle \sigma_i^{\mathrm{EMP}} \rangle = \frac{1}{G} \sum_{g=1}^{G} \sigma_i^{g,\mathrm{EMP}} .$$

In the above formula $G$ is the number of observed genomes in the data set. For example, in the case of prokaryotes in the SUPERFAMILY database, $G = 327$ and, calling $\Xi_i$ the function plotted in figure 1C, we have simply $\langle \sigma_i^{\mathrm{EMP}} \rangle = 2\Xi - 1$.

For the analitycal treatment, we considered the case $M = 1$, $q = 2$, where at each iteration, one genome is selected from a population of two. Starting from configuration $g(n)$, in the proliferation step genomes

$g^{'}, g^{''}$ are generated with CRP rules, and the selection step chooses $g(n+1) = \mathrm{argmax}(\mathcal{F}(g^{'}), \mathcal{F}(g^{''}))$. In this case, since the selection rule chooses strictly the maximum, it is able to distinguish the sign of $\langle \sigma_i^{\mathrm{EMP}} \rangle$ only. For this reason, it is sufficient to account for the positivity (which we label by "+") and negativity ("-") of this function for a given domain index $i$. The genomes $g^{'}$ and $g^{''}$ proposed by the CRP proliferation step can have the same (labeled by "1"), higher ("$1_+$") or lower ("$1_-$") fitness than their parent, depending on $p_O$, $p_N$ and by the probabilities to draw a universal or contextual domain family, $p_+$ and $p_-$ respectively. Using these labels, the scheme of the possible states and their outcome in the selection step is given by the table below.

| proliferation $(g^{'}, g^{''})$ | probability | selection |
|---|---|---|
| $(1,1)$ | $p_O^2$ | old |
| $(1,1_-)$ | $2\,p_O\,p_N\,p_-$ | old |
| $(1,1_+)$ | $2\,p_O\,p_N\,p_+$ | new+ |
| $(1_+,1_+)$ | $p_N^2\,p_+^2)$ | new+ |
| $(1_+,1_-)$ | $2\,p_N^2\,p_-\,p_+$ | new+ |
| $(1_-,1_-)$ | $p_N^2\,p_-^2$ | new− |

From this table, it is straightforward to derive the modified probabilities $\hat{p}_O$ and $\hat{p}_N$ of the complete iteration:

$$\hat{p}_O = p_O\,(p_O + 2\,p_N p_-)$$

$$\hat{p}_N = p_N\,(p_N + 2\,p_O p_+) = p_{N+} + p_{N-} \ ,$$

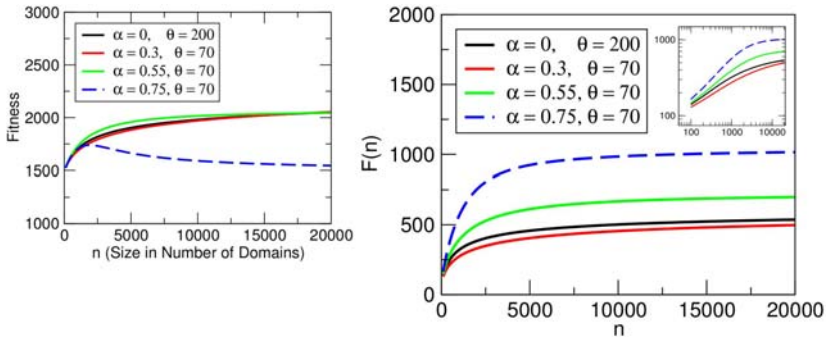
where $p_{N+} = p_N p_+ (2 - p_N p_+)$ and $p_{N-} = p_N^2 (1-p_+)^2$ are the probabilities that the new domain is drawn from the universal or contextual families respectively.

We now write the macroscopic evolution equation for the number of domain families using the same procedure as in the main text. Calling $k^+(n)$ and $k^-(n)$ the number of domain classes that have positive or negative $\langle \sigma_i^{\mathrm{EMP}} \rangle$ and are *not* represented in $g(n)$,

$$\begin{cases} \partial_n F(n) = \quad \hat{p}_N \\ \partial_n k^+(n) = -\hat{p}_{N+} \\ \partial_n k^-(n) = -\hat{p}_{N-} \end{cases} .$$

Now, $p_+ = k^+/(k^- + k^+) = k^+/(D - F(n)$, so that we can rewrite

$$\begin{cases} \partial_n F(n) = & (\frac{\alpha F(n)+\theta}{n+\theta}) \left[\frac{\alpha F(n)+\theta}{n+\theta} + \frac{2k^+(n)}{D-F(n)}(\frac{n-\alpha F(n)}{n+\theta})\right] \\ \partial_n k^+(n) = & -(\frac{\alpha F(n)+\theta}{n+\theta}) \frac{k^+(t)}{D-F(n)} \left[2 - (\frac{\alpha F(n)+\theta}{n+\theta}) \frac{k^+(n)}{D-F(n)}\right] \\ \partial_n k^-(n) = & -(\frac{\alpha F(n)+\theta}{n+\theta})^2 (\frac{k^+(n)}{D-F(n)})^2 \end{cases} \tag{1}$$




Supporting Figure S3.6: Numerical solutions of the mean-field equations of the CRP model with selection. Left panel: fitness $\mathcal{F}(n)$ for different values of $\alpha$. Right panel: F(n) plotted in linear and logarithmic (inset) scales.

The above equations have the following consistency properties

- $\partial_n\Big(k^+ + k^- + F\Big) = 0$, hence $k^+ + k^- + F = D \quad \forall n$.
- $\partial_n F \leq 1$, hence $F(n) \leq n$.
- $\partial_n F \geq 0$, $\partial_n k^+ \geq 0$ and $\partial_n(F + k^+) \geq 0$ so that $F$ grows faster than $k^+$ decreases.

Choosing the initial conditions from empirical data $n_0, F(n_0)$ size and number of domain classes of the smallest genome, we have, since $F(n_0) < n_0$ and $\alpha \leq 1$,

$$\frac{\alpha F(n_0) + \theta}{n_0 + \theta} < 1 \ .$$

It is simple to verify that under this condition the system always has solutions that relax to a finite value $F_\infty < D$. Indeed, after the time $n^*$ where $k^+(n*) = 0$, the equations reduce to $\partial_n k^+ = 0$, $k^- = D - F$ and

$$\partial_n F(n) = \left(\frac{\alpha F(n) + \theta}{n + \theta}\right)^2$$

immediately giving our result.

Numerical solutions of Eq. (1) give the same behavior for $F(n)$ as the direct simulations (figures S3.6A, figure 1B). In particular, while this function grows as a power law for small genome sizes, it saturates at the relevant scale, giving good agreement with the data. This behavior is connected to the finite size of the pool of universal domain families, which we can interpret as the effect of a certain optimality in the core functions of the different organisms. The internal laws of domain usage of this model were obtained from direct simulations only, and, as discussed in the main text, give a more quantitative agreement with the data (figure 2B). Finally, one interesting point can be made about the dynamics of the fitness. Figure S3.6B, shows that, for large values of $\alpha$ (above 0.7) this function reaches a maximum at sizes between 2000 and 4000. This is also where most of the genomes in the data set are found, indicating that this range of genome sizes allows the optimal usage of universal and contextual domain families.



## S4. OTHER VARIANTS OF THE CRP

We discuss here mean-field arguments for the robustness of our results on the asymptotics of $F(n)$ for two variants of the original model, including a small domain loss rate and global duplications.

*a. Global Duplications.* One can consider the presence of global duplication moves. At each time step, if duplication is chosen, a number of domains selected with $q > 1$ trials from a binomial distribution with parameter $p_O^i$ is duplicated in the same time step. The innovation step remains the same. In this case, it is not possible to measure time with the size $n$ of the genome, but this observable follows the evolution equation

$$\dot{n} = qp_O + p_N \ , \tag{2}$$

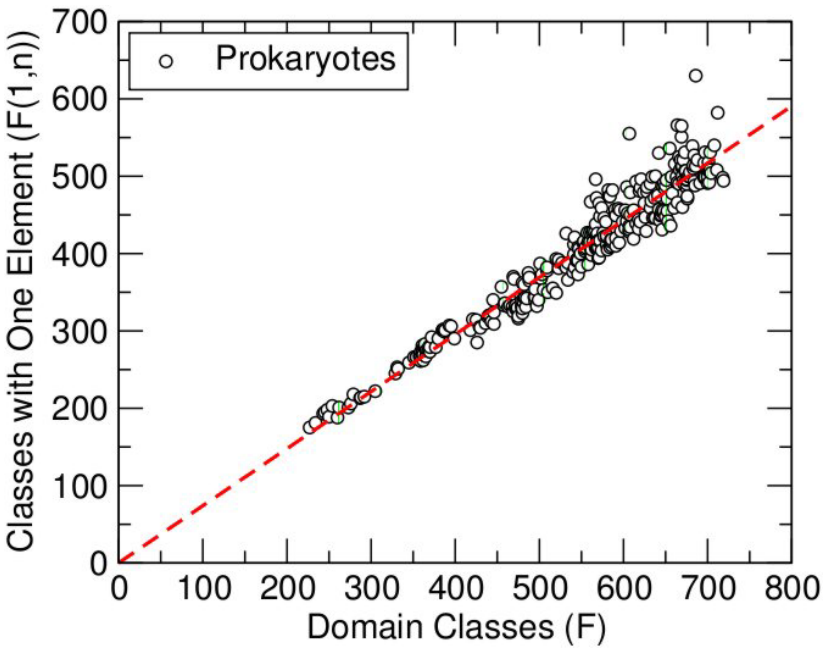
where $\dot{}$ indicates the derivative with respect to time $t$. In terms of $t$, our mean field equations are worked out simply as $\dot{F}(t) = p_N$ and $\dot{K}_i(t) = qp_O^i$. Using Eq. (2), they can be simply converted in terms of $n$, yielding

$$\partial_n F(n) = \frac{\alpha F(n) + \theta}{qn + (q-1)\alpha F(n) + \theta} \ ,$$

and

$$\partial_n K_i(n) = \frac{K_i - \alpha}{n + \frac{\theta}{q}} \ .$$

The first equation gives as leading scaling $F(n) \sim n^{(\alpha/q)}$, showing that the growth of $F$ is pushed towards effectively lower values of $\alpha$ by global duplications, as a consequence of the rescaling of time by the global moves. The dynamics for $K_i$, instead, is affected only by a renormalization of the parameter $\theta$. The qualitative results of the model are therefore stable to the introduction of a global duplication rate, in the hypothesis that the extent of these duplications does not scale with $n$.




Supporting Figure S4.7: The number of domain classes with one member, related to $F(1, n)$, as observed from the prokariote data set for superfamilies. The linear scaling is evident. A fit yields $\gamma \simeq 0.7$.

*b. Domain Loss.* A second interesting variant of the model considers the introduction of a homogeneous domain deletion, or loss rate. Domain loss is known to occur in genomes. However, it is not considered in our basic model for simplicity and economy of parameters. In order to introduce it in the CRP, we define a loss probability $p_L = \delta$. This is equally distributed among domains, so that the *per class* loss probability is $p_L^i = \delta \frac{K_i}{n}$. Consequently, the duplication and innovation probability $p_O$ and $p_N$ are rescaled by a factor $(1 - \delta)$. The mean-field evolution equation for the number of domain classes becomes

then

$$\dot{F}(t) = (1-\delta)\frac{\alpha F + \theta}{n+\theta} - \delta\frac{F(1,n)}{n} \ ,$$

where the sink term for $F$ derives from domain loss in classes with a single element, quantified by $F(1,n)$.

In order to solve this equation, one needs an expression for $F(1,n)$. Here, we report an argument based on the fact that *empirically*, $F(1,n) = \gamma F(n)$, with $0 < \gamma < 1$ (figure S4.7). Using this experimentally motivated ansatz, we can show that for small $\delta$, the scaling of $F(n)$ is subject only to a small correction. Again, since time does not count genome size, one has to consider the evolution of $n$ with time $t$, given in this model simply by $\dot{n} = 1 - 2\delta$. Using this equation it is possible to obtain the evolution equation for $F(n)$. Considering an expansion in small $\delta$ and large $n$, this reads to first order

$$\frac{\partial_n F(n)}{F(n)} = \frac{\alpha}{n}\left[1 + \delta\left(\frac{\alpha-\gamma}{\alpha}\right)\right] \ .$$

The above equation gives the conventional scaling for $F(n)$, with the aforementioned correction. Note that the correction could be positive or negative, depending on the relative values of $\alpha$ and $\gamma$. An analogous argument holds for $\alpha = 0$.