Divergent Evolution of a Structural Proteome: Phenomenological Models

C. Brian Roland* and Eugene I. Shakhnovich†
*Chemical Physics Program, and †Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts

ABSTRACT We develop models of the divergent evolution of genomes; the elementary object of sequence dynamics is the protein structural domain. To identify patterns of organization that reflect mechanisms of evolution, we consider the individual genomes of many procaryote species, studying the arrangement of protein structural domains in the space of all polypeptide structures. We view the network of structural similarities as a graph, called the organismal Protein Domain Universe Graph (oPDUG); vertices represent types of structural domains and edges represent strong structural similarity. As observed before, each oPDUG is a highly nonrandom graph, as evidenced in the vertex degree distribution, which resembles a Pareto law (which has a power-law asymptotic). To explain this and other peculiar properties of the oPDUGs, we construct an evolving-graph model for the long-timescale evolutionary dynamics of oPDUGs, containing only divergent mechanisms of domain discovery. The model generates degree distributions (resembling Pareto laws) and clustering-coefficient distributions that are characteristic of the oPDUGs. In the infinite-graph limit, we analytically compute the exponent for specific biological parameters, as well as the complete phase diagram of the model, finding two distinct regimes of domain innovation dynamics. Thus, divergent evolutionary dynamics quantitatively explains the nonrandom organization of oPDUGs.

INTRODUCTION

Within the field of molecular biophysics, there are unanswered questions regarding the molecular evolution of proteins. In particular: if two protein structural domains have similar structure, do they necessarily have a common ancestor (1)? The answer to this question concerns the relative importance of divergent and convergent mechanisms of protein domain fold discovery. In this work, we describe a divergent model of protein evolution in structure space. We compute the properties of this model, and compare the results to the properties of proteomes of real organisms. Before we delineate our model, we first present the experimental observations that check different explanations of protein evolution.

One of the striking observations made in the structural genomics effort is the uneven fold usage within the genomes of individual organisms: the number of SCOP-folds with a given number of genes follows a power law (2–6). For several organisms, the distribution of domain genes among folds was captured by a simple divergent model (3) with a ''rich get richer'' mechanism. This mechanism presents one explanation of the uneven distribution of domains among folds.

However, it is accepted that alternative models, employing convergent hypotheses, could also explain the uneven fold distribution (7). Dokholyan et al. (8) suggested that in addition to the fold distribution, the structural similarities between domains within a fold also held discriminating evolutionary information. This suggestion is motivated by earlier observations: ''Although the definition of discrete fold types is useful for counting purposes, the Dali Domain Dictionary also makes explicit the graded similarities that exist between the members of the same fold type and that may extend beyond the borders of fold categories'' (9). The arrangement of the set of Dali domains in structure space (space of all C_α-traces) is clarified by viewing the set as a network in which the nodes are the domains and the binary interactions between nodes correspond to the pairwise structural similarities (Dali Z-scores). The graph representation of this network is called the protein domain universe graph (PDUG).

In a subsequent work, Deeds et al. (10,11) studied how individual genomes cover structure space through the analysis of organismal PDUGs (oPDUGs). For each of many fully sequenced procaryote genomes, putative structural domain sequences were identified by, and assigned to, Dali domains according to high sequence similarity with a Dali-domain sequence. Thus, the list of Dali domains ''present'' in the genome was assembled. In an oPDUG, each Dali domain present in the genome is represented as a labeled vertex, and two vertices are connected by an edge if their Dali Z-score exceeds a cutoff. At an organism-independent value of the cutoff, each oPDUG exhibited a nonrandom global connectivity, visible in the degree distribution; in graph terminology, the number of edges emanating from a vertex is called ''the degree of a vertex'' (see Appendix, The degree in the mathematics of graphs, for further explanation), and the fraction of vertices with a certain degree is called ''the degree distribution''. Specifically, it was found that the degree distribution of each oPDUG differed strongly from that of a classical random graph with the same number of vertices and edges (in this type of graph, edges are placed between randomly chosen pairs of vertices). This striking feature of the oPDUGs provides a stringent experimental measurement, even more discerning than the uneven fold distribution, against which to compare divergent and convergent models of protein evolution.

Submitted January 17, 2006, and accepted for publication August 28, 2006.

Address reprint requests to E. I. Shakhnovich, Dept. of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138. Tel.: 617-495-4130; Fax: 617-384-9228; E-mail: eugene@belok.harvard.edu.

0006-3495/07/02/701/16 $2.00

doi: 10.1529/biophysj.106.081265

The nonrandom degree distribution of the oPDUGs was captured by a divergent model of oPDUG evolution (8,10). In the following, we attempt to resolve the mechanisms at work in this previous model. In so doing, we formulate and characterize a phenomenological model of oPDUG evolution that 1), reproduces the nonrandom connectivity of real oPDUGs (which have finite size); and 2), allows analytical calculation of the behavior of infinite graphs. Specifically, simulation results suggest that in the limit of large graphs (long evolutionary times), the model generates graphs that are well fit by a power law at high degree, i.e., the graphs are asymptotically scale-free. We analytically compute the scale-free exponent in the large-graph, high-degree limit. Our primary result is that as the graph size increases in our model, the nonrandom connectivity of finite graphs develops into the asymptotically scale-free connectivity of infinite graphs.

First, we attempt to identify important features, in addition to the nonrandom degree distribution, of the studied oPDUGs. In particular, we calculate the distribution of clustering coefficients for four importantly different organisms. Second, we present two models of the time development of oPDUGs, each isolating mechanisms at work in previous models. Specifically, the second of the two models has a mechanism with a type of memory not present in the first. Third, we computer-simulate the models in a finite time regime. We find that, in this time regime, the memory-full mechanism does not outperform the memory-less mechanism with respect to reproducing course features of the real oPDUGs. Fourth, we discuss analytical results for the long-time behavior of the memory-less model, presenting the phase diagram. Last, we discuss the successes and failures of the models, and the implication for the evolution of organisms.

BIOLOGICAL DATA: ARRANGEMENT OF DALI DOMAINS IN STRUCTURE SPACE

Genomes under study

A phylogenetic analysis of many procaryote genomes, in which the characters correspond to the presence or absence of the Dali domains, revealed that the genomes had many domains in common (12). Thus, to study the features generic to all of the oPDUGs, we need only study a small number of examples that are significantly different from each other. We consider a single example from each of four major clades identified in the neighbor-joining phylogeny (12). We study *Agrobacterium tumefaciens* C58 from the proteobacteria-like clade, *Streptococcus pneumonia* R6 from the gram-positive-like clade, *Campylobacter jejuni* from the ϵ-proteobacteria-like clade, and *Archaeoglobus fulgidus* from the archae-like clade.

Failure of the classical random-graph model (CR) for the oPDUG

Our first aim is to evaluate the evolutionary information content of the oPDUGs. Under the null hypothesis that the oPDUG does not hold signals about the nature of evolution (dynamics or driving forces), we define the CR model for the architecture of an oPDUG. In the CR model, which stands for classical random-graph model, the statistics of an oPDUG is similar to that of a classical random graph (13), with the same number of vertices and edges. If we define a graph ensemble as a collection of graphs such that each graph is prepared according to certain probabilistic rules, a single graph in the CR graph ensemble is made by assigning edges to pairs of vertices at random (uniformly). We call each graph in the CR ensemble a shuffled oPDUG, and compute 50 such shufflings. The assembly of the CR ensemble washes away correlations in the edges shared by any group of three vertices; in this sense, this graph ensemble is devoid of nontrivial evolutionary information.

To compare the CR model to a real oPDUG, we compute the degree distribution and clustering-coefficient distribution as follows. For a given oPDUG, each domain (vertex) has some number of structural neighbors (edges); in graph terminology, the number of edges emanating from a vertex is called the degree (see Appendix, The degree in the mathematics of graphs, and Albert and Barabasi (13)). The degree distribution,

$$n_k[oPDUG] = \text{fraction of vertices with degree } k, \quad (1)$$

is plotted in Fig. 1. For each oPDUG, the number of vertices and edges defines a CR model, for which we compute the ensemble-averaged degree distribution, which is the degree distribution averaged over all graphs in the ensemble (Fig. 1). Specifically, for any ensemble of graphs $\{G^m\}$, $m \in \{1, \ldots, M\}$, the ensemble-averaged degree distribution is denoted by

$$\langle n_k[G] \rangle = \frac{1}{M} \sum_{m=1}^{M} n_k[G^m]. \quad (2)$$

In Fig. 1, N is the number of vertices and $D = 2E/N$ is the graph degree (see Appendix, The degree in the mathematics of graphs), i.e., the average-over-vertices of the degree of a vertex, where E is the total number of edges in the oPDUG. Additionally, for a given oPDUG, we compute the clustering coefficient for each vertex t with degree $k_t \geq 2$ according to the standard definition

$$C_t = \frac{2E_t}{k_t(k_t - 1)}, \quad (3)$$

where E_t is the number of edges between neighbors of vertex t (13); C_t is not defined for vertices with $k = 0$ or $k = 1$. The clustering-coefficient distribution is the histogram of these values, normalized to the total number of $k \geq 2$ vertices, N_C (Fig. 2). For the corresponding CR model, we compute the clustering-coefficient distribution for each graph in the ensemble, normalizing each distribution by the value of N_C for that particular graph. The normalized distribution is averaged over all graphs in the CR ensemble, giving equal weight to

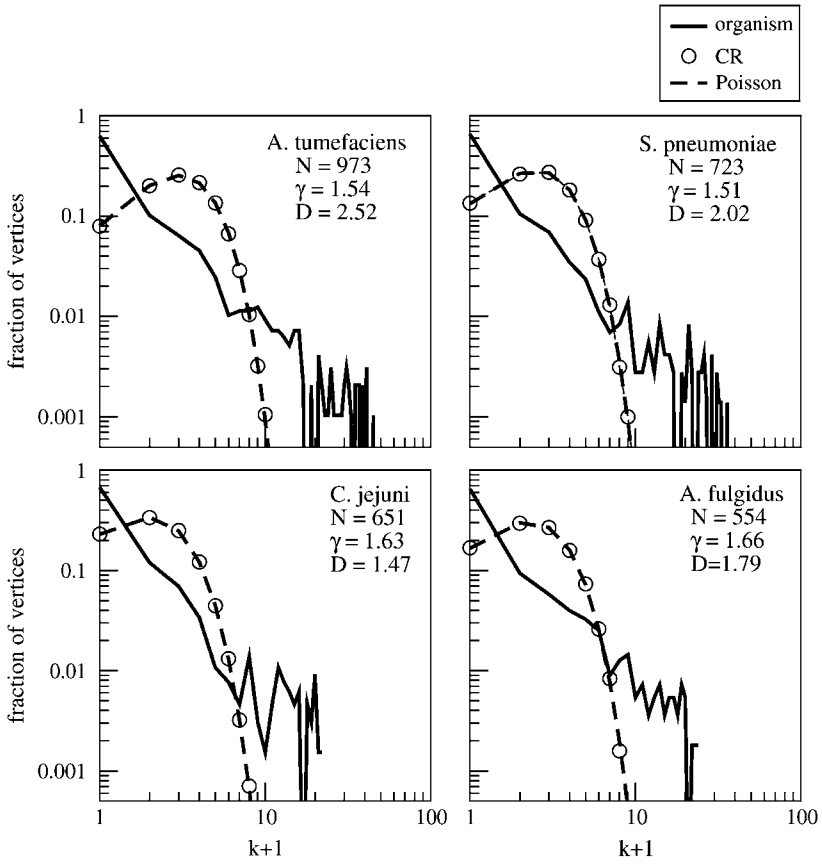

FIGURE 1 oPDUG degree distributions (*solid line*) for a representative from each of four major clades. N is the total number of vertices in a graph. γ is the exponent of the fit to the Pareto law $A/(k+1)^{\gamma}$; the fit was performed on the entire interval spanned by the data (10). D is the graph degree. The curve intersects the abscissa where the fraction of vertices is zero. Also shown is the ensemble-averaged degree distribution for the CR model (*circles*). Each graph in the CR ensemble has the same number of vertices and edges as the corresponding oPDUG, but the edges are assigned to pairs of vertices at random. The Poisson distribution (*dashed line*) fits the CR model.

each graph; this results in the ensemble-averaged clustering-coefficient distribution shown in Fig. 2.

For each oPDUG studied, the degree distribution differs strongly from the prediction of the CR model (Fig. 1). By comparison, the degree distribution of a real oPDUG shows 1), many vertices with degree $k = 0$ (orphans); and 2), many vertices with degree $k \gg D$. Since CR represents the null hypothesis, we purport that features 1 and 2 are signatures of evolution (either dynamics or driving forces). To highlight these features and to put them on equal footing, we view the degree distribution in log-log scale with the k axis shifted. Features 1 and 2 result in an approximately straight line in this scale. The simplest function to satisfy this observation is the Pareto law

$$\frac{A}{(k+1)^{\gamma}}, \tag{4}$$

where parameters A and γ are coupled through the normalization condition. Thus, it has proven convenient to characterize the degree distribution of each oPDUG with a fitted value of γ (10). We do not claim, however, that the data is better fit to a Pareto law than every other analytical function. We claim only that the Pareto law is a convenient shorthand for the presence of features 1 and 2, and that γ conveniently gives the relationship between the two.

Additionally, features 1 and 2 can be quantified in terms of the standard deviation (spread) of the degree. Consistent with the generic properties of classical random graphs, CR is well fit by a Poisson distribution

$$n_{\mathrm{k}}^{\mathrm{CR}} \cong e^{-\mathrm{D}} \frac{D^{\mathrm{k}}}{k!}, \tag{5}$$

with parameter D, where the mean value of the distribution is D and the standard deviation is $\sqrt{D}$ (13). For *A. tumefaciens*, $D = 2.52$ is the graph degree, so for the corresponding CR degree distribution, the mean is 2.5 and the standard deviation is 1.3. Thus, the CR degree distribution is relatively well localized about the mean value, i.e., it has a small spread. However, whereas the oPDUG degree distribution has a mean value of 2.5 as well, the standard deviation is computed to be 6.3. The oPDUG degree distribution has a large spread compared to the corresponding CR degree distribution.

For each oPDUG studied, the clustering-coefficient distribution differs strongly from the prediction of the CR model (Fig. 2). In the oPDUGs shown, of the vertices with $k \geq 2$, 25–30% have $C = 1$ and 5–15% have $C = 0$ (*stars*). Fig. 3, which summarizes gross properties of all 59 proteomes, shows that most organisms have a fraction of $C = 1$ and $C = 0$ vertices, consistent with the four clade examples. These results are in strong disagreement with the CR model, suggesting that the clustering-coefficient distribution also contains information about evolution (dynamics or driving forces).

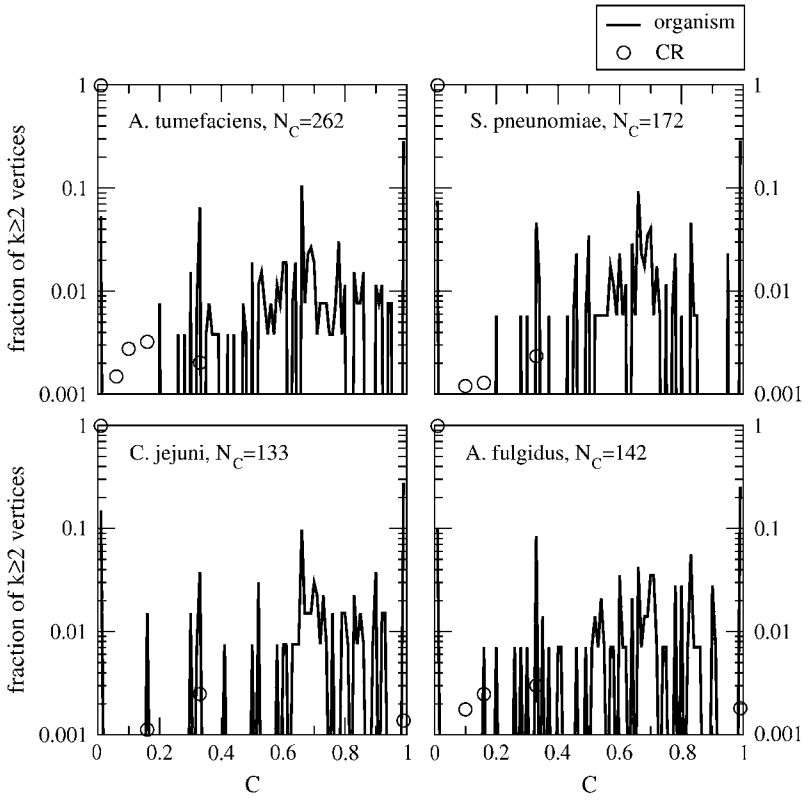

FIGURE 2 oPDUG clustering-coefficient distributions (*solid line*) for a representative from each of four major clades. The bin size is 0.01. For easy viewing, the peaks at $C = 1$ and $C = 0$ are artificially shifted away from plot boundaries. N_C is the total number of vertices with $k \geq 2$ in the real oPDUG. Also shown is the ensemble-averaged clustering-coefficient distribution for the CR model (*circles*).

In summary, the network statistics of the oPDUGs is nonrandom in the sense that the degree and clustering-coefficient distributions compare poorly with the predictions of the CR model.

Criteria for model development

In the next section, we complement the null hypothesis by proposing a model for the evolutionary dynamics that generated the oPDUGs. To compare the oPDUGs with this model, we restrict our focus to *A. tumefaciens*, because it has the largest number of domains. From the key features of the *A. tumefaciens* oPDUG, we list the following set of criteria for a stochastic-dynamic model of the evolution of an oPDUG. Within the ensemble of graphs generated by the model, there must be a significant fraction of individual graphs with the following features:

1. the degree distribution is well fit by the Pareto law at low degree
2. the degree distribution is well fit by the Pareto law at high degree
3. the exponent of the Pareto law is ~1.6

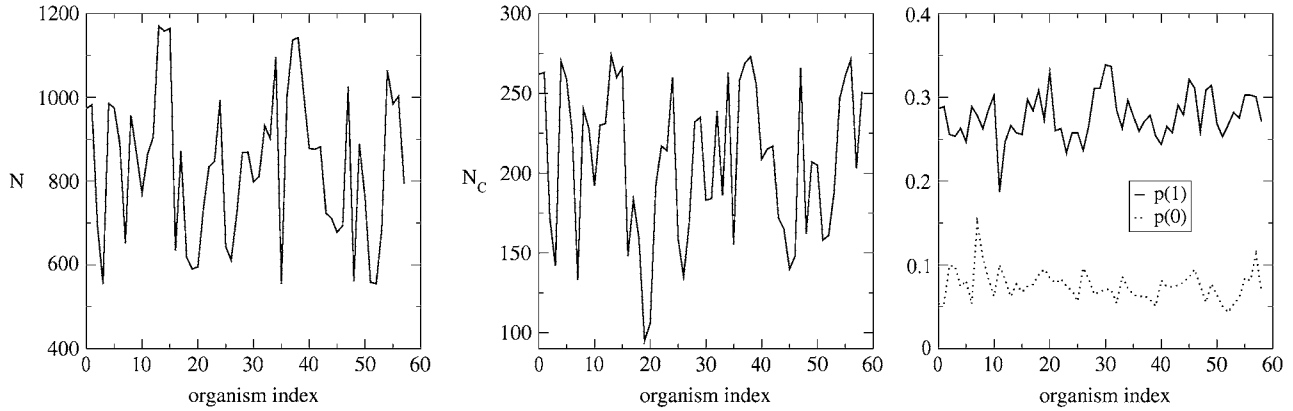

FIGURE 3 Gross properties of the 59 procaryote oPDUGs. For a given oPDUG, N is the total number of vertices, N_C is the number of vertices with degree $k \geq 2$, and $p(0)$, $p(1)$ are the values of the clustering-coefficient distribution at $C = 0$ and $C = 1$. The organism index is defined in Deeds et al. (10).

4. the distribution of clustering coefficients has a strong peak at $C = 1$
5. the distribution of clustering coefficients has a strong but subdominant peak at $C = 0$.

MODEL: THE DUPLICATION AND DIVERGENCE OF PARALOG FAMILIES

Sequence pockets and the structural proteome

The architecture of polypeptide sequence space motivates our models of how a genome's Dali domains are arranged in structure space, i.e., of its oPDUG. We thus recount the bioinformatic method described by Mirny et al. (14), which attempts to identify the pattern of sequence positions involved in physical interactions important to stabilizing a sequence in its native-state fold. For a given fold, e.g., immunoglobulin, it was found that the positions that are highly conserved within any given family of sequences (high sequence identity) match up, after structural alignment, with the conserved positions in any other family of proteins having the same fold. This matching up appeared in high values of the conservatism-of-conservatism (CoC) for certain sequence positions (14). For a given fold, each family differed in types of residues at the high CoC positions, but had the pattern of high CoC positions in common. These quantitative results are consistent with earlier qualitative observations: ''The map [of fold space] further reveals a small number of densely populated regions where the common features are topological motifs at the core of the domains'' (9).

Based on this observation, we model the space of all polypeptide sequences by assuming the existence of evolutionarily stable regions called sequence pockets (see Fig. 4) (15). Sequences in a pocket are similar in the sense that there is a pattern of key sequence positions, i.e., the high CoC positions, at which the residue type (in a reduced alphabet) is the same for each sequence. We call this pattern the fold pattern. The residue types at non-key sequence positions provide the degrees of freedom that give volume to the pocket. Each sequence is evolutionarily stable in two respects. First, each sequence in the pocket is useful to the cell because it folds as an independent unit to a well defined native conformation, i.e., each sequence corresponds to a structural domain (for definitions of a structural domain, see Holm and Sander (9)). Second, if we make a single residue-type substitution (reduced alphabet) at any one of the positions in the fold pattern, the resulting sequence does not fold independently. A sequence pocket is defined both by the fold pattern and the residue type at each position in the fold pattern.

The pocket of structural domain sequences corresponds to a set of native-state structures that forms a localized cluster in the space of all polypeptide structures (Fig. 5). Thus, a sequence pocket can be labeled with, and well represented by, any member sequence and its associated structure. We take each sequence pocket to precisely define a type of structural domain, i.e., to define a domain type.

If we define an organism's complete proteome as the collection of all structural domain sequences contained in its genome, we imagine the evolution of its complete proteome as a dynamic process in which previously unoccupied sequence pockets become occupied. In other words, evolution is described as a timeline of domain-type discoveries. Initially, a single seed sequence occupies each of a set of pockets. On some short evolutionary timescale $t_\circ$, each seed sequence gives rise to a steady-state population of descendants that are similar in the sense that all are confined to the parent's pocket (Fig. 6). We call this monophyletic population of sequences in a stable pocket belonging to the same genome a paralog family (group of paralogous sequences), and consider it to be the elementary unit of evolution. On some longer evolutionary timescale, each paralog family gives rise to a sequence that seeds a distinct pocket, thus overcoming the single-residue-substitution barrier that confines the family on short timescales (Fig. 7). In accord with the scenario of divergent evolution, we assume that the distinct pocket is typically unoccupied, i.e., that a new domain type is discovered.

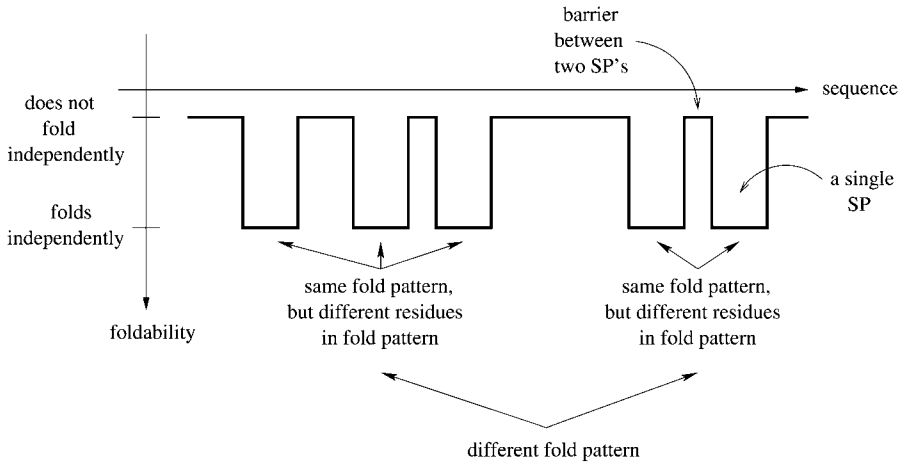

FIGURE 4 Our model for the organization of polypeptide sequence space. There are stable regions called sequence pockets (*SP*); inside an SP, all sequences share both the same fold pattern and the same residue type at each position within the pattern. SPs are separated from each other by sequences that do not fold independently.

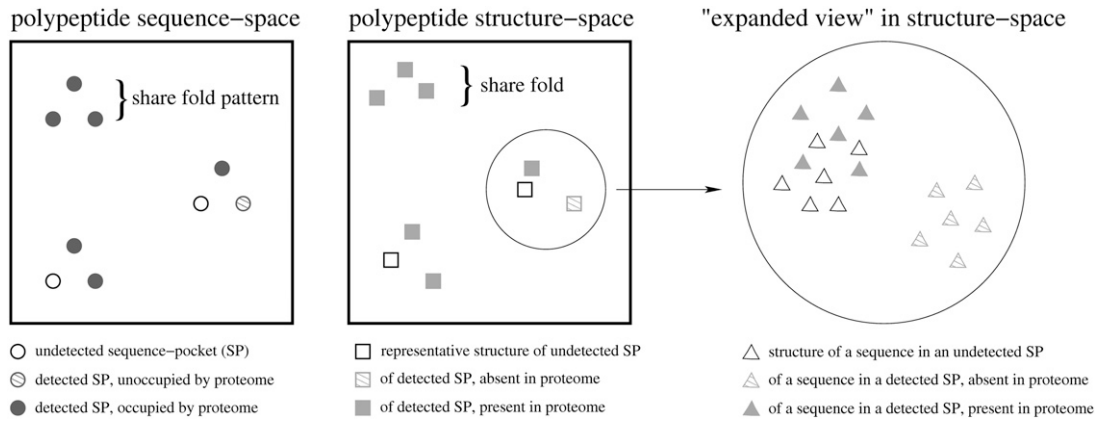

FIGURE 5 Our model for the organization of polypeptide sequence space and structure space. The cartoonized distances in sequence space correlate with the timescale for spontaneous mutation from one sequence to another. In structure space, the representative structures of two SPs are close (far) if the fold pattern is common (distinct). On the far right is an ''expanded view'' of the region of structure space corresponding to three SPs with common fold pattern; we show structures for all sequences in each SP. Here, we schematize how an organism's proteome might populate the SPs.

We now accommodate the constraint that the observer can only detect a subset of all sequence pockets, using bioinformatics. Thus, we define the long-timescale evolutionary state of an organism by the occupation states (filled or empty) of detected sequence pockets, i.e., by its structural proteome—which we define as the list of detected domain types found in the organism's genome (Fig. 5). We now identify the sequence-structure pairs in the Dali Domain Dictionary as an experimental proxy for the set of detected domain types in our caricatured model of protein sequence space.

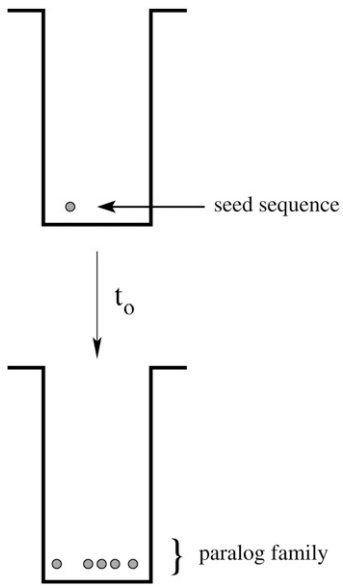

FIGURE 6 A paralog family is a lineage (having common ancestor) of structural domain sequences confined to a sequence pocket. A single seed sequence undergoes duplication-and-divergence events to produce a paralog family on timescale $t_\circ$ (15), which is the timescale on which a given sequence spontaneously mutates into another sequence within the same SP.

Although the organization of sequence space into stable pockets is a reasonable model for the high-sequence-identity region of structural-similarity versus sequence-similarity plots (16,17), and the picture of evolution through domain-type discovery is highly plausible, the following question arises: what kinds of domain-type discoveries explain the current evolutionary state (structural proteome) of organisms?

Divergent versus convergent evolution

At this point, we distinguish between three kinds of domain-type discovery. An occupied sequence pocket may seed an unoccupied sequence pocket with a fold that is 1), the same as the parent fold—we call this ''neutral fold discovery''; 2), different from the parent fold and unoccupied by any paralog family—''divergent fold discovery''; and 3), different from the parent fold but already occupied by paralog families—''convergent fold discovery''.

One explanation of the uneven fold population is the convergent evolutionary hypothesis: there have been many convergent fold discoveries, and the organism either 1), selects domain types corresponding to folds that accommodate many functions; or 2), exhibits a larger number of domain types for folds that can intrinsically accommodate more

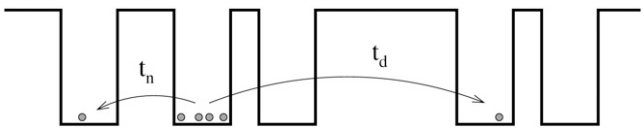

FIGURE 7 Sequence-pocket discovery events. In a neutral fold discovery, a sequence in some SP spontaneously mutates into a sequence in a different SP with the same fold pattern on some intermediate timescale $t_n \gg t_o$. In a divergent fold discovery, a sequence in some SP spontaneously mutates into a sequence in an SP with a different fold pattern on some long timescale $t_d \gg t_n \gg t_o$ (see Model, Spontaneous versus fixed mutations).

distinct domain types (high designability). Another explanation, the divergent evolutionary hypothesis, is that convergent fold discoveries are rare or nonexistent. In this work, we attempt to explain the network statistics (degree and clustering-coefficient distributions) of structural proteomes using only neutral and divergent fold discovery.

Spontaneous versus fixed mutations

A physically reasonable assumption for the organization of sequence space is that the spontaneous mutations that cause neutral fold discovery occur much more frequently than the spontaneous mutations that cause divergent fold discovery. That is, the timescale for divergent fold discovery, t_d, is much longer than the timescale for neutral fold discovery, t_n, i.e., $t_d \gg t_n$ (see Fig. 5 and Fig. 7). We take this inequality to be a key property of our model of protein sequence space (see Model, Sequence pockets and the structural proteome).

However, spontaneous mutations may not survive generations of reproduction of a population of organisms if they do not provide a substantial fitness advantage. It is possible that the spontaneous mutations causing neutral fold discovery are fixed in a population very infrequently, whereas the spontaneous mutations causing divergent fold discovery are fixed very frequently (personal communication 2005, E. J. Deeds). For simplicity, we assume that fixation effects restore the balance between neutral- and divergent-discovery rates. Thus, in the models that follow, there is no explicit distinction between neutral and divergent fold discovery.

Evolving-graph models

M0: model without memory

According to the evolutionary model above (Sequence pockets and the structural proteome), the evolutionary state of each procaryote is synonymous with its structural proteome. At this point, we choose to reduce the detail of our description by tracking the oPDUG representation of the evolutionary state.

Our first model for the time development of an oPDUG is described as follows. Preliminarily, we define a set of discrete time points $t \in \{1, \ldots, N\}$, each interval of time between $t-1$ and t is a time step labeled with the terminal time point t. We imagine that during the first time step, $t = 1$, of the existence of the structural proteome, a single seed sequence fills some sequence pocket. This sequence pocket is represented by a single vertex on the oPDUG, labeled $t = 1$. At the start of each subsequent time step t, a randomly-chosen occupied sequence pocket spawns an intrepid sequence (due to gene duplication) that is destined to discover a new sequence pocket. At the time of duplication, the intrepid sequence resides in the parent sequence pocket; thus, the structural similarity to the parent's representative structure is high.

On the graph, the birth of the intrepid sequence is represented by adding a ''baby'' vertex, with label b indicating the time step $b = t$ of creation, which shares an edge (*dashed*) with a randomly-chosen parent vertex p (Fig. 8). The baby vertex also shares an edge (*dotted*) with each neighbor vertex n_i, where a neighbor is a vertex that shares an edge with the parent. Subsequently, the intrepid sequence undergoes mutations that place it in a new sequence pocket, where it seeds a new paralog family. The representative structure for the new sequence pocket will sometimes be similar to—or different from—the parent pocket's structure, and thus also to the structures of the parent's structurally neighboring pockets. This variation in the magnitude of structural divergence is motivated by the diversity of structural similarities in the ''twilight'' and ''midnight'' zones of protein sequence alignments (17,18).

Thus, on the graph, at the end of its first time step of life, the baby vertex has retained the baby-parent edge with retention probability r_p (Fig. 8). Given that the baby-parent edge is retained, each baby-neighbor edge is retained with probability r_n. If the baby-parent edge is not retained, then all of the baby-neighbor edges are lost as well; we call this baby vertex an orphan. The diverged baby vertex corresponds to a newly occupied sequence pocket.

We call this model M0, for zero memory. This designation will be given meaning in the next section.

M1: a model with memory

We now comment on a particularly relevant feature of the model M0. To do so, we define a special type of memory associated with the baby-neighbor edge retention mechanism of duplication-and-divergence models. Consider a particular vertex that parents a baby at time step s and at time step $t > s$. Of the parent's neighbors that exist at time s, the subset that retain edges with baby s may be correlated with the subset that retain edges with baby t. A correlation between the two subsets means that they typically have an unusually large number of vertices in common. If the correlation between these two subsets is nonzero, then we say that the baby-neighbor edge retention mechanism has memory.

Note that the baby-neighbor edge retention mechanism of M0 has zero memory. This feature is the most fundamental

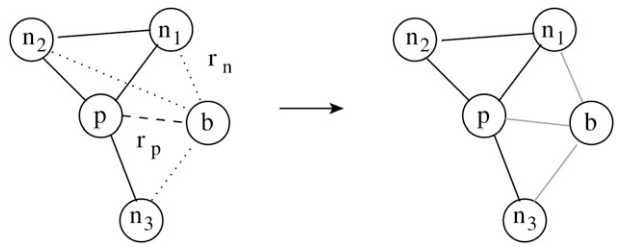

FIGURE 8 Schematic of M0, showing the divergence of a baby vertex during its first time step of existence. At birth, the baby vertex b shares edges with the parent p (*dashed*) and each of the neighbors n_i (*dotted*). After a single time step, the $b-p$ edge is retained with probability r_p. If the $b-p$ edge is retained, then each $b-n_i$ edge is independently retained with probability r_n. The edges that survive the divergence (*grey*), for one particular realization of the probabilistic process, are shown on the right.

distinction between M0 and previous models (8). To evaluate the importance of memory, we considered a second model called M1—for full memory—that contains a mechanism with extremely strong memory. In the Supplementary Material, we show that the memory-full mechanism does not outperform the memory-less mechanism with respect to mimicking the real oPDUGs; we compute the degree and clustering-coefficient distributions of M1. We conclude that memory mechanisms are not essential to explaining the data. Therefore, we present M0 as a simple representation of the evolutionary dynamics that generated the real oPDUGs.

RESULTS

Finite-graph ensembles of M0

To fit M0 to the *A. tumefaciens* oPDUG, the ''phase diagram'' of the model was surveyed. At each of a large number of points in the parameter space—the (r_p, r_n)-square, a graph ensemble was generated by computer simulation. The graph ensemble, called $\Gamma^0(r_p, r_n; N)$, is a collection of M graphs in which each individual graph is independently evolved with parameters (r_p, r_n) for $t \in \{1 \ldots, N\}$. For the results shown, $M = 10^3$ and $N = 10^3$.

According to the criteria for agreement with the *A. tumefaciens* oPDUG listed in Biological Data, Criteria for model development, the parameter values $(0.6, 0.8)$ provide best fit for $\Gamma^0(r_p, r_n; N)$. The parameters were chosen such that the ensemble-averaged degree distribution has a Pareto-fit exponent that approximates the exponent obtained for *A. tumefaciens*. Fig. 9 shows that $\Gamma^0(0.6, 0.8; N)$ contains individual graphs that mimic the oPDUG degree distribution. Additionally, we compute the Pareto-fit exponent for each member of the ensemble, and plot their distribution (Fig. 10). Thus, $\Gamma^0(0.6, 0.8; N)$ contains a significant fraction of individual graphs with exponents similar to that of the real oPDUG. M0 fails, however, to generate the high fraction of orphans observed in the real oPDUG.

In Fig. 9, we plot the clustering-coefficient distributions of $\Gamma^0(0.6, 0.8; N)$. M0 over-represents N_C, which is the total number of $k \geq 2$ vertices. This over-representation is probably related to the under-representation of orphans. M0 succeeds in generating a dominant peak at $C = 1$, and a subdominant peak at $C = 0$.

We summarize these results by saying that M0, for finite-graph sizes, captures the nonrandom statistics (degree and clustering-coefficient distributions) of the real oPDUGs. In the next section, we demonstrate that the nonrandom degree distribution of the finite-graph ensemble becomes scale-free at high k for large N.

Infinite-graph ensembles of M0

We study the asymptotic behavior, with increasing N, of the M0 phase diagram. First, we map a nonanalytic transition of the model, using the ensemble-averaged graph degree as a global order parameter. In particular, we find a surface in parameter space at which the graph degree exhibits a divergence for infinite N (Eq. 7). Second, we obtain the

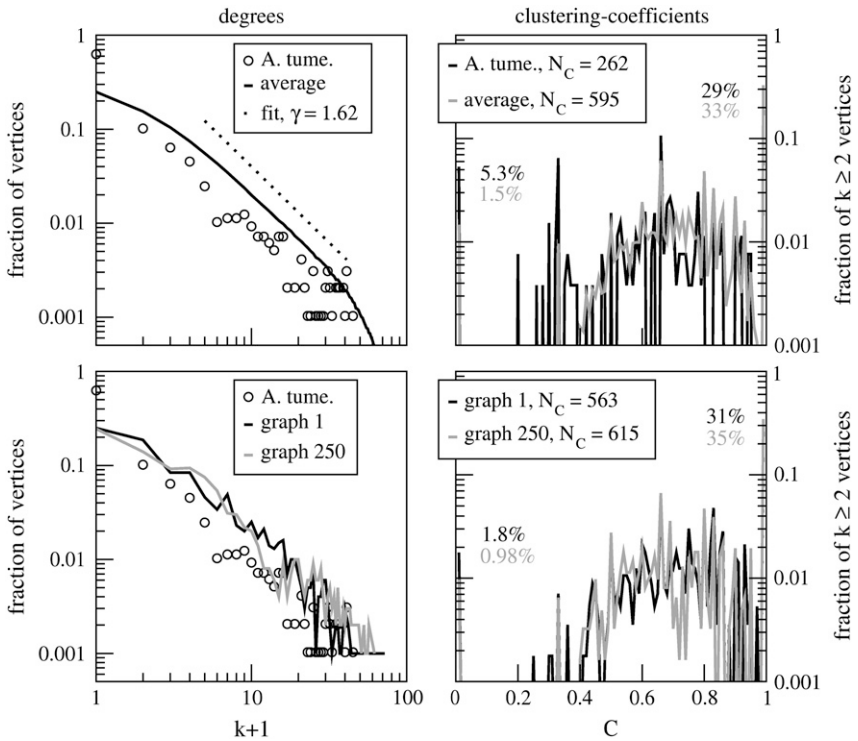

FIGURE 9 The statistics of the model M0 compared to the oPDUG of *A. tumefaciens*. For the ensemble-averaged degree distribution of M0, the Pareto-law fit with vertical shift is shown, and indicates the interval on which the fit is performed. The lower limit of the fitting interval is $k = 4$ and the upper limit is the ensemble average of $k_{max}[G]$ minus the standard deviation in $k_{max}[G]$; $k_{max}[G]$ is the degree of the most highly connected vertex in graph G.

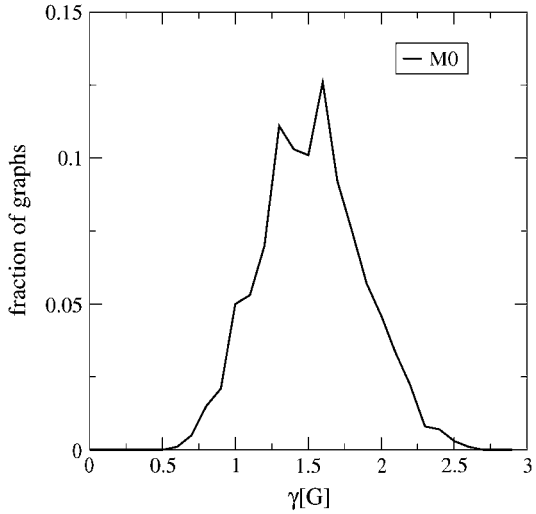

FIGURE 10 The distribution of single-graph degree distribution exponents within $\Gamma^0(0.6, 0.8; N)$, for $M = 1000$ and $N = 1000$. Each exponent is obtained from a Pareto fit on the interval $k \in \{4, \ldots, k_{\max}[G]\}$. Bin size is 0.1.

analytical solution of the high-k behavior of the ensemble-averaged degree distribution in the $N \to \infty$ limit, using a power-law ansatz.

Ensembled-averaged graph degree

We consider the $\Gamma^0(r_p, r_n; N)$ ensemble in which each graph G^m—labeled $m \in \{1 \ldots .M\}$ with $M \gg N$—is independently evolved under the rules of M0 (see Model, M0: model without memory). We consider $r_p, r_n \in [0, 1]$, where r_p and r_n are the baby-parent and baby-neighbor edge retention probabilities, respectively. For graph G^m at time $t = N$, $E[G^m(N)]$ is the total number of edges and $D[G^m(N)]$ is the graph degree, i.e., the average-over-vertices of the single-vertex degree. We note the general relation $D[G^m] = 2E[G^m]/N$.

We define the ensemble-average of $E[G^m]$ at time $t = N$ as

$$E(N) \equiv \frac{1}{M} \sum_{m=1}^{M} E[G^m(N)]$$

and similarly define $D(N)$ as the ensemble average of $D[G^m]$ (see Appendix, Ensemble-averaged graph-degree). It can be shown that $E(N)$ obeys the equation

$$\frac{dE(N)}{dN} = r_p + \frac{2r_p r_n}{N} E(N), \tag{6}$$

where we have treated N as a continuous variable. The solutions of this equation can be obtained by an isomorphism with the edge-growth equation for the model of Redner et al. (19); additionally, we give a derivation that includes correction terms in Appendix, Ensemble-averaged graph-degree). The large-N behavior of the solution, given in terms of average graph degree $D(N)$, is

$$D(N) \cong \begin{cases} \dfrac{r_p}{0.5 - r_p r_n}, & \text{if } r_p r_n \in [0, 0.5) \\ 2r_p \log N, & \text{if } r_p r_n = 0.5 \\ \dfrac{r_p}{r_p r_n - 0.5} N^{2r_p r_n - 1}, & \text{if } r_p r_n \in (0.5, 1] \end{cases} \tag{7}$$

for $r_p \in [0, 1]$. We call the $r_p r_n < 0.5$ region of the phase diagram the linear regime, because $E(N)$ grows linearly with N [Fig. 11]. In like manner, we call the line $r_p r_n = 0.5$ the linear-log growth regime, and the region $r_p r_n > 0.5$ the super-linear regime. We summarize these results by noting that in the linear regime, the ensemble-averaged graph degree approaches a finite value as N increases, whereas in the super-linear regime, it grows without bound (goes to infinity). We call the transition between the two regimes a nonanalytic transition,n because for infinite N, $D(N)$ diverges as the line $r_p r_n = 0.5$ is approached from below.

Ensemble-averaged degree distribution

Our computer simulations suggest that, as N increases, the ensemble-averaged degree distribution of M0 develops a stable power-law regime at high k (Fig. 12, and we include Fig. 13 for completeness). To characterize M0, we would like to calculate the value of the exponent in the large-N limit. Thus, we again consider the ensemble $\Gamma^0(r_p, r_n; N)$, but restrict our attention to $r_p, r_n \in (0, 1)$. Given graph G^m at time $t = N$, we call $n_k[G^m(N)]$ the fraction of vertices with degree k; this is the degree distribution of graph G^m. The ensemble-averaged degree distribution is then

$$n_k(N) \equiv \frac{1}{M} \sum_{m=1}^{M} n_k[G^m(N)]. \tag{8}$$

To compute how $n_k(N)$ changes from time $t = N$ to $t = N+1$, we derive in Appendix, Ensemble-averaged degree distribution, the rate equation

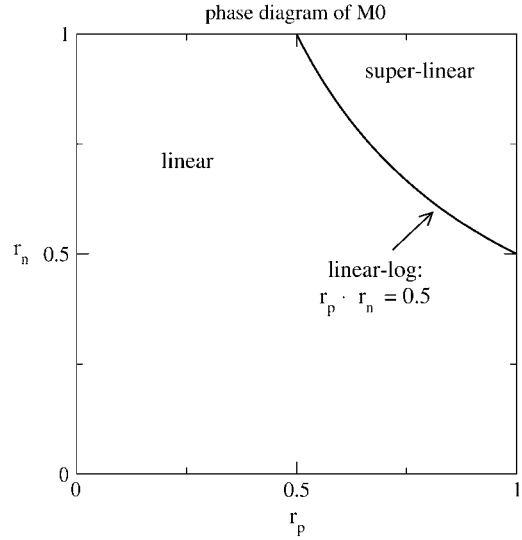

FIGURE 11 The phase boundary between the linear and superlinear regimes of M0 in the infinite-graph limit.

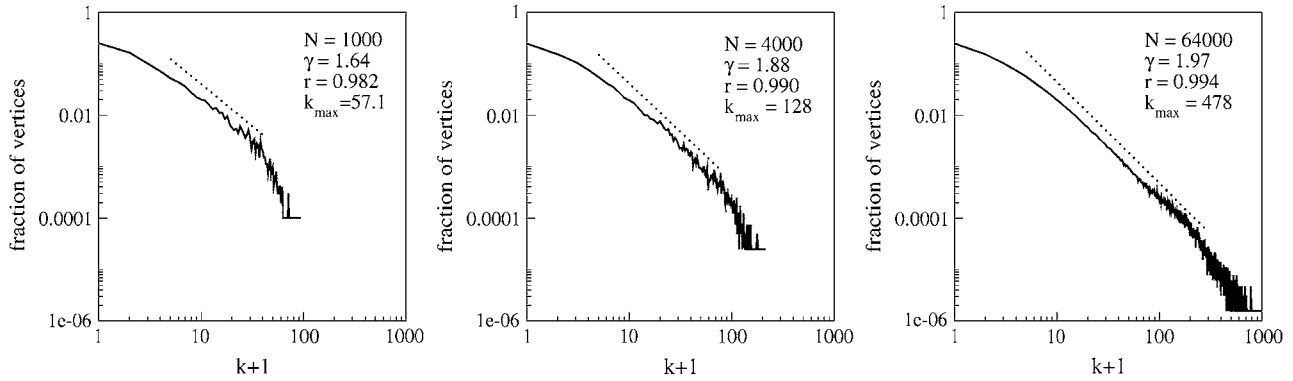

FIGURE 12 The high-k region of the degree distribution (*solid line*) of M0 is well-fit by a Pareto law $A/(k+1)^{\gamma}$; the fit (*dotted line*) is shown with vertical shift and indicates the interval on which the fit is performed. Both the quality of fit (r is the correlation coefficient) and the size of the fitting window increase with N, suggesting that the degree distribution is scale-free at high k for large N. Each plot is the ensemble-averaged degree distribution with $M = 10$ and $(r_p, r_n) = (0.6, 0.8)$. The lower limit of the fitting interval is $k = 4$ and the upper limit is the ensemble average of $k_{max}[G]$ minus the standard deviation in $k_{max}[G]$; $k_{max}[G]$ is the degree of the most highly connected vertex in graph G. The value k_{max} in each plot is the ensemble average of $k_{max}[G]$.

$$\Delta N_k(N+1) = A_{k-1}\, n_k(N) - A_k\, n_k(N) + r_p \sum_{\ell=k-1}^{N} \binom{\ell}{k-1} r_n^{(k-1)} (1-r_n)^{\ell-(k-1)} n_\ell(N), \tag{9}$$

valid for $k \geq 1$, where

$$\Delta N_k(N+1) \equiv (N+1)\, n_k(N+1) - N\, n_k(N) \tag{10}$$

is the ensemble-averaged change in the number of vertices with degree k, and $A_k = r_p + r_p r_n\, k$ is termed the attachment kernel in studies of preferential attachment models of complex networks (20). Equation 9 involves no approximations for the model. We look for a solution n_k° that is stationary, meaning that it does not change in time, at $N \gg k \gg 1$. In this parameter regime, the solution n_k° satisfies the algebraic equation

$$n_k^{\circ}[1 + A_k] = n_{k-1}^{\circ} A_{k-1} + \frac{r_p}{r_n} n_{\ell^*}^{\circ}, \text{ where } \ell^* = \frac{k-1}{r_n}. \tag{11}$$

As motivated by our simulation results (Fig. 12), we assume that at $N \gg k \gg 1$, our stationary solution has a power-law form, i.e., $n_k^{\circ} = A/k^{\gamma}$. This assumption results in an equation for γ as a function of r_p and r_n:

$$\gamma = 1 + \frac{1}{r_p r_n} - r_n^{\gamma-2}, \quad \text{for} \quad r_p, r_n \in (0,1). \tag{12}$$

Fig. 14 compares the results of the numerical solution of Eq. 12 to simulation. At moderately low r_n, Eq. 12 appears to have two solutions. Comparison with simulation suggests that the physical solution is the larger of the two. For the physical solution, the transition between the $\gamma > 2$ and $\gamma < 2$ regime coincides with the transition between the linear and superlinear growth regimes. The simulation results suggest that the location of the boundary between the two regimes of γ depends on N.

We emphasize that although Eq. 12 may be solved anywhere on the r_p, r_n square, we corroborate the power-law

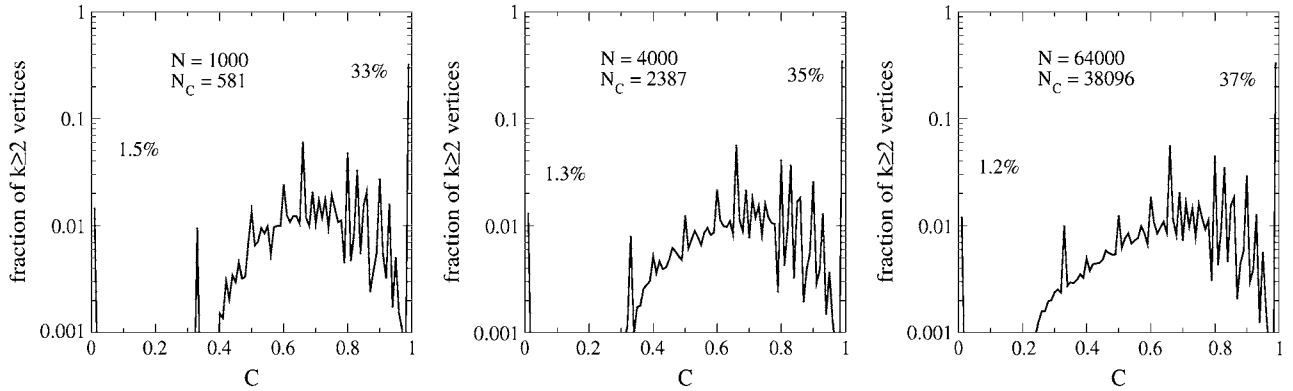

FIGURE 13 The clustering-coefficient distribution of M0 at increasing values of N. Each plot is the ensemble-averaged distribution with $M = 10$ and $(r_p, r_n) = (0.6, 0.8)$. N_C is the ensemble average of the number of $k \geq 2$ vertices.

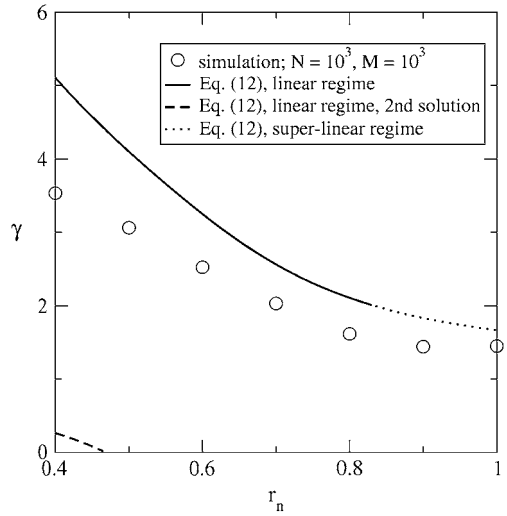

FIGURE 14 The Pareto-fit exponent of the ensemble-averaged degree distribution of $\Gamma^0(0.6, r_n; N)$ at $M \gg N \gg k \gg 1$, obtained by numerical solution of Eq. 12, and at $M = 10^3$, $N = 10^3$, intermediate k, obtained by simulation.

ansatz with simulation results only in the specific region near $r_p = 0.6, r_n = 0.8$ (Fig. 12).

CONCLUSIONS

Our primary result is the development of an asymptotically scale-free model (M0) that is consistent with the statistics of finite procaryote oPDUGs. By comparison with the null model CR, we demonstrated that procaryote oPDUGs hold information about evolution (dynamics or driving forces), and that this information appears in the nonrandom shapes of the degree and clustering-coefficient distributions. We then countered the null model with a dynamical model (M0) that explained the nonrandom statistics in terms of the mechanism of duplication and divergence. Simulation results demonstrated that the high-k (degree) region of the ensemble-averaged degree distribution develops into a power law as N (graph size) increases. We then computed analytically the exponent of the power-law regime in the infinite-graph limit. So, by asymptotically scale-free, we mean that in the large-graph limit, the ensemble-averaged degree distribution is power law at high degree. This work suggests that the statistical features of oPDUGs are consistent with an asymptotically scale-free dynamical process whereby new domains are discovered via duplication and divergence.

We note that, as with any area in scientific research, we cannot prove that models presented in this work, M0 and/or M1, are the only ones that satisfactorily describe the oPDUG. We cannot rule out the existence of alternative models, including convergent ones, which could also explain the peculiarities of the oPDUG. However, this possibility seems somewhat academic at the moment, since no convergent model has been proposed to describe nonrandom, asymptotically scale-free organization of the oPDUGs. In our earlier study (21), we attempted to develop a convergent PDUG model for a much simpler model—lattice proteins—of a protein universe, without apparent success. The analysis presented by Deeds et al. (21) suggests that inventing a satisfactory convergent model of oPDUG is a challenging task. As regards alternative divergent models, they are perhaps possible, but, as this study shows, the simplest memory-less model M0 performs reasonably well, especially in the high-k regime.

Specific results for M0, for finite-graph sizes, are the following. When M0 is fit to the oPDUG of *A. tumefaciens*, the parameters generate an ensemble of graphs in which the degree distributions of individual graphs are well fit by a Pareto law (at high k) with exponents in the neighborhood of 1.6. Additionally, the normalized clustering-coefficient distributions have a strong peak at $C = 1$ and a strong but subdominant peak at $C = 0$. These results are consistent with our observations of real organisms (see Biological Data).

Specific results for M0, in the infinite-graph limit, are the following. The ensemble of graphs with an asymptotically large number of vertices has two regimes of behavior separated by a sharp phase boundary. In the linear growth regime, the degree $D(N)$ approaches a limit as N increases and for the Pareto-fit exponent of the ensemble-averaged degree distribution, $\gamma > 2$ (for definition of $D(N)$, see Appendix, The degree in the mathematics of graphs). In the linear-log regime, $D(N)$ grows logarithmically with N and $\gamma = 2$. In the superlinear regime, $D(N)$ grows algebraically with N and $\gamma < 2$. Fitting *A. tumefaciens* to M0 for $N = 1000$ results in parameters $r_p = 0.6$ and $r_n = 0.8$, placing this procaryote genome in the linear growth regime. If we assume that these parameters are independent of time, then for the Pareto-fit exponents at time $N = 1000$, $\gamma_{1000} \approx 1.6$, and at time $N \to \infty$, $\gamma_\infty > 2$, we have $\gamma_{1000} < \gamma_\infty$ (Fig. 14). The structural proteome of *A. tumefaciens* has not reached its asymptotic behavior because the Pareto-fit exponent of the degree distribution is far from the long-time value. In this sense, *A. tumefaciens* is young. Fig. 12 confirms that the Pareto-fit exponent γ increases with the graph size.

Our secondary result is that memory mechanisms, as defined in M1: a model with memory, are unnecessary to model the oPDUGs. This result is the key step in distilling certain aspects of the previous model (8) to a form simple enough to extract analytical results about the long-time behavior. The simulation results in the Supplementary Material section shows that this distillation does not compromise accuracy.

Our third main finding, is that our models, both M0 and the memory-full model (M1) discussed in Supplementary Material, fail to quantitatively reproduce the orphan fraction of real procaryote oPDUGs. In both models, only 30% of the vertices are orphans, compared to 60% for the real oPDUG of *A. tumefaciens* (Fig. 9 and Supplementary Fig. S2). This is the most significant shortcoming of the model. We understand this shortcoming as follows. It is a simple fact that each

oPDUG may have some orphan Dali domains that share a Dali fold with other domains, and other orphans that are the sole occupants of folds (singlet folds). Some structural orphans may have been generated by the neutral-fold-discovery mechanism and others by the divergent mechanism (see Model, Divergent versus convergent evolution). That these mechanisms may have distinct timescales may have consequences for the oPDUGs (see Model, Spontaneous versus fixed mutations). Our models do not distinguish between the two kinds of discoveries and we speculate that this is the source of the failure. Work is in progress to evaluate the validity of this conjecture.

APPENDIX

The degree in the mathematics of graphs

The single-vertex degree is the number of edges emanating from a particular vertex. The graph degree, $D[G]$, of a graph G is the average-over-vertices, in the graph, of the single-vertex degree. The ensemble-averaged graph degree, $D(N)$, is the average of $D[G]$ over all graphs in the ensemble, where the ensemble has been time-evolved for N time steps. These definitions can be found in Albert and Barabasi (13).

Ensemble-averaged graph-degree

Equation-of-motion for E(N)*: derivation*

We consider the $\Gamma^0(r_p, r_n; N)$ ensemble in which each graph G^m—labeled $m \in \{1, \ldots, M\}$ with $M \gg N$—is independently evolved under the rules of M0 (see Results). We consider $r_p, r_n \in [0, 1]$, where r_p and r_n are the baby-parent and baby-neighbor edge retention probabilities, respectively. We would like an estimate of the degree of a typical vertex in a typical graph in this ensemble. Therefore, we compute the ensemble-averaged graph degree, $D(N)$, which we define as follows. Graph G^m at time N is composed of N vertices, each labeled $\{1, \ldots, N\}$, indicating the time step during which that vertex was created and underwent divergence from the parent vertex; if we define a lattice of time points as $t \in \{1, \ldots, N\}$, each interval of time between $t-1$ and t is a time step labeled with the terminal time point t. Each vertex t has some degree that we call $k_t[G^m(N)]$, and for the entire graph we define the graph degree, $D[G^m]$, as

$$D[G^m(N)] \equiv \frac{1}{N}\sum_{t=1}^{N} k_t[G^m(N)], \tag{13}$$

which is the average-over-vertices of the single-vertex degree. We now define the ensemble average for some property $f[G]$ of a graph as the average over the M graphs $\{G^m\}$ in the ensemble at time $t = N$, using the notation

$$\langle f[G(N)]\rangle \equiv \frac{1}{M}\sum_{m=1}^{M} f[G^m(N)]. \tag{14}$$

Thus, the ensemble-average of $D[G(N)]$ is

$$D(N) \equiv \langle D[G(N)]\rangle = \frac{1}{M}\sum_{m=1}^{M} D[G^m(N)], \tag{15}$$

and to compute this, we first consider the total number of edges in a graph.

For graph G^m at time $t = N$, $E[G^m(N)]$ is the total number of edges. The change in $E[G^m]$ from time $t = N$ to $t = N+1$ must be

$$\begin{aligned}\Delta E[G^m(N+1)] &\equiv E[G^m(N+1)] - E[G^m(N)] \\ &= k_{N+1}[G^m(N+1)],\end{aligned} \tag{16}$$

because the only mechanism by which the edge number changes is by the addition of the baby vertex during time-step $N+1$. The ensemble average of ΔE is

$$\begin{aligned}\Delta E(N+1) \equiv \langle \Delta E[G(N+1)]\rangle &= \frac{1}{M}\sum_{m=1}^{M} \Delta E[G^m(N+1)] \\ &= \frac{1}{M}\sum_{m=1}^{M} k_{N+1}[G^m(N+1)].\end{aligned} \tag{17}$$

We define $\Gamma^r(N+1)$ as the subensemble of $\Gamma^0(r_p, r_n; N+1)$ in which the baby-parent $(b-p)$ edge is retained during time-step $N+1$. We relabel all graphs so that $m \in \{1, \ldots, M_r\}$, with $M_r = Mr_p$, labels the graphs in $\Gamma^r(N+1)$. We can write

$$\begin{aligned}\Delta E(N+1) &= \frac{1}{M}\sum_{m=1}^{M_r} \left\{1 + k^n_{N+1}[G^m(N+1)]\right\} \\ &= r_p + r_p \frac{1}{M_r}\sum_{m=1}^{M_r} k^n_{N+1}[G^m(N+1)],\end{aligned} \tag{18}$$

where $k^n_{N+1}[G^m(N+1)]$ is the number of edges that are retained between the baby and the neighbors of the parent. We must average $k^n_{N+1}[G^m(N+1)]$ over $\Gamma^r(N+1)$.

We further subdivide $\Gamma^r(N+1)$ into ''structure groups'' $\{\Gamma^s_{\mu_s}\}$, where $\mu_s \in \{1, \ldots, M_r/M_s\}$ and M_s is the number of graphs in each group. Each group μ_s contains graphs $\{G^{m(\mu_s, m_s)}(N+1)\}$, where $m_s \in \{1, \ldots, M_s\}$, and the function $m(\mu_s, m_s)$ gives the graph label m. In $\Gamma^s_{\mu_s}$, the subgraphs $\{G^{m(\mu_s, m_s)}(N)\}$ are identical to each other in the sense that for any vertex labeled with birth date t in the first graph, the neighbor-vertex labels are the same as the neighbor-vertex labels for vertex t in the second graph. Note that different structure groups need not have different structure.

Because M is large, M_s is also large, and within each structure group, there will be many graphs in which the parent vertex has the same label t_p (giving its birth date). Thus, we further partition each structure group $\Gamma^s_{\mu_s}(N+1)$ into ''parent groups'' $\{\Gamma^{s,p}_{\mu_s,\mu_p}\}$, where $\mu_p \in \{1, \ldots, M_s/M_p\}$ and M_p is the number of graphs in each group. Each parent group (μ_s, μ_p) contains graphs $\{G^{m(\mu_s,\mu_p,m_p)}(N+1)\}$, where $m_p \in \{1, \ldots, M_p\}$, such that subgraphs $\{G^{m(\mu_s,\mu_p,m_p)}(N)\}$ are identical and baby $N+1$ is parented by the same vertex in each $G^{m(\mu_s,\mu_p,m_p)}(N+1)$; $m(\mu_s, \mu_p, m_p)$ is the function that relates labels (μ_s, μ_p, m_p) to label m

$$\begin{aligned}\Delta E(N+1) &= r_p + r_p \frac{M_s}{M_r}\sum_{\mu_s=1}^{M_r/M_s} \frac{M_p}{M_s}\sum_{\mu_p=1}^{M_s/M_p} \frac{1}{M_p}\sum_{m_p=1}^{M_p} k^n_{N+1} \\ &\quad \times [G^{m(\mu_s,\mu_p,m_p)}(N+1)] \\ &= r_p + r_p \frac{M_s}{M_r}\sum_{\mu_s=1}^{M_r/M_s} \frac{M_p}{M_s}\sum_{\mu_p=1}^{M_s/M_p} r_n k_p[G^{m(\mu_s,\mu_p,1)}(N)] \\ &= r_p + r_p \frac{M_s}{M_r}\sum_{\mu_s=1}^{M_r/M_s} \frac{1}{N}\sum_{t=1}^{N} r_n k_t[G^{m(\mu_s,1,1)}(N)] \\ &= r_p + r_p r_n \frac{M_s}{M_r}\sum_{\mu_s=1}^{M_r/M_s} D[G^{m(\mu_s,1,1)}(N)] \\ &= r_p + r_p r_n \frac{1}{M_r}\sum_{m=1}^{M_r} D[G^m(N)].\end{aligned} \tag{19}$$

Finally, we use the fact that the average of $D[G^m(N)]$ over the subensemble $\Gamma^r(N+1)$ of M_r is the same as the average over all graphs in the entire ensemble $\Gamma^0(r_p, r_n; N)$ (here, special attention must be paid to $t = N$ versus $t = N+1$). We obtain

$$\begin{aligned}\Delta E(N+1) &= r_p + r_p r_n \frac{1}{M}\sum_{m=1}^{M} D[G^m(N)] \\ &= r_p + \frac{2r_p r_n}{N}\frac{1}{M}\sum_{m=1}^{M} E[G^m(N)] \\ &= r_p + \frac{2r_p r_n}{N}E(N), \end{aligned} \tag{20}$$

where in the second line we used the general relation $D[G^m] = 2E[G^m]/N$, and in the third line we used the definition of the ensemble average $E(N)$ of $E[G(N)]$. In the continuous N approximation, this equation becomes

$$\frac{dE(N)}{dN} = r_p + \frac{2r_p r_n}{N}E(N). \tag{21}$$

This is Eq. 6 from Ensemble-averaged graph degree.

where $\delta(k,k')$ is the Kronecker δ with the above meaning. Thus, the total number of vertices with degree k is

$$N_k[G^m(N)] \equiv \sum_{t=1}^{N} \delta(k, k_t[G^m(N)]), \tag{26}$$

where t runs over all vertices in the graph. Thus, for graph G^m, the change in the number of vertices with degree k, from time $t = N$ to time $t = N+1$, is

$$\begin{aligned}\Delta N_k[G^m(N+1)] &\equiv N_k[G^m(N+1)] - N_k[G^m(N)] \\ &= \delta(k, k_{N+1}[G^m(N+1)] \\ &\quad + \sum_{t=1}^{N} \phi(k, k_t[G^m(N+1)], k_t[G^m(N)]), \end{aligned} \tag{27}$$

where

$$\begin{aligned}\phi(k, k_t[G^m(N+1)], k_t[G^m(N)]) &\equiv \delta(k, k_t[G^m(N+1)]) - \delta(k, k_t[G^m(N)]) \\ &= \begin{cases} 1 & \text{if vertex } t \text{ aquires an edge and } k_t[G^m(N)] = k-1 \\ 0 & \text{if vertex } t \text{ does not aquire an edge} \\ -1 & \text{if vertex } t \text{ aquires an edge and } k_t[G^m(N)] = k, \end{cases} \end{aligned} \tag{28}$$

Equation of motion for E(N): solution

The solutions of the equation

$$\frac{dE(N)}{dN} = a + \frac{b}{N}E(N) \tag{22}$$

for arbitrary real parameters a,b are obtained using the method of integrating factors, resulting in

$$E(N) = \begin{cases} \frac{a}{1-b}N + \left[E_0 - \frac{a}{1-b}N_0\right]\left(\frac{N}{N_0}\right)^b, & \text{if b} \neq 1 \\ aN\log\left(\frac{N}{N_0}\right) + E_0\frac{N}{N_0}, & \text{if b} = 1 \end{cases}, \tag{23}$$

where N_0 and E_0 are the initial values. For our model, we have $N_0 = 1$, $E_0 = 0$, $a = r_p$, and $b = 2r_p r_n$. With these parameter values, our solutions may be expressed in terms of $D(N)$ (defined in Ensemble-averaged degree) to obtain the exact result

$$D(N) = \begin{cases} \frac{r_p}{0.5 - r_p r_n}(1 - N^{2r_p r_n - 1}), & \text{if } r_p r_n \neq 0.5 \\ 2r_p \log N, & \text{if } r_p r_n = 0.5 \end{cases}, \tag{24}$$

for any N. In the large-N limit, we obtain Eq. 7 from Ensemble-averaged graph degree.

Ensemble-averaged degree distribution

We again consider the ensemble $\Gamma^0(r_p, r_n; N)$ as described in Appendix, Ensemble-averaged graph-degree, but restrict our attention to $r_p, r_n \in (0,1)$. We would like to compute a degree distribution, giving the fraction of vertices with degree k, that estimates the degree distribution of a typical graph in $\Gamma^0(r_p, r_n; N)$. Thus, we now derive an equation for how the ensemble-averaged degree distribution, called $n_k(N)$ and defined below, changes between time points $t = N$ and $t = N+1$.

For graph G^m at time N, $k_t[G^m(N)]$ is the degree of vertex t and

$$\delta(k, k_t[G^m(N)]) = \begin{cases} 1, & \text{if vertex } t \text{ has degree } k \\ 0, & \text{otherwise} \end{cases}, \tag{25}$$

is the contribution of vertex t to the flux of vertices into the population of degree k between time points $t = N$ and $t = N+1$.

To compute the ensemble-average of $\Delta N_k[G^m(N+1)]$, we use the notation $\langle \ldots \rangle$ defined in Appendix, Equation of motion for $E(N)$: derivation, and write

$$\Delta N_k(N+1) \equiv \langle \Delta N_k[G(N+1)] \rangle \equiv \Delta N_k^b(N+1) + \Delta N_k^{pn}(N+1), \tag{29}$$

where

$$\Delta N_k^b(N+1) \equiv \langle \delta(k, k_{N+1}[G(N+1)]) \rangle \tag{30}$$

is the contribution to the ensemble-averaged flux due to the baby (in each G^m), and

$$\Delta N_k^{pn}(N+1) \equiv \left\langle \sum_{t=1}^{N} \phi(k, k_t[G(N+1)], k_t[G(N)]) \right\rangle \tag{31}$$

is the contribution to the flux due to the parent and neighbors of the parent.

Contribution of the "baby"

To compute

$$\Delta N_k^b(N) = \frac{1}{M}\sum_{m=1}^{M} \delta(k, k_{N+1}[G^m(N+1)]), \tag{32}$$

we follow the development in the Appendix, Equation of motion for $E(N)$: derivation, and define $\Gamma^r(N+1)$ as the subensemble of graphs in which the baby-parent edge is retained during time step $N+1$. We immediately have

$$\Delta N_k^b(N) = \begin{cases} \frac{1}{M}\sum_{m=1}^{M_r} \delta(k, k_{N+1}[G^m(N+1)]), & \text{for } k \geq 1 \\ (1 - r_p), & \text{for } k = 0 \end{cases}, \tag{33}$$

where graphs $m \in \{1, \ldots, M_r\}$ are those in which the baby-parent edge is retained. Since we are interested in high k, we focus on $k \geq 1$.

As before, $\Gamma^{r}(N+1)$ is further partitioned into ''structure groups'' $\{\Gamma^{s}_{\mu_s}\}$, where $\mu_s \in \{1,\ldots,M_r/M_s\}$ and M_s is the number of graphs in each group. We further partition each structure group $\Gamma^{s}_{\mu_s}(N+1)$ into ''parent groups'' $\{\Gamma^{s,p}_{\mu_s,\mu_p}\}$, where $\mu_p \in \{1,\ldots,M_s/M_p\}$ and M_p is the number of graphs in each group. Each parent group (μ_s,μ_p) contains graphs $\{G^{m(\mu_s,\mu_p,m_p)}(N+1)\}$, where $m_p \in \{1,\ldots,M_p\}$, such that subgraphs $\{G^{m(\mu_s,\mu_p,m_p)}(N)\}$ are identical and baby $N+1$ is parented by the same vertex in each $G^{m(\mu_s,\mu_p,m_p)}(N+1)$. We can write

$$\Delta N^{b}_{k}(N+1) = \frac{1}{M}\sum_{\mu_s=1}^{M_r/M_s}\sum_{\mu_p=1}^{M_s/M_p}\sum_{m_p=1}^{M_p}\delta(k,k_{N+1}[G^{m(\mu_s,\mu_p,m_p)}(N+1)]) = \frac{M_s}{M}\sum_{\mu_s=1}^{M_r/M_s}\frac{M_p}{M_s}\sum_{\mu_p=1}^{M_s/M_p}\frac{1}{M_p}\sum_{m_p=1}^{M_p}\delta(k,k_{N+1}[G^{m(\mu_s,\mu_p,m_p)}(N+1)]), \quad (34)$$

and focus on a single parent group (μ_s,μ_p), evaluating

$$\frac{1}{M_p}\sum_{m_p=1}^{M_p}\delta(k,k_{N+1}[G^{m(\mu_s,\mu_p,m_p)}(N+1)]) = \begin{cases}\text{fraction of divergence events in which the baby}\\ \text{retains } k-1 \text{ edges with neighbors of the parent}\end{cases} = \Theta(k_{t_p}-(k-1))\binom{k_{t_p}}{k-1}r_n^{(k-1)}(1-r_n)^{k_{t_p}-(k-1)}, \quad (35)$$

and

$$\binom{k}{k'} = \frac{k!}{k'!(k-k'!)} \quad (37)$$

is the usual binomial coefficient. We insert the second line of Eq. 35 into Eq. 34 to obtain

$$\begin{aligned}\Delta N^{b}_{k}(N+1) &= \frac{M_s}{M}\sum_{\mu_s=1}^{M_r/M_s}\frac{M_p}{M_s}\sum_{\mu_p=1}^{M_s/M_p}\Theta(k_{t_p}-(k-1))\binom{k_{t_p}}{k-1}r_n^{(k-1)}(1-r_n)^{k_{t_p}-(k-1)}\\ &= \frac{M_s}{M}\sum_{\mu_s=1}^{M_r/M_s}\frac{1}{N}\sum_{t=1}^{N}\Theta(k_t-(k-1))\binom{k_t}{k-1}r_n^{(k-1)}(1-r_n)^{k_t-(k-1)}\\ &= \frac{M_s}{M}\sum_{\mu_s=1}^{M_r/M_s}\sum_{\ell=k-1}^{N}\frac{N_\ell[G^{m(\mu_s,1,1)}(N)]}{N}\binom{\ell}{k-1}r_n^{(k-1)}(1-r_n)^{\ell-(k-1)}\\ &= \sum_{\ell=k-1}^{N}\binom{\ell}{k-1}r_n^{(k-1)}(1-r_n)^{\ell-(k-1)}\frac{M_s}{M}\sum_{\mu_s=1}^{M_r/M_s}\frac{N_\ell[G^{m(\mu_s,1,1)}(N)]}{N}. \end{aligned} \quad (38)$$

where t_p is the time step of creation (i.e., label) of the parent, k_t is shorthand for $k_t[G^{m(\mu_s,1,1)}(N)]$, and

$$\Theta(k-k') \equiv \begin{cases}1, & \text{if } k \geq k'\\ 0, & \text{if } k < k'\end{cases} \quad (36)$$

We can simplify this equation by noting that all graphs in a structure group have the same degree distribution, so that

$$\begin{aligned}\Delta N^{b}_{k}(N+1) &= \sum_{\ell=k-1}^{N}\binom{\ell}{k-1}r_n^{(k-1)}(1-r_n)^{\ell-(k-1)}\frac{1}{M}\sum_{m=1}^{M_r}\frac{N_\ell[G^{m(\mu_s,1,1)}(N)]}{N}\\ &= \sum_{\ell=k-1}^{N}\binom{\ell}{k-1}r_n^{(k-1)}(1-r_n)^{\ell-(k-1)}r_p\frac{1}{M}\sum_{m=1}^{M}\frac{N_\ell[G^{m(\mu_s,1,1)}(N)]}{N}\\ &= r_p\sum_{\ell=k-1}^{N}\binom{\ell}{k-1}r_n^{(k-1)}(1-r_n)^{\ell-(k-1)}n_\ell(N). \end{aligned} \quad (39)$$

The second equality is achieved by noting that for the graphs $G^{m}(N+1) \in \Gamma^{r}(N+1)$, the subgraphs $G^{m}(N)$ have the same distribution of structures as the graphs in $\Gamma^{0}(r_p, r_n; N)$. Therefore, if we define the ensemble-averaged degree distribution as the average of the fraction of vertices with degree k, i.e.,

$$n_k(N) \equiv \begin{cases} \dfrac{\langle N_k[G(N)]\rangle}{N}, & \text{for} \quad k \geq 0 \\ 0, & \text{for} \quad k < 0 \end{cases}, \tag{40}$$

we obtain

$$\Delta N_k^{b}(N+1) = r_p \sum_{\ell=k-1}^{N} \binom{\ell}{k-1} r_n^{(k-1)} (1-r_n)^{\ell-(k-1)} n_\ell(N), \quad \text{for} \quad k \geq 1. \tag{41}$$

Contribution of the parent and neighbors

Again, we consider how each graph G^{m} may change from time-point $t = N$ to $t = N+1$, i.e., during time step $N+1$. Again, we define the subensemble $\Gamma^{r}(N+1)$as in the Appendix, Equation of motion for $E(N)$: derivation, and further subdivide $\Gamma^{r}(N+1)$ into ''structure'' groups $\{\Gamma^{s}_{\mu_s}\}$, $\mu_s \in \{1, \ldots, M_r/M_s\}$. Each graph in $G^{s}_{\mu_s}$ is given a label $m_s \in \{1, \ldots, M_s\}$ and $m(\mu_s, m_s)$ is the function that relates these labels to the original label m. With this partitioning, we have

$$\begin{aligned} \Delta N_k^{pn}(N+1) &= \frac{1}{M} \sum_{m=1}^{M_r} \sum_{t=1}^{N} \phi(k, k_t[G^{m}(N+1)], k_t[G^{m}(N)]) \\ &= \frac{1}{M} \sum_{\mu_s=1}^{M_r/M_s} \sum_{m_s=1}^{M_s} \sum_{t=1}^{N} \phi(k, k_t[G^{m(\mu_s,m_s)}(N+1)], k_t[G^{m(\mu_s,m_s)}(N)]) \\ &= \frac{M_s}{M} \sum_{\mu_s} \sum_{t} \frac{1}{M_s} \sum_{m_s} \phi(k, k_t[G^{m(\mu_s,m_s)}(N+1)], k_t[G^{m(\mu_s,m_s)}(N)]). \end{aligned} \tag{42}$$

We use the shorthand k_t for $k_t[G^{m(\mu_s,1)}]$, in combination with Eq. 28, to obtain

$$\begin{aligned} &\frac{1}{M_s} \sum_{m_s=1}^{M_s} \phi(k, k_t[G^{m(\mu_s,m_s)}(N+1)], k_t[G^{m(\mu_s,m_s)}]) \\ &= \left[\frac{1}{N} + \frac{1}{N} r_r k_t\right] [\delta(k, k_t+1) - \delta(k, k_t)], \end{aligned} \tag{43}$$

where the factor $(1/N + k_t r_n/N)$ is the fraction of the M_s graphs in which the baby is parented by either vertex t *or* any neighbor of vertex t. Thus,

Using Eq. 40, we obtain

$$\begin{aligned} \Delta N_k^{pn}(N+1) = n_{k-1}(N)[r_p + r_p r_n (k-1)] \\ - n_k(N)[r_p + r_p r_n k]. \end{aligned} \tag{45}$$

In studies of preferential attachment models of complex networks (20), the coefficient of $n_k(N)$ is termed the attachment kernel, here defined $A_k \equiv r_p + r_p r_n k$. So we write

$$\Delta N_k^{pn}(N+1) = A_{k-1} n_{k-1}(N) - A_k n_k(N). \tag{46}$$

Finally, we sum $\Delta N_k^{b}(N+1)$ and $\Delta N_k^{pn}(N+1)$ to obtain an exact formula for the ensemble-averaged flux of vertices, during a single time step, into the population with degree $k \geq 1$,

$$\begin{aligned} \Delta N_k(N+1) = A_{k-1} n_{k-1}(N) - A_k n_k(N) \\ + r_p \sum_{\ell=k-1}^{N} \binom{\ell}{k-1} r_n^{(k-1)} (1-r_n)^{\ell-(k-1)} n_\ell(N). \end{aligned} \tag{47}$$

This is called the rate equation. The analogous formula for $k = 0$ results by replacing the last term with $(1-r_p)$.

Stationary solution at high k

We attempt a stationary (doesn't change in time) solution to Eq. 47 for $N \gg k \gg 1$, calling it n_k^{o}. Setting $\Delta N_k(N+1) = n_k^{o}$, we obtain an equation algebraic in the n_k^{o}:

$$n_k^{o}[1 + r_p + r_p r_n k] = n_{k-1}^{o}[r_p + r_p r_n (k-1)] + \Delta N_k^{b}(N+1), \tag{48}$$

where, from Eq. 39,

$$\begin{aligned} \Delta N_k^{pn}(N+1) &= \frac{M_s}{M} \sum_{\mu_s=1}^{M_p/M_s} \sum_{t=1}^{N} \left[\frac{1}{N} + \frac{1}{N} k_t r_n\right] [\delta(k-1, k_t) - \delta(k, k_t)] \\ &= \frac{M_s}{M} \sum_{\mu_s=1}^{M_p/M_s} \sum_{\ell=0}^{N} N_\ell[G^{m(\mu_s,1)}(N)] \left[\frac{1}{N} + \frac{1}{N} \ell r_n\right] [\delta(k-1, \ell) - \delta(k, \ell)] \\ &= \frac{M_s}{M} \sum_{\mu_s=1}^{M_p/M_s} \frac{N_{k-1}[G^{m(\mu_s,1)}(N)]}{N} [1 + (k-1) r_n] - \frac{N_k[G^{m(\mu,1)}(N)]}{N} [1 + k r_n] \\ &= \frac{1}{M} \sum_{m=1}^{M_p} \frac{N_{k-1}[G^{m}(N)]}{N} [1 + (k-1) r_n] - \frac{N_k[G^{m}(N)]}{N} [1 + k r_n] \\ &= r_p \frac{1}{M} \sum_{m=1}^{M} \frac{N_{k-1}[G^{m}(N)]}{N} [1 + (k-1) r_n] - \frac{N_k[G^{m}(N)]}{N} [1 + k r_n]. \end{aligned} \tag{44}$$

$$\Delta N_{\mathrm{k}}^{\mathrm{b}}(N+1) = r_{\mathrm{p}} \sum_{\ell=\mathrm{k}-1}^{\mathrm{N}} \binom{\ell}{k-1} r_{\mathrm{n}}^{(\mathrm{k}-1)} (1-r_{\mathrm{n}})^{\ell-(\mathrm{k}-1)} n_{\ell}^{\mathrm{o}}. \quad (49)$$

If we write the summand as $g_{\mathrm{k}-1}^{\ell} n_{\ell}^{\mathrm{o}}$, the factor $g_{\mathrm{k}-1}^{\ell}$ is well approximated by a Gaussian, so we can write

$$\Delta N_{\mathrm{k}}^{\mathrm{b}}(N+1) \cong \frac{r_{\mathrm{p}}}{r_{\mathrm{n}}} \sum_{\ell=\mathrm{k}-1}^{\mathrm{N}} G_{\sigma}(\ell-\ell^{*}) n_{\ell}^{\mathrm{o}} \quad (50)$$

where $G_{\sigma}(\ell-\ell^{*})$ is a normalized Gaussian with mean $\ell^{*} = (k-1)/r_{\mathrm{n}}$ and variance $\sigma^2 = (k-1)(1-r_{\mathrm{n}})/r_{\mathrm{n}}^2$; the approximation gets better with increasing k. If we assume that n_{ℓ}^{o} varies slowly in the neighborhood of ℓ^{*}, in particular, that it changes by an amount $\mathrm{O}(k^0)$ over an interval that grows faster than $k^{1/2}$, then an appropriate rescaling of ℓ shows that G_{σ} approaches Dirac's delta (times the measure of the rescaled ℓ), whereas n_{ℓ}^{o} is relatively smooth. Upon integration, we obtain

$$\Delta N_{\mathrm{k}}^{\mathrm{b}}(N+1) \cong \frac{r_{\mathrm{p}}}{r_{\mathrm{n}}} n_{\ell^{*}}, \quad (51)$$

with corrections that vanish with k. We can then write

$$n_{\mathrm{k}}^{\mathrm{o}}[1 + r_{\mathrm{p}} + r_{\mathrm{p}} r_{\mathrm{n}} k] = n_{\mathrm{k}-1}^{\mathrm{o}}[r_{\mathrm{p}} + r_{\mathrm{p}} r_{\mathrm{n}}(k-1)] + \frac{r_{\mathrm{p}}}{r_{\mathrm{n}}} n_{\ell^{*}}^{\mathrm{o}},$$

$$\text{where} \quad \ell^{*} = \frac{k-1}{r_{\mathrm{n}}}. \quad (52)$$

We additionally assume that the stationary solution is a power law at high k, i.e., we set $n_{\mathrm{k}}^{\mathrm{o}} = Ak^{-\gamma}$; this form satisfies our assumption of slow variation. We insert the power-law ansatz into Eq. 52 and solve for γ to obtain

$$\gamma = 1 + \frac{1}{r_{\mathrm{p}} r_{\mathrm{n}}} - r_{\mathrm{n}}^{\gamma-2}, \quad \text{for} \quad r_{\mathrm{p}}, r_{\mathrm{n}} \in (0,1), \quad (53)$$

where we ignore $\mathrm{O}(1/k)$ corrections. Note that at no point did we treat N as continuous. We discuss solutions of Eq. 53 in the main text, but here we note that the asymptotically low-r_{n} behavior is obtained by assuming $\gamma > 1$; thus, as $r_{\mathrm{n}} \to 0$, the solution has a power-law divergence.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

We thank Eric Deeds for help at the initial stages of this work and Julius Lucks, Jason Donald, and Edo Kussell for helpful comments on the article.

This work was supported by the National Institutes of Health. C.B.R. also acknowledges support from the Harvard Merit Fellowship.

REFERENCES

1. Koonin, E. V., Y. I. Wolf, and G. P. Karev. 2002. The structure of the protein universe and genome evolution. *Nature*. 420:218–223.
2. Gerstein, M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archael genomes in terms of protein structure. *J. Mol. Biol.* 274:562–576.
3. Qian, J., N. M. Luscombe, and M. Gerstein. 2001. Protein family and fold occurance in genomes: power-law behavior and evolutionary model. *J. Mol. Biol.* 313:673–681.
4. Murzin, A. G., S. E. Brenner, T. J. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.
5. Lo Conte, L., B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. 2000. SCOP: a structural classification of proteins database. *Nucl. Acids Res.* 28:257–259.
6. Holm, L., and C. Sander. 1996. Mapping the protein universe. *Science*. 273:595–602.
7. Li, H., R. Tellig, C. Tang, and N. Wingreen. 1996. Emergence of preferred structures in a simple model of protein folding. *Science*. 273: 666–669.
8. Dokholyan, N. V., B. Shakhnovich, and E. I. Shakhnovich. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci. USA*. 99:14132–14136.
9. Holm, L., and C. Sander. 1998. Dictionary of recurrent domains in protein structures. *Proteins*. 33:88–96.
10. Deeds, E. J., B. Shakhnovich, and E. I. Shakhnovich. 2004. Proteomic traces of speciation. *J. Mol. Biol.* 336:695–706.
11. Shakhnovich, B. E., J. M. Harvey, S. Comreau, D. Lorenz, C. Delisi, and E. I. Shakhnovich. 2003. ELISA: Structure-function inferences based on statistically significant and evolutionarily inspired observations. *BMC Bioinformatics*. 4:34.
12. Deeds, E. J., H. Hennessey, and E. I. Shakhnovich. 2005. Procaryotic phylogenies inferred from protein structural domains. *Genome Res.* 15:393–402.
13. Albert, R., and A.-L. Barabasi. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74:47–97.
14. Mirny, L. A., and E. I. Shakhnovich. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics, and function. *J. Mol. Biol.* 291:177–196.
15. Dokholyan, N. V., and E. I. Shakhnovich. 2001. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* 312: 289–307.
16. Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.
17. Wilson, C. A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relationships between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* 297:233–249.
18. Rost, B. 1999. Twighlight zone of protein sequence alignments. *Protein Eng.* 12:85–94.
19. Kim, J., P. L. Krapivsky, B. Kahng, and S. Redner. 2002. Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev. E.* 66:055101 (R).
20. Krapivsky, P. L., and S. Redner. 2001. Organization of growing random networks. *Phys. Rev. E.* 63:066123.
21. Deeds, E. J., N. V. Dokholyan, and E. I. Shakhnovich. 2003. Protein evolution within a structural space. *Biophys. J.* 85:2962–2972.