Conservation and topology of protein interaction networks under duplication-divergence evolution

Kirill Evlampiev and Hervé Isambert*

Physico-chimie Curie, Centre National de la Recherche Scientifique Unité Mixte de Recherche 168, Institut Curie, Section de Recherche, 11 rue P. & M. Curie, 75005 Paris, France

Communicated by M. Gromov, Institut des Hautes Études Scientifiques, Bures-sur-Yvette, France, April 30, 2008 (received for review March 22, 2007)

Genomic duplication-divergence processes are the primary source of new protein functions and thereby contribute to the evolutionary expansion of functional molecular networks. Yet, it is still unclear to what extent such duplication-divergence processes also restrict by construction the emerging properties of molecular networks, regardless of any specific cellular functions. We address this question, here, focusing on the evolution of protein–protein interaction (PPI) networks. We solve a general duplication-divergence model, based on the statistically necessary deletions of protein–protein interactions arising from stochastic duplications at various genomic scales, from single-gene to whole-genome duplications. Major evolutionary scenarios are shown to depend on two global parameters only: (*i*) a protein conservation index (*M*), which controls the evolutionary history of PPI networks, and (*ii*) a distinct topology index (*M′*) controlling their resulting structure. We then demonstrate that conserved, nondense networks, which are of prime biological relevance, are also necessarily scale-free by construction, irrespective of any evolutionary variations or fluctuations of the model parameters. It is shown to result from a fundamental linkage between individual protein conservation and network topology under general duplication-divergence evolution. By contrast, we find that conservation of network motifs with two or more proteins cannot be indefinitely preserved under general duplication-divergence evolution (independently from any network rewiring dynamics), in broad agreement with empirical evidence between phylogenetically distant species. All in all, these evolutionary constraints, inherent to duplication-divergence processes, appear to have largely controlled the overall topology and scale-dependent conservation of PPI networks, regardless of any specific biological function.

evolutionary constraint | scale-free graph | functional motif | orthology | statistical model

The primary source of new protein functions is generally considered to originate from *duplication* of existing genes followed by functional *divergence* of their duplicate copies (1–3). In fact, duplication-divergence events have occurred and continue to occur at a wide range of genomic scales, from many independent duplications of individual genes[†] [10^{-3} fixed events per gene per million years (MY) (4)] to rare but evolutionary dramatic duplications of entire genomes [one fixed event per 100–200 MY (5)]. For instance, there have been between two and four *consecutive* whole-genome duplications in all major eukaryote kingdoms in the past 300–500 MY (5). This actually amounts to a more-or-less similar contribution of new genes from whole-genome duplication as from individual gene duplications [i.e., one fixed event per 100–200 MY $\simeq 10^{-3}$ fixed events per gene per MY, assuming a 10% fixation rate after a whole-genome duplication with ≈10,000 genes (5)].

This succession of whole-genome duplications, together with the accumulation of individual gene duplications, must have greatly contributed to shaping the global structure of large biological networks, such as protein–protein interaction (PPI) networks, that control cellular activities. In fact, concordant empirical evidence reveals the evolutionary persistence of duplication-derived protein–protein interactions. For instance, there are clear enrichments of recent protein duplicates around common protein partners compared with randomly picked pairs of proteins (5, 6), although the fraction of proteins identified as having undergone a (recent) duplication (<200 MY) remains typically small in absolute terms, for example, 10% (4). Similarly, protein residues implicated in protein–protein interaction are generally the most conserved at the surface of proteins (7), revealing their duplication-derived origin,[‡] with typically little more than one conserved binding interface per protein-binding domains.[§]

Ispolatov *et al.* (10) proposed an interesting local duplication-divergence model of PPI network evolution based on (*i*) the *statistical* deletion of individual, duplication-derived interactions and (*ii*) a *time-linear* increase in genome and PPI network sizes. Clearly, the deletion of redundant interactions arising from duplication is necessary to avoid the emergence of biologically irrelevant, densely connected PPI networks, lacking low-degree connectivities. Yet, we expect that *independent* local duplications and, *a forteriori*, partial- or whole-genome duplications all lead to *exponential*, not time-linear, evolutionary dynamics of PPI networks. In the long time limit, exponential dynamics should outweigh all time-linear processes that have been assumed in earlier PPI network evolution models (10–15). Models based on time-linear processes also assume that local evolutionary dynamics remain essentially frozen, as long as they are not directly affected by a local modification of the network. Yet, in reality, sequence mutations and environmental changes continue to affect the evolution of whole PPI networks, not just in the immediate surroundings of recently duplicated proteins.

In this article, we propose and asymptotically solve a general duplication-divergence model based on prevailing exponential dynamics[¶] of PPI network evolution under local, partial, or global genome duplications. The only interaction changes that are considered are *deletions* of duplication-derived interactions. In particular, the rewiring dynamics of PPI networks by *de novo* creation of protein-binding interfaces (4) is neglected (10), as suggested by the empirical evidence mentioned earlier (see also

Author contributions: H.I. designed research; K.E. and H.I. performed research; K.E. and H.I. analyzed data; and K.E. and H.I. wrote the paper.

The authors declare no conflict of interest.

*To whom correspondence should be addressed. E-mail: herve.isambert@curie.fr.

[†]Duplicated protein domains or subdomains are also quite common even within ancestral proteins, such as the ubiquitous aquaporin membrane proteins in eubacteria, archaea, and eukaryotes, or the TATA-binding protein from archaea and eukaryotes.

[‡]Except for a few interesting cases of protein-binding mimicry, typically found in virus–host protein–protein interactions (8).

[§]Except for domains that self-assemble into homo-oligomers, which *must* have at least two binding interfaces, see table 2 in ref. 9.

[¶]Results from the time-linear duplication-divergence model (10) are recovered as a special limit, see supporting information (SI) *Appendix*.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804119105/DCSupplemental.

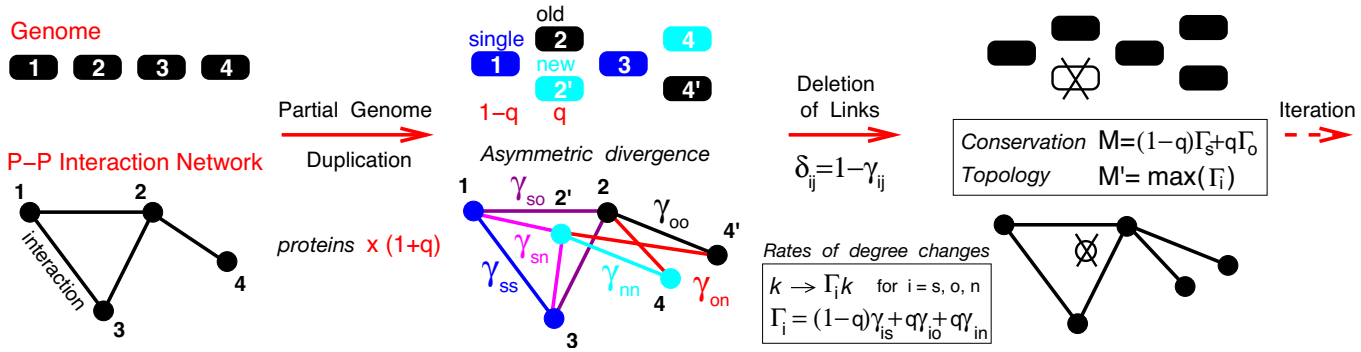

Fig. 1. General duplication-divergence model for protein–protein interaction network evolution. Successive duplications of a fraction q of genes are followed by an asymmetric divergence of gene duplicates (e.g., 2 vs. 2′). New duplicates (n) are left essentially free to accumulate neutral mutations with the likely outcome of becoming nonfunctional and eventually deleted unless some new, *duplication-derived* interactions are selected; old duplicates (o), however, are more constrained to conserve old interactions already present before duplication. Interactions on the locally ($q \ll 1$), partially ($q < 1$) or fully ($q = 1$) duplicated network are then preserved stochastically with different probabilities γ_{ij} ($0 \leq \gamma_{ij} \leq 1$, $i, j = s, o, n$) reflecting the recent history of each interacting partners, that are either singular, nonduplicated genes (s) or recently duplicated genes undergoing asymmetric divergence (o/n). Two effective parameters, M and M', that depend on the rates of connectivity change, Γ_i, and underlying parameters q and γ_{ij}, control the evolutionary history or conservation (M) and resulting structure or topology (M') of PPI networks (see text).

Evolution of PPI network motifs). Indeed, our aim here is to establish a theoretical baseline from which other evolutionary processes beyond strict gene duplication and interaction loss events, such as shuffling of protein domains (5) or horizontal gene transfers, can then be considered.

A visual overview of the model is shown in Fig. 1 including its two main effective parameters, M and M', that control, respectively, the evolutionary history or *conservation* (M) and resulting structure or *topology* (M') of PPI networks under duplication-divergence evolution. In this article, we demonstrate a fundamental relation between protein conservation (M) and network topology (M'), that is, $M \leq M'$, that is strictly independent from *any* evolutionary variations or fluctuations of the model parameters. We then discuss simple consequences in terms of evolutionary *linkage* between individual protein conservation and PPI network topology. The approach is also extended to outline the evolutionary statistics of small-network motifs including two or more proteins. In particular, we show that network motifs, unlike individual proteins, cannot be indefinitely conserved under general duplication-divergence evolution, regardless of any network-rewiring dynamics. Throughout the article, theoretical assumptions and results are commented on with brief discussions highlighting their biological relevance.

Results

General Duplication-Divergence Model. The general duplication-divergence (GDD) model is designed to capture PPI network properties caused by evolutionary constraints, inherent to duplication-divergence processes and independent of selective adaptation (3) or any specific biological function. Concretely, the GDD model analyzes the deletion statistics of protein–protein interactions that arise from stochastic duplications at various genomic scales, from single-gene to whole-genome duplications. This deletion statistics of duplication-derived interactions is indeed a necessary "background" dynamics of PPI network evolution to prevent the emergence of biologically irrelevant, densely connected PPI networks, lacking low-degree connectivities.

In practice, a fraction q of extant genes is randomly *duplicated* at each time step of the GDD model. The divergence of both duplicated and nonduplicated genes then leads to the stochastic deletion or conservation of their related interactions, before another round of duplication-divergence occurs (Fig. 1). In the following, we first solve the GDD model assuming that q is constant over evolutionary time scales. We then study more realistic scenarios combining, for instance, rare whole-genome duplications ($q = 1$) with more frequent local duplications of individual genes ($q \ll 1$), and including also stochastic fluctuations in *all* microscopic parameters of the GDD model (see Fig. 1 and below). To analyze the deletion statistics of duplication-derived interactions, we assume that ancient and recently duplicated interactions are stochastically conserved with distinct probabilities γ_{ij}'s, depending only on the recently duplicated or nonduplicated state of each protein partners, as well as on the *asymmetric* divergence between "old" and "new" (or more "conserved" and more "divergent") gene duplicates (5), see the Fig. 1 legend ("s" for "singular," nonduplicated genes and "o"/"n" for old/new asymmetrically divergent duplicates). Here, we consider nonoriented PPI networks, that is, $\gamma_{ij} = \gamma_{ji}$, for $i, j = s, o, n$.

The first effective parameters derived from these microscopic evolutionary parameters are the average rates of connectivity change Γ_i (i.e., $k \rightarrow k\Gamma_i$) for each type of node $i = s, o, n$, where $\Gamma_i = (1 - q)\gamma_{is} + q(\gamma_{io} + \gamma_{in})$ is independent from node connectivity k. In the following, we assume $\Gamma_o \geq \Gamma_n$ by definition of old and new duplicates caused by asymmetric divergence. Note that self-interacting proteins, corresponding to self-link loops, are not taken into account, for simplicity, in the main text, because they can be shown to have little effect on the asymptotic evolutionary regimes of the connectivity distribution (see *SI Appendix*, Fig. S3 and *SI Text*, for details).

We study the GDD evolutionary dynamics of PPI networks in terms of ensemble averages $\langle Q^{(n)} \rangle$ defined as the mean value of a feature Q over all realizations of the evolutionary dynamics after n successive duplications. This does not imply, of course, that all network realizations "coexist," but only that a random selection of them is reasonably well characterized by the theoretical ensemble average. Although it is generally not the case for exponentially growing systems, here, we can show that ensemble averages over all evolutionary dynamics indeed reflect the properties of typical network realizations for biologically relevant regimes (see *Statistical Properties of GDD Models* in *SI Appendix*).

In the following, we focus on the number of proteins (or "nodes") N_k of connectivity k in PPI networks, while postponing the analysis of GDD models for simple network motifs to the end of the article and the *SI Appendix*. The total number of nodes in the network is noted $N = \Sigma_{k \geq 0} N_k$ and the total number of interactions (or "links") $L = \Sigma_{k \geq 0} kN_k/2$. The dynamics of the ensemble averages $\langle N_k^{(n)} \rangle$ after n duplications is analyzed by using a generating function,

$$F^{(n)}(x) = \sum_{k\geq 0} \langle N_k^{(n)} \rangle x^k. \quad [1]$$

The evolutionary dynamics of $F^{(n)}(x)$ corresponds to the following recurrence deduced from the microscopic definition of the GDD model (see *SI Appendix*),

$$F^{(n+1)}(x) = (1-q)F^{(n)}(A_{s(x)}) + qF^{(n)}(A_{o(x)}) + qF^{(n)}(A_{n(x)}) \quad [2]$$

where we note for $i = s, o, n$,

$$A_i(x) = (1-q)(\gamma_{is}x + \delta_{is}) + q(\gamma_{io}x + \delta_{io})(\gamma_{in}x + \delta_{in}) \quad [3]$$

where $\delta_{ij} = 1 - \gamma_{ij}$ are deletion probabilities ($i, j = s, o, n$) and $A_i'(1) = (1-q)\gamma_{is} + q(\gamma_{io} + \gamma_{in}) = \Gamma_i$, average rates of connectivity change for each type of nodes i = s, o, n (Fig. 1).

Network Expansion (Γ) and Protein Conservation (*M*). The total number of nodes generated by the GDD model, $F^{(n)}(1)$, growths exponentially with the number of partial duplications, $F^{(n)}(1) = C\cdot(1+q)^n$, where C is the initial number of nodes, as a constant fraction of nodes q is duplicated at each time step. Yet, some nodes become completely disconnected from the rest of the graph during divergence and rejoin the disconnected component of size $F^{(n)}(0)$. From a biological point of view, these disconnected nodes represent genes that have presumably lost all biological functions and become pseudogenes before being simply eliminated from the genome. We neglect the possibility for nonfunctional genes to reconvert to functional genes again after suitable mutations, and remove them at each round of partial duplication[‖] focusing solely on the connected part of the graph.

In particular, the link growth rate $\Gamma = (1-q)\Gamma_s + q\Gamma_o + q\Gamma_n$ obtained by taking the first derivative of Eq. **2** at $x = 1$, controls whether the connected part of the graph is exponentially growing ($\Gamma > 1$) or shrinking ($\Gamma < 1$).

Let us now introduce another rate of *prime* biological interest, $M = (1-q)\Gamma_s + q\Gamma_o$. It is the *average rate of connectivity increase (M >1) of decrease (M < 1) for the most conserved duplicate lineage*, which corresponds to a stochastic alternance between singular (s) and most conserved (o) duplicate descents. In particular, we have by construction, $M < \Gamma = M + q\Gamma_n$, independently from any evolutionary parameters, q and $\gamma_{ij} > 0$. This implies three main evolutionary regimes from the perspective of network expansion (Γ) and protein conservation (M) (Fig. 2):

- If $M < \Gamma < 1$. PPI networks are vanishing in this regime with seemingly little biological relevance.
- If $M < 1 < \Gamma$. PPI networks are expanding, in this case, but their proteins are *not* conserved over long evolutionary time scales. This implies that the networks forget their evolutionary history exponentially fast, as most nodes eventually disappear and, with them, all traces of network evolution. These networks are *not* preserved over time, but instead are continuously renewed from duplication of the (few) most connected nodes (Fig. 2). Individual proteins of a given network realization are thus more similar to one another than to any protein of other network realizations, which can be seen, from a speciation perspective, as PPI networks of phylogenetically distant organisms. This is in sharp contrast to the widespread structural orthology observed across all extant life forms, even though functions of orthologs often differ (see *Evolution of PPI Network Motifs*).
- If $1 < M < \Gamma$. By contrast, PPI networks remember their past evolution from the very beginning, in this case, as proteins statistically keep on increasing their connectivity once they have emerged from a duplication-divergence event. This implies that most proteins are conserved *throughout* the evolution process and preserve some interaction partners. This is indeed in broad agreement with empirical evidence, because traces of protein conservation are even observed within the core transcriptional and translational machineries across all three major living kingdoms (16).

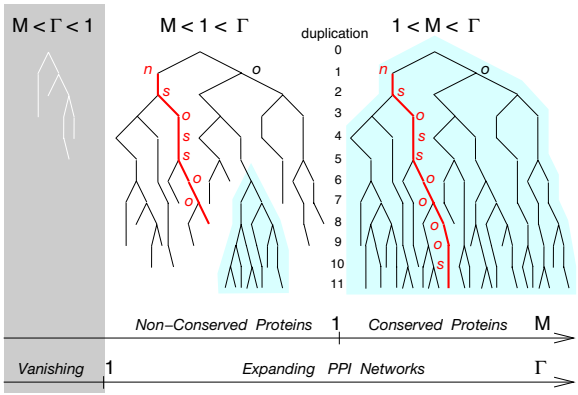

Fig. 2. Evolutionary growth (Γ) and protein conservation (M) of PPI networks. The constitutive constraint, $M < \Gamma$, defines three evolutionary regimes discussed in the text.

Evolution of PPI Network Degree Distribution. We now turn to the evolution of the degree distribution and other topological properties of PPI networks, which correspond to the technical core of the GDD model. To this end, we rescale the exponentially growing connected graph by introducing a normalized generating function for the average degree distribution,

$$p^{(n)}(x) = \sum_{k\geq 1} p_k^{(n)} x^k \quad \text{with} \quad p_k^{(n)} = \frac{\langle N_k^{(n)} \rangle}{\langle N^{(n)} \rangle}, \quad [4]$$

where $\langle N^{(n)} \rangle = \Sigma_{k\geq 1} \langle N_k^{(n)} \rangle$, that is, after removing $\langle N_0^{(n)} \rangle$. $F^{(n)}(x)$ can be reconstructed from the shifted degree distribution, $\tilde{p}^{(n)}(x) = p^{(n)}(x) - 1$, as

$$F^{(n)}(x) = \langle N^{(n)} \rangle \tilde{p}^{(n)}(x) + C\cdot(1+q)^n, \quad [5]$$

which yields the following recurrence for $\tilde{p}^{(n)}(x)$,

$$\tilde{p}^{(n+1)}(x) = \frac{(1-q)\tilde{p}^{(n)}(A_{s(x)}) + q\tilde{p}^{(n)}(A_{o(x)}) + q\tilde{p}^{(n)}(A_{n(x)})}{\Delta^{(n)}} \quad [6]$$

where $\Delta^{(n)}$ is the ratio between two consecutive graph sizes in terms of connected nodes, that is, $\Delta^{(n)} = \langle N^{(n+1)} \rangle / \langle N^{(n)} \rangle$,

$$\Delta^{(n)} = -(1-q)\tilde{p}^{(n)}(A_{s(0)}) - q\tilde{p}^{(n)}(A_{o(0)}) - q\tilde{p}^{(n)}(A_{n(0)}) > 0 \quad [7]$$

Although $\Delta^{(n)}$ is not known *a priori* and should, in general, be determined self-consistently with $\tilde{p}^{(n)}(x)$ itself, it is directly related to the evolution of the mean degree $\bar{k}^{(n)} = \Sigma_{k\geq 1} k p_k^{(n)}$ obtained by taking the first derivative of Eq. **6** at $x = 1$,

$$\frac{\bar{k}^{(n+1)}}{\bar{k}^{(n)}} = \frac{(1-q)\Gamma_s + q\Gamma_o + q\Gamma_n}{\Delta^{(n)}} = \frac{\Gamma}{\Delta^{(n)}}. \quad [8]$$

[‖]Note, however, that pseudogenes may still have a critical role in evolution by providing functional domains that can be fused to adjacent genes. This supports a view of PPI network evolution in terms of protein domains instead of entire proteins (*SI Appendix*, Fig. S6*B*, and ref. 5). Yet, we showed in ref. 5 that extensive domain shuffling does not change the resulting network topology from duplication-divergence models.

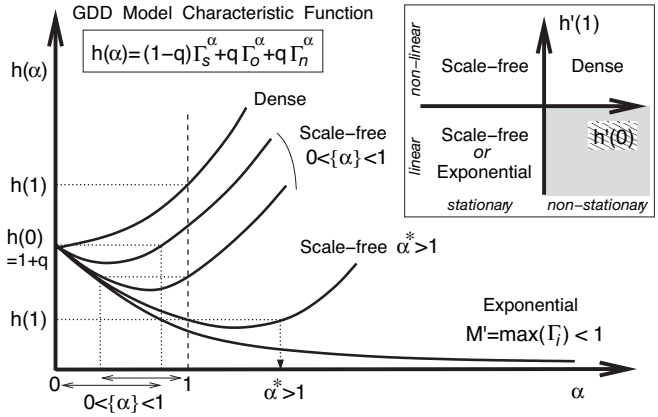

Fig. 3. Asymptotic degree distribution for GDD models. Asymptotic regimes are deduced from the *convex* characteristic function $h(\alpha)$ and its derivatives h′(0) and h′(1) (see text).

Hence, although connected networks grow exponentially both in terms of number of links (link growth rate Γ) and number of connected nodes (node growth rate $\Delta^{(n)}$), features normalized over these growing networks, such as node mean connectivity (Eq. **8**) or distributions of node degree (or simple network motifs, see below), exhibit richer evolutionary dynamics in the asymptotic limit $n \to \infty$, as we will now discuss.

Asymptotic Analysis of Node Degree Distribution (*M*′). The node degree distribution can be shown (see *SI Appendix*) to converge toward a limit function $p(x)$, with $\tilde{p}(x) = p(x) - 1$ solution of the functional Eq. **6**.

$$\tilde{p}(x) = \frac{(1-q)\tilde{p}(A_{s(x)}) + q\tilde{p}(A_{o(x)}) + q\tilde{p}(A_n(x))}{\Delta} \qquad [9]$$

where $\Delta = \lim_{n\to\infty}\Delta^{(n)}$ with both $\Delta \leq 1 + q$, the maximum node growth rate, and $\Delta \leq \Gamma$, the link growth rate, because the number of connected nodes cannot increase faster than the number of links. Asymptotic regimes with $\Delta = \Gamma$ correspond to the same exponential growth of the network in terms of connected nodes and links, and will be referred to as *linear* regimes, hereafter, whereas $\Delta < \Gamma$ corresponds to *nonlinear* asymptotic regimes, which imply a diverging mean connectivity $\bar{k}^{(n)} \to \infty$ in the asymptotic limit $n \to \infty$ (Eq. **8**).

To determine Δ and $p(x)$ self-consistently, we first express successive derivatives of $p(x)$ at $x = 1$ in terms of lower derivatives by using Eq. **9**,

$$\partial_x^k p(1)\left[1 - \frac{(1-q)\Gamma_s^k + q\Gamma_o^k + q\Gamma_n^k}{\Delta}\right] = \sum_{l=[k/2]}^{k} \alpha_{k,l}\partial_x^l p(1), \qquad [10]$$

where $\alpha_{k,l}$ are positive functions of the $1 + 6$ parameters. Inspection of this expression readily defines two classes of asymptotic regimes, *regular* and *singular* regimes, depending on the value of a *topology index* $M' = \max_i(\Gamma_i)$, for $i = s, o, n$. The detailed analysis relies on the "characteristic function" $h(\alpha) = (1-q)\Gamma_s^\alpha + q\Gamma_o^\alpha + q\Gamma_n^\alpha$, as outlined below and in Fig. 3 (see *SI Appendix*, *Asymptotic Methods*, for proof details).

Regular regimes, if $M' = \max_i(\Gamma_i) < 1$, for $i = s, o, n$. In this case, the only possible solution is $\Delta = h(1)$ (i.e., linear regime). Hence, since $M' < 1$, $h(1) > h(k)$, and successive derivatives $\partial_x^k p(1)$ are thus finite and positive for all $k \geq 1$. This corresponds to an exponential decrease of the node degree distribution for $k \gg 1$, $p_k \propto e^{-\mu k}$ with a power law prefactor. The limit average connectivity (Eq. **8**) is finite in this case, $\bar{k} < \infty$.

Singular regimes, if $M' = \max_i(\Gamma_i) > 1$, for $i = s, o, n$. In this case, Eq. **10** suggests that there exists an integer $r \geq 1$ for which the rth derivative is negative, $\partial_x^r p(1) < 0$, which is impossible by definition. This simply means that neither this derivative nor any higher ones exist (for $k \geq r$). We thus look for self-consistent solutions of the "characteristic equation" $h(\alpha) = \Delta$ (with $r - 1 < \alpha \leq r$) corresponding to a singularity of $p(x)$ at $x = 1$ and a power law tail of p_k, for $k \gg 1$ (17),

$$p(x) = 1 - \cdots - A_\alpha(1-x)^\alpha + \cdots \text{and} \quad p_k \propto k^{-\alpha-1} \qquad [11]$$

where the singular term $(1-x)^\alpha$ is replaced by $(1-x)^r \ln(1-x)$ for $\alpha = r$ exactly. Several asymptotic behaviors are predicted from the convex shape of $h(\alpha)$ ($\partial_\alpha^2 h \geq 0$), depending on the signs of its derivatives $h'(0)$ and $h'(1)$ (Fig. 3 *Inset*).

- If $h'(0) < 0$ and $h'(1) < 0$. There exists an $\alpha^\star > 1$ so that $h(\alpha^\star) = h(1)$ and the condition $\Delta \leq h(1)$ implies that $\alpha^\star \geq \alpha \geq 1$. The solution $\alpha = 1$ requires $h'(1) = 0$ and should be rejected in this case. Hence, because $\bar{k} < \infty$ for $\alpha > 1$, we must have $\Delta = h(1)$ (linear regime) and a scale-free limit degree distribution with a *unique* $\alpha = \alpha^\star > 1$, $p_k \propto k^{-\alpha^\star - 1}$ for $k \gg 1$.
- If $h'(0) < 0$ and $h'(1) = 0$. $\alpha = 1$, $\Delta = h(1)$, and $p_k \propto k^{-2}$ for $k \gg 1$ ($\bar{k}^{(n)} \to \infty$ as $n \to \infty$).
- If $h'(0) < 0$ and $h'(1) > 0$. The general condition $\Delta \leq \min(h(0), h(1))$ leads *a priori* to a whole range of possible $\alpha \in]0, 1]$ corresponding to stationary scale-free degree distributions with diverging mean degrees $\bar{k}^{(n)} \to \infty$. Yet, numerical simulations suggest that there might still be a unique asymptotic node growth rate Δ regardless of initial conditions or evolutionary trajectories, although convergence is extremely slow (see *SI Appendix*, *Numerical simulations*).
- If $h'(0) \geq 0$ and $h'(1) > 0$. $\Delta = h(0) = 1 + q$, implying that all duplicated nodes are selected in this case. No suitable α exists as the node degree distribution is exponentially shifted toward higher and higher connectivities. This is a dense, nonstationary regime with seemingly little relevance to biological networks.

Finally, note that the characteristic equation $\Delta = h(\alpha)$ can be recovered directly from the average change of connectivity $k \to k\Gamma_i$ and the following continuous approximation (by using $N^{(n)} = \sum_k N_k^{(n)} \simeq \int_u N_u^{(n)}\, du$ and $\langle N_k^{(n)}\rangle \propto k^{-\alpha-1}$),

$$\frac{\langle N^{(n+1)}\rangle}{\langle N^{(n)}\rangle} \simeq \frac{\int\langle (1-q)N_{k\Gamma_s}^{(n)}\Gamma_s + qN_{k\Gamma_o}^{(n)}\Gamma_o + qN_{k\Gamma_n}^{(n)}\Gamma_n\rangle dk}{\int_u \langle N_u^{(n)}\rangle du} = h(\alpha)$$

Local ($q \ll 1$) and Global ($q = 1$) Duplication Limits. The asymptotic degree distribution of the GDD model can be conveniently mapped into the (Γ, M) plane for two limit regimes of prime biological relevance: (*i*) for local duplication events ($q \ll 1$ and $\gamma_{ss} = 1$; Fig. 4*A*) and (*ii*) for whole-genome duplication events ($q = 1$; Fig. 4*B*). See *SI Appendix* for details.

The local duplication-divergence limit leads to scale-free limit degree distributions for both conserved and nonconserved networks, with power law exponents $1 < \alpha + 1 \leq 3$ if $\gamma_{so} \simeq 1$ (i.e., which ensures that most previous interactions are conserved in at least one copy after duplication). By contrast, the whole-genome duplication-divergence limit leads to a wide range of asymptotic behaviors from nonconserved, exponential regimes to conserved, scale-free regimes with arbitrary power law exponents. Conserved, nondense networks require, however, an asymmetric divergence between old and new duplicates ($\gamma_{oo} \neq \gamma_{nn}$) (5) and lead to scale-free limit degree distributions with the same range of exponents $1 < \alpha + 1 \leq 3$ for maximum divergence asymmetry ($\gamma_{oo} \simeq 1$ and $\gamma_{nn} \simeq 0$).

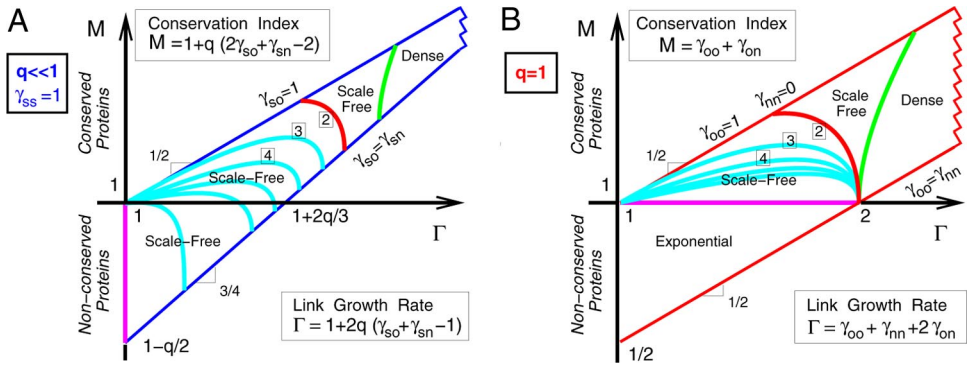

Fig. 4. Asymptotic phase diagram of PPI networks under the GDD model. (*A*) Local duplication-divergence limit ($q \ll 1$ and $\gamma_{ss} = 1$). (*B*) Whole-genome duplication-divergence limit ($q = 1$). Boxed figures are power law exponents ($\alpha + 1$) of scale-free regimes (Eq. **11**).

Evolutionary Variations of Model Parameters. The previous analysis with fixed parameters $\{q, \gamma_{ij}\}$ can be readily extended to combine local and global PPI network duplications (Fig. 4 *A* and *B*) or even include *any* evolutionary variations and stochastic fluctuations of the GDD model parameters with arbitrary series $\{q^{(n)}, \gamma_{ij}^{(n)}\}_R$ (see *SI Appendix*). Protein conservation is then found to be controlled by the cumulated product of connectivity growth/decrease rates following the most conserved, old duplicate lineage,

$$M = \left(\prod_n^R [(1 - q^{(n)})\Gamma_s^{(n)} + q^{(n)}\Gamma_o^{(n)}] \right)^{1/R} \quad [12]$$

with conserved (resp. nonconserved) protein evolutionary regimes corresponding to $M > 1$ (resp. $M < 1$).

A similar geometric average also controls the nature of the asymptotic degree distribution as the network topology index now reads,

$$M' = \left(\prod_n^R \max_i(\Gamma_i^{(n)}) \right)^{1/R} \quad [13]$$

with $M' < 1$ corresponding to exponential networks and $M' > 1$ to scale-free (or dense) networks with an effective node degree exponent α and effective node growth rate Δ that are self-consistent solutions of the generalized characteristic equation,

$$h(\alpha) = \left(\prod_n^R h^{(n)}(\alpha) \right)^{1/R} = \Delta, \quad [14]$$

where $h^{(n)}(\alpha) = (1 - q^{(n)})\Gamma_s^{(n)\alpha} + q^{(n)}\Gamma_o^{(n)\alpha} + q^{(n)}\Gamma_n^{(n)\alpha}$, as before. This leads to *exactly* the same discussion for singular regimes as with constant q and Γ_i (Fig. 3) because of the convexity of the generalized function $h(\alpha)$ ($\partial_\alpha^2 h(\alpha) \geq 0$; see *SI Appendix* for details and discussion on the $R \to \infty$ limit).

In particular, because $(1 - q^{(n)})\Gamma_s^{(n)} + q^{(n)}\Gamma_o^{(n)} \leq \max_i(\Gamma_i^{(n)})$ for *all* $q^{(n)}$ and $\Gamma_i^{(n)}$ ($i = s, o, n$), we *always* have $M \leq M'$. This relation implies a fundamental linkage between protein conservation and network topology under general duplication-divergence evolution, regardless of all possible evolutionary variations of the model parameters, $q^{(n)}$ and $\Gamma_i^{(n)}$. We expect, in particular, that *all conserved networks are necessarily scale-free* (or dense) ($1 < M \leq M'$), whereas *exponential networks can never be conserved* ($M \leq M' < 1$), under general duplication-divergence evolution.

Evolution of PPI Network Motifs. The generating function approach, introduced for the one-node degree distribution $p_k^{(n)}$ (Eqs. **1–6**), can be generalized to analyze the evolutionary statistics of multinode correlation functions and related clustering coefficient, distribution of first-neighbor average connectivity g_k (18) (see Fig. 6) and small-network motifs. Yet, although M' also controls transitions between major evolutionary regimes for multinode correlation functions, their analysis remains technically involved (*SI Appendix*).

By contrast, the conservation property of network motifs under general duplication-divergence evolution turns out to be remarkably simple, as outlined in Fig. 5. We derive conservation indices for specific network motifs by summing over all possible combinations of s nodes (with probability $\kappa_s = 1 - q$) or o nodes (with probability $\kappa_o = q$) and the corresponding γ_{ij} ($i, j = s, o$) (Fig. 5). Clearly, network motifs with a larger number of interactions, $p \geq 1$, have lower conservation indices, $M_p \simeq O(\gamma_{ij}^p)$ (Fig. 5). Moreover, because the probability to conserve a specific interaction γ_{ij} cannot be exactly 1, because of deleterious mutations (i.e., $\gamma_{ij} < 1$), motif conservation indices M_p must *all* be <1, regardless of any parameter variations, $q^{(n)}$ and $\gamma_{ij}^{(n)}$.

Hence, network motifs *cannot* be indefinitely conserved under duplication-divergence evolution, even though their individual proteins *are* typically conserved in the network (if $M > 1$) (Fig. 2). This implies that *structural* orthology between individual proteins from phylogenetically distant species *cannot* indefinitely coincide with *functional* orthology at the level of protein interactions and complexes, in broad agreement with empirical evidences (19). The resulting turnover toward more and more divergent interaction partners is a simple evolutionary consequence of the GDD model, regardless of any network-rewiring dynamics (that have been neglected here). In particular, even the most conserved orthologous proteins (*s*/*o* descents) must eventually perform different functions, but conserved ancestral functions are inevitably passed down to less conserved protein complexes (*s*/*o*/*n* descents) in phylogenetically distant species. This inherent evolutionary constraint of the GDD model sets

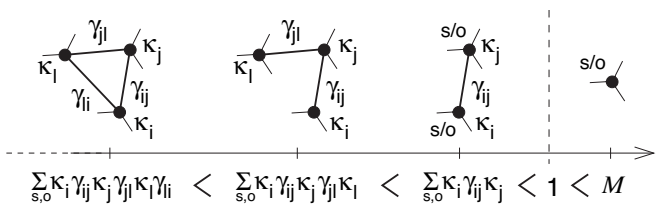

Fig. 5. Motif conservation indices. Although individual proteins are typically conserved (if $M > 1$), network motifs including two or more proteins *cannot* be indefinitely conserved under general duplication-divergence evolution ($\kappa_o = 1 - \kappa_s = q$; see text).

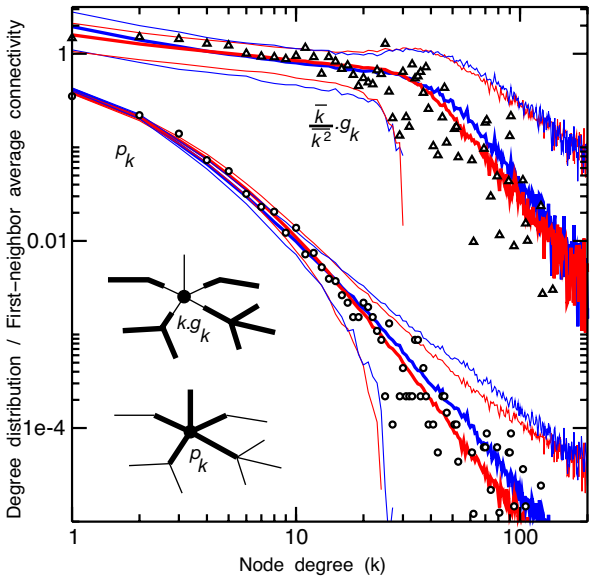

Fig. 6. Comparison between empirical PPI network data (20) and finite-size, duplication-divergence simulations. Protein physical interaction data for yeast are taken from the Biomolecular Interaction Network Database (20) (4,576 proteins, 9,133 physical interactions, $\bar{k} = 3.99$, $\bar{k}^2 = 106$). Both connectivity distribution p_k (open circles) and first-neighbor average connectivity g_k (18) (open triangles) are shown. Best one-parameter fit of the data (blue curves) ($\pm 2\sigma$ for expected deviations) for the most asymmetric genome duplication-divergence model with $\gamma_{on} = 0.26$ (and $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$) (5). Best two-parameter fit of the data (red curves) ($\pm 2\sigma$ for expected deviations) with the same most asymmetric whole-genome duplication limit but at the level of protein domains instead of entire proteins, with $\gamma_{on} = 0.1$ and $\lambda = 0.3$, which corresponds to $1/(1 - \lambda) = 1.5$ average number of binding domains per proteins. See *SI Appendix*, Fig. S6 and ref. 5 for details about this model including domain shuffling.

conceptual and practical limitations on functional annotation transfer between genomes (19).

Comparison with Empirical PPI Network Data. Although the GDD model can, in principle, accommodate $1q + 6\gamma_{ij}$ fitting parameters, and any evolutionary variations or fluctuations $q^{(n)}$ and $\gamma_{ij}^{(n)}$, we have restricted all quantitative comparisons with the empirical data (20) to biologically relevant regimes, limited to one or two effective fitting parameters, only.

The results, corresponding to 10^3 to 10^4 protein nodes and low network growth rates in the conserved, scale-free regime ($1 < M \leq M'$), are compared with Yeast PPI network data (20) (Fig. 6), both for the connectivity distribution p_k and the first-neighbor average connectivity g_k (18) (see Fig. 6 legend and *SI Appendix* for model and fitting details).

Extending these numerical simulations to larger PPI network sizes ($>10^5$ nodes) shows little changes in these distributions for small $k \leq 20$ (*SI Appendix*, Figs. S5–S7). This suggests that the available empirical PPI network for yeast (20) has essentially reached its asymptotic degree distribution, for $k \leq 20$. By contrast, global convergence of the GDD model is typically quite slow for large growth rate regimes with little biological relevance, as shown in *SI Appendix*, *Numerical simulations*, Figs. S5–S7.

Conclusions

Although conservation and network topology have *a priori* no reason to be related features of emerging PPI networks in the course of evolution, we demonstrate in this article that they are, in fact, linked properties under general duplication-divergence evolution, because of a fundamental linkage between protein conservation and network topology indices, that is, $M \leq M'$, regardless of any variations or fluctuations of the model parameters.

By contrast, we also showed that conservation of network motifs cannot be indefinitely preserved under general duplication-divergence evolution (independently from any network rewiring dynamics). This underlies intrinsic limitations on functional annotation transfer across phylogenetically distant genomes (19).

All in all, these evolutionary constraints, inherent to duplication-divergence processes and independent from selective adaptation (3), appear to have largely controlled the overall topology and scale-dependent conservation of PPI networks, regardless of any specific biological function.

ACKNOWLEDGMENTS. We thank U. Alon, R. Bruinsma, M. Cosentino-Lagomarsino, T. Fink, R. Monasson, and M. Vergassola for discussion. This work was supported by the Ministry of Higher Education and Research, Centre National de la Recherche Scientifique, and Institut Curie.

1. Ohno S (1970) *Evolution by Gene Duplication* (Springer, New York).
2. Li WH (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
3. Lynch M (2007) *The Origins of Genome Architecture* (Sinauer, Sunderland, MA).
4. Berg J, Lässig M, Wagner (2004) A Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol* 4:51.
5. Evlampiev K, Isambert H (2007) Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst Biol* 1:49.
6. Maslov S, Sneppen K, Eriksen KA, Yan KK (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* 4:9.
7. Jones S, Thornton JM, (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 93:13–20.
8. Henschel A, Kim WK, Schroeder M (2006) Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics* 22:550–555.
9. Kim WK, Henschel A, Winter C, Schroeder M, (2006) The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput Biol* 2:e124.
10. Ispolatov I, Krapivsky PL, Yuryev (2005) A Duplication-divergence model of protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys* 71:061911.
11. Albert R, Barabási A-L (2002) Statistical mechanisms of complex networks. *Rev Mod Phys* 74:47–97.
12. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5:101–113.
13. Raval A (2003) Some asymptotic properties of duplication graphs. *Phys Rev E Stat Nonlin Soft Matter Phys* 68:066119.
14. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. *ComPlexUs* 1:38–44.
15. Ispolatov I, Krapivsky PL, Mazo I, Yuryev A (2005) Cliques and duplication-divergence network growth. *N J Phys* 7:145.
16. Hartman H, Favaretto P, Smith TF (2006) The archaeal origins of the eukaryotic translational system. *Archaea* 2(1):1–9.
17. Flajolet P, Sedgewick R (2006) *Analytic Combinatorics.* Available at http://algo.inria.fr/flajolet/Publications/books.html. Accessed March 22, 2007.
18. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910.
19. Yu H, *et al.* (2004) Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. *Genome Res* 14:1107–1118.
20. Alfarano C, *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 33:D418–D424.