Logistic map analysis of biomolecular network evolution

R. R. Stein and H. Isambert[*]

Institut Curie, CNRS-UMR168, UPMC, 11 rue P. et M. Curie, F-75005 Paris, France and
Fondation Pierre-Gilles de Gennes pour la recherche, 29 rue d'Ulm, F-75005 Paris, France

(Received 3 February 2011; revised manuscript received 22 August 2011; published 10 November 2011)

We study the expansion of biomolecular networks from the view point of first evolutionary principles based on the duplication and divergence of ancestral genes. The expansion of gene families and subnetworks is analyzed in terms of logistic map compositions, which capture the varying functional constraints of individual genes in the course of evolution. Using a mean-field approach, we then demonstrate the existence of spontaneous growth-rate variations between gene families and discuss the relevance of such heterogeneous expansions for the emergent properties of actual biomolecular networks.

DOI: 10.1103/PhysRevE.84.051904 PACS number(s): 87.18.−h, 05.40.−a, 87.16.ad, 87.23.Kg

I. INTRODUCTION

In the course of evolution, biomolecular networks have experienced heterogeneous expansions of different gene families. In particular, regulatory and signaling genes have undergone a more rapid expansion than other gene families in typically genomes [1]. Such heterogeneous expansions have been proposed to arise through horizontal gene transfers in prokaryotes [2]. Yet horizontal gene transfers are less prevalent in eukaryotes, suggesting that the heterogeneous expansions of their gene families arose, instead, from actual variations in the gene duplication-divergence events, which have occurred repeatedly and independently along each eukaryote lineage [3].

The importance of duplication-divergence processes in evolution has long been recognized [4] and motivated convergent theoretical efforts to analyze its impact on the emergent properties of biomolecular networks [5–14].

Here we analyze the expansions of gene families and subnetworks, which cannot be studied by standard generating function analysis [14]. To this end, we first demonstrate that duplication-divergence processes can be analyzed in terms of logistic map compositions, which capture the varying functional constraints of individual genes in the course of evolution. We then establish that, even beyond explicit distinctions between gene families, heterogeneities in their expansion rates arise in fact spontaneously under duplication-divergence evolution and are directly coupled to the emergent properties of their biomolecular subnetworks.

II. GENERAL DUPLICATION-DIVERGENCE MODEL

The General Duplication-Divergence (GDD) model, introduced in Ref. [14], aims at capturing the statistical properties of biomolecular networks evolving through stochastic duplication and divergence events at various genomic scales, from single-gene to whole-genome duplications [12]. At each evolutionary step, depicted in Fig. 1, a fraction q of genes is randomly duplicated, and the interactions of the resulting network are stochastically conserved with distinct probabilities, γ_{ij}. These probabilities reflect the different functional constraints between interacting partners, depending on their recent duplication history. Indeed, "single" nonduplicated genes ('s') are more evolutionary constrained than recently duplicated gene pairs ('o'/'n'), which typically undergo asymmetric divergence, i.e., $\gamma_{nj} \ll \gamma_{oj} \simeq \gamma_{sj}$, $j = s,o,n$.

We note, N_k, the number of genes (or "nodes") of connectivity k in the overall network, $N = \sum_{k\geqslant 0} N_k$, the total number of nodes and, $L = \sum_{k\geqslant 0} kN_k/2$, the total number of interactions (or "links"). We then study the evolutionary dynamics of the ensemble averages $\langle N_k\rangle^{(n)}$ after n duplications using the generating function,

$$F^{(n)}(x) = \sum_{k\geqslant 0} \langle N_k\rangle^{(n)} x^k. \tag{1}$$

The evolutionary dynamics of $F^{(n)}(x)$ corresponds to the following recurrence deduced from the microscopic definition of the GDD model [14]:

$$F^{(n+1)}(x) = \sum_{i}^{s,o,n} \epsilon_i F^{(n)}(A_i(x))$$
$$A_i(x) = 1 - \Gamma_i(1-x) + D_i(1-x)^2, \tag{2}$$

where $\epsilon_s = 1-q$, $\epsilon_o = \epsilon_n = q$, $\Gamma_i = (1-q)\gamma_{is} + q(\gamma_{io} + \gamma_{in})$, and $D_i = q\gamma_{io}\gamma_{in}$ for $i = s,o,n$. In the following, we note $\Gamma = \sum_i \epsilon_i \Gamma_i$, the network growth rate in terms of number of links and $D = \sum_i \epsilon_i D_i$. The exponentially growing network is rescaled by introducing a normalized generating function for the average degree distribution:

$$p^{(n)}(x) = \sum_{k\geqslant 1} p_k^{(n)} x^k \quad \text{with} \quad p_k^{(n)} = \frac{\langle N_k\rangle^{(n)}}{\langle N\rangle^{(n)}}, \tag{3}$$

where $\langle N\rangle^{(n)} = \sum_{k\geqslant 1} \langle N_k\rangle^{(n)}$, after removing $\langle N_0\rangle^{(n)}$.

Then the complement generating function for degree distribution, $\tilde{p}^{(n)}(x) = 1 - p^{(n)}(x)$, follows the recurrence:

$$\tilde{p}^{(n+1)}(x) = \frac{\sum_i^{s,o,n} \epsilon_i \tilde{p}^{(n)}(A_i(x))}{\Delta^{(n)}}, \tag{4}$$

where $\Delta^{(n)} = \langle N\rangle^{(n+1)}/\langle N\rangle^{(n)} = \sum_i \epsilon_i \tilde{p}^{(n)}(A_i(0))$ is the network growth rate in terms of number of genes.

A characteristic equation relating the asymptotic growth rate $\Delta = \lim_{n\to\infty} \Delta^{(n)}$ and the exponent $\alpha + 1$ of the

[*] herve.isambert@curie.fr

1539-3755/2011/84(5)/051904(7)

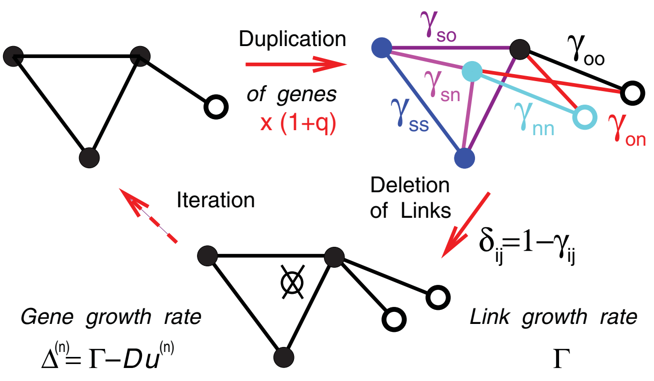

FIG. 1. (Color online) Schematic of the general duplication-divergence model. The expansion of a gene family (white genes) arising from a single gene of initial connectivity $d \geqslant 1$ is analyzed in the text..

power-law degree distribution of the resulting networks ($p_k \sim 1/k^{\alpha+1}$) was established in Ref. [14]:

$$\Delta = \sum_i \epsilon_i \Gamma_i^{\alpha}. \tag{5}$$

Yet, this characteristic equation relating the two unknowns Δ and α could not be solved in general, except in simple cases for which Δ is independently known, namely, when gene and interaction growth rates are the same, i.e., $\Delta = \Gamma$.

Here we demonstrate that the network size $\langle N\rangle^{(n)}$ and expansion rate $\Delta^{(n)}$ can in fact be evaluated independently from α in *all* evolutionary regimes ($\Delta \leqslant \Gamma$) by analyzing the duplication-divergence dynamics in terms of compositions of *logistic maps* (Fig. 2), which are simple quadratic forms known to exhibit complex dynamics.

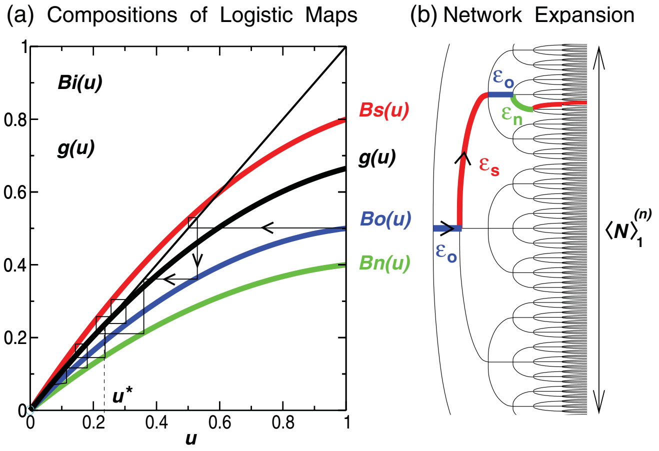

FIG. 2. (Color online) Logistic map analysis of network expansion. (a) Example of stochastic composition of the three logistic maps, $B_i(u) = \Gamma_i u - D_i u^2$, $i = s,o,n$, and mean-field weighted composition $g(u) = u\ \exp[\sum_i \epsilon_i (\Gamma_i - D_i\ u) \ln(\Gamma_i - D_i\ u)/(\Gamma - D\, u)]$ with an effective fixed point $u^{\star}$; see text. Here $\Gamma_s = 1.3$, $D_s = 0.5$, $\Gamma_o = 0.9$, $D_o = 0.4$, $\Gamma_n = 0.7$, $D_n = 0.3$, $\epsilon_o = \epsilon_n = 1 - \epsilon_s = 0.3$, $\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i) \simeq 0.1 > 0$ and $u^{\star} \simeq 0.23$. (b) Corresponding evolutionary history of individual genes and resulting network expansion.

III. LOGISTIC MAP ANALYSIS OF BIOMOLECULAR NETWORK EXPANSION

A. Explicit expansion of (sub)networks $\langle N\rangle^{(n+1)}$

We first note that $\Delta^{(n)}$ can be expressed through successive calls to the three functions $A_i(x)$, $i = s,o,n$:

$$\begin{aligned}
\Delta^{(n)} &= \sum_{i_0} \epsilon_{i_0} \tilde{p}^{(n)}\big(A_{i_0}(0)\big) \\
&= \frac{\sum_{i_1} \sum_{i_0} \epsilon_{i_1} \epsilon_{i_0} \tilde{p}^{(n-1)}\big(A_{i_1}\big(A_{i_0}(0)\big)\big)}{\Delta^{(n-1)}} \\
&= \frac{\sum_{i_1} \sum_{i_0} \epsilon_{i_1} \epsilon_{i_0} \tilde{p}^{(n-1)}\big(A_{i_1}\big(A_{i_0}(0)\big)\big)}{\sum_{i_0} \epsilon_{i_0} \tilde{p}^{(n-1)}\big(A_{i_0}(0)\big)} \\
&= \frac{\sum_{i_0,i_n}^{s,o,n} \big(\prod_k^n \epsilon_{i_k}\big)\ \tilde{p}^{(0)}\big(A_{i_n}\big(A_{i_{n-1}}\big(\cdots\big(A_{i_0}(0)\big)\big)\big)\big)}{\sum_{i_0,i_{n-1}}^{s,o,n} \big(\prod_{k=0}^{n-1} \epsilon_{i_k}\big)\ \tilde{p}^{(0)}\big(A_{i_{n-1}}\big(\cdots\big(A_{i_0}(0)\big)\big)\big)}.
\end{aligned}$$

Then the expansion of a biomolecular network under duplication-divergence evolution, $\langle N\rangle^{(n+1)}$, can be explicitly expressed as

$$\begin{aligned}
\langle N\rangle^{(n+1)} &= \Delta^{(n)} \langle N\rangle^{(n)} = N_o \prod_{k=0}^{n} \Delta^{(k)} \\
&= N_o \sum_{i_0,i_n}^{s,o,n} \left(\prod_k^n \epsilon_{i_k}\right) \tilde{p}^{(0)}\big(A_{i_n}\big(A_{i_{n-1}}\big(\cdots\big(A_{i_0}(0)\big)\big)\big)\big),
\end{aligned}$$

where N_o and $\tilde{p}^{(0)}(x) = 1 - p^{(0)}(x)$ refer to the number of nodes N_o and the node degree distribution $p^{(0)}(x)$ of the initial network.

Note, however, that the approach is equally valid to describe the expansion of a subnetwork within the overall network, starting, for instance, from a single gene ($N_o = 1$) with initial degree $d \geqslant 1$. We will study such simple initial conditions, below, as the expansion of subnetworks derived from a single gene can then be linearly combined to describe the expansion of biomolecular networks starting from arbitrary initial graphs.

B. Subnetwork generated by a single node of initial degree $d = 1$

1. Subnetwork expansion as composition of logistic maps

We will first discuss the expansion of a subnetwork generated by a single node of initial degree 1, which corresponds to $\tilde{p}^{(0)}(x) = 1 - x$. Then $\langle N\rangle^{(n+1)} = \langle N\rangle_1^{(n+1)}$ simply becomes

$$\begin{aligned}
\langle N\rangle_1^{(n+1)} &= \sum_{i_0,i_n}^{s,o,n} \left(\prod_k^n \epsilon_{i_k}\right) \big[1 - A_{i_n}\big(A_{i_{n-1}}\big(\cdots\big(A_{i_0}(0)\big)\big)\big)\big] \\
&= \sum_{i_0,i_n}^{s,o,n} \left(\prod_k^n \epsilon_{i_k}\right) B_{i_n}\big(B_{i_{n-1}}\big(\cdots\big(B_{i_0}(1)\big)\big)\big) = Z^{(n)},
\end{aligned}$$

where $A_i(x)$, $i = s,o,n$, are transformed into the three logistic maps, $B_i(u) = u\ (\Gamma_i - D_i\ u)$, $u = 1 - x$, with similar composition rules:

$$\begin{aligned}
B_i(u) &= 1 - A_i(1-u) = u\ (\Gamma_i - D_i\ u), \\
B_i(B_j(u)) &= 1 - A_i(1 - B_j(u)) = 1 - A_i(A_j(x)).
\end{aligned}$$

So the expansion of the subnetwork, in terms of averaged number of nodes $\langle N\rangle_1^{(n+1)}$ corresponds to a partition function $Z^{(n)}$ over all possible *weighted* compositions of the three logistic maps, $B_i(u)$ (Fig. 2). In particular, each weighted composition [Fig. 2(a)] corresponds to the probability for a unique duplication-divergence history of individual genes in the course of evolution, as depicted in Fig. 2(b).

Interestingly, the duplication-divergence dynamics of biomolecular networks combines three logistic maps, $B_i(u) = u\,(\Gamma_i - D_i u)$, that are each individually in simple monotonic convergence regimes ($\Gamma_i \leqslant 2$) with a single stable fixed point, either at $u_i^\star = 0$ if $\Gamma_i \leqslant 1$ or at $u_i^\star = (\Gamma_i - 1)/D_i$ if $1 < \Gamma_i \leqslant 1 + q \leqslant 2$. While the stochastic compositions of logistic maps have been shown to affect their stability in complex dynamical regimes [15], the analysis of *weighted* compositions of logistic maps remains apparently an open problem, even when each map is in a monotonic convergence regime.

We propose below a mean-field approach that addresses this question and leads to two evolutionary scenarios: (1) a linear expansion regime with an "effective" fixed point at $u^\star = 0$ and $\Delta = \Gamma$ and (2) a nonlinear expansion regime with $u^\star > 0$ and $\Delta \simeq \Gamma - D\,u^\star < \Gamma$.

2. *Mean-field approach*

Noting $u_{i_{0,n}}^{(n)} = B_{i_n}(B_{i_{n-1}}(\cdots(B_{i_0}(1))))$, we have

$$u_{i_{0,n}}^{(n)} = \left(\Gamma_{i_n} - D_{i_n} u_{i_{0,n-1}}^{(n-1)}\right) u_{i_{0,n-1}}^{(n-1)} = \Delta_{i_{0,n}}^{(n)}\; u_{i_{0,n-1}}^{(n-1)}$$
$$= \Delta_{i_{0,n}}^{(n)} \Delta_{i_{0,n-1}}^{(n-1)} \cdots \Delta_{i_{0,1}}^{(1)} \Delta_{i_{0,0}}^{(0)} u_0 = \prod_k^n \Delta_{i_{0,k}}^{(k)},$$

where $\Delta_{i_{0,k}}^{(k)} = \Gamma_{i_k} - D_{i_k} u_{i_{0,k-1}}^{(k-1)}$ and $u_0 = 1$. Hence,

$$\langle N\rangle_1^{(n+1)} = \sum_{i_0,i_n}\left(\prod_k^n \epsilon_{i_k}\right) u_{i_{0,n}}^{(n)} = \sum_{i_0,i_n}\left(\prod_k^n \epsilon_{i_k}\Delta_{i_{0,k}}^{(k)}\right).$$

Yet, using $B_{i_k}(u) = u(\Gamma_{i_k} - D_{i_k} u)$, we can also directly derive $Z^{(n)} = \prod_k^n \Delta^{(k)} = \langle N\rangle_1^{(n+1)}$ as

$$Z^{(n)} = \prod_k^n \left(\sum_{i_k} \epsilon_{i_k}\Delta_{i_k}^{(k)}\right) = \sum_{i_0,i_n}\left(\prod_k^n \epsilon_{i_k}\Delta_{i_k}^{(k)}\right),$$

$$\text{with}\quad \Delta_{i_k}^{(k)} = \Gamma_{i_k} - D_{i_k} u^{(k-1)}$$

$$\text{and}\quad u^{(k-1)} = \frac{\sum_{i_0,i_{k-1}}^{s,o,n}\left(\prod_{l=0}^{k-1}\epsilon_{i_l}\right)\left(u_{i_{0,k-1}}^{(k-1)}\right)^2}{\sum_{i_0,i_{k-1}}^{s,o,n}\left(\prod_{l=0}^{k-1}\epsilon_{i_l}\right) u_{i_{0,k-1}}^{(k-1)}},$$

where $u^{(k-1)}$ is independent of the trajectory history as it is effectively averaged over all trajectories $u_{i_{0,k-1}}^{(k-1)}$.

Hence, although all compositions $u_{i_{0,n}}^{(n)}$ contribute to the partition function $Z^{(n)} = \langle N\rangle_1^{(n+1)}$, only a subset of trajectories contributes significantly and becomes eventually independent of their history in the asymptotic limit, i.e., $u_{i_{0,n}}^{(n)} \simeq u^{(n)}$ and $\Delta_{i_{0,n}}^{(n)} \simeq \Delta_{i_n}^{(n)}$.

Now, the average number of times $n_{i_k}^{(k)}$ each factor $\Delta_{i_k}^{(k)}$ is used in the partition function $Z^{(n)} = \sum_{i_k}(\prod_k^n \epsilon_{i_k}\Delta_{i_k}^{(k)})$ is given by the mean-field result:

$$n_{i_k}^{(k)} = \Delta_{i_k}^{(k)} \frac{\partial \ln(Z^{(n)})}{\partial \Delta_{i_k}^{(k)}} = \frac{\epsilon_{i_k}\Delta_{i_k}^{(k)}}{\Delta^{(k)}}. \tag{6}$$

So, averaging over all trajectories $u_{i_{0,k}}^{(k)} = \Delta_{i_{0,k}}^{(k)}\; u_{i_{0,k-1}}^{(k-1)}$, we expect that $u^{(k)}$ follows:

$$\begin{aligned} u^{(k)} &= \left(\prod_{i_k} {\Delta_{i_k}^{(k)}}^{n_{i_k}^{(k)}}\right) u^{(k-1)} \\ &= u^{(k-1)} e^{\sum_{i_k}\epsilon_{i_k}\Delta_{i_k}^{(k)}\ln(\Delta_{i_k}^{(k)})/\Delta^{(k)}} \end{aligned} \tag{7}$$

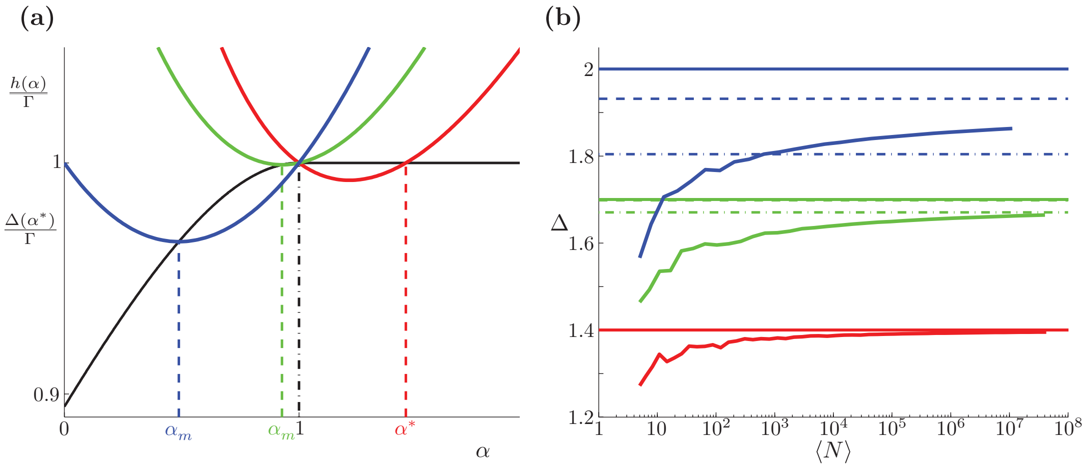

FIG. 3. (Color online) Network expansion rates $\Delta^{(n)}$ under duplication-divergence evolution. (a) Examples of characteristic functions $h(\alpha)$ and geometrical constructs giving Δ and α in linear regimes (red curve, right) [$\Delta = \Gamma = h(\alpha^\star)$ and $\alpha = \alpha^\star > 1$] and nonlinear regimes (green curve, middle, and blue curve, left) [$\Delta = \Delta_{\min} = \Gamma - D\,u_m = h(\alpha_m)$ and $\alpha = \alpha_m < 1$]. (b) Numerical estimates of network expansion rates $\Delta^{(n)}$ corresponding to the expansion regimes in (a). Evolutionary parameters: $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$, $\gamma = \gamma_{on} = \gamma_{no} = 0.2$ (red lower curves), 0.35 (green middle curves), 0.5 (blue upper curves). For $\gamma = 0.2$ (red lower curves), the network expansion is linear, $\Delta = \Gamma$ (continuous line). For $\gamma > 0.32$ (green middle curves and blue upper curves), the network expansion is nonlinear, $\Delta = \Delta_{\min} = \Gamma - D\,u_m$ (dashed line) $\simeq \Gamma - D\,u^\star$ (dotted-dashed line)$< \Gamma$ (continuous line).

in the mean-field approximation, which leads to

$$\begin{aligned} u^{(k)} &= g(u^{(k-1)}), \\ g(u) &= u\ e^{\sum_i \epsilon_i (\Gamma_i - D_i\ u) \ln(\Gamma_i - D_i\ u)/(\Gamma - D\,u)}, \\ g'(0) &= e^{\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i)/\Gamma}, \end{aligned}$$

where $g(u)$ has a single fixed point either at $u^\star = 0$ if $g'(0) \leqslant 1$ or at $u^\star > 0$, if $g'(0) > 1$. In that case, it is the unique positive solution $u^\star > 0$ of $m(u^\star) = \sum_i \epsilon_i (\Gamma_i - D_i\ u^\star) \ln(\Gamma_i - D_i\ u^\star) = 0$ [Fig. 2(a)].

3. *Linear* (**Δ = Γ**) *versus nonlinear* (**Δ < Γ**) *expansions*

Hence, we have two regimes, $u^\star = 0$ ($\Delta = \Gamma$) or $u^\star > 0$ ($\Delta < \Gamma$), for the expansion of biomolecular (sub)networks under duplication-divergence evolution, starting from node(s) of connectivity $d = 1$. Their asymptotic node degree distribution p_k depends on the value of α_m, which minimizes the convex characteristic function $h(\alpha) = \sum_i \epsilon_i \Gamma_i^\alpha$, $h'(\alpha_m) = \sum_i \epsilon_i \Gamma_i^{\alpha_m} \ln(\Gamma_i) = 0$,

Linear regime $d = 1$ ($\Delta = \Gamma$): If $\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i) \leqslant 0$, $\Delta^{(n)} \to \Delta = \Gamma$ (i.e., $u^\star = 0$) implies that the networks eventually expand at the same rates in terms of links (Γ) and nodes (Δ). This result ($\Delta = \Gamma$) and its domain of validity [$\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i) \leqslant 0$] were independently derived in Ref. [14], where the resulting network degree distributions are also established for linear regimes ($\alpha_m \geqslant 1$):

(1) $\alpha_m = \infty$ corresponds to an exponential degree distribution, $p_k \propto x_0^{-k}$, $x_0 = \min[1 + (1 - \Gamma_i)/D_i] > 1$.

(2) $1 \leqslant \alpha_m < \infty$ leads to a scale-free node-degree distribution, $p_k \propto k^{-\alpha^\star - 1}$, where $\alpha^\star > \alpha_m$ is the unique solution >1 of the characteristic equation $h(\alpha) = \sum_i \epsilon_i \Gamma_i^\alpha = \Delta = \Gamma$, as depicted in Fig. 3(a) (red construct).

Nonlinear regime $d = 1$ ($\Delta < \Gamma$): If $\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i) > 0$, $\Delta^{(n)} \to \Delta = \Gamma - D\,u^\star < \Gamma$, where $u^\star > 0$ is the solution of $\sum_i \epsilon_i (\Gamma_i - D_i\ u^\star) \ln(\Gamma_i - D_i\ u^\star) = 0$; see Fig. 2. $\Delta < \Gamma$ implies that the networks expand more slowly in terms of nodes than links in this case, leading to a diverging mean connectivity $\bar{k}^{(n)} \sim (\Gamma/\Delta)^n \to \infty$. Nonlinear regimes present either stationary or nonstationary asymptotic node-degree distributions depending on the value $\alpha_m < 1$:

(3) $0 < \alpha_m < 1$ leads to stationary scale-free degree distributions, $p_k \propto k^{-\alpha - 1}$, where α is the solution of the characteristic equation $h(\alpha) = \sum_i \epsilon_i \Gamma_i^\alpha = \Delta \simeq \Gamma - D\,u^\star$, while $u^\star$ is simultaneously solution of the mean-field equation $m(u^\star) = \sum_i \epsilon_i (\Gamma_i - D_i\ u^\star) \ln(\Gamma_i - D_i\ u^\star) = 0$, in the mean-field approximation. The latter can be differentiated as $\partial_\alpha m(u^\star) = \partial_\alpha u^\star m'(u^\star) = 0$ to yield $h'(\alpha) = -D\partial_\alpha u^\star = 0$ [as $m'(u^\star) < 0$[1]], implying that $\alpha = \alpha_m$ and, hence, $\Delta = \Delta_{\min} = \Gamma - D\,u_m = h(\alpha_m)$ and $p_k \propto k^{-\alpha_m - 1}$, as depicted in Fig. 3(a) (green and blue constructs).

(4) $\alpha_m \leqslant 0$ leads to nonstationary degree distributions corresponding to increasingly dense networks with $\Delta = 1 + q < \Gamma$.

These mean-field estimates of Δ in linear and nonlinear regimes are in good agreement with the corresponding numerical simulations of network expansion [Fig. 3(b)], although nonlinear regimes are found to converge more slowly toward their asymptotic limits.

C. Subnetwork generated by a single node of initial degree *d* > 1

We now discuss the expansion of a subnetwork starting initially with a single node of degree $d > 1$, which corresponds to $\tilde{p}^{(0)}(x) = 1 - x^d$. Then $\langle N \rangle^{(n+1)} = \langle N \rangle_d^{(n+1)}$ simply becomes

$$\begin{aligned} \langle N \rangle_d^{(n+1)} &= \prod_k^n \Delta_d^{(k)} = \sum_{i_0,i_n}^{s,o,n} \left(\prod_k^n \epsilon_{i_k} \right) \left[1 - v_{i_{0,n}}^{(n)\,d}\right] \\ &= \sum_{i_0,i_n}^{s,o,n} \left(\prod_k^n \epsilon_{i_k} \right) u_{i_{0,n}}^{(n)} \left[1 + v_{i_{0,n}}^{(n)} + \cdots + v_{i_{0,n}}^{(n)\,d-1}\right], \end{aligned}$$

where $v_{i_{0,n}}^{(n)} = A_{i_n}(A_{i_{n-1}}(\cdots(A_{i_0}(0)))) = 1 - u_{i_{0,n}}^{(n)}$.

Hence, for large initial degrees, $d \gg 1$, we find for the early rounds of duplication, $v_{i_{0,n}}^{(n)^d} \ll 1$ and $\Delta_d^{(k)} \simeq \sum_{i_k} \epsilon_{i_k} = 1 + q$, corresponding to the maximum expansion rate. Then, as for $d = 1$, two asymptotic regimes should be distinguished for the evolutionary expansion of subnetworks starting from a single node with $d > 1$:

Linear regime $d > 1$ ($\Delta = \Gamma$): If $\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i) \leqslant 0$, $u^{(n)} \to u^\star = 0$ and $v^{(n)} \to v^\star = 1$ implying that

$$\frac{\langle N \rangle_d^{(n)}}{\langle N \rangle_1^{(n)}} \to d, \tag{8}$$

where $\langle N \rangle_1^{(n)} = \prod_{k=0}^{n-1} \Delta^{(k)}$. This corresponds to a heterogeneous expansion of subnetworks, for which nodes with d initial neighbors generate on average d times more nodes in their descent than nodes with one neighbor only. As a consequence, any part of the initial network eventually generates an expanding subnetwork of size $\langle N \rangle_{\text{sub}}^{(n)}$ proportional to its initial mean degree, $\langle N \rangle_{\text{sub}}^{(n)} \propto \bar{d}^{(o)} \langle N \rangle_1^{(n)}$, where $\bar{d}^{(o)} = \sum_d d\ p_d^{(o)}$. This leads to biomolecular subnetworks with gene families of different sizes but exhibiting, ultimately, the same degree distributions independent of $\bar{d}^{(o)}$ (see p_k above for $d = 1$). The emergence of different stationary degree distributions within the same biomolecular network requires, in fact, explicit distinctions between gene families and their evolutionary parameters (γ_{ij}).

This heterogeneous expansion of simple subnetworks in linear regimes, $\langle N \rangle_d^{(n)} \sim d\ \langle N \rangle_1^{(n)}$, is in good agreement with the corresponding numerical simulations (Fig. 4) (red curves), although $\langle N \rangle_d^{(n)} / \langle N \rangle_1^{(n)}$ is found to converge more slowly toward d (red dash lines) for large initial degree, $d \gg 1$.

Nonlinear regime $d > 1$ ($\Delta < \Gamma$): If $\sum_i \epsilon_i \Gamma_i \ln(\Gamma_i) > 0$, $u^{(n)} \to u^\star > 0$ and $v^{(n)} \to v^\star = 1 - u^\star < 1$ where $u^\star > 0$ is the solution of $\sum_i \epsilon_i (\Gamma_i - D_i\ u^\star) \ln(\Gamma_i - D_i\ u^\star) = 0$; see Fig. 2(a). It implies that

$$\frac{\langle N \rangle_d^{(n)}}{\langle N \rangle_1^{(n)}} \to \frac{1 - (1 - u^\star)^d}{u^\star} \sim \frac{1}{u^\star} \ll d \tag{9}$$

[1] Noting that $m(u) = \sum_i \epsilon_i (\Gamma_i - D_i\ u) \ln(\Gamma_i - D_i\ u)$ is convex [i.e., $m''(u) = \sum_i \epsilon_i D_i^2 / (\Gamma_i - D_i u) > 0$] and in nonlinear regimes $m(0) = \sum_i \epsilon_i \Gamma_i \ln \Gamma_i > 0$, $m(u^\star) = 0$ and $m(1) = \sum_i \epsilon_i (\Gamma_i - D_i) \ln(\Gamma_i - D_i) < 0$ [as $\Gamma_i - D_i = 1 - A_i(0) < 1$], we deduce that $m'(u^\star) < 0$.

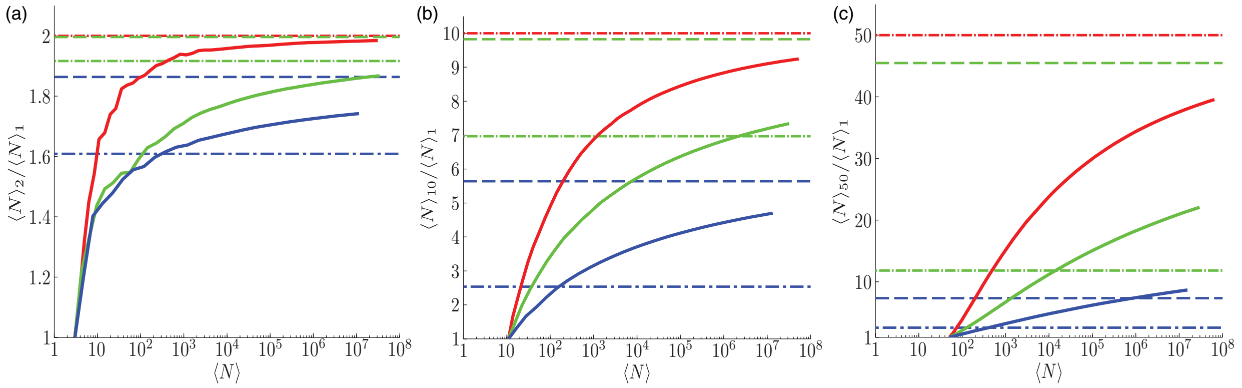

FIG. 4. (Color online) Heterogeneous expansion of subnetworks. Numerical expansion of subnetworks $\langle N\rangle_d^{(n)}$ and (a) $d = 2$, (b) $d = 10$, and (c) $d = 50$. Same evolutionary parameters as in Fig. 3: $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$, $\gamma = \gamma_{on} = \gamma_{no} = 0.2$ (red upper curves), 0.35 (green middle curves), 0.5 (blue lower curves). For $\gamma = 0.2$ (i.e., linear regimes, red upper curves) the network expansion is homogeneous, $\langle N\rangle_d^{(n)}/\langle N\rangle_1^{(n)} \to d$ (red dashed lines, top) with a slower convergence for large initial degree, $d = 50$ (c). For $\gamma > 0.32$ (i.e., nonlinear regimes, green middle curves and blue lower curves), the network expansion is more homogeneous, $\langle N\rangle_d^{(n)}/\langle N\rangle_1^{(n)} \ll d$ (red dotted-dashed lines, top) for large initial degree, $d \gg 1$. Mean-field estimates of asymptotic expansion $\langle N\rangle_d^{(n)}/\langle N\rangle_1^{(n)} \to (1-(1-u^\star)^d)/u^\star$ (dotted-dashed lines) $\sim (1-(1-u_m)^d)/u_m$ (dashed lines), where $u_m = (\Gamma - \Delta_{\min})/D$ corresponds to $\Delta = \Delta_{\min} = \Gamma - D\,u_m = h(\alpha_m)$ (Fig. 3).

for $d \gg -1/\ln(1-u^\star)$, where $\langle N\rangle_1^{(n)} = \prod_{k=0}^{n-1}\Delta^{(k)}$. This corresponds to a more homogeneous expansion across gene families, which exhibit distinct and diverging average connectivities and lead altogether to a long-tailed scale-free degree distribution $p_k \propto k^{-1-\alpha_m}$ with $0 < \alpha_m < 1$ and $\bar{k}^{(n)} \to \infty$ as for the case of $d = 1$. Such nonlinear expansion regimes, converging toward small exponent power-law degree distributions, are typically observed for the out-degree distributions of transcription networks and signal transduction networks [16].

This more homogeneous expansion of simple subnetworks in nonlinear regimes is in good agreement with the corresponding numerical simulations (Fig. 4) (blue curves), showing that $\langle N\rangle_d^{(n)}/\langle N\rangle_1^{(n)} \ll d$ (red dash lines) for large initial degree, $d \gg 1$.

Hence, in both evolutionary regimes, nodes with high connectivity, $d \gg 1$, lead to large local expansion rates, $\Delta_d^{(k)} \simeq 1 + q$, and, ultimately, to variable expansions of gene families in relation to the emergent properties of their biomolecular subnetworks.

Yet, in spite of these larger local expansion rates of highly connected nodes ($\Delta_d^{(k)} \simeq 1+q$, $d \gg 1$), we find that the global expansion of the overall network, $\langle N\rangle^{(n)} = \sum_{d\geqslant 1}\langle N\rangle_d^{(n)} p_d$ (with $p_d \propto 1/d^{1+\alpha}$ for $d \gg 1$), remains, in fact, dominated by the slower expansion of the more numerous low-connectivity nodes, in both linear ($\langle N\rangle_d^{(n)} = \langle N\rangle_1^{(n)} d$, $\alpha = \alpha^\star > 1$) and nonlinear ($\langle N\rangle_d^{(n)} = \langle N\rangle_1^{(n)}/u^\star$, $0 < \alpha = \alpha_m < 1$) regimes.

IV. SIZE VARIANCE OF EXPANDING NETWORKS

We now show that the logistic map analysis can also be used to estimate the size variance of expanding (sub)networks in the mean-field approximation. Recalling $\langle N\rangle^{(n+1)} = N_o \prod_{k=0}^{n}(\sum_{i_k}\epsilon_{i_k}\Delta_{i_k}^{(k)})$, we can estimate $\langle N(N-1)\rangle^{(n+1)}$, by summing over all possible pairs of genes after n duplications of a random fraction q of extant genes, using a logistic map composition analysis including duplication splitting events (Fig. 5).

This yields the following recurrence for $\langle N(N-1)\rangle^{(n)}$:

$$\begin{aligned}\langle N(N-1)\rangle^{(n+1)} \simeq& \left[\prod_{k=0}^{n}\left(\sum_{i_k}\epsilon_{i_k}\Delta_{i_k}^{(k)}\right)\right]^2 N_o(N_o-1)\\ &+\sum_{\ell=0}^{n}\left\{\prod_{k=0}^{\ell-1}\left(\sum_{i_k}\epsilon_{i_k}\Delta_{i_k}^{(k)}\right)2q\;\Delta_o^{(\ell)}\Delta_n^{(\ell)}\right.\\ &\left.\times\left[\prod_{k=\ell+1}^{n}\left(\sum_{i_k}\epsilon_{i_k}\Delta_{i_k}^{(k)}\right)\right]^2\right\}N_o\end{aligned}$$

$$\langle N(N-1)\rangle^{(n+1)} \simeq \Delta^{(n)2}\langle N(N-1)\rangle^{(n)} + 2q\;\Delta_o^{(n)}\Delta_n^{(n)}\langle N\rangle^{(n)},$$

(a) Compositions of Logistic Maps (b) Network 2nd Moment

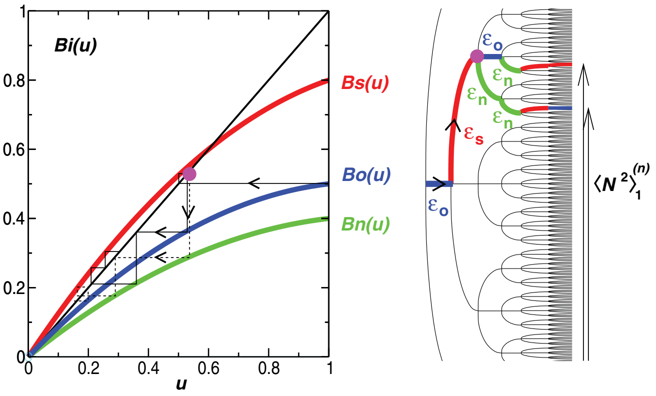

FIG. 5. (Color online) Logistic map analysis of network expansion second moment. (a) Example of stochastic composition of *pairs* of logistic maps, $B_i(u) = \Gamma_i u - D_i u^2$, $i = s,o,n$, with an early joint duplication-divergence history followed by a duplication splitting event ('o'/'n') and further disjoint duplication-divergence events for each gene duplicate. Here $\Gamma_s = 1.3$, $D_s = 0.5$, $\Gamma_o = 0.9$, $D_o = 0.4$, $\Gamma_n = 0.7$, $D_n = 0.3$, $\epsilon_o = \epsilon_n = 1 - \epsilon_s = 0.3$. (b) Corresponding evolutionary history of a specific pair of genes and resulting network second moment $\langle N^2\rangle_1^{(n)}$ obtained by summing over the partially coupled histories of all possible gene pairs.

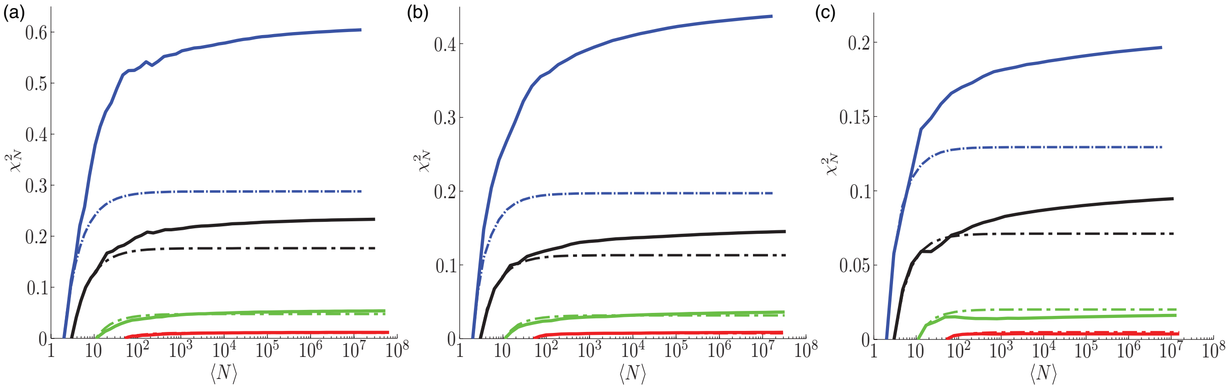

FIG. 6. (Color online) Expansion variance χ_N^2 of subnetworks. Numerical expansion variance χ_N^2 of subnetworks $\langle N \rangle_d^{(n)}$ in linear (a) and nonlinear (b and c) expansion regimes (continuous lines). Same evolutionary parameters as in Figs. 3 and 4: $q = 1$, $\gamma_{oo} = 1$, $\gamma_{nn} = 0$, $\gamma = \gamma_{on} = \gamma_{no} = 0.2$ (a), 0.35 (b), 0.5 (c), and for initial degree connectivity $d = 1$ (blue curves, top), 2 (black curves, upper middle), 10 (green curves, lower middle), and 50 (red curves, bottom). Numerical integration of ${\chi_N^2}^{(n)}$ from Eq. (10) is also plotted for the corresponding regimes (dotted-dashed lines).

where we have assumed that $\Delta_{i_n}^{(n)}$ becomes eventually independent of any particular duplication-divergence series $\{i_k\} = s,o,n$, in the mean-field approximation. This leads to the following recurrences for the second moment $\langle N^2 \rangle^{(n)}$ and variance ${\chi_N^2}^{(n)} = (\langle N^2 \rangle - \langle N \rangle^2)^{(n)} / \langle N \rangle^{2(n)}$,

$$\begin{aligned}\langle N^2 \rangle^{(n+1)} &\simeq \Delta^{(n)^2} \langle N^2 \rangle^{(n)} \\ &\quad + \big[2q\ \Delta_o^{(n)} \Delta_n^{(n)} + \Delta^{(n)}(1 - \Delta^{(n)})\big] \langle N \rangle^{(n)} \\ {\chi_N^2}^{(n+1)} &\simeq {\chi_N^2}^{(n)} + \frac{2q\ \Delta_o^{(n)} \Delta_n^{(n)} + \Delta^{(n)}(1 - \Delta^{(n)})}{\langle N \rangle^{(n)} \Delta^{(n)^2}},\end{aligned}$$

which yields

$${\chi_N^2}^{(n)} \simeq \frac{1}{N_o} \sum_{\ell=0}^{n-1} \frac{2q\ \Delta_o^{(\ell)} \Delta_n^{(\ell)} + \Delta^{(\ell)}(1 - \Delta^{(\ell)})}{\Delta^{(\ell)} \prod_{k=0}^{\ell} \Delta^{(k)}}. \qquad (10)$$

Hence, in the GDD model with $\Delta^{(n)} \to \Delta > 1$ and $\Delta_{o,n}^{(n)} \to \Delta_{o,n}$ one finds, that ${\chi_N^2}^{(n)}$ is always *finite*, in the asymptotic limit:

$${\chi_N^2}^{(n)} \to C_1 + C_2 \frac{2q \Delta_o \Delta_n + \Delta(1 - \Delta)}{N_o \Delta(\Delta - 1)} < \infty.$$

This means that, the variance $(\langle N^2 \rangle - \langle N \rangle^2)^{(n)}$ is typically of the same order as $\langle N \rangle^{2(n)}$, implying that the fluctuations remain of the same order as the means, thereby justifying *a posteriori* the validity of the ensemble average approach to model the emergent properties of expanding biomolecular networks ($\Delta^{(n)} > 1$) under duplication-divergent evolution.

This finite asymptotic variance, ${\chi_N^2}^{(n)} < \infty$, of biomolecular subnetworks under duplication-divergence expansion is in good agreement with the corresponding numerical simulations, both in linear [Fig. 6(a)] and in nonlinear [Figs. 6(b) and 6(c)] regimes.

Note, however, that in the asymptotic limit of a vanishing expansion rate, i.e., $\Delta^{(n)} \to \Delta \sim 1$, ${\chi_N^2}^{(n)}$, is found to diverge as, ${\chi_N^2}^{(n)} \sim 1/(\Delta^{(n)} - 1)$. In particular, the ensemble average framework of the model becomes inappropriate to describe the evolutionary dynamics of fixed-sized networks with $\Delta^{(n)} = 1$ and $\langle N \rangle^{(n)} = N_o$. While eukaryote genomes and biomolecular networks do not appear to be under such size constraints, as they span nearly 10^5 folds in size [3], biomolecular networks of prokaryotes have been suggested to be under simultaneous duplication-divergence evolution and genome size constraint, which was shown to imply a need for horizontal gene transfer [3].

In this limit of fixed-sized networks with $\Delta^{(n)} = 1$ and $\langle N \rangle^{(n)} = N_o$, ${\chi_N^2}^{(n)}$ is found to diverge as ${\chi_N^2}^{(n)} \sim n$, indicating that the GDD model does not correspond to a collection of networks of fixed size N_o, but instead, to the unbounded expansion of a few atypical networks exactly compensated by the vanishing of most other network instances, resulting, overall, in a constant mean size N_o for the ensemble average of the network collection. This failure of the ensemble average framework in the limit of nonexpanding biomolecular networks ($\Delta^{(n)} = 1$) is reminiscent of the emergence of similar spontaneous heterogeneities in population dynamics models under population size constraints [17].

In summary, we have studied analytically the expansions of gene families under duplication-divergence evolution and analyzed the emergent properties of their biomolecular subnetworks in terms of logistic map compositions. In particular, we showed that, beyond explicit distinctions between gene families, their heterogeneous expansion arises in fact spontaneously under duplication-divergence evolution and is directly coupled to their subnetwork connectivity.

ACKNOWLEDGMENT

We thank Christophe Soulé, Tsvi Tlusty, Philippe Marcq, Thomas Risler, and Luis Dinis for discussion.

[1] E. van Nimwegen, Trends Genet. **19**, 479 (2003).

[2] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen, Proc. Natl. Acad. Sci. USA **106**, 9743 (2009).

[3] H. Isambert and R. R. Stein, Biol. Direct **4**, 28 (2009).

[4] S. Ohno, *Evolution by Gene Duplication* (Springer, New York, 1970).

[5] A. Raval, Phys. Rev. E **68**, 066119 (2003).

[6] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, ComPlexUs **1**, 38 (2003).

[7] R. Pastor-Satorras, E. Smith, and R. V. Solé, J. Theor. Biol. **222**, 199 (2003).

[8] J. Berg, M. Lässig, and A. Wagner, BMC Evol. Biol. **4**, 51 (2004).

[9] I. Ispolatov, P. L. Krapivsky, and A. Yuryev, Phys. Rev. E **71**, 061911 (2005).

[10] D. V. Foster, S. A. Kauffman, and J. E. S. Socolar, Phys. Rev. E **73**, 31912 (2006).

[11] J. Enemark and K. Sneppen, J. Stat. Mech. (2007) P11007.

[12] K. Evlampiev and H. Isambert, BMC Syst. Biol. **1**, 49 (2007).

[13] R. V. Solé and S. Valverde, J. R. Soc. Interface **5**, 129 (2008).

[14] K. Evlampiev and H. Isambert, Proc. Natl. Acad. Sci. USA **105**, 9863 (2008).

[15] J. M. Gutierrez, A. Iglesias, and M. A. Rodriguez, Phys. Rev. E **48**, 2507 (1993).

[16] X. Zhu, M. Gerstein, and M. Snyder, Genes Dev. **21**, 1010 (2007).

[17] B. Houchmandzadeh, Phys. Rev. E **66**, 052902 (2002).