蛋白质相互作用网络特征的理论再现*

万茜[1)] 周进[1)] 刘曾荣[2)†]

1) (上海大学上海市应用数学与力学研究所, 上海 200072)

2) (上海大学系统生物技术研究所, 上海 200444)

(2010 年 3 月 20 日收到; 2011 年 3 月 29 日收到修改稿)

无标度性、小世界性、功能模块结构及度负关联性是大量生物网络共同的特征. 为了理解生物网络无标度性、小世界性和度负关联性的形成机制, 研究者已经提出了各种各样基于复制和变异的网络增长模型. 在本文中, 我们从生物学的角度通过引入偏爱小复制原则及变异和非均匀的异源二聚作用构建了一个简单的蛋白质相互作用网络演化模型. 数值模拟结果表明, 该演化模型几乎可以再现现在实测结果所公认的蛋白质相互作用网络的性质: 无标度性、小世界性、度负关联性和功能模块结构. 我们的演化模型对理解蛋白质相互作用网络演化过程中的可能机制提供了一定的帮助.

关键词: 蛋白质相互作用网络, 偏爱小, 非均匀的异源二聚作用, 功能模块结构

PACS: 02.50.–r, 87.15.km

1 引 言

从分子水平上看, 生命现象产生于生物分子之间的相互作用. 生物分子之间的这种相互作用关系可以通过网络来表示, 其中节点表示基因、蛋白质或代谢物, 而边表示蛋白质之间的物理相互作用, 转录调控或代谢反应. 常见的网络有基因调控网络、蛋白质相互作用网络及新陈代谢网络. 正确地构建这些网络是在分子水平上理解生命的关键步骤之一. 因此, 关于生物网络的构建成为了系统生物学的研究热点之一.

细胞各种重要生理过程的实现, 包括信号的传导、对外界环境及内部环境变化的响应等都是以蛋白质之间相互作用为基础. 近年来, 蛋白质相互作用网络已经得到了广泛的研究. 随着酵母双杂交, 基于质谱的串联亲和纯化等高通量实验技术的发展和生物信息学在蛋白质相互作用预测领域的广泛应用, 人们得到越来越多可利用的蛋白质相互作用数据来构建蛋白质相互作用网络 [1−5]. 对这些蛋白质相互作用网络结构特征的研究发现蛋白质相互作用网络具有如下有趣的拓扑性质: (1) 蛋白质相互作用网络是稀疏的, 节点的平均度都比较小; (2) 蛋白质相互作用网络是小世界网络, 具有较短的平均路径长度及与纯粹的随机图相比有较大的平均聚类系数 [4−7]; (3) 蛋白质相互作用网络是无标度网络 [6,8], 也就是说, 这些网络的度分布服从幂律分布 $P(k) \sim k^{-\theta_1}, 2 < \theta_1 < 3$. 这意味着蛋白质相互作用网络是高度异质 (heterogeneous) 的, 即存在大量的拥有少量边的节点和小数目的拥有大量边的枢纽 (hubs) 节点; (4) 蛋白质相互作用网络具有功能模块结构 [9,10], 即度为 k 的节点的平均聚类系数 $C(k)$ 以幂率 $C(k) \propto k^{-\theta_2}$, $\theta_2 \to 1$ 衰减. 这表明拥有少量边的节点具有大的聚类系数属于稠密连接的小子网络, 而枢纽节点具有小的聚类系数连接着不同的子网络 [12], 这样就显示出了功能模块结构; (5) 蛋白质相互作用网络表现出度负关联的性质 [12,13], 其中所有度为 k 的节点的邻居

* 国家自然科学基金 (批准号:10832006, 10972129, 11172158) 和上海市重点学科 (批准号:S30104) 资助的课题.

† E-mail: zrongliu@126.com

的平均度 $K_{nn}(k)$ 遵循 $K_{nn}(k) \sim k^{-\theta_3}$, $\theta_3 \to 0.5$. 因此, 度大与度小节点之间的连接被累积, 而那些度大与度大以及度小与度小节点之间的连接被抑制 [12,13].

为了再现蛋白质相互作用网络的上述拓扑特征, 人们已经提出了各种各样基于复制和变异的网络增长模型. 文献 [7,14] 提出了两个简单的蛋白质演化模型来再现蛋白质相互作用网络的无标度和小世界性质. Chung 等 [15] 分析了节点的完全复制和部分复制的图演化模型, 并指出节点的部分复制可以产生幂指数小于 2 的幂律特征图. 基于基因的复制和新生成基因的重连, Pastor-Satorras 等 [16] 研究了蛋白质相互作用的演化并指出该机制可以再现人们熟知的生物的蛋白质组的统计特征. Ispolatov 等 [17] 提出了一个单参数的复制变异网络模型来描述蛋白质相互作用网络的演化, 并考虑了当参数变化时网络的自平均 (self-average) 性质. 另外, Ispolatov 等 [18] 还研究了网络中团 (clique) 的分布. Liu 等 [19,20] 讨论了复制和变异对生物网络的度负关联性的影响. 然而, 据我们所知, 所有上述的理论模型都没有说明网络的功能模块结构. 最近, Takemoto 等 [21] 完全从理论上提出了一种通过合并全连通子图来演化网络模型, 并表明该演化模型可以生成幂律度分布和功能模块结构. 进一步地, 他们对该模型进行了改进, 用适应性驱使的择优连接 (preferential attachment) 来选择 n 个节点, 结果表明改进的模型生成的网络除了具有幂律度分布和功能模块结构外, 还具有度负关联性质 [22]. 通过对上述现有工作的分析, 我们发现现有的蛋白质相互作用网络演化模型绝大部分都不能全面地反映网络的无标度性、小世界性、功能模块结构及度负关联性的拓扑性质, 一般都是局部反映其中的某些性质. 文献 [22] 给出的演化网络模型基本具有了所有结构特点, 但为此作者提出了以全连通子图作为一个复制单元的机制, 这种复制机制难以与生物学过程对应起来, 因而生物含义难以解释.

在这样的基础上, 我们从生物学角度提出了一个基于生物进化自然选择的更合理的蛋白质相互作用网络演化模型. 从一个单连通的随机图开始, 通过引入偏爱小复制原则及变异和非均匀的异源二聚 (heterodimerization) 作用, 构建了一个简单的蛋白质相互作用网络演化模型. 数值分析表明, 该演化模型除了可以再现真实的蛋白质相互作用网络的一系列, 诸如无标度性、小世界性和度负关联性的拓扑性质外, 还能表现出功能模块结构. 这个结果对理解蛋白质相互作用网络演化过程的可能机制提供了一定的帮助.

2 网络增长模型

从自然选择和生物进化论的角度来看, 复制和变异是生物网络进化的内在基本机制. 复制是生物网络增长的最主要来源, 而变异是功能多样性产生的起源. 考虑到蛋白质网络中那些度大的节点, 由于生物功能的保守性, 不易发生变异; 而那些在相互作用周边的组分则容易发生变异 [23], 这样我们可以合理地假设每个节点被选取复制的概率与它的度成反比. 另外, 文献 [24,32] 也报道了蛋白质的度和它们的可复制性之间的负相关性. 因此, 在模型中, 我们采取偏爱小选取复制节点的方案, 度为 k_i 的节点 i 以概率 $P(i) = k_i^{-1}/\sum_j k_j^{-1}$ 被选取复制, 即度小的节点被选取复制的概率更大.

在每个节点复制完成之后, 根据经典的模型 [25], 新生成的节点与被复制的节点拥有完全相同的功能, 随后, 新生成的节点与被复制的节点两者之一或者由于退化 (degenerative) 突变变成非功能性或者获得一些新的功能而被自然选择保留下来. 由于功能的重要性, 蛋白质之间的相互作用是保守的, 它们的进一步改变受到了限制, 然而这种限制对新生成的节点会极大地减弱 [7,16], 也就是说新生成的节点更容易发生变异. 另外, 文献 [16] 指出被复制节点和新生成的节点都发生变异得到的结果和实测的数据是不符合的. 因此假定删边和加边的变异只发生在新产生的节点上. 这个变异过程模拟如下: 对连接到新生成的节点的每一条边, 以概率 α 删除, 且以概率 γ 在新生成的节点与网络中所有的其他节点之间添加新边. 另外, Hase 等指出了异源二聚作用, 即新生成的节点与被复制的节点之间以一定的概率添加一条新边, 这个假设在酵母蛋白质相互作用网络的演化模型中是不满意的, 所以进一步提出了非均匀的异源二聚作用的新模型, 其中异源二聚作用偏爱连接那些共享

很多共同邻居的新生成与被复制的节点对之间, 并表明了该模型可以再现酵母蛋白质相互作用网络的三角形的偏态分布 (skewed distribution). 为此, 在我们的模型中, 我们也假设在新生成和被复制节点之间异源二聚作用的概率 β 正比于它们共同的邻居数目 ($n_{\rm cn}$), 也就是 $\beta = \beta_0 n_{\rm cn}$, 其中 β_0 是常数. 若 $\beta \geqslant 1$, 则取 $\beta = 1$.

综上所述, 我们所考虑的演化模型由下面步骤所确定. 我们从一个具有 N_0 个节点的单连通的随机图开始, 在每一个时间步上执行以下的步骤直到达到所期望的网络规模:

1) 网络中度为 k_i 的节点 i 以概率 $P(i) = k_i^{-1}/\sum_j k_j^{-1}$ 被选取复制, 产生的新节点记为 i', 它连接到 i 的所有邻居上;

2) 对连接到新生成的节点 i' 的每一条边, 以概率 α 删除; 另外, 被复制节点 i 和新生成的节点 i' 以概率 β 连接, 这里 β 正比于它们共同的邻居数目 ($n_{\rm cn}$), 也就是 $\beta = \beta_0 n_{\rm cn}$, 其中 β_0 是常数; 若 $\beta \geqslant 1$, 则取 $\beta = 1$;

3) 新生成的节点 i' 以概率 γ 连接到网络的所有其他节点上.

在每一个时间步结束的时候, 如果产生了孤立节点, 我们把它从网络中删除掉. 从生物的观点来看, 这些孤立节点表示那些丧失了所有生物功能但还没有从基因组中清除的伪基因. 我们忽略非功能性基因经过适当的变异后恢复到功能性基因的概率, 在每一个时间步结束后删除它们, 留下的网络仍是连通的 [26].

3 结果和讨论

下面将直接通过数值模拟来分析上述的演化模型. 不失一般性, 假设初始网络是由 100 个节点组成的平均度为 3.88 单连通随机图, 且要求网络最终的节点数为 4000.

模型中有 3 个自由的参数, 即删边概率 α, 异源二聚作用概率 β 和加边概率 γ, 所以我们先通过可利用的实证数据来粗略估计这 3 个参数的可能的取值. 文献 [12] 中实证数据表明节点数为 3891 的酵母蛋白质相互作用网络的平均度 $< K >$ = 3.74, 平均聚类系数 $< C >$ = 0.066 和平均最短路径长度 $< L >$ = 4.85. 另外, 文献 [6] 估计了酵母蛋白质相互作用网络中加边和删边的概率的比值为 $\gamma/\alpha \leqslant 0.001$. 这些实证数据为我们在下面的数值模拟过程中选取 3 个参数的取值范围提供一定的依据和参考.

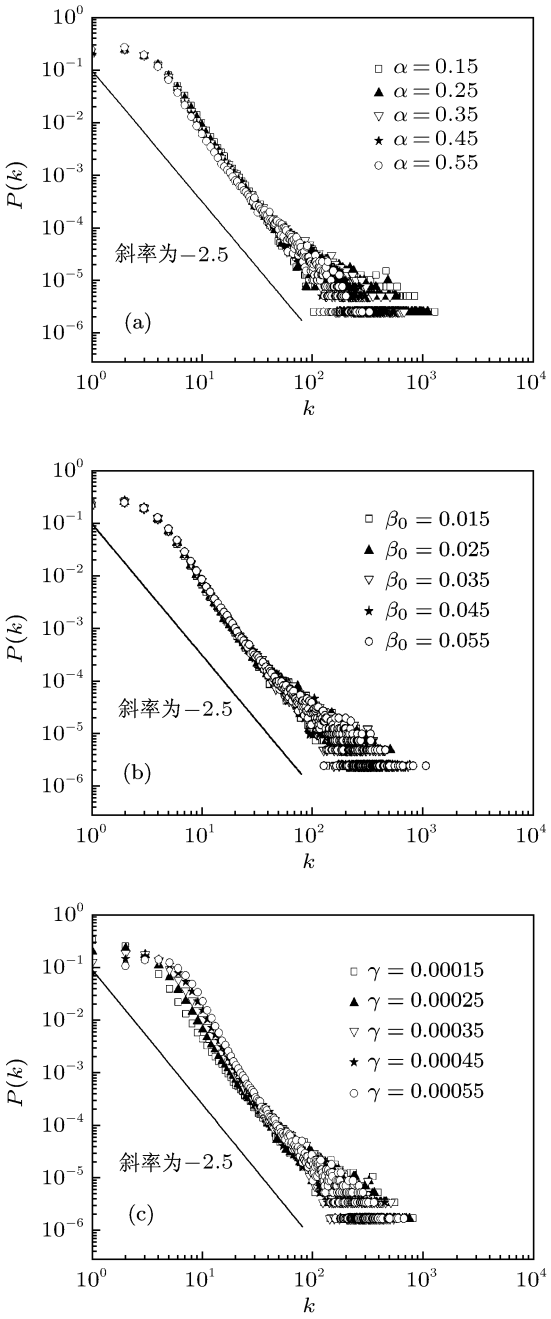

图 1 度分布 $P(k)$ (a) 表示模型在 $\beta_0 = 0.03$, $\gamma = 0.00025$ 和 α 从 0.15 到 0.55 参数下的度分布; (b) 表示模型在 $\alpha = 0.35$, $\gamma = 0.00025$ 和 β_0 从 0.015 到 0.055 参数下的度分布; (c) 表示模型在 $\alpha = 0.35$, $\beta_0 = 0.035$ 和 γ 从 0.00015 到 0.00055 参数下的度分布. 所有的数据都是对 100 次独立的仿真求平均的结果

在下面的数值模拟中, 图中每一数据点都是通

过如下过程取得. 在给定参数后, 取 100 个随机网络作为初始网络, 然后用上述过程生成网络, 再取平均, 得到了数据点. 同时, 我们还对这些样本所得的数据进行了线性回归分析, r 是用来判定一个线性回归直线拟合优度的好坏, p 为结果可信程度的一个递减指标. 以下统计分析将进一步的证实我们数值模拟所得结论是可靠的.

3.1 度分布

度分布是刻画网络的一个重要的统计特性, 定义为 $P(k)=(\sum_i \delta(k_i-k))/N$ 表示任意选取一个节点, 它的度是 k 的概率, 其中 $\delta(\cdot)$ 和 N 分别表示 Kronecker δ 函数和网络的总节点数 [22].

图 1 表示在各种参数条件下网络模型的度分布. 数值模拟结果表明, 在各种参数条件下网络模型的度分布呈现出幂律分布, 反映了网络的无标度特性; 并且可以看出不同演化模型得到的网络的度分布的幂指数都趋近于 2.5, 这和实测的蛋白质相互作用网络中的特性是相符合的. 同时, 我们对数据进行了统计分析, 以图 1 中 $\alpha=0.45, \beta_0=0.03, \gamma=0.00025$ 网络模型所得的数据为例, 把所有度为 k 和相应的 $p(k)$ 两组数据进行线性回归分析得到 $r=0.976$, $p<0.001$, 相应的幂指数为 2.467, 此结果表明数值模拟所得的结论是可靠的, 对其他参数条件下的数据也进行了相同的统计分析, 结果都被证实. 另外, 从图 1(a) 观察到随着 α 的增加, 幂指数是基本保持不变的, 图 1(b) 中也有相同的趋势, 但在图 1(c) 中随着 γ 的增加, 幂指数有轻微减少的趋势. 这表明, 无标度性对于模型参数而言是鲁棒的, 这种现象在生物网络中是相当普遍的.

3.2 聚 - 度系数

大量的真实生物网络具有功能模块结构, 这个特性可以通过所有度为 k 的节点的平均聚类系数 $C(k)$ 与 k 的关系来刻画. 对规则复制有模块结构的网络可以证明有 $C(k)\sim k^{-1}$, 后来从实测生物网络中也发现了类似结果, 因而目前一般用这个统计量来说明网络有功能模块结构的特征 [9]. 所有度为 k 的节点的平均聚类系数 $C(k)$ 定义为 $C(k)=(\sum_i C_i\delta(k_i-k))/(\sum_i \delta(k_i-k))$, 其中 $C_i=2e_i/(k_i(k_i-1))$ 是节点 i 的聚类系数, 度量了网络中节点的邻居的凝聚力, k_i 是节点 i 的度, e_i 是节点 i 的 k_i 个邻居之间的相连的边数 [27].

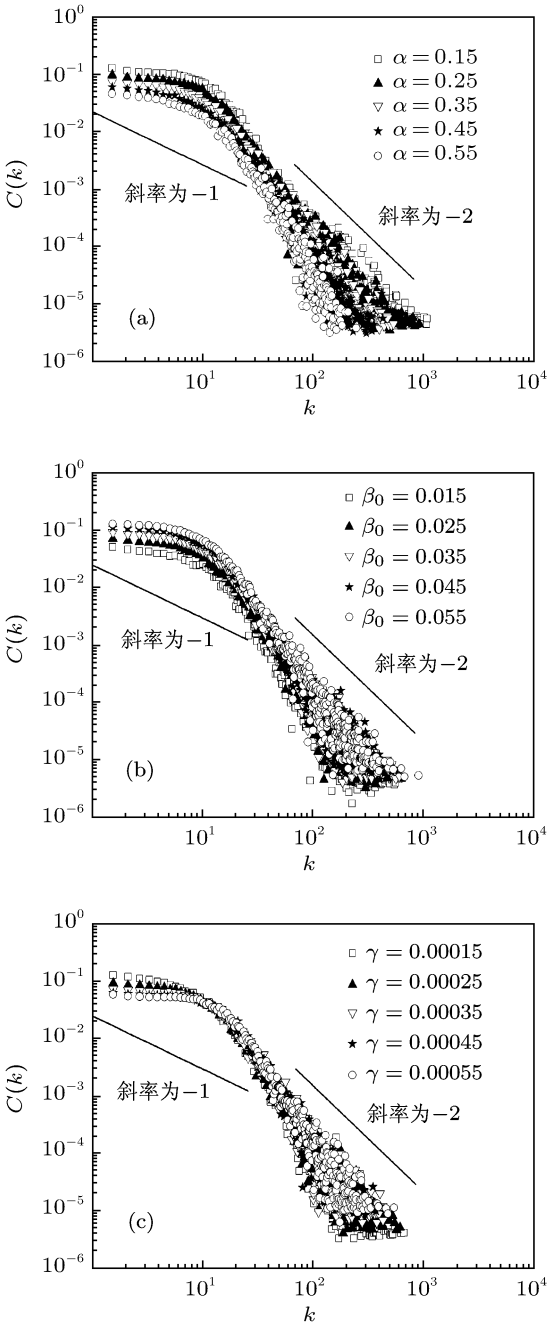

图 2 $C(k)$ 分布 (a) 表示模型在 $\beta_0=0.03$, $\gamma=0.00025$ 和 α 从 0.15 到 0.55 参数下的 $C(k)$ 分布; (b) 表示模型在 $\alpha=0.35$, $\gamma=0.00025$ 和 β_0 从 0.015 到 0.055 参数下的 $C(k)$ 分布; (c) 表示模型在 $\alpha=0.35$, $\beta_0=0.035$ 和 γ 从 0.00015 到 0.00055 参数下的 $C(k)$ 分布. 所有的数据都是对 100 次独立的仿真求平均的结果

实证结果表明, $C(k)$ 是关于 k 的递减函数, 在度大的范围内出现胖尾 (fat tail) 分布. 这种现象表

明度小的节点一般属于内部高度连通的小模块, 而度大的节点的邻居则分别属于不同的模块, 也就是层次模块结构 [28]. 图 2 表示在各种参数条件下网络模型的 $C(k)$ 分布. 我们可以看到, $C(k)$ 与 k 在双对数坐标下显示出平头 (flat head) 胖尾布, 这和实测数据是基本相似的, 呈现出了功能模块结构 [9]. 众所周知, 当 $C(k)$ 分布的幂指数分接近 1 时, 模块结构特征最明显. 我们的数值结果表明在 $\alpha = 0.35$, $\beta_0 = 0.035$ 和 $\gamma = 0.00025$ 附近, $C(k)$ 与 k 之间的这种近似关系被保持. 此外, 我们可以观察到 $C(k)$ 分布的幂指数主要依赖于删边概率 α 和异源二聚作用概率 β. 更确切地说, 幂指数随着 α 的增加而减少, 随着 β_0 的增加而增大. 另外, 从图 2 我们可以看到当 k 小时, $C(k) \sim k^{-1}$, 而当 k 大时, $C(k) \sim k^{-2}$. $C(k)$ 的这种性质已经在实测数据中得到了证实, 并给出了解释 [1−3,11,29]. 最后, 我们对数据进行了统计分析, 以图 2 中 $\alpha = 0.45, \beta_0 = 0.03, \gamma = 0.00025$ 网络模型所得的数据为例, 我们取前段, 即 k 小时的数据, 进行线性回归分析得到 r = 0.965, $p < 0.001$, 相应的幂指数为 1.052, k 大时的数据分析结果为 r = 0.856, $p < 0.001$, 相应的幂指数为 2.186, 对其他参数条件下的数据也进行类似的分析, 结果都表明数值模拟所得的结论具有可靠性.

3.3 度 - 度关联

网络的度 - 度关联, 即节点度之间的关系, 是刻画相邻的节点之间连接的趋势. 若度大的节点倾向于连接度大的节点, 网络称为正相关 (assortative) 网络, 而度大的节点倾向于连接度小的节点时, 网络就称为负相关 (disassortative) 网络 [31,32]. 这里我们利用所有度为 k 的节点的平均邻居度 $K_{nn}(k)$ 来度量度 - 度关联, 定义为 $K_{nn}(k) = (\sum_i k_{nn,i}\delta(k_i - k))/(\sum_i \delta(k_i - k))$, 其中 $k_{nn,i} = (\sum_{j \in O(i)} k_j)/k_i$ 是节点 i 邻居的平均邻居度, $O(i)$ 对应于节点 i 的邻居的集合. 如果 $K_{nn}(k)$ 是关于 k 的递增函数, 那么网络是正相关的; 如果是 k 的递减函数, 那么网络是负相关的.

表 1 各种不同参数下网络模型的平均度 $< K >$, 平均聚类系数 $< C >$, 最短路径长 $< L >$ 的值

	$\beta_0 = 0.03 \quad \gamma = 0.00025$				
	$\alpha = 0.15$	$\alpha = 0.25$	$\alpha = 0.35$	$\alpha = 0.45$	$\alpha = 0.55$
$< K >$	4.5699	4.2731	4.0182	3.7757	3.5490
$< C >$	0.1073	0.0810	0.0620	0.0488	0.0368
$< L >$	3.8792	4.2551	4.6610	4.9493	5.3838
	$\alpha = 0.35 \quad \gamma = 0.00025$				
	$\beta_0 = 0.015$	$\beta_0 = 0.025$	$\beta_0 = 0.035$	$\beta_0 = 0.045$	$\beta_0 = 0.055$
$< K >$	3.9339	3.9984	4.0472	4.1060	4.1622
$< C >$	0.0382	0.0535	0.0690	0.0846	0.0947
$< L >$	4.6222	4.6502	4.5982	4.6521	4.6406
	$\alpha = 0.35 \quad \beta_0 = 0.035$				
	$\gamma = 0.00015$	$\gamma = 0.00025$	$\gamma = 0.00035$	$\gamma = 0.00045$	$\gamma = 0.00055$
$< K >$	3.4830	4.0472	4.6663	5.2800	5.9475
$< C >$	0.0801	0.0690	0.0610	0.0550	0.0517
$< L >$	4.7655	4.5982	4.5282	4.3982	4.2716

图 3 表示在各种参数条件下网络模型的 $K_{nn}(k)$ 与 k 的关系. 数值仿真结果表明, 在各种参数下 $K_{nn}(k)$ 都是随 k 递减的, 说明模型所生成的网络可以产生度负关联. 我们可以看到 $K_{nn}(k)$ 分布的幂指数主要依赖于删边概率 α 和加边概率 γ. 更具体地, 幂指数随着 α 的增加而减少, 随着 γ 的增加而稍微减少. 另外, 数值结果表明在 $\alpha = 0.35$, $\beta_0 = 0.035$ 和 $\gamma = 0.00025$ 附近, 当 k

小时, $K_{nn}(k) \sim k^{-0.5}$, 而当 k 大时, $K_{nn}(k) \sim k^{-1}$. $K_{nn}(k)$ 的这种性质已经在实测数据中得到了证实, 并给出了解释 [12,13]. 最后, 我们对数据进行了统计分析, 以图 3 中 $\alpha = 0.45, \beta_0 = 0.03, \gamma = 0.00025$ 网络模型所得的数据为例, 我们取前段, 即 k 小时的数据, 进行线性回归分析得到 $r = 0.998$, $p < 0.001$, 相应的幂指数为 0.460, k 大时的数据分析结果为 $r = 0.933$, $p < 0.001$, 相应的幂指数为 1.551, 对其他参数条件下的数据也进行类似的分析, 结果都表明数值模拟所得的结论是可靠的.

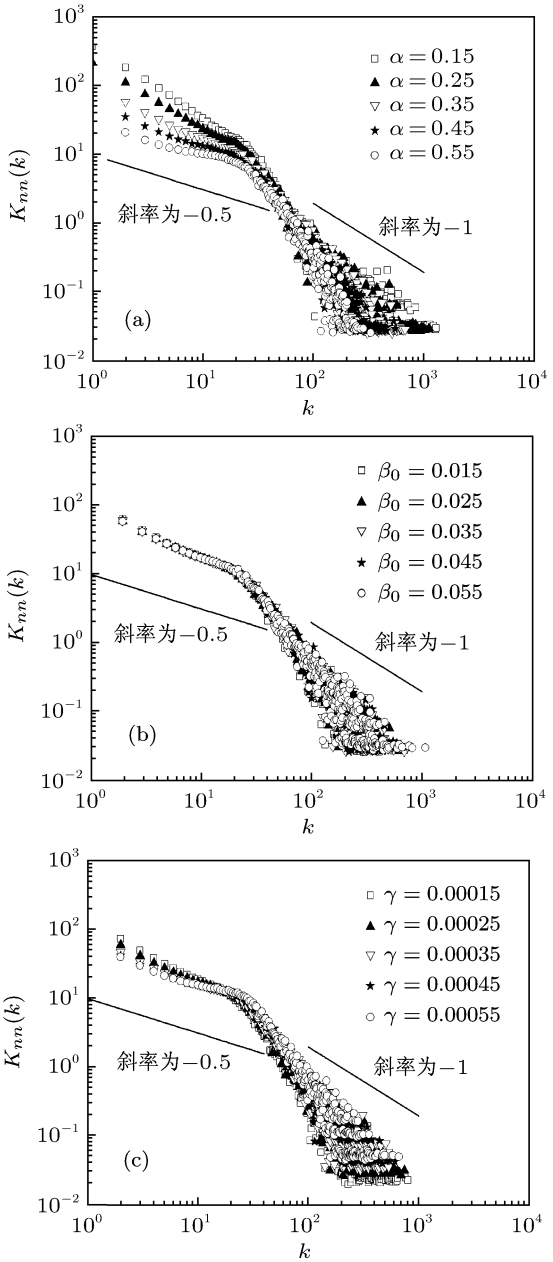

图 3 $K_{nn}(k)$ 分布 (a) 表示模型在 $\beta_0 = 0.03, \gamma = 0.00025$ 和 α 从 0.15 到 0.55 参数下的 $K_{nn}(k)$ 分布; (b) 表示模型在 $\alpha = 0.35$, $\gamma = 0.00025$ 和 β_0 从 0.015 到 0.055 参数下的 $K_{nn}(k)$ 分布; (c) 表示模型在 $\alpha = 0.35$, $\beta_0 = 0.035$ 和 γ 从 0.00015 到 0.00055 参数下的 $K_{nn}(k)$ 分布. 所有的数据都是对 100 次独立的仿真求平均的结果

3.4 稀疏性和小世界性

在各种参数情况下计算了网络模型平均度 $<K>$, 平均聚类系数 $<C>$, 最短路径长度 $<L>$, 表 1 给出了计算结果. 这些结果与文献 [12] 给出的实证数据基本相符合. 同时, 与具有相同大小的随机网络相比, 网络模型的平均聚类系数 $<C>$ 比随机网络的要大得多. 因此, 构建的网络具有稀疏性和小世界性.

4 结 论

从自然选择的进化论出发, 构建了一个蛋白质相互作用网络的演化模型. 在模型中, 引进了反映生物进化过程出现的复制, 聚合和变异的 3 个参数, 并考虑了生物上有证据表明存在的偏爱小的特征. 数值结果表明这个模型能够很好地再现出实测蛋白质相互作用网络的稀疏性、小世界性、无标度性、度负关联性和功能模块结构各方面的统计特征. 与已有工作不同之处是, 不仅比其他工作更全面的反映网络的拓扑结构, 而且在构建模型中, 每一个步骤都有清楚的生物含义和以生物实验作为佐证.

在我们构建的模型中, 反映随机性的 3 个参数和反映确定性的偏爱小原则, 这两者都有明确的生物含义, 我们发现这两者对于生成合理的生物网络似乎都不可缺少. 这个结果可能预示生物进化过程是由确定性和随机性这两个因素决定, 生物系统的形成是二者相互作用达到某种平衡的结果.

另外, 我们详细地分析删边概率 α, 异源二聚作用概率 β 和加边概率 γ 这三个参数对无标度性, 功能模块结构和度负关联性的影响. 结论如下：(1) 我们网络模型的无标度性对模型的参数是鲁棒的, 这种现象在生物中是很普遍的；(2) 功能模块结构主要依赖于删边概率 α 和异源二聚作用概率 β. 更确切地说, 幂指数随着 α 的增加而减少, 随着 β_0 的增加而增大；(3) 度负关联主要依赖于删边概率 α 和加边概率 γ. 更具体地, 幂指数随着 α 的增加而

减少, 随着 γ 的增加而稍微减少. 相对而言, 这些结果仅给出了生物网络进化论机制的初步探索. 我们相信这些结果可能会对生物网络构建的机制研究具有参考价值.

[1] Uetz P, Giot L, Cagney G, Mansfield T A, Judson R S, Knight J R, Lockshon D, Narayan V, Srinvasan M, Pochart P, Emili Q A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg J M 2000 *Nature* **403** 623

[2] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y 2001 *Proc. Natl. Acad. Sci. USA* **98** 4569

[3] Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes H W, Stumpflen V 2006 *Nucleic Acids Res.* **34** 436

[4] Li S, Armstrong C M, Bertin N, Ge H, Milstein S, Boxem M, Vidalain P O, Han J J, Chesneau A, Xu L, Tewari M, Wong S L, Zhang L V, Berriz G F, Jacotot L, Vaglio P, Reboul J, Kishikawa T, Li Q, Gabel H W, Elewa A, Baumgartner B, Rose D J, Yu H, Bosak S, Sequerra R, Fraser A, Mango S, Saxton W M, Strome S, Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Stamm L, Cunsalus K, Harper J W, Cusick M E, Roth F P, Hill D E, Vidal M 2004 *Science* **303** 540

[5] Giot L, Bader J S, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao Y L, Ooi C E, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, Mcdaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Loime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, Dasilva A, Zhong J, Stanyon C A, Finley R L, White K P, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets R A, Mckenna M P, Chant J, Rothberg J M 2003 *Science* **302** 1727

[6] Wanger A 2001 *Mol. Biol. Evol.* **18** 1283

[7] Sole R V, Pastor-Satorras R, Smith E D, Kepler T 2002 *Adv. Comp. Syst.* **5** 43

[8] Jeong H, Mason S P, Barabasi A L, Oltvai Z N 2001 *Nature* **411** 41

[9] Ravasz E, Barabasi A L 2003 *Phys. Rev. E* **67** 026112

[10] Williams R J, Martinez N D, Berlow E L, Dunne J A, Barabasi A L 2002 *Science* **297** 1551

[11] Yook S H, Oltvai Z N, Barabasi A L 2004 *Proteomics* **4** 929

[12] Hase T , Niimura Y, Kaminuma T, Tanaka H 2008 *PloS One* **3** e1667

[13] Maslov S, Sneppen K 2002 *Science* **296** 910

[14] Vazquez A, Flammini A, Maritan A, Vespignani A 2003 *ComPlexUs* **1** 38

[15] Chung F, Lu L, Dewey T G, Galas D J 2003 *J. Comput. Biol.* **10** 677

[16] Pastor-Satorras R, Smith E, Sole R V 2003 *J. Theor. Biol.* **222** 199

[17] Ispolatov I, Krapivsky P L, Yuryev A 2005 *Phys. Rev.* E **71** 061911

[18] Ispolatov I, Krapivsky P L, Yuryev A 2005 *New J. Phys.* **7** 145

[19] Dan Z, Liu Z R, Wang J Z 2007 *Chin. Phys. Lett.* **24** 2766

[20] Xu C S, Liu Z R, Wang R Q 2010 *Physica* A **389** 643

[21] Takemoto K, Oosawa C 2005 *Phys. Rev.* E **71** 046116

[22] Takemoto K, Oosawa C 2007 *Math. Bios.* **208** 454

[23] Kellis M, Patterson N, Endrizzi M, Birren B, Lander E S 2003 *Nature* **423** 241

[24] Prachumwat A, Li W H 2006 *Mol. Biol. Evol.* **23** 30

[25] Ohno S 1970 *Evolution by Gene Duplication* (New York: Springer-Verlag) pp100–110

[26] Evlampiev K, Isambert H 2008 *Proc. Natl. Acad. Sci. USA* **105** 9863

[27] Watts D J, Strogatz S H 1998 *Nature* **393** 440

[28] Zhou Y B, Cai S M, Wang W X, Zhou P L 2009 *Physica* A **388** 999

[29] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701

[30] Costa L F, Rodrigues F A, Vieso G T, Boas P R V 2007 *Adv. Phys.* **56** 167

[31] Farid N, Christensen K 2006 *New J. Phys.* **8** 212

[32] Li L, Huang Y, Xia X, Sun Z 2006 *Mol Biol Evol*. **23** 12

Emergence of features in protein-protein interaction networks *

Wan Xi [1)] Zhou Jin[1)] Liu Zeng-Rong[2)†]

1) (*Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, Shanghai* 200072*, China*)

2) (*Institute of System Biology, Shanghai University, Shanghai* 200444*, China*)

(Received 20 March 2010; revised manuscript received 29 March 2011)

Abstract

Scale-free connectivity, small-world pattern , hierarchical modularity and disassortative mixing are prominent features shared by most biological networks. Up to now, various network growth models invoking gene duplication and divergence have been proposed to understand the evolutionary mechanisms shaping the scale-free connectivity, small-world pattern and disassortative mixing . In this paper, we present an evolutionary model by introducing a rule of small preference duplication of a node and sequently divergence plus non-uniform heterodimerization meaning that the probability that an heterodimerization link is added between duplicated nodes is proportional to the number of common neighbors shared by these nodes based on biological background, showing that our model can almost display series of topological characteristics of real protein interaction networks, such as scale-free connectivity, small-world pattern, disassortativity of degree-degree correlation and hierarchical modularity. Our model may yield relevant insights into the evolutionary mechanism of protein interaction networks behind it.

Keywords: protein interaction networks, small-preference duplication, non-uniform heterodimerization, hierarchical modularity

PACS: 02.50.–r, 87.15.km

∗ Project supported by the National Natural Science Foundation of China (Grant Nos. 10832006, 10972129, 11172158), and the Key Disciplines of Shanghai Municipality, China (Grant No. S30104).

† E-mail: zrongliu@126.com