基于复杂网络模型的基因调控网络的计算模拟

张律文[1,2] 谢 江[1,2] 陈建娇[1] 张 武[1]
(上海大学计算机工程与科学学院 上海 200072)[1] (上海大学系统生物技术研究所 上海 200444)[2]

摘 要 随着大通量基因芯片数据的产生,基因调控机制的网络化研究需求日趋迫切。提出了基于复杂网络理论的基因调控网络的模拟方法,构建了基因调控网络模拟器GN-Simulator。通过分析真实基因调控网络的拓扑特性,提出了对应的矩阵模型,并充分考虑了网络的生物学鲁棒性和动力学稳定性,给出了人工基因网络的生成过程和计算模拟方法。计算实验表明,GN-Simulator能高效地模拟出与真实基因调控网络高度相似的大规模人工网络,并可为不同算法提供无偏验证的多样化人工模拟数据。

关键词 复杂网络,无标度特性,动力学稳定性,模拟,基因调控网络

Artificial Gene Regulatory Networks Construction Based on Complex Network

ZHANG Lu-wen[1,2] XIE Jiang[1,2] CHEN Jian-jiao[1] ZHANG Wu[1]
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)[1]
(Institute of Systems Biology, Shanghai University, Shanghai 200444, China)[2]

Abstract Along with the explosion of high-throughput genomic information produced from such as DNA microarrays, modeling the mechanism of gene regulations and constructing gene regulatory network become an urgent task. We proposed a novel method, GN-Simulator, to simulate gene regulatory networks based on complex network theory. The real gene networks were modeled explicitly in matrixes based on the scale-free topology. According to the robust biological mechanism, we constructed artificial gene regulatory networks related topology to the dynamic stability. We considered the features of real gene network structures into the artificial system. The computing experiments illustrate that GN-Simulator can simulate large-scale gene networks efficiently with high confidence level. Moreover, GN-Simulator generates various synthetic networks for testifying different algorithms performance and provides reasonable estimation on them.

Keywords Complex network, Scale-free topology, Dynamic stability, Simulation, Gene regulatory network

1 引言

复杂生命现象是大量基因互相调控、协同作用的结果。随着基因组测序的完成、基因组学的快速发展,生物学积累了大量的基因调控实验数据,如何挖掘出海量数据所蕴藏的生命现象和生物规律已成为生命科学的研究热点。如今,复杂生物系统的研究开始从对细胞内个别基因功能的局部性描述发展到对复杂基因调控网络(Gene Regulatory Network)的定量刻画。由于基因与其他生物小分子倾向于成组地通过网状的相互作用而影响生物系统的功能,因此对基因功能的研究必须分析其调控网络。

关于基因调控机制的系统性研究,最行之有效的方法就是结合计算机科学和数学等理论把基因之间的复杂相互作用进行整合,并简化成网络,然后建立理论模型,从而研究基因调控网络的结构、网络模块及动力学性质。自2000年Nature上发表了利用复杂网络就理论研究生物网络拓扑特性[1]的研究成果之后,复杂网络被广泛地应用于生物网络,包括基因调控网络的构建和模拟等各方面,使得基因网络的研究取得了极大进展。

用来模拟真实基因调控的复杂网络模型经历了几个阶段的发展。最初,简单的随机网络模型[2]被用来描述基因网络的结构,但在1998年,Watts和Strogats在结合规则网络和随机网络特点的基础上,建立了小世界(Small-world)网络模型[3],接着1999年,Barabase和Albert发现了真实网络的无标度(Scale-free)性质[4]。在对大规模基因网络进行数据采集和统计分析后,无标度被认为是最接近真实基因网络的一种拓扑结构[5],基因相互作用网络的结构特性可能与其他复杂系统网络(比如Internet网)在很大程度上是一致的。

在构建人工基因网络的过程中,仅仅关注网络的大体拓扑结构是远远不够的。为了尽可能地模拟出接近真实基因网络的人工网络,构建了基因调控网络模拟器Gene network simulator(GN-Simulator),从以下几个方面系统地考虑了网

到稿日期:2009-02-20 返修日期:2009-05-08 本文受上海市重点学科建设项目(项目编号:J50103),上海大学系统生物研究基金(2008-5)资助。

张律文 博士生,主要研究方向为高性能计算、生物信息学等,E-mail:zhanglvwen@shu.edu.cn;谢 江 博士,副教授,主要研究方向为高性能计算、生物信息学等;陈建娇 博士生,主要研究方向为生物信息学等;张 武 教授,博士生导师,主要研究方向为高性能计算、生物信息学等。

络特征。

1)网络度分布:在随机一致的原则下挑选出的节点的度数为 k 的概率为 $p(k)$。

2)网络拓扑结构:无标度网络模型虽然最符合现实复杂网络的论断仍有争议,但依然被认为是目前最接近现实网络特性的模型,因此我们的模拟器采用的是无标度模型。

3)人工网络的鲁棒性:模拟出的网络必须符合动力学稳定性。

GN-Simulator 的意义在于,理论上将复杂网络模型与经过大量实验证明的真实基因调控网络特征很好地结合起来,使得模拟网络具有较高的可信度;技术上它为基因调控网络甚至生物网络的模拟提供了有效的一体化解决途径。在更为重要的实际应用中,许多构建基因网络的算法需要得到详细全面的测试,虽然在真实基因网络上可以进行直接而简单的测试实验,但由于已知真实网络的规模有限、类型很少、动力学机制不明等缺点,很容易对特定算法产生偏袒,那么由 GN-Simulator 模拟产生的大规模、多类型、参数可控的高可信基因调控网络无疑成为公平系统地验证算法的金标准(gold standard)。本文将开展构建人工基因调控网络的研究,并对基因调控网络的模拟结果进行动力学稳定性分析,给出其应用于算法评估的步骤。

2 基因调控网络模拟器构建

2.1 人工网络的矩阵表示

采用网络来描述基因调控关系是一种简便有效的方法,即把基因简化为节点(vertex),其调控作用简化为节点之间的连线(edge),基因调控网络就是由节点集合 V 及其节点间边的集合 E 构成的:

$$Network=(V,E)$$

由于邻接矩阵可以用来描述一个网络中节点与节点之间的关系,因此网络的拓扑结构用邻接矩阵 A 表示:

$$A_{N\times N}=\begin{bmatrix} 0 & a_{11} & \cdots & a_{1N} \\ a_{21} & 0 & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & 0 \end{bmatrix}$$

其中,A 中的每一个元素 a_{ij} 代表基因之间的调控关系。考虑到真实调控网络中每个基因的状态改变主要取决于其他基因对它的作用,为了简化分析,基因自身的自调控可以被忽略,即 A 中对角元素均为 0。基因之间的调控关系有两种:激活和抑制,真实调控网络中基因之间的抑制作用通常强于激活作用,且 a_{ij} 的具体值并不影响网络的整体动力学性质,因此以 $a_{ij}=-\infty$ 表示抑制作用,$a_{ij}=1$ 表示激活作用。

此外,在网络的存储方面,由于真实的生物网络大多是稀疏网络,如芽殖酵母的蛋白质相互作用网络的平度连结度仅约为 2[6],在储存对应的稀疏矩阵时,如果直接用邻接矩阵表示法,对于一个有 n 个节点的网络,将占用 n^2 的存储空间,当模拟大规模的全局基因网络时,将产生过高的空间复杂度;因此,采用边列表示法建立了一个以边为元素的数组来记录每条边的出节点和入节点,从而使得整个网络需要占用的存储空间下降为 $2n$。

2.2 给定度分布及基于无标度拓扑模型的网络构建

为了定义一个基因网络模型,必须建立网络的拓扑结构,或是连线图。这包括两方面的工作:一是决定网络中节点(基因)的数量,二是确定对每个节点产生激活或抑制效应的是哪些节点,即每个节点的出入度。在对 43 个物种的基因网络的度分布进行统计分析[7]后发现,这些网络的入度和出度分布均服从幂率分布:$p(k)\sim k^{-r}$,幂指数 r 约为 2.2,为基因网络的无标度拓扑特性提供了又一有力证据。同时,研究表明,当网络的平均连接度在 2~3 时,网络中的大部分具有生物学意义的动力学行为在此区间产生[8]。如果仅关注某一执行具体功能的子模块,其平均连接度相对高于整个网络的平均连接度,约在2~4之间。因此,在构建基因网络模拟器(GN-Simulator)时,平均连接度的默认区间为 $k=2\sim4$。另外,人工基因网络的构建也考虑了无标度网络演化的 BA 模型(Barabase & Albert model)的内在机制:网络规模的增长和择优连结。根据这种无标度特性,网络不断有新节点加入,且倾向于与度数较高的节点连接,那么,网络中连接度高的节点往往是网络形成过程中较早加入的节点。基于真实基因网络以上特性的分析,模拟真实基因网络的过程概括如下:

步骤 1 基因网络初始状态:m_0 个节点,e_0 条边,m_0 个节点被随机分为 l 个子集,那么每个子集中各有 $m_{01},m_{02},\cdots,m_{0l}$ 个节点,$e_{01},e_{02},\cdots,e_{0l}$ 条边。

步骤 2 增加新节点。一个新节点首先被随机分到一个子集中,随着新节点的加入将产生 m 条入边或 m 条出边,其中 $m\leqslant \mathrm{Min}(m_{01},m_{02},\cdots,m_{0l})$。

步骤 3 产生新的入边。当新节点以概率 p 产生 m 条入边时,系统首先随机为新节点在初始网络中选择一个子集,然后在该子集中偏好选取 $m-1$ 个节点与之相连。原则为:择优选择连接节点的概率取决于该节点的入度。与此同时新节点的另一条边也在择优连接的原则下与另一个子集的节点相连。

步骤 4 产生新的出边。当新节点以概率 q 产生 m 条出边时,也随机为新节点在初始网络中选择一个子集,并在该子集中偏好选取 $m-1$ 个节点与之相连。原则为:择优选择连接节点的概率取决于该节点的出度。同时新节点的另一条边也在择优连接的原则下与另一个子集的节点相连。

步骤 5 重复步骤 2—4,同时统计节点数,直到网络增长到需要的规模为止。

通过以上步骤,初步构建了 GN-Simulator。

2.3 人工网络的鲁棒性和动力学稳定性

生物系统和调控网络的基本性质是鲁棒性和稳定性[9],这两种重要的性质有利于生物应对复杂多变的外界环境和不断受到扰动的内部环境,甚至在某些基因缺失的情况下,整体的生物状态和重要的基本功能还能够保持稳定。因此,如何构建出具有动力学稳定性的人工网络是一项重要的研究内容。我们必须首先清楚地分析生物调控网络是如何实现鲁棒性和稳定性的。一方面通过网络的结构性质[10],例如基因的冗余性、功能模块化、网络中的负反馈机制,可以实现稳定性。另一方面,生物系统可以通过网络的整体结构和动力学性质来实现鲁棒性[11]。这种由网络的整体性质产生的鲁棒性是本文关注的重点。

在建立好基因调控网络的理论模型后,将进一步研究其网络的动力学稳定性,包括状态稳定性和结构稳定性。状态稳定性指研究网络中不同节点所代表的基因数量发生变化

时，网络所执行的生物学过程能否继续。结构稳定性指网络中不同节点的具体数值所代表的基因相互作用的强度发生改变时生物学状态的稳定性。

在模拟基因调控网络的过程中，为了构建具有动力学稳定性的人工网络，我们进行了具有全局稳定不动点的骨架网络的设计。在一个网络中，具有最少的边数且对整个网络的整体动力学性质起决定性作用的一个连通子集被称为该网络的骨架网。为了保证网络的整体动力学性质，这个特殊的子网是不能随意改变的。在网络的除骨架网之外的拓扑结构上进行随机扰动，例如，随机删除或增加一条边，抑或改变一条边的属性，如果网络仍然定性地保持其整体动力学性质不变，就成功构建了一个拓扑结构上具有鲁棒性的人工网络[12]。

此外，为验证模拟网络的稳定性，可以计算其相关矩阵的特征根，如果所有特征根实部为负或最大特征根实部为负，系统必将趋于稳定。同时，最大负特征根的绝对值越大，系统在经历微小的扰动后趋于稳定的速度将越快。

3 数值试验及应用分析

如前所述，GN-simulator 建立的一个重要目的在于为全面评估基因调控关系的预测算法提供“金标准”人工网络。首先，在应用于算法评估之前我们验证了人工基因调控网络的动力学稳定性。基于 GN-simulator，本文模拟产生了 10 个有着不同平均连接度的基因调控网络样本，其平均连接度 $k=3.9$，抑制作用关系与激活作用关系之比 $r=0.7$，网络规模为节点数 $n=100$，网络拓扑为具有小世界特性的无标度结构。图 1 给出了一个人工网络的动力学稳定性示例，在加给系统一个微小的扰动后，可以看出整个系统在 10 个时间步内快速收敛并达到平衡状态。

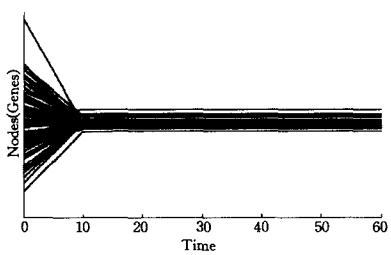

图 1 微扰后网络中所有节点随时间收敛图

同时，从著名的基因调控和基因组数据库 KEGG[13] 及 NCBI[14] 数据库中收集了人类转录因子调控网络、酵母基因调控网络和蛋白质组网络，将它们的网络拓扑性质与 GN-Simulator 模拟的网络进行对比，如表 1 所列，本文模拟的基因调控网络与真实网络相比具有高相似性。

表 1 拓扑性质比较

	GN-Simulator	HTFN	Yeast	Proteome
Average degree	3	3.7	3.7	2.4
Average clustering	0.13	0.17	0.15	0.7
Average path length	4.7	4.5	4.15	6.81
Assortative mixing	−0.16	−0.18	−0.05	−0.15

将 GN-Simulator 产生的并经过动力学稳定性测试和相似性实验证明的人工基因调控网络用于算法评估，以比较在不同的网络参数下各个预测算法的整体表现，从而公平地反映出算法的优缺点以及对于特定网络的预测偏好。应用 GN-Simulator 产生的人工基因调控网络进行算法评估的步骤如下。

1）网络参数设置：根据用户选择，生成 n 个具有不同网络规模、平均度（average-degree）、平均路径长度（average path length）、群聚系数（clustering coefficient）和介数（vertex betweenness and edge betweenness）的人工基因调控网络。

2）网络模拟：根据每个网络生成相应的邻接矩阵，记为 $A_1, A_2, \cdots, A_n$，以稀疏矩阵的形式表示，并以边列表的形式储存。

3）稳定性检验：随机选取矩阵中一定数量的节点，将其数值增加 0.1（保证扰动足够微小，以免使整个系统偏离平衡点过大），检测整体网络在一定时间步内的收敛性，如果不符合动力学稳定性，可进行修正：将对角元素减去矩阵的最大特征值，这样既不会改变网络的拓扑结构和度分布等特性，同时又保证了系统的稳定。

4）结果比对：用户以矩阵形式提交其算法预测出来的基因调控网络，记为 $A_1', A_2', \cdots, A_n'$，将 A 与 A' 进行相似性比对，给出正确率、召回率和覆盖率等统计性比较结果。表 2 给出了不同网络参数下几种典型算法的表现情况。

表 2 不同人工网络测试样本下的算法表现

Algorithms	Coverage(Number of nodes, average degree)				
	(50,2)	(200,2)	(500,2)	(500,3)	(500,4)
SWNI[15]	65%	95%	97%	91%	85%
ARACNE	84%	77%	63%	64%	68%
BANJO	76%	72%	58%	58%	59%
Clustering	63%	55%	53%	57%	60%

从表 2 可以看出，不同的算法均对网络规模和网络平均度敏感，但敏感趋势和敏感度不同。SWNI 在大规模稀疏网络的预测方面相对其它算法表现较优，但其预测覆盖率却随着网络稠密化而降低。另外，可以明显看出 ARACNE 等 3 种算法的预测覆盖率随着网络规模的增大而下降，但似乎在对较稠密网络的预测中有较好的表现。

结束语 本文提出了专门针对基因调控网络的人工网络建立模型。与其他的复杂网络分析软件 Pajek，Netdraw 和 Ucient 等比较，GN-Simulator 具有很强的针对性和生物信息学应用性，为系统生物领域的研究人员提供了基因网络模拟工具。数值试验结果显示，GN-Simulator 产生的网络与真实基因调控网络相比，在整体结构和动力学性质上有很高的相似性，能弥补真实网络规模不足等缺点，为各种基因调控关系预测算法提供了大量多种类的测试网络数据集；同时，它集成了算法评估、比较和分析功能，对算法在不同网络数据集上的表现进行打分，为生物网络结构及其预测算法提供了一个高效、合理的计算实验平台。

参 考 文 献

[1] Jeong H, Tombor B, Albert R, et al. The large-scale organization of metabolic networks[J]. Nature, 2000, 407: 651-654

[2] Erdös P, Rényi A. On random graphs[J]. Publ Math Debrecen, 1959, 6: 290-297

[3] Strogatz S H. Exploring complex networks[J]. Nature, 2001, 410: 268-276

[4] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286: 509-512

（下转第 221 页）

$0, x_i \in X_{II}$，X_{III}最优解为 $x_i=-1, x_i \in X_{III}$。正如尚荣华等人认为[6]，该问题的理论 Pareto 面满足 $f_2 = 1 - f_1^{(H(t)+15*(1+H(t))^2)^{-1}}$，而不是由凸变为非凸。为了观察方便，将目标函数值 f_1 和 f_2 都平移 $0.5t$。在该测试问题中，$|X_I|=1, |X_{II}|=|X_{III}|=15$。

从图 4(a)、图 4(b)和图 4(c)可知算法 DMEIA 所搜索到的最优解与其他两种算法相比分布更均匀。DNSGAII-A 好于 DBM，但在 f_2 接近 1 时很难找到最优解。从图 4(d)可知 DMEIA 相对其他两种算法其收敛的最小值不如其他两种算法小，但 DMEIA 随时间变化波动较小，相对平稳。

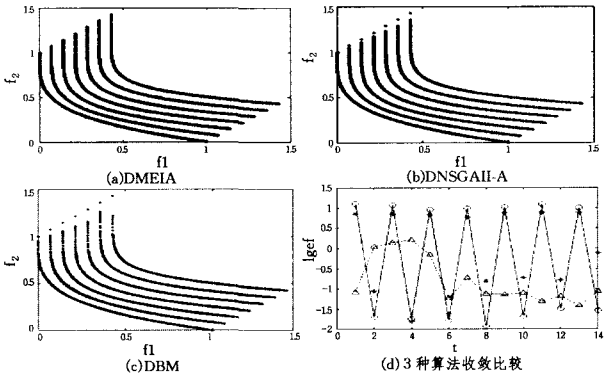

图 4　DMOP2 的 3 种算法的仿真结果图

5.2.3　DMOP3

DMOP3 是第 2 种类型，且该问题的理论 Pareto 面满足 $f_2=(1+G(t))*(1-\sqrt{f_1})$，且 $0 \leqslant G(t) \leqslant 1, 1-\sqrt{f_1} \leqslant 1$，由此可知 $f_2 \leqslant 2$。在该测试问题中，$|X_I|=5, |X_{II}|=25$。

从图 5(a)、图 5(b)和图 5(c)可知算法 DMEIA 所得最优解明显好于其他两种算法，并且 DNSGAII-A 和 DBM 均不能找到 Pareto 面上的全部解。从图 5(d)可知 DMEIA 相对其他两种算法其随时间变化波动较小，相对平稳。DBM 收敛的最小值最小。

结束语　本文提出了一种新的动态多目标人工免疫系统模型，并给出了模型的核心元素——DMEIA 算法。算法结合了进化算法与免疫算法的优良特性，适时地调用模型规则集 R 中的环境追踪规则，并采用基于 Parzen 窗估计法进行信息熵的密度估计。仿真实验表明 DMEIA 具有很强的环境追踪能力及全局搜索能力，在收敛性、分布性方面也体现了良好的效果，这些充分验证了模型的综合性能。

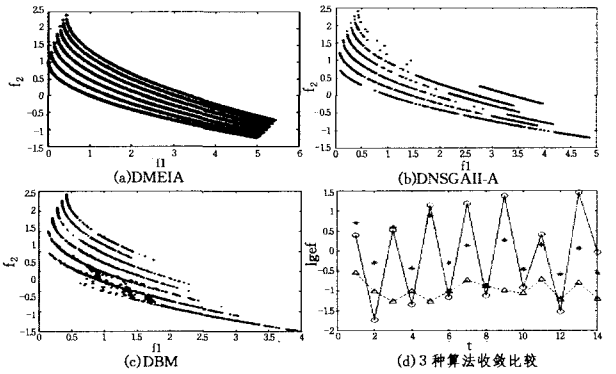

图 5　DMOP3 的 3 种算法的仿真结果图

参 考 文 献

[1] Wang Yuping, Dang Chuangyin. An evolutionary algorithm for dynamic multi-objective optimization[J]. Applied Mathematics and Computation, 2008, 205(1): 6-18

[2] Jin Y, Branke J. Evolutionary Optimization in Uncertain Environments—A Survey[J]. IEEE Transactions Evolutionary Computation, 2005, 9(3): 303-317

[3] 尚荣华，焦李成，公茂果，等. 免疫克隆算法求解动态多目标优化问题[J]. 软件学报，2007，11(8)：2700-2711

[4] 郑金华. 多目标进化算法及其应用[M]. 北京：科学出版社，2007：1-10

[5] Farina M, Deb K, Amato P. Dynamic multiobjective optimization problems: test case, approximations, and applications[J]. IEEE Transactions on Evolutionary Computation, 2004, 8(5): 425- 442

[6] Deb K, Uda YA B R N, Karthik S. Dynamic multi-objective optimization and decision-making using modified NSGA-II: a case study on hydro-thermal power scheduling bi-objective optimization problems[R]. KanGAL Report. 2006

(上接第 213 页)

[5] Soranzo N, Bianconi G, Altafini C. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data[J]. Bioinformatics, 2007, 23: 1640-1647

[6] Lee T I, Rinaldi N J, Robert F, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae[J]. Science, 2002, 298: 799-804

[7] Tyers M. Cell cycle goes global[J]. Curr Opin Cell Biol, 2004, 16: 602-613

[8] Li F T, Long T, Lu Y, et al. The yeast cell-cycle cutwork is robustly designed[J]. Proc. Natl. Acad. Sci. USA, 2004, 101: 4781-4786

[9] Kitano H. Biological rebustness[J]. Nature Reviews Genetics, 2004, 5: 826-837

[10] Fengli R, Jinde C. Asymptotic and robust stability of genetic regulatory networks with time-varying delays[J]. Nerocomputing, 2008, 71: 834-842

[11] Crombach A, Hogeweg P. Evolution of Evolvability in Gene Regulatory Networks[J]. PLoS Comput Biol, 2008, 4

[12] Xiaohua L, Hongyan G, Yuanwei J. Robust decentralized connective stabilization for expanding construction of large-scale systems[C]// Automation and Logistics, 2008. ICAL 2008. Sept 2008: 1121-1125

[13] National Center for Biotechnology Institute[OL]. http://www.ncbi.nlm.nih.gov/

[14] Kyoto Endyclopedia of Genes and Genomes[OL]. http://www.genome.jp/kegg/

[15] Luwen Z, Wu Z, Mei X, et al. Reverse engineering large-scale gene networks using linear model and robust regression synthetic versus real data. 待发表