

DOI:10.3724/SP.J.1008.2008.01106

• 综 述 •

基于时间序列表达数据基因调控网络模型的研究进展

金志超, 吴 骋, 高青斌, 姜 洋, 贺 佳*

第二军医大学卫生勤务学系卫生统计学教研室, 上海 200433

[摘要] 基因间的调控是随时间、环境变化的动态事件, 基因调控网络是一个连续而复杂的动态系统, 基于时间序列的基因组 DNA 微阵列为研究者提供了构建动态调控网络的工具。本文介绍几种基于时间序列基因表达数据调控网络模型(时序布尔网络、微分方程、动态贝叶斯等), 分析几种模型的优缺点, 并展望未来的研究趋势。

[关键词] 时间序列; 基因调控网络; 时序布尔网络; 微分方程; 动态贝叶斯

[中图分类号] R 394.36 **[文献标志码]** A **[文章编号]** 0258-879X(2008)09-1106-04

Gene regulatory network models based on time series gene expression data: recent progress

JIN Zhi-chao, WU Cheng, GAO Qing-bin, JIANG Yang, HE Jia*

Department of Health Statistics, Faculty of Military Health Services, Second Military Medical University, Shanghai 200433, China

[ABSTRACT] Regulation between genes is a dynamic event associated with changes of time and circumstances. Gene regulatory network is a complicated and dynamic system. Time series gene microarray provides a tool for creating dynamic gene regulatory network. In this paper, we review several models of dynamic gene regulatory network based on time series gene expression data, including temporal Boolean network, differential equation, dynamic Bayesian networks, etc.. The advantages and disadvantages of the models were analyzed and the future of the research is predicted.

[KEY WORDS] time series; gene regulatory networks; temporal Boolean networks; differential equation; dynamic Bayesian networks

[Acad J Sec Mil Med Univ, 2008, 29(9): 1106-1109]

基因调控网络是一个连续而复杂的动态系统, 基因间的调控是随时间、环境变化的动态事件, 基因组 DNA 微阵列为研究者提供了较好的认识调控网络的工具。时间序列(time-series)基因表达数据包含着丰富的基因调控信息, 可以想象, 位于基因调控网络上游基因的变化, 理应处于下游基因的前面。当上游基因(比如转录因子 TF)的表达发生变化以后, 这种变化会沿着基因调控网络传播。当测量了每个基因在不同时间的表达水平以后, 就能从这种时间序列数据中反推得到关于基因之间调控顺序以及调控对象的重要信息。一般认为, 时间序列基因表达数据与同样大小的静态基因表达数据相比, 含有更大的推导基因调控网络的信息量^[1]。本文对基于时间序列基因表达数据构建调控网络的研究现状进行了综述, 并对未来的研究趋势进行展望。

1 时间序列基因表达的特点

时间序列表达实验是在不同的时间点对细胞周期以及

特定状态条件下的基因表达进行测定(比如 Geier 的基因敲除实验, Boldrick 的疫苗暴露实验, Cho 的细胞有丝分裂周期实验等^[2-4]), 以观测各个时间点上或细胞周期中基因表达的变化情况, 寻找调控基因和目标基因。它和一般静态表达实验的主要区别有两点^[5]: 一是静态表达实验是对不同样本在同一时间点上的测量, 而时间序列表达实验则是对同一样本在几个时间点进行连续测量; 二是时间序列表达实验的测量值在连续的时间点上有较强的自相关, 而静态表达实验要求样本的测量值服从独立同分布的条件。

由于实验费用的昂贵, 大多数实验室测量的时间点(样本)相对于庞大的基因数(观测)来说是远远不够的^[4, 6-8], 过少的时间点将影响模型参数的估计, 因此构建的模型必须简单, 参数尽可能少。尽管实验技术取得了很大的进展, 但是实验噪声和转录翻译时延问题始终没有得到很好解决。高通量的基因表达实验普遍存在实验噪声(noise), 在普通的微阵列实验中, 噪声要占整个信号的 20%~30%^[9]; 由于转录

[收稿日期] 2008-01-10 **[接受日期]** 2008-04-10

[基金项目] 国家自然科学基金(30671821), 上海市自然科学基金(04ZR14049). Supported by National Natural Science Foundation of China (30671821) and Natural Science Foundation of Shanghai(04ZR14049).

[作者简介] 金志超, 硕士生. E-mail: jzc109@hotmail.com

* 通讯作者(Corresponding author). Tel: 021-25070418, E-mail: hejia@smmu.edu.cn

翻译时间,两个基因之间的调控存在一定的时延^[10],这是构建动态网络调控必须考虑的关键问题,尽管可以用一些算法减少时延产生的误差,但是最根本的还是实验方法的改进。

2 基于时间序列基因表达数据的基因调控网络模型

2.1 时序布尔网络模型 布尔网络是一种离散时间模型,每个基因在表达水平上都有一个时间函数,用 $X_i(t) \in \{0,1\}$ 表示, $i \in \{1,2,3,\dots,N\}$, $t \in \{0,1,2,3,\dots,T\}$, 基因 i 的 k 个调控基因用 $r_i^1 \dots r_i^k$ 表示,每个基因有一个调控函数或者动态函数 $f_i: \{0,1\}^k \rightarrow \{0,1\}$, 基因 i 的动态变化由下式给出^[11]: $X_i(t+1) = f_i[X_{r_i^1}(t), \dots, X_{r_i^k}(t)]$ 。从上式可得基因 i 在 $t+1$ 时刻的状态由 t 时刻的状态决定,这样得到的调控网络存在两个缺点:一是转录翻译时延可能会大于一个单位时间,由此造成网络精确性较差;二是并不是所有影响基因表达水平的变量都必须都是可观测的,上述布尔网络由此会造成假阴性(false negative)。

Silvescu 等^[12]提出了时序布尔网络模型(Temporal Boolean networks model),处理多于一个单位时间跨度的基因间的依赖性,允许每个基因在 $t+1$ 时刻的表达受至多 k 个基因在时刻 $t, \dots, t-(T-1)$ (T 是一个基因可以影响另一个基因的时限)表达水平的布尔函数控制,而不仅仅是 t 时刻(图 1)。

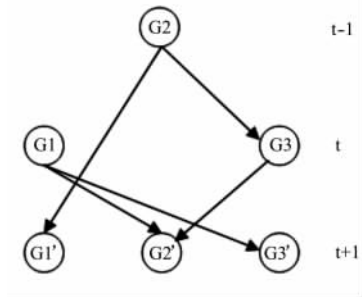


图 1 时序布尔网络模型
Fig 1 Temporal Boolean networks model

Silvescu 等^[12]运用决策树(decision tree)描述时序布尔网络,调控基因和目标基因作为节点(node)出现在决策树的相应位置,并利用模拟数据对模型的特异性、敏感性和精确性进行评价,获得了比较满意的结果。时序布尔网络不是一种量化的生化网络,相比真实的基因调控网络显得比较简单和粗糙。

Gupta 等^[13]将 Pearson 相关系数和布尔网络相结合,首先通过聚类得到相应的基因归类以及类间系数,然后从每个类中选取具有代表性的基因(可以是假定的一个和类有相似平均表达量的基因),将这些基因间的相关系数作为基因网络中相互影响的权重。根据 Pearson 相关系数定义的权重,为布尔网络基因间的调控关系定义如下:当 Pearson 相关系数在 $(0.8,1]$ 时定义为 1,在 $[-0.8, +0.8]$ 时定义为 0,在 $[-1, -0.8)$ 时定义为 -1。在运用布尔网络构建调控网络时,考虑时延、基因的降解以及蛋白质的半衰期,对 3 个时序基因表达数据进行较好的模拟。

布尔网络和马氏链(Markov chains)结合起来可处理概率框架下的不确定性,由此引进 PBNs(Probabilistic Boolean Networks)模型^[14]。PBNs 模型是在布尔网络模型的基础上增加了对父代基因集合的概率选择,它合并了基于规则的基因之间的依赖关系,可进行全局网络动态的系统研究,可在数据和模型选择上处理不确定性,PBNs 为量化基因相互作用过程中的相互影响和敏感度提供了一种方法^[15]。

2.2 微分方程模型 Chen 等^[16]提出运用微分方程对时间序列基因表达资料建模。模型建立了许多假设,包括每个基因线性转录函数和翻译产物对转录率的反馈机制,将基因的转录、翻译和 RNA、蛋白的降解都加以考虑,具体模型如下:

$\frac{dx}{dt} = f(p) - Vr, \frac{dp}{dt} = Lr - Up$, n 为基因数, r 为 mRNA 浓度, p 为蛋白质浓度, r 和 p 都是以 t 为函数的 n 维向量, $f(p)$ 为转录函数, L 为翻译常数, V 为 mRNA 的降解率, U 为蛋白质的降解率。这个模型对于周期表达的数据适用,同时假设所有的基因表现出不同的表达模式。该模型运用线性转录和稳定网络的假设,有效地减少了参数搜索空间,优于其他简单线性模型;其对转录翻译时延也进行了探讨,但没有得出很好的结果。这个模型存在着以下缺陷,一是实际生物学实验中蛋白质的含量往往难以获得;二是忽略了其他调控物如代谢产物的影响。

de Hoon 等^[17]对 Chen 模型^[16]进行了改进,用基因之间的相互作用代替调控蛋白质的作用,在 Chen 模型^[16]的基础上引入误差项 $\epsilon(t)$,在选择基因时,通过设计一种新的统计学方法,从枯草杆菌 3 823 个基因中选择有统计学意义的 684 个基因,并用 k -means 对这 684 个基因进行聚类,减少模型参数个数,用 Akaike's Information Criterion(AIC)决定矩阵中哪些元素为零,避免了模型的过度拟合(overfitting),最小 AIC 的模型被看作是最优化的模型,但是作者并未对模型作出明确的解释。

由于测量时间点很少,用来推断网络的样本数据量也比较少,为了增加样本数据量,D'haeseleer 等^[18]用非线性插值对时间点进行拟合。数据来源于 3 个不同的试验,间隔一定的时间,使用表达水平的 3 次插值将数据拼接(增加测量的时间点),然后用微分方程来推导交互矩阵(interaction matrix)。基因网络模型也是一组线性微分方程,其中基因表达量的改变取决于其他基因输入的权重,互作的权重用最小二乘法来拟合内插的时间序列,由于数据空间分布不均匀,分布密度大的基因得到的权重大,引起了一定的偏性。该方法较精确地拟合了数据,但可能存在一定程度上的过度拟合,同时该模型没有减少基因的搜索空间,计算量比较大。

Kim 等^[19]综合考虑了芯片的噪声以及基因表达时延,利用普通微分方程结合数字图解构建调控网络,在较小的网络中(4 个节点),推测网络和参考网络一致性较好,但当网络节点增加至 8 个时,网络结构的精确性随之降低,网络的稳定性较差。

2.3 动态贝叶斯模型 动态贝叶斯模型(Dynamic Bayesian Networks, DBNs)可表示具有时间序列性质的随机变量 X 的联合概率分布。通常将其结构分为 2 个部分:(1)初始状

态: $B_0 = (G_0, \Theta_0)$, 表示在初始状态随机变量 $X(0)$ 的联合概率分布; (2) 转移过程 $B_1 = (G_1, \Theta_1)$, 表示所有时间点 t 上的转移概率 $P_r\{X(t) \mid X(t-1)\}$ 。因此 DBNs 模型通常也表示为 (B_0, B_1) 。一组随机时间序列变量 $X(0), X(1), \dots, X(T)$ 的联合分布可表示为:

$$P_r\{X(0), X(1), \dots, X(T)\} = P_r\{X(0)\} \prod_{t=1}^T P_r\{X(t) \mid X(t-1)\} = \prod_{i=1}^n P_r\{X_i(0) \mid pa[X_i(0)]\} \times \prod_{t=1}^T \prod_{j=1}^n P_r\{X_j(t) \mid pa[X_j(t)]\}$$

DBNs 非常适用于处理非线性关系和由随机现象引起的不确定性, 具有强大的知识表达能力和概率推理能力。DBNs 最大的优点在于它可以克服普通贝叶斯有向无环 (acyclic) 的缺点, 能表示出基因之间的负反馈调控, 负反馈调控在基因调控网络中是非常普遍的; 其次, DBNs 将数据的时序特点纳入模型, 能够改进因果关系的学习能力, 提高模型的预测精度。

Murphy 等^[20] 根据 2 个基因之间的调控存在一定的时延, 首次提出用动态贝叶斯网络模型分析时间序列基因表达数据。Ong 等^[21] 利用大肠埃希菌生理学上的改变建立相应的动态贝叶斯模型, 得到网络节点的因果关系以及反馈环, 利用先验知识将同一操纵子下的基因归类作为共调控, 可以大大减少模型的参数, 并且改进了基因相互作用网络的建模精度。Perrin 等^[22] 利用机器学习 (machine learning) 方法, 提出一种能够处理缺失变量的随机调控模型, 该模型适合处理基因调控和基因表达随机性质, 是一种动态贝叶斯网络, 模型的参数学习采用有罚分机制的最大似然法, 由改进的 EM 算法实现。作者采用大肠杆菌 S. O. S. DNA 修复网络有关的实验数据集进行测试, 在对网络的建模过程中, 发现了一个缺失变量, 证明了学习方法对未知数据的良好预测性能。Zou 等^[10] 针对 DBNs 在时延和计算时间两个方面的缺点, 利用离散 DBNs 模型构建具有多时延的调控网络, 为每个基因选择潜在的调控基因, 减少搜索空间, 考虑不同调控目标基因间不同的时延, 精确估计生物学转录时间, 但是该算法复杂, 离散化的同时引起了信息的丢失。崔光照等^[23] 在 Zou 的基础上对模型进行了改进, 提出了用连续 DBNs 构建多时延的基因调控网络模型, 用互信息模型估计调控基因和目标基因之间的时延, 根据时延对时间序列表达数据进行重新组织, 最后利用 DBNs 重构网络, 这种算法避免了离散化带来的信息丢失, 提高了网络构建的精度。

2.4 其他模型 除了上述传统的模型外, 还有一些模型和理论用以分析基因表达数据, 如信息理论^[24], 传统的互信息关联模型用熵和互信息描述基因和基因的关联, 但是这种关联缺少方向性, 即无法说明调控和被调控的关系。Tung 等^[25] 利用转移熵理论很好的解决了这个问题, 转移熵用来量化从一个时间点到另一时间的信息。

$$T_{Y \rightarrow X} = \sum_{x(t+1), x^k(t), y^l(t)} p[x(t+1), x^k(t), y^l(t)] \log \frac{p[x(t+1) \mid x^k(t), y^l(t)]}{p[x(t+1) \mid x^k(t)]}$$

此外, 徐肖江等^[26] 利用线性网络模型, 并结合调控元件识别和基因聚类等方法在细胞周期和环境胁迫响应过程的

基因调控网络进行预测, 结果与已知的实验结果相吻合。

3 基因调控网络模型的发展瓶颈

3.1 各模型的固有弱点 布尔网络的缺点在于其本质上的确定性, 它把内部的遗传功能和相互作用理解为逻辑规则, 但是在生物体内则是关系复杂, 具有自组织性, 与真实网络相比, 布尔网络是一种粗造的模型; 微分方程是一种精细的数学模型, 可以描述基因调控的因果关系, 且基因间相互可以明确表示, 但模型对数据的要求较高, 而目前微阵列数据质量较低, 信噪比低, 测量噪声大等, 而且数据缺失比较严重, 再者时间列表表达资料一般样本量远远小于其维数, 模型的参数设置和估计较为困难, 目前的研究主要集中于简单和前期研究较多的网络; 动态贝叶斯是近年来研究的比较多的一种模型, 具有强大的知识表达能力和概率推理能力, 其精细化程度介于微分方程和布尔网络之间, 它的负反馈性、非线性以及随机不确定性与真实的生物系统相似, 但也存在计算量较大的问题。

3.2 系统缺陷 基因调控网络经过近十年的研究, 以及实验技术的发展, 网络模型的精度有了较大的改进, 但是基于时间序列基因表达数据的调控网络发展缓慢, 主要是由以下几个原因造成的。

(1) 模型的局限性。目前大多数工作假设调控网络结构恒定, 不随时间发生变化, 各个节点之间的条件转移概率代表了系统的动态特征。这个假设是与真正生物系统的调控过程不吻合的。

(2) 微阵列数据质量的问题。样本数量过少, 因为随着网络基因数的增加, 所需的样本也呈几何级数的增加, 但由于实验费用昂贵, 得到充分的样本几乎不大可能; 实验的噪声以及表达时延问题始终没有得到较好的解决, 而噪声对于小样本网络构建的影响是很大的。因此, 现在对于动态网络调控研究最关键的是增加样本量, 减少实验噪声。

(3) 算法理论和生物学实验脱节。由于实验花费巨大, 现在大多数模型都是在几个比较经典的数据库上进行模拟, 所得的结论局限性比较大, 而且得到的模型没有经过实验的验证, 缺少建模—模拟—实验循环逼近真实生物学基因调控网络的机制。

4 展 望

研究基因调控机制是生物学的中心任务之一, 推断基于时间序列基因表达数据的基因调控网络是达成这一中心任务的有效方法, 尽管现有网络模型存在着不足, 但是研究前景相当广阔。算法研究结合实验研究, 如微扰实验的动态调控研究和细胞周期基因调控网络研究已经越来越受到关注。结合生物学信息建模将是该领域今后发展的主要趋势, 如利用启动子 (promoter) 信息^[27-28], 结合基因本体论 (Gene Ontology)^[29], 利用蛋白质-蛋白质相互作用 (protein-protein interactions)、蛋白质-基因相互作用 (protein-DNA interactions)^[2] 等。

[参 考 文 献]

[1] 李 瑶, 贺 佳, 陈一东, 曹勇伟, 许 可. 基因芯片数据分析与

- 处理[M]. 北京:化学工业出版社,2006:266-267.
- [2] Geier F, Timmer J, Fleck C. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge[J]. BMC Syst Biol, 2007, 1: 11.
- [3] Boldrick J C, Alizadeh A A, Diehn M, Dudoit S, Liu C L, Belcher C E, et al. Stereotyped and specific gene expression programs in human innate immune responses to bacteria[J]. Proc Natl Acad Sci USA, 2002, 99: 972-977.
- [4] Cho R J, Campbell M J, Winzler E A, Steinmetz L, Conway A, Wodicka L, et al. A genome-wide transcriptional analysis of the mitotic cell cycle[J]. Mol Cell, 1998, 2: 65-73.
- [5] Bar-Joseph Z. Analyzing time series gene expression data[J]. Bioinformatics, 2004, 20: 2493-2503.
- [6] Spellman P T, Sherlock G, Zhang M Q, Iyer V R, Anders K, Eisen M B, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization[J]. Mol Biol Cell, 1998, 9: 3273-3297.
- [7] Zhu G, Spellman P T, Volpe T, Brown P O, Botstein D, Davis T N, et al. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth[J]. Nature, 2000, 406: 90-94.
- [8] Pramila T, Miles S, GuhaThakurta D, Jemiole D, Breeden L L. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G₁ phase of the cell cycle[J]. Genes Dev, 2002, 16: 3034-3045.
- [9] Rocke D M, Durbin B. A model for measurement error for gene expression arrays[J]. J Comput Biol, 2001, 8: 557-569.
- [10] Zou M, Conzen S D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data[J]. Bioinformatics, 2005, 21: 71-79.
- [11] Perkins T J, Hallett M, Glass L. Dynamical properties of model gene networks and implications for the inverse problem [J]. Biosystems, 2006, 84: 115-123.
- [12] Silvescu A, Honavar V. Temporal Boolean network models of genetic networks and their inference from gene expression time series[J]. Complex Systems, 1997, 13: 54-70.
- [13] Gupta R R, Achenie L E. A network model for gene regulation [J]. Computer and Chemical Engineering, 2007, 31: 950-961.
- [14] Lähdesmäki H, Hautaniemi S, Shmulevich I, Yli-Harja O. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks[J]. Signal Processing, 2006, 86: 814-834.
- [15] Shmulevich I, Dougherty E R, Kim S, Zhang W. Probabilistic Boolean Networks; a rule-based uncertainty model for gene regulatory networks[J]. Bioinformatics, 2002, 18: 261-274.
- [16] Chen T, He H L, Church G M. Modeling gene expression with differential equations[J]. Pac Symp Biocomput, 1999, 4: 29-40.
- [17] de Hoon M J, Imoto S, Kobayashi K, Ogasawara N, Miyano S. Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations [J]. Pac Symp Biocomput, 2003, 8: 17-28.
- [18] D'haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury [J]. Pac Symp Biocomput, 1999, 4: 41-52.
- [19] Kim S, Kim J, Cho K H. Inferring gene regulatory networks from temporal expression profiles under time-delay and noise [J]. Comput Biol Chem, 2007, 31: 239-245.
- [20] Xiong H, Choe Y. Structural systems identification of genetic regulatory networks[J]. Bioinformatics, 2008, 24: 553-560.
- [21] Ong I M, Glasner J D, Page D. Modelling regulatory pathways in *E. coli* from time series expression profiles[J]. Bioinformatics, 2002, 18 Suppl 1: S241-S248.
- [22] Perrin B E, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché-Buc F. Gene networks inference using dynamic Bayesian networks[J]. Bioinformatics, 2003, 19 Suppl 2: II 138-II 148.
- [23] 崔光照, 张勋才, 曹祥红, 董亚非, 王延峰. 基于动态贝叶斯网络的多时延基因调控网络构建[J]. 科学技术与工程, 2005, 5: 1247-1251.
- [24] 张国伟, 邵世煌, 齐金鹏, 张颖. 基于信息度量的基因网络建模[J]. 生物信息学, 2006, 4: 170-172.
- [25] Tung T Q, Taewoo R, Lee K H, Lee D. Inferring gene regulatory network from microarray time series data using transfer entropy[J]. CBMS, 2007, (7): 383-388.
- [26] 徐肖江, 王连水, 丁达夫. 从酵母表达时间序列估计基因调控网络[J]. 生物化学与生物物理学报, 2003, 35: 707-716.
- [27] Wang W, Cherry J M, Nochomovitz Y, Jolly E, Botstein D, Li H. Inference of combinatorial regulation in yeast transcriptional networks; a case study of sporulation[J]. Proc Natl Acad Sci USA, 2005, 102: 1998-2003.
- [28] Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, et al. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection[J]. Bioinformatics, 2003, 19 Suppl 2: II 227-II 236.
- [29] Ferrazzi F, Magni P, Sacchi L, Nuzzo A, Petrovic U, Bellazzi R. Inferring gene regulatory networks by integrating static and dynamic data[J]. Int J Med Inform, 2007, 76 Suppl 3: S462-S475.

[本文编辑] 贾泽军