· 专题报道 ·

基因调控网络模型构建方法

虞慧婷[1],吴 骋[1],柳伟伟[1],付旭平[2],贺 佳[1*]
(1. 第二军医大学卫生勤务学系卫生统计学教研室,上海 200433;2. 复旦大学遗传学研究所遗传工程重点实验室,上海 200433)

[摘要] 基因调控网络的研究从基因之间相互作用的角度揭示复杂的生命现象,是功能基因组学研究的重要内容,也是当前生物信息学研究的前沿。基因芯片技术在生物信息学中的应用为基因调控网络的研究提供大量可供研究与分析的基础数据。本文介绍了基因调控网络的起源与近年来的发展情况,详细说明了基因调控网络构建的前提与基本原理,并分析几种经典调控网络模型:布尔网络模型、线性及非线性模型和贝叶斯网络模型,阐述各种模型构建的基本原理和算法,结合基因芯片的数据特点,探讨各种模型的优缺点及其适用情况,对各种网络模型进行分析与总结。

[关键词] 基因调控网络;布尔网络模型;线性模型;非线性模型;贝叶斯模型
[中图分类号] Q 756 [文献标识码] A [文章编号] 0258-879X(2006)07-0737-04

Modeling of gene regulatory networks

YU Hui-ting[1], WU Cheng[1], LIU Wei-wei[1], FU Xu-ping[2], HE Jia[1*] (1. Department of Health Statistics, Faculty of Health Services, Second Military Medical University, Shanghai 200433, China; 2. State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Science, Fudan University, Shanghai 200433)

[**ABSTRACT**] Gene regulatory networks (GRN), which focuses on the complex interactions of genes in life, is an important part in the study of the functional genomics and is the frontier of bioinformatics research. Application of gene-chip technique in bioinformatics provides a great number of basic data for the research of GRN. This paper reviews the origin and recent development of GRN, explicates the preconditions and rationales for construction of GRN, and analyzes several classic GRN models: Boolean networks, linear models, non-linear models and Bayesian networks. The rationales, basic algorithms, advantages, disadvantages and applicability of the models are reviewed based on the characteristics of gene-chip data.

[**KEY WORDS**] gene regulatory networks; Boolean networks; linear models; non-linear models; Bayesian networks
[Acad J Sec Mil Med Univ,2006,27(7):737-740]

分子生物学的深入研究揭示了复杂的生命现象是大量基因相互作用的结果,在这个过程中,基因调控起了重要作用。利用 DNA 微阵列研究系统生物学,在系统的层次上理解基因调控是目前生物信息学中最富挑战性的工作。

基因网络的研究始于 20 世纪 60 年代,Rater 描述了控制原核生物的分子基因系统组织的特点,随之,Kaufman 通过简单的逻辑规则研究基因网络动力学,构造了自主的随机布尔网络模型。20 世纪 90 年代实验数据的增加加速了基因网络理论的研究,到目前为止,已发展了大量的研究基因网络的方法。本文将详细介绍几种经典的网络模型,并对其算法和优缺点进行比较分析。

1 基因调控网络构建的前提与基本原理、方法

生物网络的构建根据不同的研究对象可分为 3 类:代谢网络、蛋白质作用网络和调控网络。其中代谢网络最基础、最保守,蛋白质相互作用网络构建较为复杂,调控网络成为目前研究的热点。需要注意的是这 3 种生物网络并不是相互独立的,而是相互重叠、相互作用影响的。

调控网络可在分子水平上分为 3 个层次:DNA 水平、RNA 水平和蛋白质水平。DNA 水平主要是研究基因在空间上的关系影响基因的表达;RNA 水平上,也就是转录水平上的调控,主要研究代谢或者是信号转导过程决定转录因子浓度的调控过程;蛋白质水平主要研究蛋白质翻译后修饰,从而影响基因产物的活性和种类的过程。目前在没有确切明白基因间的相互作用关系时,网络调控研究的进行是建立在一定的假设前提下的。

基因网络构建的前提假设和基本原理如下:(1)

[基金项目] 国家自然科学基金(30471502);上海市自然科学基金(04ZR14049). Supported by National Natural Science Foundation of China(30471502) and Natural Science Foundation of Shanghai Municipal Government(04ZR14049).
[作者简介] 虞慧婷,硕士生.
* Corresponding author. E-mail:hejia@smmu.edu.cn

如果 2 条基因序列谱相似，则这 2 条基因协作调控，并可能功能相似[1]；(2)具有相同表达模式的基因可能有同样的表达过程[2]；(3)整体的基因表达模式是局部调控催化反应的联合作用的结果。在这些前提假定下可对基因芯片数据进行数学上的处理与分析，进一步建立基因调控网络。通过建立基因调控网络模型，可以对某一物种或组织中的全部基因的表达关系进行整体的模拟分析和研究，在系统的框架下认识生命现象，特别是信息流动的规律，识别和推断基因调控网络结构特征和调控关系，认识复杂的分子调控过程，用以理解支配基因表达和功能的基本规则，揭示基因表达过程中的信息传输规律，在整体的框架下研究基因的功能。

通常基因网络调控的研究方法主要有：(1)通过聚类分析建立模型；(2)进行反复的微扰分析重构模型[3]；(3)通过“反向技术”来推断网络。聚类分析建立模型是一种常用的构建基因调控网络，探索未知功能基因的方法。聚类分析是一种常用的构建基因调控网络、探索未知功能基因的方法，常用的聚类方法有：层次聚类、K-means 聚类、自组织图聚类、人工神经网络等。目前各种算法都广泛运用于基因芯片的数据分析中，但是算法间尚缺乏系统性，需要从数据出发，进行系统的理论探讨。“微扰分析”主要通过反复的微扰试验来实现网络重构(network reconstruction)。它根据已经建立的基因调控网络指导实验设计，获得高通量实验数据，然后根据实验结果优化网络模型的结构和参数[4]。但是目前成功建立的网络调控模型比较少，很难做进一步的调控网络模型研究。通过“反向技术”来推断网络，就是从基因表达的数据反向推断未知的或隐含的基因网络拓扑结构的技术，它需要选定合适的参数模型，并用适当的算法推断网络参数，确定输入输出规则，预测网络随时间的变化规律，这也是目前研究较多的构建基因调控网络模型的方法。

2　基因调控网络模型

目前，研究基因网络的模型很多，分类方法不尽相同，通常有：离散网络和连续网络、确定型网络和随机网络、定性网络和定量网络等等。以下介绍几种常见的经典网络模型。

2.1　布尔网络模型　布尔网络模型[5,6]是一种简单的离散型基因调控网络。1998 年 Yuh 等[7]综合以往的研究结果，详细分析了海胆 Stronglocentrotus Purpuratus 基因 Endol16，对这一基因转录水平的基因调控网络进行了逻辑描述。在布尔网络中，每个基因所处的状态或者是“开”，或者是“关”。状态“开”表示一个基因转录表达，形成基因产物，而状态“关”则代表一个基因未转录。基因之间的相互作用关系由布尔表达式来表示，例如：A and not $B \rightarrow C$ 表示“如果 A 基因表达，且 B 基因不表达，则 C 基因表达”。以有向图 $G=(V, F)$ 表示布尔网络，其中 V 是图的节点集合(图 1 中的 A、B、C)，每个节点代表 1 条基因，或者代表 1 个环境刺激，F 表示转录表达路径，即图 1 中的有向边。

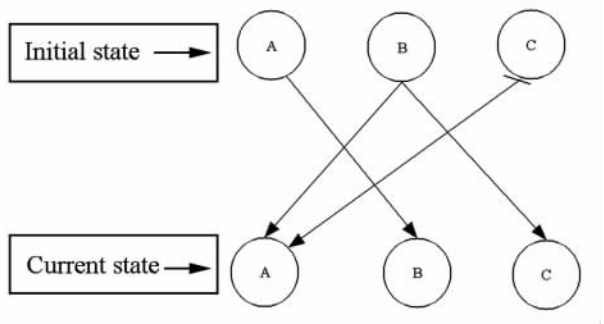

图 1　布尔网络模型

Fig 1　Boolean networks model

Rules for function: (1) A stimulates B; (2) B stimulates A and C; (3) C restrains A

布尔网络强调的是基本的全局网络而不是一种定量的生化模型，相比于真实的基因网络，布尔网络模型比较简单粗糙。它把内部的遗传功能和相互作用理解为逻辑规则，但是由于基因间相互作用的复杂性，用每条基因的一条逻辑规则来做推断常常会导致错误规则。

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. 从而引进 PBN (probabilistic Boolean network)模型[8]。PBN 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。

2.2　线性组合模型　线性组合模型[9]是一种连续型网络模型，在这种模型中，一个基因的表达值通常描述为若干个其他基因表达值的加权和。基本表示形式为：$X_i(t+\Delta t)=\Sigma\omega_{ij}X_j(t)$，其中，$X_i(t+\Delta t)$ 是基因 i 在 $t+\Delta t$ 时刻的表达水平，$X_j(t)$ 是基因 j 在 t 时刻的表达水平，而 ω_{ij} 代表基因 j 的表达水平对基因 i 的影响。在这种基因相互关系表示形式中，还可以增加其他数据项，以逼近基因调控的实际情况。将上述表达式转换为线性差分方程，描述一个基因

表达水平的变化趋势。这样,在给定一系列基因表达水平的实验数据之后,即给定每个基因的时间序列 $X_i(t)$,就可以利用最小二乘法或者多重分析法求解整个系统的差分方程组,从而确定方程中的所有参数,即确定 ω_{ij}。最终,利用差分方程分析各个基因的表达行为。

2001 年,D'haeseleer 等[10]已用这种方法分析了大鼠脊髓和海马回的基因表达数据,建立了一个包含有 65 个基因的模型,精确地复制出网络调控轨迹,包括脊髓发育、海马回发育到海马回损伤。但是由于对哺乳动物中枢神经系统的调控作用所知不多,许多预测行为并不能被准确地证实。在实际应用中,由于芯片数据常常是成千上万的基因,这意味着巨大的运算量,且芯片所测的时间点数较少,通常还不能求出解析解。而且线性模型将基因间的相互作用都近似地看成一种线性关系,这与实际中基因间的相互作用关系往往是非线性的不相符,因此模型尚需进行改进。

2.3 *加权矩阵模型* 加权矩阵模型[10]与线性组合模型相似,在该模型中,一个基因的表达值是其他基因表达值的函数。含有 n 条基因的基因表达状态用 n 维空间中的向量 $u(t)$ 表示,$u(t)$ 的每一个元素代表 1 条基因在时刻 t 的表达水平。以加权矩阵 ω 表示基因之间的相互调控作用,ω 的每一行代表 1 条基因的所有调控输入,ω_{ij} 代表基因 j 的表达水平对基因 i 的影响。在时刻 t,基因 j 对基因 i 的净调控输入为 j 的表达水平[即 $u_j(t)$]乘以 j 对 i 的调控影响程度 ω_{ij}。基因 i 的总调控输入 $r_i(t)$ 为:$r_i(t)=\sum_j \omega_{ij}u_j(t)$ 这一形式与线性组合模型相似,若 ω_{ij} 为正值,则基因 j 激发基因 i 的表达,负值表示基因 j 抑制基因 i 的表达,0 表示基因 j 对基因 i 没有作用。与线性组合模型不同的是,基因 i 最终表达响应还需要经过一次非线性映射:

$$u_i(t+1)=\frac{1}{1+e^{[\alpha_j r_j(t)+\beta_j]}}$$

该函数是神经网络中常用的 *Sigmoid* 函数,其中 α 和 β 是 2 个常数,规定非线性映射函数曲线的位置和曲度。通过上式,计算出 $t+1$ 时刻基因 i 的表达水平。在最初阶段,加权矩阵的值是未知的。但是可以利用机器学习方法,根据基因表达数据估计加权矩阵中各个元素的值。对于这样的模型,可以利用线性代数方法和神经网络方法进行分析。实验表明,该模型具有稳定基因表达水平,与实际生物系统相一致。在这种模型中还可以加入新的变量,模拟环境条件变化对基因表达水平的影响。

Weaver 等[10]首先将加权矩阵应用于环境的相关因素研究,描述了转录调控网络关系。构造了 5 个环境因素在不同情况下的调控网络关系。并找到与环境变化有关和无关的两类基本的基因簇,对环境的治理提供了重要的线索。目前,加权矩阵模型已经成功运用于模拟小型生化通路,但是用于基因组的模型进行模拟尚受限制。因为该模型的运算条件较为苛刻,模型中的参数设置还是一个较难问题。另外加权线性组合模型是一种精细化模型,鉴于目前基因芯片数据还存在许多缺陷,如数据缺失严重、系统误差较大、数据的样本量通常小于其维数,当数据质量不高时,通常不提倡采用精确的模型来模拟调控网络[8],因此加权矩阵模型的运用尚需将数据质量进一步提高。

2.4 *贝叶斯网络模型* 布尔网络模型是一种粗放的定性方法,而加权矩阵模型又是通过精细的数学分析来量化描述生物过程,贝叶斯网络模型则可以看成是这两种模型的一种折衷。Murphy[14]和 Friedman 等[12]分别于 1998 和 1999 年提议建立基于贝叶斯网络[11~13]的基因调控网络模型。Murphy 等[14]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(dynamic bayesian networks, DBNs)模型分析时序基因表达数据。Friedman 等[12]利用贝叶斯网络构建了一个包含 800 个基因的基因调控网络。

贝叶斯网络引入有向无圈图模型[15,16]和隐马尔可夫链来描述变量间的联系与相互作用,构建调控网络模型,通常贝叶斯网络可用数对 $B=(G,\Theta)$ 表示[8]。其中,G 为一有向无圈图,图中结点对应随机变量 $X=\{X_1,\cdots,X_n\}$,在微阵列数据中表示基因的表达向量。B 中另一部分 Θ,表示一组条件概率分布。根据马尔可夫假设:每个变量 X_i 在给定 G 中的父结点前提下各变量之间相互独立,于是得到随机变量 X 的联合概率分布:

$$P_r\{X_1,\cdots,X_n\}=\prod_{i=1}^{n}P_r\{X_1 \mid ancestors(X_1)\}$$

其中 $ancestors(X_i)$ 表示 X_i 的父结点集合。为了确定 X 的联合概率分布,需要确定上式中出现的各个条件概率,所有这些条件概率构成了 Θ。贝叶斯网络的核心就是通过将这种条件独立关系解释为因果关系,并用来表示基因间的因果调控关系。

DBNs 模型,可表示具有时序性的随机变量 X(即一组时间序列)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)。一组随机时序变量 $X(0), X(1), L, X(T)$的联合分布可表示为：

$$P_r\{X(0), X(1), \cdots, X(T)\} = P_r\{X(0)\}\prod_{t=1}^{T} P_r\{X(t) \mid X(t-1)\} = P_r\{X_i(0) \mid ancestors\{X_i(0)\}\times \prod_{t=1}^{T}\prod_{j=1}^{n} P_r\{X_j(t) \mid ancestors\{X_j(t)\}$$

DBNs 模型[15]的优点可概括为：(1)贝叶斯网络引入图模型，揭示了基于统计假设的基因表达水平中的因果关系；(2)把线性、非线性模型和隐马氏模型作为特例涵括在内，考虑了随机元和隐变量的引入，能很好地处理隐变量；(3)可以对数据采集过程进行清晰建模，能处理数据缺失问题及数据噪声，还能估计网络不同特征的置信度。但是在基因网络结构的研究中，贝叶斯网络方法仍有其局限性：有向无环结构的假设与生物体的生命周期现象并不符合。其次，在贝叶斯网络的学习过程中，网络的结构会非常的复杂，计算量非常之大。

3　结　语

生命科学近年来获得突破性进展，数据挖掘以其丰富、灵活的分析功能和强大的分析能力为生物数据的分析解决了瓶颈之难，在生物信息学领域具有良好的研究与应用前景。基因的网络分析是生命信息挖掘的重要手段之一，但目前在许多方面尚处于尝试和探索阶段。大量模型不断涌现，各种数学工具不断引进，这为网络调控模型的构建创造了良好的数学理论基础。随着基因数据的不断扩展以及数据质量的进一步提高，基因调控网络建模的准确性将得到进一步提高。随着后基因组学的不断发展，基因网络必然会在生命科学的研究中发挥巨大的作用。

[参 考 文 献]

[1] Styczynski MP, Stephanopoulos G. Overview of computational methods for the inference of gene regulatory networks, computers and chemical engineering [J]. 2005, 29: 519-534.

[2] D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering [J]. Bioinformatics, 2000, 16: 707-726.

[3] Wagner A. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps [J]. Bioinformatics, 2001, 17: 1183-1197.

[4] Bansal AK, Koradia V. The role of reverse engineering in the development of generic formulations, pharmaceutical technology [J]. Pharmaceat Technol, 2005, 29: 50-54.

[5] Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets [J]. J Theor Biol, 1969, 22: 437-467.

[6] Kauffman S. The lager scale structure and dynamics of gene control circuits: an ensemble approach [J]. J Theor Biol, 1974, 44: 167-190.

[7] Yuh C, Bolouri H, Davidson EH. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene [J]. Science, 1998, 279: 1896-1902.

[8] Lhdesmki H, Hautaniemi S, Shmulevich I, et al. Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks [J]. Signal Processing, 2006, 86: 814-834.

[9] Kato M, Tsunoda T, Takagi T. Inferring genetic networks from DNA microarray data by multiple regression analysis [J]. Genome Informatics, 2000, 11: 118-128.

[10] Quackenbush J. Computational analysis of microarray data [J]. Nat Reviews Genetics, 2001, 2: 418-427.

[11] 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.

[12] Friedman N. Inferring cellular networks using probabilistic graphical models [J]. Science. 2004, 303: 799-805.

[13] Rice JJ, Tu Y, Stslsvitzky G. Reconstructing biological networks using conditional correlation analysis [J]. Bioinformatics, 2005, 21: 765-773.

[14] Bickel DR. Probabilities of spurious connections in gene networks: application to expression time series [J]. Bioinformatics, 2005, 21: 1121-1128.

[15] Albert I, Albert R. Conserved network motifs allow protein-protein interaction prediction [J]. Bioinformatics, 2004, 20: 3346-3352.

[16] Baldi P, Brunak S. 生物信息学-机器学习方法 [M]. 张东辉，黄　颖，蔡　军 等译. 北京：中信出版社，2003：263-282.

[收稿日期] 2006-01-19　　**[修回日期]** 2006-05-25

[本文编辑] 尹　茶

致 读 者

因世界卫生组织(WHO)正在对国际临床试验注册平台(ICTRP)公告中文版作最后修订，本刊原拟第 7 期刊出的 ICTRP 公告中文版延至下期刊出，敬请读者关注。