临界布尔网络的函数泛化问题研究

于雄香[1],沈良忠[2],尚学群[3],刘文斌[1]
(1.温州大学物理与电子信息工程学院,浙江温州 325035;2.温州大学城市学院,浙江温州 325035;
3.西北工业大学计算机科学与技术学院,陕西西安 710072)

摘 要: 布尔网络是研究基因调控网络的一种非常重要的模型,通过时序数据推理基因之间的调控关系是研究网络动态行为和干预策略的基础.现有的预测研究主要集中在基因之间的调控关系,而对调控基因与目标基因之间的布尔函数的作用方式研究甚少.由于基因调控网络是一种处于有序和无序之间的临界网络,本文研究了众数规则、基于偏斜和基于互信息的三种泛化方法对临界布尔网络的稳态分布距离和灵敏度误差的影响.结果表明合理的泛化能够明显提高预测网络的稳态分布距离和灵敏度误差指标.三种泛化方法中,基于互信息的泛化方法的总体性能最好.

关键词: 基因调控网络, 布尔网络, 动态行为, 泛化
中图分类号: TN911 **文献标识码:** A **文章编号:** 0372-2112 (2015)10-2076-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.10.029

Study on the Generalization of Boolean Functions in Critical Boolean Networks

YU Xiong-xiang[1], SHEN Liang-zhong[2], SHANG Xue-qun[3], LIU Wen-bing[1]
(1. *College of Physics & Electronic Information Engineering, Wenzhou University, Wenzhou, Zhejiang* 325035, *China*;
2. *City College, Wenzhou University, Wenzhou, Zhejiang* 325035, *China*;
3. *School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an* 710072, *China*)

Abstract: Boolean network has been a major model to study gene regulatory networks. Lots of work have been focused on inferring networks from time-series data and designing potential intervention policies. However, one important problem still remains unsolved, that is the generalization of Boolean function. In general, the inference algorithms always assume a random Boolean value for the unobserved states. As many theoretical and experimental results support that gene regulatory networks lie between the boundary of ordered and disordered regimes, we studied three generalization methods: the majority rule, bias-based and mutual information-based methods. Results both on simulation networks and melanoma network show that reasonable generalization can improve both the steady-state distribution distance and the sensitivity error. And among the three methods, the mutual information-based method performs better than the other two.

Key words: gene regulatory network; Boolean network; dynamic behavior; generalization

1 引言

基因之间的调控关系与生物体的生长发育、疾病与衰老等生命现象密切相关.近年来,随着各种高通量生物技术如微阵列,CHIP-chip 等发展,基因调控网络的预测、动态行为分析及干预的研究取得了很大的进展.常见的基因调控网络模型主要有:布尔网络、贝叶斯网络,及微分方程模型等[1].由于高通量技术的数据中往往存在大量噪声,加之网络中通常包含成千上万个基因.以微分方程为代表的细粒度模型仅适合于研究几个基因之间动态变化过程.1960 年,Kauffman 提出了一种粗粒度的模型——布尔网络研究基因之间调控关系[2].他将基因的表达状态简化为"0"(抑制或不表达)和"1"(激活或表达)二种状态.这种模型的优点是模型简单,对噪声不敏感,同时适合于研究比微分方程模型更大的系统.

布尔网络的预测是一个极具挑战性的逆向工程问题,特别是在小样本情况下.在网络预测方面,人们已经

收稿日期:2014-06-06;修回日期:2015-01-16;责任编辑:郭游
基金项目:国家自然科学基金(No.61572367,No.61272018,No.61174162,No.61332014);浙江省自然科学基金(No.R1110261,LY13F010007)

提出了各种预测方法,如基于互信息的 Reveal[3],ARACNE[4],基于模型复杂度约束的最小描述长度(MDL)[5-9],基于信号处理的确定性系数(CoD)[10,11],以及基于一致性最佳扩展(Best-fit Extension)[12,13]. Qian 等对已有各种预测方法的比较发现,Best-fit 方法在所有算法中的总体性能最好[14]. Shumlevich 在布尔网络模型基础上加入了外部噪声以及系统不确定性,提出了更一般的概率布尔网络模型,从而为基因调控网络的干预研究奠定了理论基础[15]. 人们先后提出了基于平均首次到达时间(MFPT)[16,17],基于网络吸引域(BOA)[16],以及基于稳态分布距离(SSD)[16,18]的干预方法及线性最优干预方法(UC)[19]. 这些方法都能够改变系统的期望状态和不期望状态的概率分布.

布尔网络的预测主要包括目标基因的调控基因集合(即调控关系)及作用的布尔函数(即调控方式)两方面. 当目标基因的调控基因确定后,就必须确定相应的布尔函数. 在小样本数据情况下,调控基因集合的状态可以分为二种:已观察状态和未观察状态. 前者的输出可以根据样本数据直接确定,除在文献[20]中使用众数规则泛化之外,已有的各种预测方法都是随机泛化未观察状态. 在布尔网络中,布尔函数直接决定网络的演化方式和动态行为. 因此,函数的泛化直接决定系统的动态行为. 稳态分布和灵敏度是反映布尔网络动态行为的两个指标. 前者刻画在一个较长的时间,每个状态的出现概率;后者则是反映网络运行过程对扰动的稳定性. 已有的研究表明:基因调控网络处于有序和无序之间的临界状态,这种网络在大部分噪声环境下能保持一定的稳定性,同时又能对某些环境变化做出适当的调整,具有适应性[1]. 因此,本文以 Best-fit 方法为基础,研究了众数规则、基于互信息和函数偏斜三种泛化方法对临界布尔网络的稳态分布距离和灵敏度误差的影响.

2　相关概念

2.1　布尔网络

一个布尔网络 $G(V,F)$ 由节点集 $V=\{x_1,\cdots,x_n\}$ 和函数集 $F=\{f_1,\cdots,f_n\}$ 组成,其中 $f_i:\{0,1\}^{k_i}\to\{0,1\}$ 为基因 x_i 的布尔函数,k_i 指调控 x_i 的基因的个数,通常也指 x_i 的入度. x_i 在 $t+1$ 时刻的状态完全由其调控基因 $x_{j1},x_{j2},\cdots,x_{jk_i}$ 在 t 时刻状态确定,可以写为

$$x_i(t+1)=f_i(x_{j1}(t),x_{j2}(t),\cdots x_{jk_i}(t)). \tag{1}$$

布尔网络的演化通常采取同步演化的方式,网络状态分为暂态和吸引子二种. 当系统进入到某些状态后将会无限期的停留的状态称为吸引子. 进入吸引子的(暂时)状态构成了一个吸引子的吸引域. Kauffman 认为吸引子可能对应细胞的某种显型(Phenotype)或细胞分化过程的不同阶段等. 2006 年哈佛大学的 Huang 通过在白血病细胞株 HL-60 分别加入 DMSO 和 atRA,七天后二个细胞群体达到相同的状态[21]. 在试验上证明了细胞状态演化轨迹及吸引子的存在. 如果每个状态可以以一个很小的概率 p 随机转入任何其它状态,布尔网络就变成一种最简单的概率布尔网络——有扰动的布尔网络(BNp). 这种网络的运行实质上是一个马尔可夫链,其动态行为则由其稳态分布来描述.

2.2　灵敏度

在布尔函数中,有些调控基因对目标基因的影响大,有些则较小. 变量 x_j 在函数 f_i 中的活性可以衡量这一影响的大小,其定义为

$$\alpha_j^{f_i}=\frac{1}{2^{k_i}}\sum_{x\in\{0,1\}^{k_i}}\partial f_i(x)/\partial x_j,0\leqslant\alpha\leqslant 1, \tag{2}$$

$\partial f_i(x)/\partial x_j=f_i(x^{(j,0)})\oplus f_i(x^{(j,1)})$ 表示 f_i 对变量 x_j 的偏微分,$\oplus$ 是模为 2 的加法,$x^{(j,k)}=(x_1,\cdots,x_{j-1},k,x_{j+1},\cdots,x_k)$,$k=0,1$. 在均匀分布的情况下,活性就等于偏导数的期望. 布尔函数 f_i 的灵敏度为其中每个布尔变量的活性之和

$$s^{f_i}=\sum_{j=1}^{k_i}\alpha_j^{f_i}. \tag{3}$$

整个布尔网络的灵敏度是其中所有布尔函数灵敏度的平均

$$S=\frac{1}{n}\sum_{i=1}^{n}s^{f_i}. \tag{4}$$

此外,当给定网络的平均连通度 K 和布尔函数的平均偏斜率 p 时,网络的灵敏度还可以表示为:

$$S=2Kp(1-p). \tag{5}$$

网络的灵敏度是一个全局动态行为参数,它反映了 1 比特的扰动对于网络演化的影响:当 $S<1$ 时,扰动将逐步消失,网络处于有序状态;当 $S>1$ 时,扰动将逐步扩散,网络处于无序状态;当 $S=1$ 时,网络处于有序和无序之间的临界状态[22].

2.3　最佳一致扩展

布尔网络的预测可以看作是一个一致性问题:即寻找一个与数据完全一致性的调控关系及布尔函数. 由于数据噪声及生物系统的不确定性,完全一致的布尔函数往往并不存在. 因而我们通常是将完全一致约束放松为一个最佳一致扩展的问题. 在最佳一致扩展问题中,一个部分定义的布尔函数 $g_{T,F}$ 由两个集合 T,$F\in\{0,1\}^n$ 定义,T 代表“正例”向量集合,F 代表“反例”向量集合. 如果 $T\subseteq T(f)=\{x\in\{0,1\}^n:f(x)=1\}$ 且 $F\subseteq F(f)=\{x\in\{0,1\}^n:f(x)=0\}$,则函数 f 是 $g_{T,F}$ 的一个扩展. 函数 f 的误差大小为

$$\varepsilon(f)=T\cap F(f)+F\cap T(f). \tag{6}$$

最佳扩展问题的目标就是寻找满足 $T^* \cap F^* = \varphi$ 和 $T^* \cup F^* = T \cup F$ 的子集 T^* 和 F^*，使得在某个函数类 C 上，存在部分定义函数 g_{T^*,F^*} 的一个扩展，并且 $T^* \cap F + F^* \cap T$ 最小. 显然，满足上述条件的部分定义函数 g_{T^*,F^*} 的任意一个扩展 $f \in C$ 具有最小误差[12,13]. 基于最佳一致扩展的推理的预测方法称为 Best-fit.

2.4 互信息

互信息是一种度量二个变量之间非线性依赖关系的测度，在基因调控关系的预测中有广泛的应用. 给定两个离散随机变量 X 和 Y，X 对 Y 的时延互信息(one-time-lag MI)为

$$I(Y_{t+1};X_t) = H(Y_{t+1}) - H(Y_{t+1} \mid X_t). \tag{7}$$

其中 $H(\cdot)$ 表示熵，X_t 和 Y_{t+1} 为两个等长向量. 这里时延互信息 $I(Y_{t+1};X_t)$ 具有方向性，其值越大，说明 X 对 Y 的影响越大[5,23].

3 三种泛化方法及评价准则

3.1 泛化方法

在没有任何生物学先验知识的情况下，布尔函数的泛化必须依据观测状态的数据信息，如调控基因与目标基因的相关性，数据中目标基因的表达与关闭的概率等. 临界布尔网络中的布尔函数如渠化函数、Post 类函数等通常具有保守性[24]. 这类布尔函数的一个主要特点是具有较高的偏斜度，即函数中 0、1 的比例远离 0.5. 基于这一认识，可以有二种泛化方法. 一种是众数规则(majority rule)，即根据表达数据中目标基因表达值 0、1 的比例，如果有 50% 及以上的数据目标基因表达值为 0，则将未观测状态的输出设为 0；反之，将其设为 1. 另一种方法是偏斜率(bias)，即根据表达数据中目标基因表达值 0、1 的比例，随机泛化未观测状态的输出. 此外，由于互信息(mutual information)能够反映两个基因之间相互作用的强弱，我们可以根据这一信息泛化未观测状态的输出. 具体方法为：对目标基因 x_i，计算每个调控基因 $x_{ij}(1 \leqslant j \leqslant K)$ 与目标基因的互信息. 选择与目标基因互信息最大的调控基因 $x_{ik}(1 \leqslant k \leqslant K)$，统计观测状态中 x_{ik} 为 0(或 1)时 x_i 的最有可能的输出状态. 最后，根据 x_{ik} 的状态泛化未观测状态的输出.

显然，众数规则是一种确定性的泛化方法，它对未观测状态采取的是一种“一刀切”的泛化策略. 基于互信息的泛化方法则是一种半确定性的泛化策略，它是依据其中一个互信息最大的调控基因的状态来泛化未观测状态. 而基于偏斜的泛化则是一种随机泛化的策略，它是基于数据中目标基因状态的偏斜度来随机泛化未观测状态.

3.2 评价准则

评价网络预测的性能指标有各种方法，由于函数的泛化主要影响网络的动态行为，本文主要采用了稳态分布距离和灵敏度误差两种指标.

1)稳态分布距离. 它反映预测网络在长期动态行为上与原始网络的接近程度. 其定义为

$$\mu^{ssd} = \sum_{k=1}^{2^n} \left| \pi_k - \pi'_k \right|. \tag{8}$$

其中 π_k、π'_k 分别表示原始网络和预测网络稳态分布中状态 x_k 的概率.

2)灵敏度误差. 它反映预测网络与原始网络在平均动态演化行为方面的接近度. 其定义为

$$u^s = \frac{|S - S'|}{S}. \tag{9}$$

其中 S、S' 分别表示原始网络和预测网络的灵敏度.

4 结果与讨论

为了研究临界网络的泛化问题，本文仿真网络的基因规模 $n = 10$. 我们分别产生了连通度 $K = 3,4,5$ 和 6 且灵敏度 $S = 1$ 的四种网络，每种网络的数量为 200. 然后由每个网络分别产生 N 为 10,15,20,30 的时序数据进行预测. 为了研究噪声的影响，我们分别对每个样本加入不同的噪声 η，其值分别是 0%，5%，10%. 稳态分布计算时假定网络每个基因的扰动概率 $p = 0.001$. 本文的计算使用 shumlevich 开发的 PBN 工具箱(http://code.google.com p/pbn-matlab-toolbox/).

4.1 仿真网络

图 1 和图 2 分别表示在给定连通度 K 和噪声 η 时，三种泛化方法的稳态分布距离和灵敏度误差随样本增加的变化情况. 可以看出，与原 Best-fit 方法的随机泛化相比，三种泛化方法均能够提高预测网络的二种性能指标. 首先，随着连通度 K 的增加，泛化对二种误差性能的提高越加明显. 连通度为 K 的布尔函数的所有输入状态的个数为 2^K. 在给定样本量 N 时，未观察状态的数量将随连通度 K 的增加而呈指数增加，从而泛化对二种性能指标的提高才越明显. 反之，当 K 较小时，泛化的作用较小，它们与随机泛化的差别不大(如 $K = 3$).

其次，互信息泛化方法比众数规则的总体性能稍好. 这说明基于互信息的泛化方法比众数规则这种“一刀切”的泛化方法好，因为它反映了基因之间调控关系的强度因素. 此外，这二种方法具有较强的抗噪能力. 由于众数规则根据样本 0、1 的偏斜将未观测状态统一泛化为 0 或 1，因此其对噪声不敏感. 至于互信息泛化，噪声可能会降低调控基因与目标基因之间的互信息，但对基因间的互信息的相对大小影响较小. 因而这种泛化方法也对噪声不敏感.

最后,偏斜泛化方法的性能对噪声较为敏感.当噪声 $\eta=0$ 时,偏斜泛化优于众数规则和互信息二种泛化方法;反之,随着噪声 η 的增加,其性能逐渐劣于其它二种方法.这一点可以从网络的灵敏度 $S=2Kp(1-p)$ 加以解释.在低噪声时,数据偏斜的期望就是布尔函数的偏斜 p,从而灵敏度误差就小,并进而导致稳态分布距离小.当数据噪声增大时,数据偏斜的期望就难以反映布尔函数的偏斜 p,从而导致其性能迅速恶化.

总之,通过泛化方法能够明显提高预测网络与实际网络的动态行为误差指标.以上讨论的三种方法中互信息泛化方法相对最好.

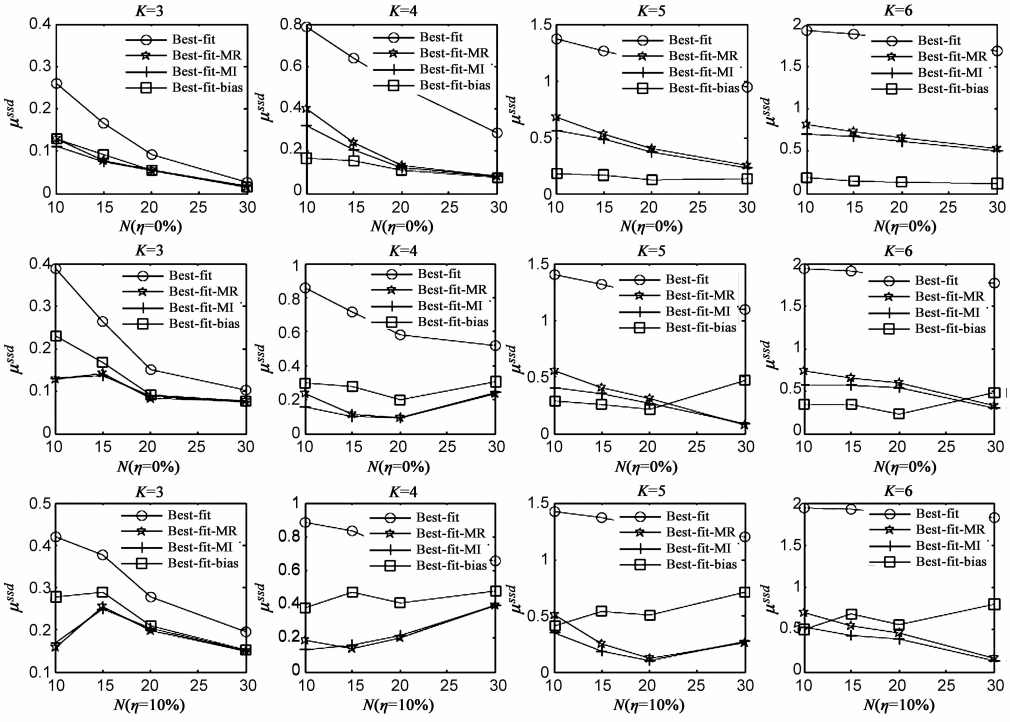

图1　仿真网络的平均稳态分布距

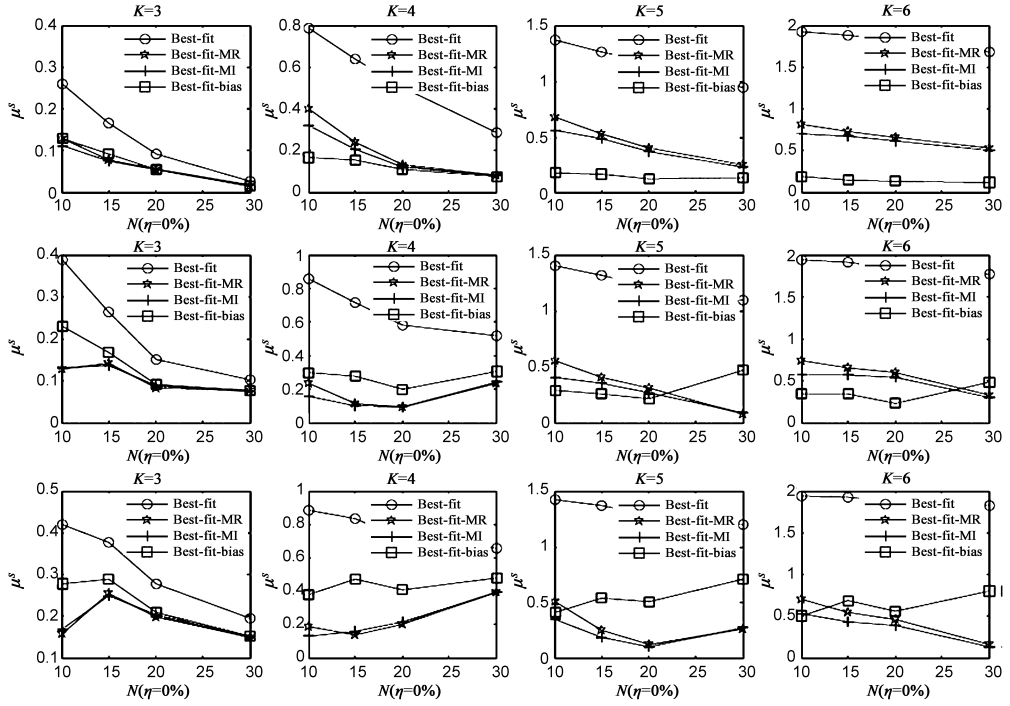

图2　仿真网络的平均灵敏度误差

4.2 黑素瘤网络

布尔网络模型已经广泛应用于各种生物系统,人们利用表达谱数据和生物知识建立了如酵母细胞周期表达、哺乳动物细胞周期表达、果蝇体节极性网络、花发育形态表达等细胞过程的调控网络模型并进行动态行为分析.下面我们以黑素瘤网络为例,研究泛化对其动态行为的影响.黑素瘤网络共包含 10 个基因,网络的平均连通度为 2.4[14].在用 Best-fit 方法预测时,黑素瘤网络的最大预测入度 K 为 3.我们产生 200 个 N 为 10,15,20,30 的时序数据,结果在图中.可以看出,两种性能与仿真网络的结果一致.由于连通度 $K=3$ 相对较小,泛化对稳态分布距离的改进相对较小.但是在噪声环境下,对灵敏度误差有比较明显的改进.

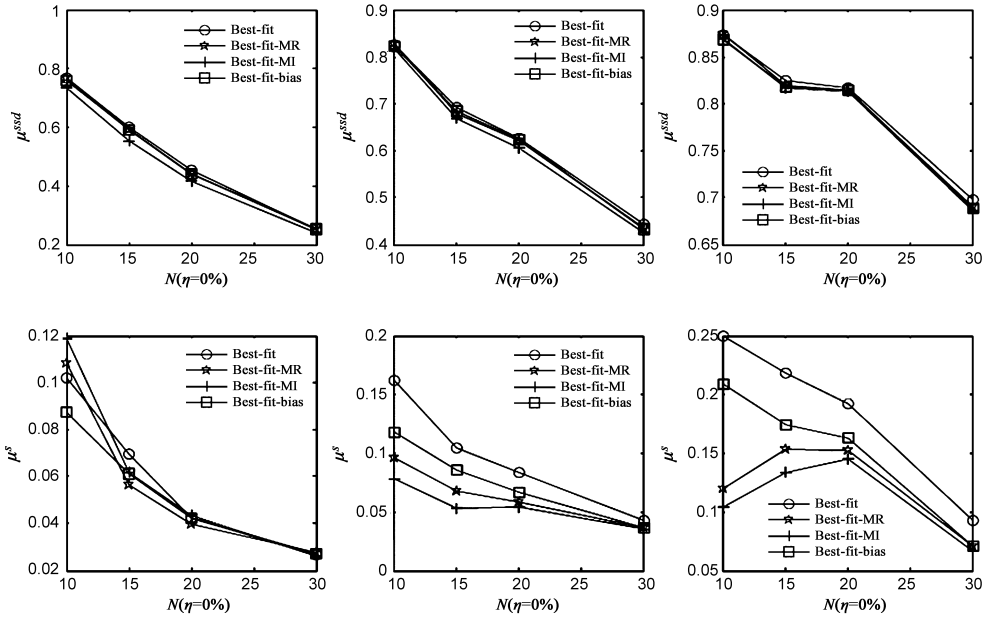

图3 黑素瘤网络的平均稳态分布距离和平均灵敏度误差

5 结束语

布尔网络已经成为研究基因调控网络的一种非常重要的模型,通过这种模型可研究系统的动态行为及干预策略.这些研究要求精确的推理出网络结构和作用方式.由于以往的研究主要集中在调控关系的推理,而对基因之间作用的布尔函数的泛化研究甚少,本文主要研究了三种泛化方法对临界网络的稳态分布距离和灵敏度误差的影响.仿真网络和黑素瘤网络的实验结果表明:众数规则、基于偏斜和互信息的三种泛化方法都能提高预测网络的两种动态行为指标.随着网络连通度的增加,泛化后的网络的平均稳态分布距离和灵敏度误差性能的改善逐渐明显.在三种泛化方法中,基于偏斜的泛化方法在没有噪声的情况下效果最好.众数规则和基于互信息的泛化方法的抗噪性能较好,且后者的总体性能在三种泛化方法中相对最好.最后,需要指出的是,在实际的生物网络中,有时还可以根据生物学先验知识来进一步提高泛化的性能,如在黑素瘤网络中,Wnt5 基因通常对其它基因具有很大的影响,我们可以利用这一信息对受其调控的基因采用众数规则泛化.本文主要研究泛化对调控网络的影响,但泛化对网络干预策略的获取会有怎样的影响呢,将是我们进一步要研究的方向.

参考文献

[1] 刘文斌,高琳.基因组信号处理[M].北京:科学出版社,2010.
Liu Wenbing, Gao Lin. Genomic Signal Processing[M]. Beijing: Science Press, 2010. (in Chinese)

[2] X Wang, Q Liu, Y Cheng, et al. Qualitative analysis of gene regulatory networks based on angular discretization[J]. Chinese Journal of Electronics, 2011, 20(4): 646 – 650.

[3] S Liang, S Fuhrman, R Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures[A]. Proceedings of International Conference on biocomputing[C]. Hawaii, United States: Pacific symposium, 1998, 18 – 29.

[4] A A Margolin, I Nemenman, K Basso, et al. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context[J]. BMC bioinformatics, 2006, 7 (Suppl 1): S7.

[5] V Chaitankar, P Ghosh, E Perkins, et al. A novel gene network inference algorithm using predictive minimum description length approach[J]. BMC Syst Biol, 2010, 4(Suppl 1): S7.

[6] V Chaitankar, C Zhang, P Ghosh, et al. Gene regulatory network inference using predictive minimum description length principle and conditional mutual information[A]. Proceedings of International Conference on Systems Biology and Intelligent Computing[C]. Oxford, England: Bioinformatics, 2009, 487 – 490.

[7] J Dougherty, I Tabus, J Astola. Inference of gene regulatory networks based on a universal minimum description length[J]. EURASIP Journal on Bioinformatics and Systems Biology, 2008, 2008: 5.

[8] W Zhao, E Serpedin, E R Dougherty. Inferring gene regulatory networks from time series data using the minimum description length principle[J]. Bioinformatics, 2006, 22(17): 2129 – 2135.

[9] I Tabus, J Astola. On the use of mdl principle in gene expression prediction[J]. EURASIP Journal on Applied Signal Processing, 2001, 2001(1): 297 – 303.

[10] S Kim, M L Bittner, K Sivakumar, et al. General nonlinear framework for the analysis of gene interaction via multivariate expression arrays[J]. Journal of biomedical optics, 2000, 5(4): 411 – 424.

[11] E R Dougherty, S Kim, Y Chen. Coefficient of determination in nonlinear signal processing[J]. Signal Processing, 2000, 80(10): 2219 – 2235.

[12] H Lähdesmäki, I Shmulevich, O Yli-Harja. On learning gene regulatory networks under the boolean network model[J]. Machine Learning, 2003, 52(1 – 2): 147 – 167.

[13] I Shmulevich, A Saarinen, O Yli-Harja, et al. Inference of genetic regulatory networks via best-fit extensions[J]. Computat Statis Approaches Genom, 2002: 197 – 210.

[14] X Qian, E R Dougherty. Validation of gene regulatory network inference based on controllability[J]. Frontiers in genetics, 2013, 4: 272 – 272.

[15] I Shmulevich, E R Dougherty, W Zhang. From boolean to probabilistic boolean networks as models of genetic regulatory networks[J]. Proceedings of the IEEE, 2002, 90(11): 1778 – 1792.

[16] X Qian, I Ivanov, N Ghaffari, et al. Intervention in gene regulatory networks via greedy control policies based on long-run behavior[J]. BMC Syst Biol, 2009, 3: 61.

[17] G Vahedi, B Faryabi, J Chamberland, et al. Intervention in gene regulatory networks via a stationary mean-first-passage-time control policy[J]. Biomedical Engineering, IEEE Transactions on, 2008, 55(10): 2319 – 2331.

[18] X Qian, E R Dougherty. Effect of function perturbation on the steady-state distribution of genetic regulatory networks: Optimal structural intervention[J]. Signal Processing, IEEE Transactions on, 2008, 56(10): 4966 – 4976.

[19] M R Yousefi, E R Dougherty. Intervention in gene regulatory networks with maximal phenotype alteration[J]. Bioinformatics, 2013, 29(14): 1758 – 1767.

[20] E R Dougherty, Y Xiao. Design of probabilistic boolean networks under the requirement of contextual data consistency[M]. Piscataway, United States: IEEE Transactions onSignal Processing, 2006: 3603 – 3613.

[21] S Huang, D E Ingber. A non-genetic basis for cancer progression and metastasis: Self-organizing attractors in cell regulatory networks[J]. Breast disease, 2007, 26(1): 27 – 54.

[22] I Shmulevich, S A Kauffman. Activities and sensitivities in boolean network models[J]. Physical Review Letters, 2004, 93(4): 048701.

[23] W Zhao, E Serpedin, E R Dougherty. Inferring connectivity of genetic regulatory networks using information-theoretic criteria[J]. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2008, 5(2): 262 – 274.

[24] I Shmulevich, H Lähdesmäki, E R Dougherty, et al. The role of certain post classes in boolean network models of genetic networks[J]. Proceedings of the National Academy of Sciences, 2003, 100(19): 10734 – 10739.

作者简介

于雄香　女,1990 年 2 月生于江苏南通,硕士研究生,主要研究方向为生物信息、数据分析. E-mail: yuxiongxiang2008@sina.com

沈良忠　男,1978 年 2 月,生于浙江海盐,硕士,副教授,主要研究方向为数据库应用、数据挖掘

刘文斌(通信作者)　男,教授,博士,1969 年出生于陕西韩城. 2004 年获华中科技大学博士学位,目前感兴趣的研究领域为生物信息学、数据挖掘、DNA 计算等. 获得省部级奖励 4 项,主持国家省部级项目 6 项,发表学术论文 40 余篇.
E-mail: wbliu6910@126.com