

生物信息前沿研究进展讲座结课论文

——基因表达调控网络研究文献综述

物理学院 张玉萍 10304830

摘要

近些年来，基因序列测序的完成、大规模测定基因表达水平的基因芯片（Microarray）技术的出现和高性能计算机的使用使得用模拟计算的方法大规模的研究基因表达调控成为可能，一些研究者已经开始绘制控制整个活细胞基因表达的调控网络。例如 λ 噬菌体的溶原/裂解活性的调控网络的数学模型已经构建出来。用数学模型的方法预测网络结构是目前研究的热点。本文对表达转录调控网络的研究现状进行综述。

基因表达调控网络

Wyrick(2002)[1] 中给出了一个基因表达调控网络的定义：一组调控因子如何调控一套基因表达的过程称为基因表达调控网络。基因表达调控网络是基因调控网络的一个重要部分。参与基因表达调控网络的元素主要包括cDNA、mRNA、蛋白、小分子等。从元素间相互联系的角度来看，基因表达调控网络是一个由节点（调控元素）、边（调控作用）组成的一个有向图结构。如图1

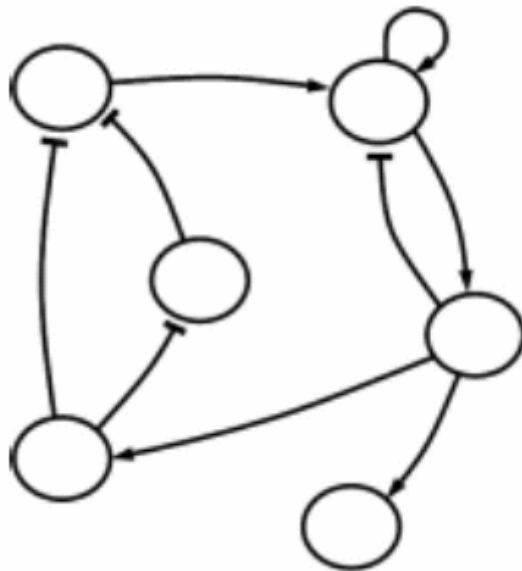


图1：简单基因网络结构示意图

图中每一个圆圈代表一个节点，也就是调控网络的元素，如基因。有向箭头表示表达增强作用，末端断线表示表达抑制作用。在基因网络中，存在基因对自身表达的自调控的现象。

总的来说表达调控网络有如下特点：

A: 网络结构复杂

网络中节点和边的数目庞大。在人体中总共有3万到4万左右的基因，而且真核生物中大多数的基因会同时被两个和两个以上的基因调控，这就使网络形成了一个非常高维的结构。

B: 网络结构变化

生物学的实验表明，相同的基因在人和动物的细胞周期中可以参加不同的生理过程，实现不同的生理功能。还有一些基因只在某些时刻和特定的外界条件下是有相互作用的，在其他条件下不会发生作用。简单的说就是两个基因间的那条边是否存在、作用的方向在不同时期是可能不一样的。

C: 相互作用类型多变

在生物体中，基因间相互作用可以有很多类型（如图1），包括了很多作用的特征：两个基因间谁影响谁、影响的方式、增强的作用还是抑制的作用、影响产生的条件、影响的强弱量级、被调控基因的表达量和调控基因的表达量直接的关系等。目前的研究表明，基因间的相互作用可能是一种非线性的作用关系。在多因子调控模式中还要考虑不同的调控因子对同一个目标调控基因产生作用时的某种逻辑关系，这种逻辑关系是由调控模式中各调控因子的相互关系决定。

D: 节点类型多样

网络节点的元素可以是DNA、mRNA、蛋白、分子、大分子、外界环境等等。

E: 节点状态变化

在细胞周期过程中，每一个基因的表达量不是固定的，会随着条件的变化而变化、蛋白质在不断的合成，同时也在不断的被降解。在不同的调控模式下，蛋白合成和降解的比率会发生变化，从而使蛋白处在不同的水平上。基因的表达量的变化会影响到相互作用的变化，会引起网络结构的变化。

F: 有向循环结构

在生物体中各种生理上的周期现象，我们很容易理解生物体中的相互作用存在周期性。至少在网络的局部上是循环的。在已经研究的比较多的低等生物E.coli的表达调控网络[2]中已经发现了循环的结构。

表达转录调控网络的研究现状

目前关于基因调控的绝大部分问题还没有解决。除了生物学家努力通过新的实验技术和生物理论来研究问题外，近几年，利用数学、统计学、神经网络、人工智能等方法在计算机上分析模拟表达调控机理，是计算分子生物学方面一个飞速发展的方向。由于分析模型的不同和采用的数据类型的差异，目前研究主要分为两个方面：基于基因芯片数据的关系推断模型和基于基因序列信息的调控因子结合位点推断模型。

下面分别就这两个方面的一些方法做一个简要介绍。

（一）基于基因芯片数据的关系推断方法

基因芯片的数据形式为：

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

图2

矩阵 X 中每一行代表一个基因，每一列代表一张芯片（样本）上基因的数据。 x_{ij} 为基因 i 在实验（条件） j 中的表达值。由基因芯片的实验原理， x_{ij} 取为相对的荧光强度的比值：

$$x_{ij} = \log_2 \frac{I_R}{I_G}$$

I_R 为芯片上样本组基因（红色荧光剂）的强度， I_G 为芯片上对照组基因（绿色荧光剂）的强度。在芯片数据的后期处理过程中可以对每张芯片内的全部基因的表达值做归一化处理，去除芯片的背景噪声。目前利用基因芯片数据做分析推断的模型不少，主要包括有向图模型、贝叶斯网络、布尔网络、微分动力模型、随机微分方程、神经网络的方法等等[3]。下面简要介绍一下研究比较多的聚类分析方法和贝叶斯网络模型。

A. 表达数据聚类分析方法

聚类是探索性数据分析和模式发现的一种基本手段，其目的是提取数据中隐含的类别结构。但是，聚类是一个模糊的概念，没有一个准确的定义。已知有几十种聚类算法和大量的专门聚类程序被用于DNA微阵列数据的分析，其类型涵盖了分级聚类、k均值聚类等[4][5][6][7]，它们没有一个显而易见的共同点。由于聚类问题的多样性和“开放性”，不大可能给出聚类的一个系统化的完备处理框架，聚类算法之间的一个重要差别在于他们是有监督的还是无监督的。在有监督聚类中，聚类基于一个给定的参考向量集或类别集。在无监督聚类中，没有一个事先给定的向量集和类别集。目前，由于基因转录的调控模式并不清楚，像k均值和自组织映射（SOM）[8]，这样的无监督聚类方法是转录关系研究中最常用的。

在聚类算法中，距离的定义非常关键，可以在很大程度上影响聚类算法的结果。根据适用情况的不同，每种距离都有自己的优缺点。Pearson相关系数能够反映表达模式形状的相似性但不强调两组测量的数值关系，对偏差比较敏感。而欧式距离可以反映两者在数量关系上的差异，不强调形状的相似性。

聚类类别数 K 的选择是非常棘手的问题，它取决于我们在什么尺度上观察数据，对于聚类问题的严格讨论，需要预先给出一种原则性的方法来比较同一数据集不同聚类结果。需要一个易于计算的全局代价/误差函数。聚类的目标就是最小化这一函数。然而，没有普遍适用的函数，代价函数必须根据具体的问题来决定，不同的代价函数会导致不同的结果。

分级聚类通过计算两两距离从数据中自动建立一棵树而非一组类别。如何从树中定义类别的方法并不明显。因为类别是通过在树的某些节点剪枝得

到的，然而并没有一种好的方法给出剪枝的标准。K均值聚类法是在固定类别数K的前提下，通过迭代计算类的成员和类的中心（代表点），直到系统收敛或涨落很小。当代价函数与一个隐含的概率混合模型[9]相对应时，K均值聚类法是经典EM算法的一种线性近似，而且一般会收敛到一个解。

每种聚类方法都有各自适用的环境和优缺点。在几千的数量级上分级聚类得到的树状结构非常复杂，很难看清楚类别的边界。而K均值聚类的主要问题是聚类结果缺乏稳健性。类别数K的选取会对结果产生很大的影响。其次是对于噪声的敏感程度，由于是以类别中所有元素的平均值作为代表点，不可避免的会受到噪声（飞值）的很大的影响，聚类结果容易出现波动。

B 贝叶斯网络方法

Friedman et al(2000) 中提出了用贝叶斯网络模型分析基因表达数据的方法。在假定整个网络结构的无环和基因表达量之间的条件独立性的前提下，以每个节点为变量，描述网络在这样一组变量上的概率分布。

考虑一个有限的随机变量集合 $X = \{X_1, \dots, X_n\}$ ，每个随机变量 X_i 可以从一个集合 $(Val(X_i))$ 中取值 x_i 。一个贝叶斯网络图描述了一个联合概率分布，用符号表示就是： $B = \langle G, \Theta \rangle$ ，其中 $G = (V, E)$ 是一个有向无环图，包含所有节点V和有向边的连接E。 Θ 代表了量化网络的参数集合。

假定各个变量 X_i 的父节点和它的非后代节点是独立的。对于每一个变量 X_i 的一个可能值 x_i 以及它在G中的父节点集合 $Pa(X_i)$ 的一个可能（向量）值 $Pa(X_i)$ ， Θ 里面包含了一族参数

$$\theta_{x_i|pa(X_i)} = P(X_i|Pa(X_i))$$

一个贝叶斯网络可以确定在X上唯一的联合概率分布，如下给出：

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i|Pa(X_i))$$

通过给定X的一个样本 $D = \{X_1, \dots, X_N\}$ ，找出一个和D最为相配的网络结构。

一般是通过一个打分函数，利用打分函数在条件独立性的条件下可分解性：

$$Score(G : D) = \sum_i Score(X_i|Pa(X_i) : N_{X_i, Pa(X_i)})$$

采用局部搜索的方法寻找使得得分增加的路径。使得最后得到的网络结构的得分是全局最大的。

在基因网络结构的研究中，贝叶斯网络方法有其局限性：有向无环结构的假设与生物体的生命周期现象并不符合。生物体中出现的周期现象如蛋白的合成到分解是周而复始循环出现的，这就说明生物体中的基因网络结构应该是一个有向、有环的网络结构。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。假设有N个节点的话，网络里面可能会有 C_N^2 条边，最少是没有边。对于

有向边的网络结构，可能的结构数目超过 2^{C_N} 之多。要在如此多的网络结构中寻找最大的得分结构，计算量是非常之大的。最后，从数据的角度看，表达数据间并不包含因果关系，只有相关关系，只有在贝叶斯网络的特定结构下，才能给出基因间因果关系的推断。

(二) 基因序列调控区的信息挖掘

基因表达调控的另一个非常重要的研究方面是调控区域的数据挖掘。调控因子通过基因启动子区域的转录结合位点TFBS结合到DNA序列上去，对基因表达的过程产生调控作用。寻找转录因子的DNA结合位点和其他的调控motif（一段具有特色功能的生物序列），对于研究基因的调控是有重要意义的。

转录调控机器需要“安装”在基因DNA上，才能对基因的表达调控起作用。在数以千计的基因中，转录因子如何识别出目标调控基因，目前还不清楚。一个比较直观的理解是，转录机识别出目标调控基因的一段有“特异性”的序列或序列的组合并结合到上面，而这段序列的特异性是其他的非目标调控基因的序列中所没有的。

一定程度上，这种有生物意义的特征序列（motif）的搜索可以在基因组的层次上用纯计算的工具进行。目前在这方面有不少的模型和算法用于实现motif的预测，包括AlignACE等[10]。其基本思想是计算每种长度为N的词（N-mers）在全基因组或基因组的一个特定区域（一般是基因转录起始点上游区域）里出现的次数。由生物学的知识可以知道，一般有意义的调控motif的长度在5~25bp范围内，太大的N会使得目标motif在大部分得基因序列中出现的次数为0，使得对于motif出现频率分布推断得有效样本量太小，缺乏统计上的可靠性。所以，在一般的问题研究中，N的值通常为6到10。出现次数超过一般水平的词成为超频词（overrepresented N-mers）。超频词构成了很多已知的调控motif。超频词的分布也包含很多的信息。如何找到一个好的统计背景模型来估计超频词也是很重要的。目前用计算的方法找到的motif，只有一小部分能在TRANSFAC数据库和文献中找到，其他部分motif有待于未来实验检验。

在表达数据的聚类方法之外，由于有了这样一些预测转录结合位点序列的方法，Blussemarker et al[11] 提出了一种调控因子结合位点序列和基因表达数据的线性关系模型：

$$A_g = C + \sum_{\mu \in M} F_{\mu} N_{\mu g}$$

A_g 是基因g的表达数据。 $N_{\mu g}$ 代表结合位点序列 M_{μ} 在基因g的上游启动子区域出现的次数。集合M代表一组显著的和某些调控因子结合位点有关的特征序列。C是一个常量， F_{μ} 代表序列 M_{μ} 对于基因表达 A_g 的贡献系数。通过逐步回归的方法可以得到基因和每个特征序列的相关系数 F_{μ} ，再通过特征序列把对应的蛋白和基因相关联，这样就可以推断基因表达和调控因子的关系。

这是较早提出结合表达数据和序列数据分析基因关系的一个模型。尽管模型中有着一些限制，如特征序列和表达数据间的线性关系以及特征序列之间作用的独立性，并没有什么可靠的依据。而且目前知道特征序列 M_{μ} 的调控因子很

少，不能给出基因和调控因子之间的关系。但是，这篇文章提出的结合表达数据和序列数据的思想是有意义的。因为表达数据有不确定性的特点。受实验的方法、条件和设备等一些外部条件的影响，表达数据的模式会有差异。而且实验中会有噪音，仅依靠表达数据并不能取得比较好的效果，而通过测序方法得到的基因序列相对于表达数据要稳定的多，调控模式体现在序列上的特征比表达模式中的特征更为稳定，但在调控因子结合位点不清楚的情况下，基因序列间的联系没有表达模式上的联系明显。结合序列和表达数据的方法不仅使预测的模型看上去更加可靠，而且可以起到相互验证的作用。

参考文献

1. Wyrick, J. J. and R.A. Young. Deciphering gene expression regulatory networks. *Curr Opin Genet Dev* 2002, 12(2): 130-6.
2. Shai S.Shen-Orr et al. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature.genetics* 2002, Vol 31, 64-68.
3. Hidde de Jong. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review, *JCB* 2002, Vol 9, 67-103
4. M.B.Eisen et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 1998, 95:14863-14868
5. U.Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 1999, 96:6745-6750
6. L.j.Heyer et al. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 1999, 9:1106-1115.
7. R.O.Duda and P.E.Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons 1973.
8. 页: 6
P.Tamayo, et al. Interpreting patterns of gene expression with selforganizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 1999, 96:2907-2912.
9. 页: 6
D.M.Titterington, et al. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York 1985.
10. 页: 6
J.D.Hughes et al. Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*. *J.Mol.Biol* 2002,296,1205-1214
11. 页: 6
H.J.Bussemaker, Regulatory element detection using correlation with expression, *Nature Genetics* 2001, Vol 27: 167-171