基因调控网络模型中的进化算法

原作：Khalid Raza， Rafat Parveen

翻译：王兵

摘要：

基因调控网络（GRN）在我们理解生物过程（Biological Process）中扮演了一个重要的角色。基因调控网络的建模在揭示基础细胞进程，测试基因功能以及理解它们之间发杂的关系中有着举足轻重的作用。而理解基因间相互作用有助于开发更有效的药物以及更精确的诊断，尤其是那些已知的跟基因的异常表达相关的疾病。在这篇论文中，我们例举了现在各种基于进化算法的基因调控网络建模并就此讨论了这个领域的挑战与机遇。

关键词:

基因调控网络，基因网络反向工程，网络推断，进化算法，遗传编程

介绍：

随着DNA微阵列技术的发展，有效而快速的测量成千上万的基因的表达量已经成为现实。并且还可以同时地在不同的条件下进行测量。这项技术现今被广泛地应用到了基因功能预测，药物开发，疾病诊断以及存活率测试等。系统生物学也因此成长为了一门独立的学科，集生物，化学，物理，数学，计算科学以及系统科学于一身对生物系统那惊人的相互作用进行系统性的研究。在过去的十年里，人们花费了大量的精力来开发各种计算方法从时序表达谱中推断基因网络。今天，从时序表达谱中反向工程(reverse-engineering)基因网络已经成为了理解生物系统相互作用及其行为的一门非常重要的技术。有非常多的理论来建模，分析，重构，推断复杂的调控关系，并提供可供实验验证的假设。比如布尔网络模型(Boolean

Network)，微分方程(differential equations)，贝叶斯网络(Bayesian networks)，Petri网络(Petri nets)，机器学习法(machine learning)，遗传计算(evolutionary computing)等等。尽管如此，绝大多数的相互作用及其参数是不知道的 。所以如何精确地推断网络对于研究者们来说还是一个非常大的挑战

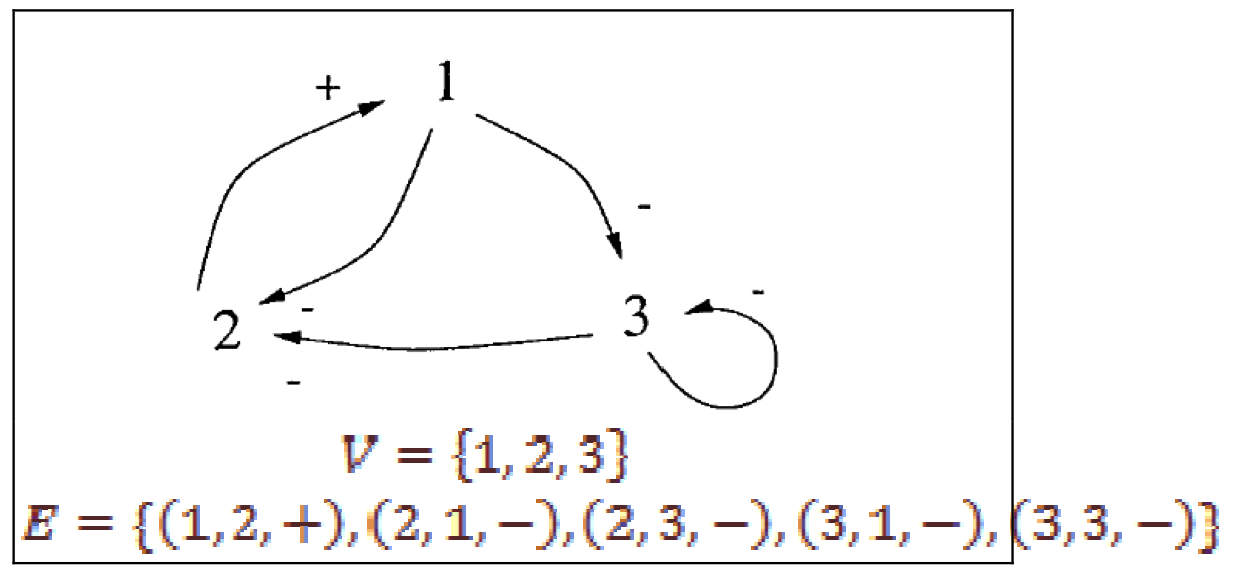
研究基因网络背后的动机在于描述生物体中基因间或者几组基因间是如何相互作用来实现复杂的调控关系的。基因网络有助与我们理解各种在各种刺激或者环境下基因间的不同相互作用。因此基因网络的能从系统的角度发现用其他方法所不能发现的生物属性。同时，基因网络能够预测网络的动态行为并与后续的实验结果相互印证，以此来测试模型的准确度和微调模型。系统生物学是一个高速发展的交叉学科，其目标在于解码细胞产物间的互相作用。

目前已经有非常多的数学模型和算法来预测基因网络。一般来说，基因调控网络的建模有四个步骤：1）选择一个合适的模型2）从数据中推断参数3）验证模型4）输入特定的环境变量预测基因调控网络行为。基因调控网络建模中基因是一个表达量随时间变化而变化的变量。这篇文章接下来的结构如下：第二部分介绍一些基本的建模技术，比如布尔网络，广义贝叶斯网络，线性和非线性微分方程，Petri网络，模糊逻辑网络，人工神经网络。第三部分介绍进化算法。第四部分介绍进化算法及其变种在基因调控网络建模中的应用。最后第五部分通过比较各种进化算法的性能做一个总结并展望未来的挑战以及发展趋势。

第二部分：建模技术

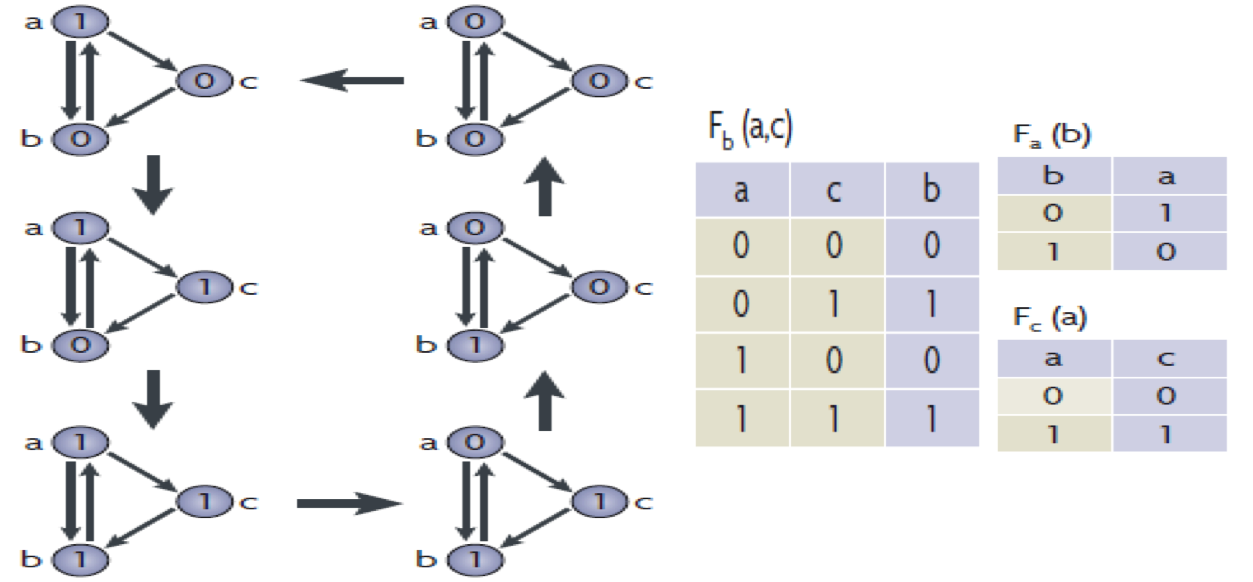
有非常多的模型从表达谱的数据来推断基因调控网络，包括有向图模型，布尔网络，广义贝叶斯网络，线性和非线性微分方程，Petri网络，模糊逻辑网络，人工神经网络等等。

有向图(Directed Graph)是所有基因网络中最简单最直接的模型。在有向图中，节点(vertices)代表基因，边(edge)代表相互作用。一条有向边可以用一个三变量的元组表示(i, j, s) i代表目标节点，j代表源节点，s是一个布尔变量(+或者-)代表i是被j激活还是抑制。有向图模型可以用来预测生物进程。典型 的有向图模型如图一：



图一：有向图

布尔网络(Boolean networks) 是基于布尔逻辑的网络，每一个节点只有两种状态：完全表达(表示为“1”)，完全不表达(表示为“0”)。节点之间的互相作用可以用布尔方程来计算，即同过输入其相关的节点的状态计算当前节点的状态，因此网络可以从一个状态进入到另一个状态，布尔网络的优势在于它的简单性和有限状态。但是其缺点也是很明显的，比如无法观察到处于中间态的基因。同时，模型假设所有的状态转化是同时的这一点也不靠谱。在布尔网络中，如果有N个基因，那么布尔网络会有 2N种可能性。这个指数爆炸的状态使得分类和计算成为一个计算上和内存上无法企及的任务。图二展示了一个三个节点的布尔网络状态，右边的转换方程描述了模型的规则。细箭头表示每个节点的调控者，粗箭头表示时间步长。布尔网络使用有向无环图(directed acyclic graph) G = (V, E) 来表示网络节点及其相关的边的转换条件。这些模型估计多变量输入的参数来估计状态变化的概率分布。这个模型主要的优点在于可以处理基因表达的随机性。缺点是无法处理基因表达的动态性。动态贝叶斯网络比较好的解决了这个问题。



图二：带有调控方程的布尔网络

微分方程(Ordinary differential equations) 是动态网络模拟中最常用的技术。尤其是基因调控网络。微分方程允许更细致的网络动态的描述，通过明确的计算浓度的变化随时间的函数。比如RNA，蛋白质，以及其他小分子。调控关系则由功能和浓度的微分函数表示。尽管微分方程可以提供非常细节化的动态网络信息，但是它需要高质量的数据，以及酶促反应参数，因此微分方程仅适用于非常小的系统。更多细节参考引文[12]及[19]。

Petri网络(Petri net) 是一个非确定性动态网络模型。Petri网络是有向图网络的一个扩展。是一个已经比较完善的基因调控网络建模技术。Petri网络成功地模拟出了基因调控在生物系统中的状态。比如动态过程的简单定量描述。缺点是不支持层次网络结构，不适用与大型的网络模型。

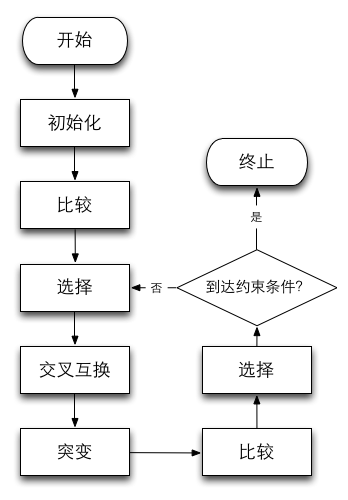
模糊逻辑(Fuzzy logic) 是一个软计算(soft computing)技术的概念，从工程界分离出来应用到科学界。模糊逻辑之所以能够适用于构建基因调控网络是因为1）模糊逻辑提取的信息是趋势而非数值，因此模糊逻辑天生具有屏蔽噪点数据的能力2）它的计算效率可以通过大规模的并行处理提升3）能够以用户友好的语言出现(比如 if-then 规则)。由于它表现非线性系统的能力以及容易理解的模糊规则使得其成功地被应用于基因调控网络建模。然而，最大的问题还是计算量，尤其是比较复杂的模糊模型。解决方案之一是预处理数据以降低计算时间，最常用的比如数据聚类(clustering)。目前已经有很多聚类的算法被提出来，包括层次聚类(hierarchical clustering)，自适应共振理论(adaptive resonance theory)，自组织图(self-organizing maps)，k-聚类(k-means)以及模糊 c-聚类(fuzzy c-means)等。

人工神经网络(Artificial neural networks) 原本是一个生物神经系统的数学模型。神经网络可以从大量数据中学习并且估计非线性方程，同时它的鲁棒性(robustness)也使得它有着较强的抗干扰性能。因此它可以被用来从表达谱数据构建基因调控网络。神经网络的结构拥有一堆节点以及相应的连接。基本上，每一个节点代表基因，但是节点还可以代表其他的因素。比如一个N维向量u(t) 可以代表N个基因组成的基因网络，而元素uj(t)是基因j在t时刻的表达状态。每一个连接代表了基因间的调控关系，同过一个权重矩阵w来定义。比如一个非零的权重矩阵wij可以表示一个j基因到i基因的连接。一个正值的权重矩阵表示一个正反馈，而一个负值的权重矩阵表示一个负反馈。一个零值的矩阵表示i和j之间没有调控关系。而i基因对j基因的控制能力则由wij和j基因的状态变量uj的乘积决定。因此，对于基因i来说，调控关系的总和为：

αi表示一个外部输入的影响，比如反应延迟(delay)。最后一个归一化函数将r(t)正规化。目前已经有各种各样的人工神经网络包括感知机(perceptrons)，自组织图(self-organizing maps)，周期性神经网络(recurrent neural networks)等被应用与基因调控网络的建模。

第三部分：进化算法基础

进化算法(Evolutionary algotithms)是一个基于达尔文进化论的优化算法。进化算法领域在最近几年高速发展。基本上，进化算法是一个模拟突变，遗传和自然选择机理的搜索算法。它结合了个体中适者生存的规律对一个结构化同时随机化的信息进行搜索。在进化算法最优化技术中，搜索完成于一个单独的点，每次迭代完成一次竞争性选择。在结果中更接近最优解的“高适应”度的点之间进行杂交，然后对一些元素给与一些单独的小突变。重组的主要目的是为了产生新的解，但这个重组又是偏向与那些已经被证明更优的解。大致上，可以分为三类进化算法的技术：遗传编程(genetic programming)，遗传算法(genetic algorithms)，进化编程(evolutionary programming)。遗传编程专注于程序本身的进化，遗传算法专注于那些重组问题的最优化，进化编程则专注于连续方程和不进行重组。进化算法是一类基于概率的算法，它们与随机算法不一样的地方在于他们的搜索结合了偏好性和随机性。基于这个原因，进化算法比定向算法鲁棒性更强。进化算法的另一个优点是它搜索的节点是一个群，而其他算法搜索的范围是一个点，因此进化算法有更大的搜索空间。进化算法的缺点是它对微阵列中常见的噪音点的影响处理不佳。图三显示了一个常规的进化算法流程。已经有不少进化算法被成功应用到基因调控网络的建模。



图三：常规的进化算法流程

第四部分： 进化算法在基因调控网络建模中的应用

有一些进化算法比如遗传编程，遗传算法，进化编程及其变种已经被应用到了基因调控模型的构建。例如Ando等的工作，结合遗传编程(Genetic Programming)和最小均方根(Least Mean Square)方法从时序表达谱中识别出最简调控网络。最小均方根方法在遗传编程中的应用是在不增加遗传编程的复杂度的前提下最小化观察值和预测值的均方根偏差(Mean Squared Error)。这个方法同过了实际数据集和人工数据集的测试，在十次运行中有大概4.21×10-3的均方根偏差，而同样情况下传统的遗传编程的均方根偏差有大约6.704×10-3。又比如Wang等的工作将遗传编程和卡尔曼过滤(Kalman filtering)相结合从时序表达谱推断基因调控网络。这个方法采用了非线性微分方程和迭代算法来构建模型。在这里遗传编程是用来产生模型的结构而卡尔曼过滤则用来估计迭代用的参数。这个模型被用来发现酵母蛋白质合成的网络构建。但是由于其采用的时序表达谱是人工整合的噪音信号较小，卡尔曼过滤器可能并不适用与分析较多噪点的数据。

Noman等开发了一个解耦的S-系统用以推断酶活系数，该系统采用了三角微分进化(Trigonometric Differential Evolution)作为优化引擎来捕捉时序表达谱中基因的动态表达。在这里，他们使用的适应函数是一个Kimaru的算法变种，用于最小化假阳性的比率。剩余的网络结构在进化算法架构下的爬山算法(hill-climbing local search)完成。在小规模人工构建的网络测试下，该算法在有噪音和没噪音的环境下表现都很不错。在现实网络，大肠杆菌DNA紧急修复网络的重构上，这个模型正确地识别出了lexA基因调控关系以及一些其他已知的调控关系。Chowdhury和Chetty等在Noman的基础上又开发了双S-系统，在这个系统中，进化算法被用来对网络的一些有用的特征进行打分使网络的推断更加精确，比如预测初始化算法(Prediction Initialization)，一个跳跃操作(Flip Operation)来匹配值，再对少数个体进行一轮爬山算法缩小范围。同时，他们提出了一个纯化技术用于加强所推断的网络的敏感性和特异性。

Maeshiro等提出了一个基于进化算法的方法预测基因调控网络。主要由两个阶段循环，每个循环由如下几步骤：1）产生M个虚拟的初始网络。2）使用信号转导研究包(Signal Transduction Advanced Research Package)对每一个网络进行高速的模拟并找出与实际实验不同的数据。3）如果达到了最大循环数或者得分最高的网络没有进步的空间则结束。4）进行选择，突变以及交叉互换以产生一个新的网络。于是第二阶段的网络较之前的网络更精确。改方法经过了5个人工设计的网络并将结果与动态贝叶斯算法相比较，发现敏感度有5%以上的提升而精确度差不多。预测出来的网络有70-80%的敏感度，70%左右的准确度。

Ho等提出了智能双阶段进化算法(intelligent two-stage evolutionary agrithm，iTEA)从时序表达谱推断由N个基因组成的S-系统模型。初始化过程中，问题被简化为N个最优化子问题，每个子问题有2(N+1)个参数。在第一阶段，每个子问题使用进化算法以及基于同源阵列和因素分析的智能交叉互换。在第二阶段，整合N个问题的结果并在基于OED的退火算法框架下排除表达谱中噪音的影响。该算法的效率在有噪音和没噪音的模拟表达下被测试。随后Huang等延伸了这一研究。他们同过扰乱已有的数据产生了额外的基因表达谱数据然后命名为扰乱的iTEA (iTEAP) 方法，该方法通过加倍原始数据噪音来搜索更多的网络，以次来获得准确性和鲁棒性。Shu等介绍了一个加强版本的iETA，称之为iETA2，同样整合进化算法预测的基因调控关系来估计大规模的基因调控网络。iETA2方法使用的杂交码包含了调控强度，被调控的基因数以及布尔调控参数的信息。其中调控强度参数的值是结合能力的顺序，布尔调控参数表示上调或者下调，用以表示调控的顺序上的节点是不是活跃的。如果I是直接调控基因数最大的节点的度(degree)，则所有基因的杂交码为5I+2，这个数独立与N。因此搜索整个网络构架的算法复杂度从O(N2)级降低到了O(N)级。

Chan等开发了基因网络搜索者(Gene Network Explorer) 使用两个阶段的算法从时序表达谱中估计基因调控网络。在第一阶段，使用遗传算法为接下来的最大期望值划分选择初始聚类中心。在第二阶段，使用卡尔曼过滤器产生一组用以描述网络动态过程的一阶微分方程，并用这些微分方程查找网络中重要的基因相互做用，预测基因表达随时间的函数。这个方法经过了人类纤维原细胞响应细胞的表达谱数据。

Tominaga和Horton开发了一个用于推断生物网络的算法，从一个时序表达谱中提取无尺度网络(scale-free)。这个S-系统利用了分布式遗传算法来最优化这个模型。这种天生的并行性使得这个模型的计算表现很出色。Ram和Chetty提出了一个基于遗传算法的一种启发式算法用于产生更好的网络。这个算法被命名为指导式的遗传算法。

Mitra等提出了一个多对象的进化双聚类算法，从微阵列数据中提取相关性并以此推断出相互作用网络。双聚类被用来发现特定环境下微阵列中拥有相似表达的一组基因。此外本地搜索策略还可以同过添加、删除基因的相关性来进行微调。预处理技术则用来保护那些已知的强相关基因不被忽视。

Maraziotis等提出了一个多层次进化算法训练下的模糊神经周期网络模型。该模型可以捕捉潜在的目标基因并讨论它们的调控类型。在这个混合策略的模型中，周期性网络，自组织结构以及进化算法训练共同作用生成了耐噪音数据的模糊调控系统。

理解基因调控关系的首要目标之一是理解转录因子(transcription factors)是如何在特定环境下控制基因表达的。Schroder等提出了多对象遗传算法(muti-objective genetic algorithms) 用以推断在一个调控因子下的一组共表达基因。他们设计了三个刺激-回馈模型，并且结合先验知识来推断网络模型。这个模型被应用与人类肝脏全基因组微阵列表达谱的调控网络建模。

Shoaib等提出了自适应模糊进化重构框架(Adaptive Fuzzy Evolutionary Reconstruction)。这个框架可以自动的决定模型的参数，比如使用模糊-PBM索引以及高斯分布估计来确定c-聚类的聚类数。这个模型经过了乳腺癌数据集测试，表现良好。其框架主要有六个步骤：1）数据预处理2）计算聚类数目3）用模糊c-聚类方法聚类4）基因选择5）基因调控网络构建6）基因调控网络比较。表1列举了多种方法组合的比较。通常基于进化算法的开发是多种方法交叉的，通过结合其他方法比如聚类，均方根，卡尔曼过滤，三角微分进化等等来获得更好的结果。

|  |  |  |
| --- | --- | --- |
| 表1：遗传算法组合推断基因调控网络 | | |
| 使用技术 | 获得的结果 | 引文 |
| 多对象遗传算法双聚类 | 使用相关性推断出的基因调控网络 | Mitra等 |
| 多对象遗传算法 | 推断调控因子及共表达基因 | Schroder等 |
| 分布式遗传算法 | 并行N问题优化 | Ho等 |
| 进化算法+先验知识 | 推断大尺度网络 | Shu等 |
| 遗传算法+聚类 | 卡尔曼过滤模型 | Chan等 |
| 遗传算法+模糊聚类 | 重构基因调控网络 | Shoaib等 |
| 有指导的遗传算法 | 发现基因调控网络 | Ram和Chetty等 |
| 多层次进化算法 | 捕捉潜在的调控类型和靶基因 | Maraziotis等 |
| 遗传编程+最小均方根 | 描述因果基因调控网络 | Ando等 |
| 遗传编程+卡尔曼过滤 | 推断基因调控网络 | Wang等 |
| 双S-系统 + TDE | 捕捉网络动态 | Noman等 |
| 分布式遗传算法+S-系统+先验知识 | 构建无尺度基因调控网络 | Tominaga和Horton等 |
| 双S-系统+初始化+跳跃 | 重构基因调控网络 | Chowdhury和Chetty等 |

第五部分：结论及挑战

基因调控模型是用来理解系统的行为的，同时它也被用来做一些预测。理解基因调控网络的必要性在于：1）它提供了以个大规模的，宏观的一个对生物体的认识，并且是在mRNA级别的2）帮助我们理解那些复杂的疾病3）药物开发中的前药和药物毒性研究4）同过比较基因调控网络有助与我们对进化的认识5）解释那些不同表型背后的基因基础。

纯进化算法只能对小规模的网络进行分析，然而同过结合其他的算法使得它有足够的效率来对付更大规模的网络。但问题是，即便如此，大规模的网络计算依然受限与计算能力。因此具有并行计算拓展能力的算法比如模糊神经网络，微分方程，人工神经网络等和进化算法的结合有助于得到更好的结果。此外，另一个挑战源于微阵列本身的噪音，因此使用比如多个来源的数据如ChIP，选择过的微阵列数据，蛋白质相互作用数据以及miRNA干涉数据等等。通过整合这些数据加上进化算法和多对象的设计，有助于加速收敛。