一种用于基因调控网络建模的CGP-WPSO混合算法

蔡昕烨 牛 耘 黄志球 范大娟
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 依靠基因调控网络来预测农作物的表现型,对于保障全球的粮食安全有着极其重要的意义。提出了一种基于笛卡尔遗传规划(Cartesian genetic programming)和线性递减惯性权重粒子群优化(linear decreasing inertia weight particle swarm optimization)的混合算法,用于基因调控网络的建模。进一步,为了验证算法的有效性,将算法应用于拟南芥开花调控系统的模型重建问题。最后通过计算机仿真实验表明,该算法能够根据农作物的基因型和环境情况,重建出能够较精确地预测农作物表现型的基因调控网络模型。

关键词 拟南芥开花调控系统,基因调控网络,基因编程,粒子群算法,CGP-WPSO混合算法
中图法分类号 TP18 **文献标识码** A

CGP-WPSO Hybrid Algorithm for Gene Regulatory Network Modeling

CAI Xin-ye NIU Yun HUANG Zhi-qiu FAN Da-juan
(College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)

Abstract The phenotype of the crops can be predicted through the gene regulatory network (GRN), which is important for the global food security. This paper proposed a Cartesian genetic programming and linear decreasing inertia weight particle swarm optimization algorithm for GRN modeling. To verify the effectiveness of the proposed algorithm, we applied it to the recovery of the Arabidopsis flowering time control system. The computer simulation indicates that our proposed algorithm is able to infer the GRN model which can predict the phenotype of the crops fairly accurately based on its genotype and environmental conditions.

Keywords Flowering time control in arabidopsis, Gene regulatory network, Genetic programming, Particle swarm optimization, CGP-WPSO hybrid algorithm

1 引言

利用基因调控网络(GRN)来预测农作物的表现型(phenotype)是当前生物信息研究中的一个重要发展趋势。近年来,随着世界人口对粮食日益增长的需求,对农作物的表现型进行有效建模(crop modeling)显得越来越迫切[1]。在传统的方法中,农作物的建模主要取决于生物学家对生物生理学的理解以及与实际经验的结合[2,3]。随着分子生物学逐渐成为生物研究的热点,目前的发展趋势是在基因的层面上,通过基因调控网络,揭示基因之间的交互作用,从而实现对农作物表现型的准确预测。

一个生物体的基因组所包含的生物信息能够控制生物分子的成长过程及其对周围环境刺激的反应。近年来,基因定序(genome sequencing)和基因芯片(micro-array)技术高速发展。截至2011年9月,3842个基因组定序项目已经完成,另外7629个基因组定序项目仍在进行中[1)]。如何挖掘这些生物数据中的有用信息,成为了生物信息领域的一个研究热点。目前的一个重要研究问题是如何从大量的观测数据,如环境(environmental)、生物表现型(phenotypical)等数据中自动重建GRN。此类问题在生物学中被称为基因型到表现型的映射关系问题(genotype to phenotype mapping problem)[4]。对基因调控网络的建模正是用来解决这种映射关系的重要方法之一。

目前已有大量的关于基因调控网络建模算法的研究结果。文献[5]提出了基于布尔网络的基因调控网络识别算法;文献[6,7]考虑了贝叶斯模型;文献[8]对重构动态贝叶斯模型的SEM算法进行了改进;文献[9]利用线性微分方程对基因表达数据进行了回归分析,提出使用一个启发式算法来构建稀疏的基因调控网络,并基于人工数据的实验验证了算法的有效性。上述方法大多应用于小规模基因调控网络的重建,而实际情况中大多基因调控网络属于高维,上述方法由于计算量过大而未能取得较好的效果[10]。与上述工作不同,本文提出一种新的混合算法CGP-WPSO,其中Cartesian Genetic Programming(CGP)用于搜索基因调控网络模型的结构,然后通过线性递减惯性权重的粒子群算法(WPSO)对基因调控网络的相关参数进行估计。PSO算法具有较快的收敛性,在粒子群中引入递减惯性权重的思想又使得算法在进行参数估计时不会过快地陷入局部最优。因此,这种新的算法综合

到稿日期:2011-10-20 返修日期:2012-02-29 本文受国家高技术研究发展计划(863计划)(2009AA010307)资助。
蔡昕烨(1983—),男,博士,讲师,主要研究领域为人工智能应用技术;牛 耘(1974—),女,博士,副教授,主要研究领域为机器学习;黄志球(1965—),男,博士,教授,主要研究领域为软件工程;范大娟(1984—),女,博士生,主要研究领域为软件工程。

1) http://genomesonline.org/cgi-bin/GOLD/bin/gold.cgi

了 Genetic Programming 的全局最优化特性和 PSO 较快的收敛性，能够快速地重建基因调控网络。

本文第 2 节介绍数据的获取和问题的定义；第 3 节着重介绍 CGP-WPSO 混合算法；第 4 节详细描述实验的设置、实验结果及其分析；最后为结论和对未来工作的展望。

2 数据和问题定义

2.1 环境和表现型数据的获取

本文所研究的环境数据由美国国家自然科学基金重大研究项目"The Evolutionary Aspects of Gene Network Pathway Signal Integration"[2)]提供。环境数据包含了光周期和温度，其分别从葡萄牙的 Coimbra 以及芬兰的 Jokioinen 等 8 个地点获取。温度选取 1971 年到 1998 年之间 3 月 1 日到 6 月 30 日的当地每日平均温度。同期的日照时间数据从美国海军天文台获得[3)]。由于植物对光的敏感性，我们使用了植物建模中通用的民用署暮光，即将太阳低于地平面 6 度作为实验的开始或/和结束时间。

构造一个具有参数的人工基因调控网络。该网络模拟了拟南芥开花调控系统的功能特征。2.2 节描述了个体基因的功能特征，相关研究工作可以参见文献[11-13]。每一个基因有一个二值型参数，其表示不同的突变型等位基因。通过这些不同的等位基因的组合构造出了 $g=100$ 个不同的基因型。每个基因型根据构造的人工基因调控网络模型，在 $pd=3$ 个不同的种植日期以及 $ps=18$ 个不同的种植地点获得相应的开花日期[13]。基因组中的每个基因都有两个等位基因，但只有存在于拟南芥开花调控系统中的等位基因才会影响最终的开花日期。

2.2 问题定义

我们的目标是利用人工数据重建出基因调控网络模型，重建后的基因调控模型所预测的开花日期应尽可能和人工数据接近，因此需要最小化人工数据和模型数据的均方根误差 E。其定义如下：

$$E=\sqrt{\frac{\sum_{i}^{n}(D_i^{data}-D_i^{model})^2}{n}} \tag{1}$$

式中，D_i^{data} 和 D_i^{model} 分别是第 i 次实验中人工开花时间和重建后的基因调控网络模型所预测的开花时间。每次实验 i 选取不同的基因型、种植时间和种植日期。

在基因调控网络模型中，采用 4 个函数来描述基因功能：(i) gain:$o=c_g \cdot i_1$；(ii) summer:$o=c_s \cdot i_1+i_2$；(iii) multiplier:$o=c_m \cdot i_1 \cdot i_2$；(iv) integrator:$o(t)=o(t-1)+c_i \cdot i_1(t)$。其中，$i_1$ 和 i_2 是输入，o 是输出。同时每一个基因对应着一个参数($c=c_g, c_s, c_m$ or c_i)。这些参数都是二值型，每一个数值对应着一个突变型等位基因。任意一个基因的输入都可以是调控网络中其它基因的输出，或者是环境数据的输入，如光周期(P)或者温度(T)。

3 CGP-WPSO 混合算法

基于笛卡尔遗传规划和线性递减惯性权重粒子群优化的混合算法(CGP-WPSO)分为 3 个步骤(见图 1)。第一步为关键基因识别，即从 100 个基因中选取以较大概率存在于基因调控网络的基因。这些基因在选取后成为候选基因，作为第二步 CGP 的输入。

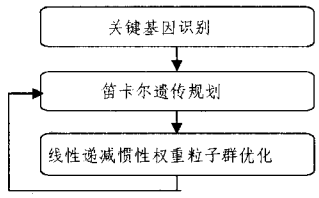

图 1 CGP-WPSO 算法流程图

CGP 初始化具有 N 个解(基因调控网络结构)的群体，并利用变异操作生成新的网络结构群体。对于每一个网络结构，使用 WPSO 算法来优化其参数，使得开花日期的均方根误差最小，如式(1)所示。第二步 CGP 和第三步 WPSO 将重复进行，直至算法收敛。

3.1 关键基因识别

如上节所述，每一个基因型都包含了 100 个基因，只有其中的一些基因是开花调控网络的一部分。基因识别的基本思想是从网络成员中排除掉那些不能改变开花时间的基因。基因识别的具体过程如下：对于每一个基因位点，基因型根据当前基因位点的变异型等位基因被划分为两组，对这两组中的每对基因型所对应的开花时间做 F-test。F-test 结果决定了对于该基因位点，不同等位基因对应的开花时间是否具有统计意义上的显著区别。

3.2 CGP 的网络结构组织形式

我们改进了 Cartesian Genetic Programming(CGP)，使之能够更加适合表示生成的网络结构模型。在 CGP 中，每一个解(网络结构)都由具有 M 个域的字符串表示，如图 2 所示，其中 M 是候选基因的个数。每一个域有 4 个条目，它们分别是 2 个基因的输入(可以是上游的其它基因或者环境数据的输入)、1 个函数(g, gain; s, summer; m, multiplier; or i, integrator)和 1 个表示该相应基因位点的索引。

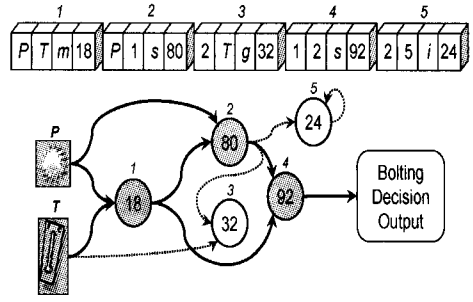

图 2 基因调控网络的字符串表达方式

图 3 给出了一个用 CGP 组织形式表示的基因调控网络拓扑结构。图中 5 个基因是从关键基因识别步骤得到的候选基因。在所有的模型中，有两个环境输入：光周期(P)或者温度(T)。在图 3 中，开花日期最终决定基因 #4 的输入分别为基因 #1 和 #2，因此这个网络结构的功能部分实际上只包含了 3 个基因。基因 #3 和基因 #4 为非功能基因，在图中用虚线表示。在这个网络模型中，当最终的开花决定基因 #4 的基因表达超过了阈值 1 时，该时间点就被预测为基因型的开

2) http://www.egad.ksu.edu

3) http://aa.usno.navy.mil/data/docs/RS_OneYear.html

花日期。需要注意的是，CGP 的组织形式只表达了网络结构，每个基因的相关参数 c_g，c_s，c_m 和 c_i 的最优化将在 3.4 节介绍。

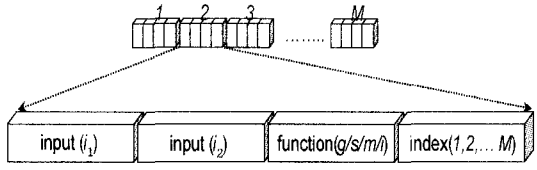

图 3　一个基因调控网络及其对应的字符串表达方式的实例

3.3　变异操作

变异操作的主要目标是对搜索空间进行有效的搜索。对调控网络结构中的每一个条目，根据预先定义的概率（称之为变异概率），在可取值范围内对其进行变异操作。另外，变异操作还受如下约束：(i) 如前所述，大部分的基因调控网络只具有前馈的网络拓扑结构，因此，我们规定变异操作不允许产生具有反馈的调控网络结构。(ii) 对于同一个网络结构，基因索引（域 4）必须是唯一的。

3.4　WPSO 算法

对于 CGP 算法产生的每一个基因调控网络结构，采用线性递减惯性权重的粒子群算法（WPSO[14]）进行参数的最优化。参数 c_1 到 c_M 在参数域空间中表示为一个向量 c，开始时 WPSO 算法将随机产生一个粒子群 $c(j)$，$j=1,\cdots,P$，其中 P 是粒子的个数。每个向量 $c(j)$ 对应 jth 粒子的位置。

用 t 来描述迭代数，通过粒子速度 $v_t(i)$ 来更新粒子的位置：

$$c_{t+1}(i)=c_t(i)+v_t(i)$$

速度将根据粒子自己记录的历史最好位置和群体的最好位置来进行更新，更新的法则是：

$$v_{t+1}(j)=w*v_t(j)+C_1\times U[0,1]\times(c_{ib}(j)-c_t(j))+C_2\times U[0,1]\times(c_{gb,t}-c_t(j))$$

$$w=(w_1-w_2)\times\frac{(MAXT-t)}{MAXT}+w_2$$

上述公式中，C_1 和 C_2 分别代表 cognitive 和 social 常数。$U[0,1]$ 是一个平均分布的随机数，取值范围是 $[0,1]$。c_{ib} 分别代表第 i 个记录的个体最优位置，$c_{gb,t}$ 表示算法记录的全局最优解。w 是非负数，称为惯性因子。w_1 和 w_2 是初始和最终的惯性权重，$MAXT$ 是可以重复的最大次数，t 是最近重复的次数。较大的 w 值有利于跳出局部极小点，而较小的 w 值有利于算法收敛，因此我们采取了自适应调整的策略，即随着迭代的进行，线性地减小 w 的值。这种改进通过自适应调整惯性因子，能兼顾搜索效率和搜索精度，故优化性能有所提高。

粒子的位置必须控制在事先定义的搜索空间中。当一个粒子超过了该空间，粒子将回到空间边界上。此外，它的速度将乘以（−1），即飞向相反的方向。每个基因调控网络模型（由调控网络结构和关联参数组成）的适应度定义为均方根误差 E，见式(1)。同时，为了更好地模拟真实的实验，对开花日期数据加入了噪声（$std=2$ 天）。

4　结果和讨论

4.1　实验设置

对于 CGP 算法，群体个数 $N=50$ 的网络结构被初始化，最大跌代数为 100；变异操作的概率为 0.12。对于 WPSO，参数向量群体数 $P=50$；最大迭代数 $MAXT$ 设为 100，最大惯性权重 w_1 和最小惯性权重 w_2 分别设为 0.9 和 0.3。cognitive 和 social 常数均被设为 2.1。运行基因识别步骤后得到候选基因数为 8。

4.2　人工基因调控网络

图 4 显示了用来生成人工开花日期的人工基因调控网络结构。表 1 分别显示了基因调控网络结构包含的基因编号、基因函数、对应参数表示符，以及对应的参数值（等位基因为 0 或 1 两种情况）。图 5 显示了 CGP-WPSO 算法执行后重建的基因调控网络结构。这个重建的基因调控网络所预测的开花日期与实际开花日期之间的方均根误差为 3 天。除了基因 #80，它包含了其它所有在人工基因调控网络中的基因。表 2 显示了图 4 中基因调控网络结构中每一个基因对应的参数和函数。但是由于人工基因调控网络和重建基因调控网络在基因数量上不同，图 4 和图 5 的比较并不能反映出两者的相似度。

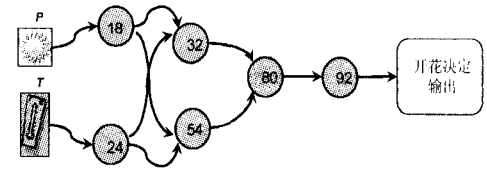

图 4　人工基因调控网络结构

表 1　人工基因调控网络的参数

基因 #	函数	参数表示符	不同等位基因对应的参数值 #	
			0	1
18	gain	$C_{g,18}$	1.3286	0.88411
24	multiplier	$C_{m,24}$	5.6e−5	8.6e−5
32	summer	$C_{s,32}$	0.54132	0.54518
54	summer	$C_{s,54}$	2.6e−5	1.16e−4
80	—	C_{80}	—	—
92	integrator	$C_{i,92}$	1.8867	2.9725

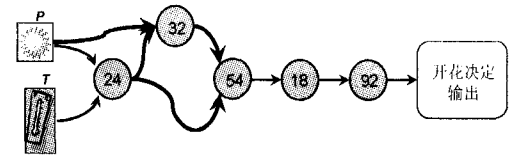

图 5　CGP-WPSO 算法重建的基因调控网络结构

表 2　CGP-WPSO 算法重建的基因调控网络的参数

基因 #	函数	参数表示符	不同等位基因对应的参数值 #	
			0	1
18	gain	$C_{g,18}$	1.3286	0.88411
24	multiplier	$C_{m,24}$	5.6e−5	8.6e−5
32	summer	$C_{s,32}$	0.54132	0.54518
54	summer	$C_{s,54}$	2.6e−5	1.16e−4
80	—	C_{80}	—	—
92	integrator	$C_{i,92}$	1.8867	2.9725

为解决上述问题，采用了数学表达式这种更加直观的比较方法。通过代入每个基因的函数和对应的参数，每个基因调控网络可以用数学表达式来表示，如表 3 所列。在参数替换以后（选择突变型等位基因为 0 时），可以看到两个基因调

（下转第 197 页）

Rough Sets, Data Mining, and Granular-Soft Computing. Berlin: Springer-Verlag, 1999: 73-81

[11] Leung Y, Li De-yu. Maximal consistent block technique for rule acquisition in incomplete information systems[J]. Information Sciences, 2003, 153(1): 85-106

[12] Guan Yan-yong, Wang Hong-kai. Set-valued information systems[J]. Information Sciences, 2006, 176(17): 2507-2525

[13] Qian Yu-hua, Liang Ji-ye, Li De-yu, et al. Approximation reduction in inconsistent incomplete decision tables[J]. Knowledge-Based Systems, 2010, 23(5): 427-433

[14] Yang Fang, Guan Yan-yong, Li Shu-jin, et al. Attributes reduct and decision rules optimization based on maximal tolerance classification in incomplete information systems with fuzzy decisions [J]. Journal of Systems Engineering and Electronics, 2010, 21 (6): 995-999

[15] 梁吉业，王宝丽，钱宇华，等. 一种不完备信息系统中极大相容块的构造算法[J]. 计算机科学，2006，33(11A)：79-82

[16] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

[17] 史忠植. 知识工程[M]. 北京：清华大学出版社，1988

(上接第 182 页)

控网络都包含最重要的 3 项"P","$T*P$"和"$O_{92}(t-1)$"。随后，对基因调控网络的数学表达式各项做敏感度分析，发现"P","$T*P$"两项前的参数对开花日期的影响远远比"T"项前的参数重要，如表 4 所列。图 6 对比了人工开花日期数据和算法重建后的基因调控网络的预测开花日期数据，同时对数据作了线形回归分析。从图中可以看出，人工开花日期数据与算法重建的基因调控网络模型预测的开花日期数据非常相近。

表 3　基因调控网络的数学表示式(参数替代前和参数替代后)

网络	最终开花决定基因的输出	参数替代后最终开花决定基因的输出
人工基因调控网络	$O(t)=C_{18}*C_{32}*C_{80}*C_{92}*T+C_{24}*C_{80}*C_{92}*P+C_{18}*C_{24}*C_{54}*C_{92}*T*P+O_{92}(t-1)$	$O(t)=5.35\times10^{-04}*T+4\times10^{-04}*P+6.2\times10^{-05}*T*P+O_{92}(t-1)$
算法重建后的基因调控网络	$O(t)=C_{18}*C_{32}*C_{54}*C_{92}*P+(C_{18}*C_{24}*C_{54}*C_{92}+C_{18}*C_{24}*C_{92})*T*P+O_{92}(t-1)$	$O(t)=3.5\times10^{-05}*P+1.4\times10^{-04}*T*P+O_{92}(t-1)$

表 4　对表 3 中数学公式各项的敏感度分析

每项的相关敏感度	预测开花日期的数学表达式	T	P	T*P
人工基因调控网络模型	$O(t)=5.35\times10^{-04}*T+4\times10^{-04}*P+6.2\times10^{-05}*T*P+O_{92}(t-1)$	1.97	7.16	8.6
算法重建的基因调控网络模型	$O(t)=3.5\times10^{-05}*P+1.4\times10^{-04}*T*P+O_{92}(t-1)$	—	7.35	6.34

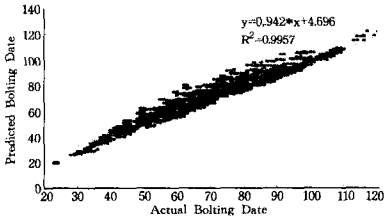

图 6　人工开花日期数据 vs 重建基因调控网络预测开花日期数据

结束语　从实验数据中自动地生成基因网络结构，是目前生物信息领域非常受关注的研究方向。在本文中，我们提出了一种能够从实验数据中直接推导出网络结构和相关参数的算法。算法运用了笛卡尔遗传规划和线性递减惯性权重粒子群优化算法。我们将其应用于拟南芥开花调控系统的模型重建问题。实验结果表明了算法的有效性。在今后的研究工作中，我们将进一步改进防止陷入局部最优的算法，并在更多真实的实验数据上对算法进行应用分析。

参 考 文 献

[1] Hanks R J, Ritchie J T. Modeling plant and soil systems[M]. Agronomy Monograph, 1991: 545

[2] Welch S M, Roe J L, Dong Z. A genetic neural network model of flowering time control in Arabidopsis thaliana[J]. Agronomy Journal, 2003, 95(1): 71-81

[3] Welch S M, Roe J L, Das S, et al. Merging genomic control networks and Soil-Plant-Atmosphere-Contium (SPAC) models[J]. Agricultural Systems, 2003, 86(3): 243-274

[4] Cooper M, Chapman S C, Podlich D W, et al. The GP problem: quantifying gene-to-phenotype relationships[J]. Silico Biology, 2002, 2(2): 151-164

[5] Bernardo D, Gardner T S, Collias J J, et al. Robust Identification of Large Genetic Networks[J]. Pacific Symposium on Biocomputing, 2004, 9: 486-497

[6] Lähdesmäki H, Shmulevich I, Yli-Harja O. On Learning Gene Regulatory Networks under the Boolean Network Model[J]. Machine Learning, 2003, 52(1/2): 147-167

[7] 刘昱昊，刘桂霞，苏兰莹，等. 边排序贝叶斯网络结构学习算法应用于基因调控网络构建[J]. 吉林大学学报：理学版，2010，48 (4)：624-630

[8] 葛玲玲，王浩和，姚宏亮. 基于改进 SEM 算法的基因调控网络构建方法[J]. 计算机应用研究，2010，27(2)：450-258

[9] Chen X W, Gopalakrishna A, Wang X K. An effective structure learning method for constructing gene networks[J]. Bioinformatics, 2006, 22(11): 1367-1374

[10] Mitra S, Das R, Hayashi Y. Genetic networks and soft computing[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(1): 94-107

[11] Baldi P, Hatfield G. DNA microarrays and gene expression[M]. Cambridge UK: Cambridge University Press, 2002

[12] Welch S M, Dong Z, Roe J L, et al. Flowering time control: gene network modeling and the link to quantitative genetics[J]. Australian Journal of Agricultural Research, 2005, 56(9): 919-936

[13] Dong Z. Incorporation of Geneic Information into the Simulation of Flowering Time in Arabidopsis Thaliana[D]. Agronomy Department, Kansas State University, 2003

[14] Shi Y, Eberhart R C. A modified particle swarm optimizer[C]// IEEE Congress on Evolutionary Computation (CEC). Anchorage, AK, 1998: 69-73