第 35 卷 第 20 期 计 算 机 工 程 2009 年 10 月
Vol.35 No.20 Computer Engineering October 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)20—0181—03 文献标识码：A 中图分类号：TP391.41

基于改进 PSO 的基因调控网络重构方法

蒋 炜，彭新一，周育人
（华南理工大学计算机科学与工程学院，广州 510006）

摘 要：提出一种基于改进粒子群优化算法的基因调控网络重构方法。该方法利用粒子群优化算法确定加权矩阵模型的最优结构及参数，从而推测出与实验数据相吻合的加权矩阵，实现利用重构的加权矩阵模型模拟基因调控网络的相互作用。实验结果表明，该方法能有效推理出复杂的基因调控网络结构。

关键词：粒子群优化算法；基因调控网络；加权矩阵模型；重构

Reconstruction Method of Gene Regulatory Network Based on Modified Particle Swarm Optimization

JIANG Wei, PENG Xin-yi, ZHOU Yu-ren
(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006)

【Abstract】 This paper presents reconstruction method of Gene Regulatory Network(GRN) based on Modified Particle Swarm Optimization (MPSO). It uses PSO to identify the optimal architecture and parameter of the weight matrices model, so that a recurrent neural network consistent with experimental data is inferred, and uses weight matrices model to simulate GRN. Experimental results show that the method is effective to infer complex interaction such as GRN.

【Key words】 Particle Swarm Optimization(PSO); Gene Regulatory Network(GRN); weight matrices model; reconstruction

1 概述

近年来，人类基因组计划的完成、DNA 微阵列技术的出现和应用使同时定量测定成千上万个基因在生物样本中的表达水平成为可能，从而为用数学计算的方法研究基因间复杂大规模的基因表达调控打下基础。目前有许多用于描述基因调控系统的数学方法和模型，加权矩阵模型[1]是一种对基因调控网络(Gene Regulatory Network, GRN)进行建模的模型形式。这种结构在基因调控作用是否有问题、以权值的形式简单描述出互作用的强度等方面具有丰富的表达描述能力，具有灵活、利于描述基因网络复杂关系的优点，在系统分析和控制设计方面有一定的优势。因此，本文综合应用加权矩阵模型和粒子群优化算法(Particle Swarm Optimization, PSO)推断基因间的相互关系，由此提出一种基于改进粒子群优化算法(Modified Particle Swarm Optimization, MPOS)的基因调控网络重构方法，并通过实验验证该方法的有效性。

2 基因调控网络加权矩阵模型

在基因调控网络加权矩阵模型中，一个基因的表达值是其他基因表达值的函数。含有 n 个基因的基因表达状态用 n 维空间中的向量 $\boldsymbol{u}(t)$表示，$\boldsymbol{u}(t)$代表一个基因在时刻 t 的表达水平。以一个加权矩阵 $\boldsymbol{W}$ 表示基因之间的相互调控作用，$\boldsymbol{W}$ 的每一行代表一个基因的所有调控输入，w_{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 w_{ij}。基因 i 的总调控输入 $r_j(t)$ 为

$$r_i(t)=\sum_{j=0}^{n} w_{ij}u_j(t) \tag{1}$$

若 w_{ij} 为正值，则基因 j 激发基因 i 的表达，而负值表示基因 j 抑制基因 i 的表达，若是 0 则表示基因 j 对基因 i 没有作用。最后基因 i 表达响应还需要经过一次非线性映射：

$$u_i(t+1)=\frac{1}{1+\mathrm{e}^{-(\alpha_i r_i(t)+\beta_i)}} \tag{2}$$

这种函数是神经网络中常用的 Sigmoid 函数，其中，α_i 和 β_i 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。在本文中 α_i =1, β_i =0，所以，Sigmoid 函数可以表示为

$$u_i(t+1)=\frac{1}{1+\mathrm{e}^{-r_i(t)}} \tag{3}$$

图 1 和表 1 是一个基因调控网络及其对应加权矩阵的例子。该模型中，基因之间的调控关系可以是从最高激励值 l 到最大抑制值-l 之间连续的任意实数。基因表达在离散状态转移中被调控，从而所有基因的表达水平可以同时被更新。

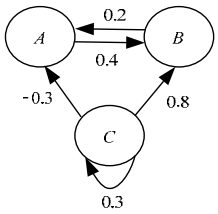

图 1 样本基因调控网络

基金项目：国家自然科学基金资助项目(60673062)；广东省自然科学基金资助项目(06025686)

作者简介：蒋 炜(1981－)，男，硕士研究生，主研方向：演化计算，遗传算法，粒子群优化算法，基因网络；彭新一，研究员；周育人，副教授、博士

收稿日期：2009-01-30 **E-mail：**great.wei@163.com

表 1　样本网络对应矩阵

	A	*B*	*C*
A	0	0.4	0
B	0.2	0	0.8
C	-0.3	0	0.3

3　粒子群优化算法

粒子群优化算法源于对简单社会系统的模拟。在 PSO 系统中，每个优化问题的解都是搜索空间中的一只鸟，称之为“粒子”，粒子代表一个候选解，具有位置和速度 2 个特征和一个由被优化的函数决定的适应值。从初始群体出发，粒子根据自己和同伴的飞行经验不断调整位置和速度，追随当前的最优粒子在解空间中不断搜索。

基本 PSO 算法采用下列公式对粒子进行操作：

$$v_{id}(k+1)=v_{id}(k)+c_1\times random()\times(pbest_{id}-x_{id}(k))+ c_2\times random()\times(gbest_d-x_{id}(k)) \tag{4}$$

$$x_{id}(k+1)=x_{id}(k)+v_{id}(k+1) \tag{5}$$

其中，i=1,2,⋯,m，m 代表粒子数；d=1,2,⋯,D，D 代表粒子的维数；k 代表迭代次数。

第 i 个粒子第 d 维迭代到目前为止具有最佳适应度的粒子被称为个体最好粒子，记为 $pbest_{id}$ ；而全部粒子第 d 维迭代到目前为止具有最佳适应度的粒子被称为全局最好粒子，标记为 $gbest_d$ 。

$x_{id}(k)$ 和 $x_{id}(k+1)$ 分别是第 i 个粒子 d 维分量在第 k 次和第 k+1 次迭代的位置；$v_{id}(k)$, $v_{id}(k+1)$ 分别是第 i 个粒子 d 维分量在第 k 次和第 k+1 次迭代的速度，为了防止粒子远离搜索空间，粒子的每一维速度 v_d 都会被限制在[$-v_{dmax}$, x_{dmax}]之间，假设将搜索空间的第 d 维定义为区间[$-x_{dmax}$, x_{dmax}]，则通常 $v_{dmax}=kx_{dmax}$ ，其中，$x\in$[-1,1]；速度 v 限定在[-2,2]之内。

c_1 , c_2 是 2 个正约束系数，本文中取 $c_1=c_2=2.05$。

$random()$ 是一个[0,1]之间的随机数。

4　改进粒子群优化算法对加权矩阵模型的重构

4.1　粒子编码

利用 PSO 重构基因调控网络加权矩阵型，需要将基因网络编码成染色体，建立位置矢量与矩阵之间的映射关系是设计 PSO 的首要问题。对于含有 N 个基因的重构问题，用 $N\times N$ 维的实向量表示粒子的位置，向量的前 N 维分量对应 1,2,⋯,N 个基因对第 1 个基因的调控输入，其后的 N 维分量对应其他 N-1 个基因对第 2 个基因的调控输入，依此类推。

4.2　适应度函数

本文把最初始时间点 t=1 到终止时间点 t_f 的真实基因调控网络的调控输出(即基因表达)与加权矩阵模型得到的实验调控输出差的绝对值作为适应度函数，适应度函数的表达式如下：

$$fitness=\sum_{t=1}^{t_f}\sum_{i=1}^{N}\left|u_i(t)-\hat{u}_i(t)\right| \tag{6}$$

其中，$u_i(t)$ 与 $\hat{u}_i(t)$ 表示基因 i 在 t 时刻真实与预测的基因表达水平，目标是适应度 $fitness$ 最小，当 $fitness$=0 时，说明真实与预测的基因网络一致。

4.3　改进粒子群优化算法

由于基因调控网络是潜在未知的，适应度函数 $fitness$ 大小只是从宏观上评价加权矩阵模型模拟的优劣，粒子群优化算法在搜索过程中易陷入早熟，而且基本 PSO 本身具有缺陷(如在理论上未能证明是绝对收敛的，有可能陷入局部最优)，因此对基本 PSO 做了一些改进，使其适应基因调控网络的重构问题，称之为 MPSO。

为了加快 PSO 的收敛速度，引入惯性因子 w，式(4)修正为

$$v_{id}(k+1)=wv_{id}(k)+c_1\times random()\times(pbest_{id}-x_{id}(k))+ c_2\times random()\times(gbest_d-x_{id}(k)) \tag{7}$$

其中，惯性因子 w 是一个与前一次速度有关的比例因子，w 控制以前的速度对当前速度的影响，较大的 w 可以加强 PSO 的全局搜索能力，而较小的 w 能加强局部搜索能力[2]。本文采用 w 在迭代过程 0.9-0.6 线性下降的方法：

$$w=0.9-\frac{0.9-0.6}{iter_{\max}}\times iter \tag{8}$$

其中，$iter_{\max}$ 为总迭代次数；$iter$ 为当前迭代次数。引入线性下降的惯性因子 w，使 PSO 在开始时搜索较大的区域，较快地定位最优解的大致位置，随着 w 逐渐减小，粒子速度减慢，开始精细地进行局部搜索(这里，w 类似于模拟退火中的温度参数)。

虽然加入惯性权重和约束因子后，收敛速度较快，但在算法后期，速度越来越小，粒子群表现出强烈的趋同性，易陷入局部极小点。针对这个问题，本文对算法进行了以下改进：

在粒子群从第 n 代向第 n+1 代“飞翔”时，粒子除追随个体极值 $pbest$ 和全局极值 $gbest$ 外，还追随从粒子群中随机选取的某个粒子的个体极值 $qbest$，式(7)可以修正为

$$v_{id}(k+1)=wv_{id}(k)+c_1\times random()\times(pbest_{id}-x_{id}(k))+ c_2\times random()\times(gbest_d-x_{id}(k))+c_3\times random()\times(qbest_{rd}-x_{id}(k)) \tag{9}$$

其中，c_3 为非负常数，且远小于 c_1 , c_2，本文通过实验调试得出 c_3 取 0.8。

在粒子的飞翔迭代公式中增加 $gbest$ 后，$pbest$, $gbest$, $gbest$ 三者共同向下一代提供信息，粒子获得的信息量增大，从而可能更快地找到优化解。同时 $qbest$ 的权重系数很小，相当于扰动信息，增加了粒子的多样性，避免算法过早收敛。

4.4　算法实现步骤

针对基因调控网络的重构问题，MPOS 的主要步骤如下：

(1)确定从最初始时间点 t_0 到最终时间点 t_f 的真实基因调控网络基因的表达水平，把基因的表达水平划分成 2 个集合——训练集和测试集。

(2)初始化参数 w , c_1 , c_2 , c_3 以及种群粒子个数 p，令计数器 $iter$ =0, m=0。

(3)随机初始化粒子群中粒子的速度和位置。

(4)运用训练集和式(6)评价种群中粒子的适应值。

(5)寻求每个粒子的个体最优解 $pbest$ 和整个种群的最优解 $gbest$。

(6)更新计数器 k=k+1。

(7)根据式(8)更新 w 。

(8)根据式(9)更新每个粒子的速度，如果 $v\geq$2，则 v=2；若 $v\leq$-2，则 v=-2。

(9)根据式(5)更新每个粒子的位置，若粒子在某一维超出搜索空间[-1,1]，则限制该粒子在搜索空间的边界。

(10)重新通过训练集和式(6)评价种群中每个粒子的适应值，根据每个粒子的适应值大小判断是否更新每个粒子的 $pbest$ 和整个种群的最优粒子 $gbest$。

(11)判断一次 PSO 是否满足最大迭代次数 $iter_{max}$，如满足，则继续执行(12)，否则，转(6)。

(12)测试算法得到的 *gbest* 在测试集中的适应值是否足够小，如足够小，则输出 *gbest*，计算终止，否则，继续执行(13)。

(13)更新计数器 $m=m+1$，判断 PSO 循环执行的次数是否满足最大迭代次数 m_{max}，若满足，则输出最好的一个解，计算终止，否则，转(3)。

5 实验

为了测试 MPSO 对基因调控网络重构的效果，将本算法和具有代表性的遗传算法[3]进行比较，分别推理图 2 的基因调控网络[4]。

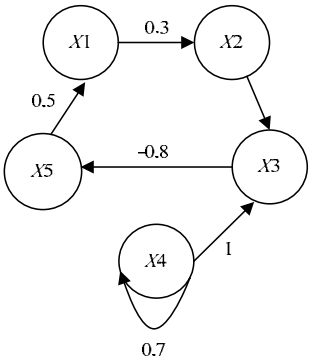

图 2　5 个节点的基因网络

MPSO 参数设置为：粒子数 p=30，t_f=7，粒子群算法的粒子迭代进化次数为 100 000 次，PSO 最大循环执行 20 次。MPSO 单独运行 10 次，表 2、表 3 是预测矩阵的样本均值及样本标准差。遗传算法的参数为：种群数为 3 000，遗传代数为 100，交叉概率为 0.99，变异概率为 0.02。

表 2　预测矩阵的样本均值

	X1	X2	X3	X4	X5
X1	0.009 5	0.001 0	-0.001 9	-0.007 5	0.521 3
X2	0.284 4	0.009 0	-0.002 5	0.000 5	0.002 5
X3	0.000 0	-0.597 5	0.012 5	0.998 6	0.000 0
X4	0.014 0	0.005 0	0.000 0	0.693 9	0.002 1
X5	0.000 0	0.001 5	-0.811 7	0.001 7	-0.000 3

表 3　预测矩阵的样本标准差

	X1	X2	X3	X4	X5
X1	0.026 60	0.004 40	0.008 50	0.023 14	0.023 73
X2	0.028 50	0.045 20	0.007 80	0.002 23	0.009 10
X3	0.000 00	0.019 00	0.045 00	0.004 70	0.000 00
X4	0.049 46	0.016 70	0.000 00	0.018 70	0.007 20
X5	0.000 00	0.006 90	0.023 70	0.005 30	0.005 59

图 3 是 MPSO 得到的基因网络，图 4 是经典遗传算法得到的基因网络。通过比较可知，改进粒子群优化算法较之传统的遗传算法能够更好地反映基因调控网络的网络结构及其调控关系。

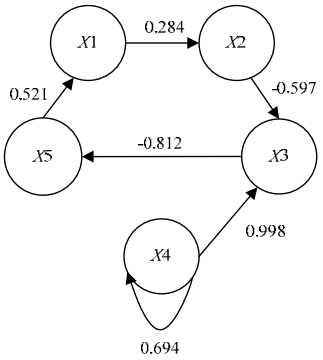

图 3　MPSO 预测的基因网络

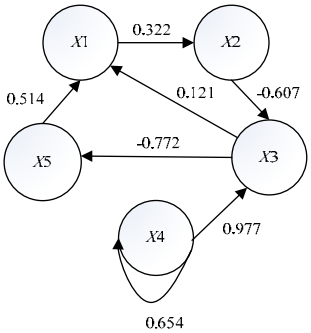

图 4　遗传算法预测的基因网络

6 结束语

本文尝试用改进的粒子群优化算法来优化重构基因调控网络，实验证明，结合使用加权矩阵模型和 MPSO 能够较准确地对基因调控网络进行重构，但该算法还存在以下不足：

(1)由于粒子群算法的随机性，并不是每次实验都能成功的预测出准确的网络结构。

(2)随着基因数目的增加，网络结构推测的准确率会逐渐下降，现阶段还不能适用于大规模的基因网络。将来的工作重点是努力提高预测的准确率，将 PSO 与遗传算法、模拟退火算法等其他算法融合，使其能够更加高效地重构基因调控网络。

参考文献

[1] Weaver D C, Workman C T, Stormo G D. Modeling Regulatory Networks with Weight Matrices[C]//Proc. of the 4th Pacific Symposium on Biocomp. Hawaii, USA: [s. n.], 1999.

[2] 黄翀鹏, 熊伟丽, 徐保国. 惯性权值对粒子群算法收敛性的影响与改进[J]. 计算机工程, 2008, 34(9): 31-33.

[3] Hitoshi I, Mimura A. Inference of a Gene Regulatory Network by Means of Interactive Evolutionary Computing[J]. Genome Informatics, 2002, 145(3): 94-103.

[4] Ressom H W, Zhang Yuji. Inferring Network Interactions Using Recurrent Neural Networks and Swarm Intelligence[C]//Proc. of EMBS'06. New York, USA: [s. n.], 2006: 4241-4244.

编辑　张　帆