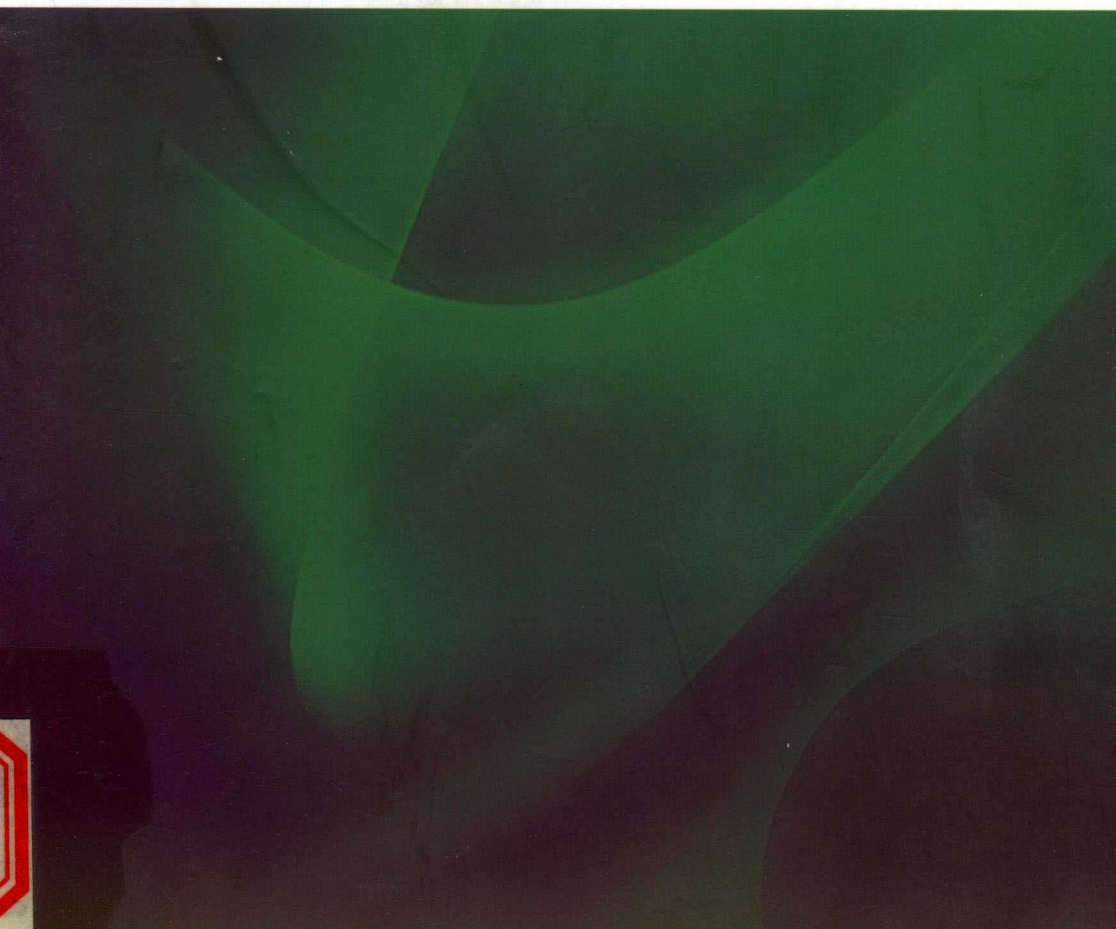
基因表达式编程算法原理及应用研究

陈青华 李亚梅 阮梦黎 编著



济南出版社

基因表达式编程算法原理及应用研究

陈青华 李亚梅 阮梦黎 编著

济南出版社

图书在版编目(CIP)数据

基因表达式编程算法及应用研究 / 陈青华，李亚梅，阮梦黎编著. —济南：济南出版社，2015. 5

ISBN 978-7-5488-1623-2

Ⅰ. ①基… Ⅱ. ①陈… ②李… ③阮… Ⅲ. ①基因表达—程序设计—算法—研究 Ⅳ. ①TP311

中国版本图书馆 CIP 数据核字（2015）第 123313 号

基因表达式编程算法及应用研究 陈青华 李亚梅 阮梦黎 编著

责任编辑 / 朱向泓		开　本	710×1000 毫米 1 / 16
封面设计 / 侯文英		印　张	15.75
		字　数	250 千
出版发行	济南出版社	版　次	2015 年 8 月第 1 版
地　址	济南市二环南路 1 号	印　次	2015 年 8 月第 1 次印刷
邮　编	250002	定　价	30.00 元
网　址	www.jnpub.com		
电　话	0531-86131726	发行电话	0531-86131730
传　真	0531-86131709		86131731
经　销	各地新华书店		86116641
印　刷	山东天马旅游印务有限公司	传　真	0531-86922073

前 言

基因表达式编程算法（GEP）是进化计算家族的最新成果之一，是由葡萄牙生物学家、计算机科学家 Ferreira 于 2001 年正式提出的。GEP 算法借鉴了生物进化“适者生存”的思想，由个体形成种群，在选择、变异等算子的作用下，以适应度函数为控制目标，种群逐渐进化，最终得到最优解。尽管 GEP 算法的基本框架与遗传算法、遗传编程类似，但由于 GEP 采取了不同的编码方式，使得其进化效率比 GP 高出 100 倍以上。

GEP 的广阔应用前景引起了国际和国内学者的普遍关注。四川大学数据库和知识工程研究所较早地将 GEP 的研究引入国内，并发表了一系列学术论文。国内关于 GEP 的专门著作，目前仅有科学出版社出版的一本（元昌安等著，2010 年）。山东管理学院（原山东省工会管理干部学院）的科研团队，自 2011 年起，先后申报了 2 项 GEP 应用方面的科研课题（1. 基因表达式编程在高血压病靶器官损害预测中的应用，山东省科技厅星火办，项目编号为 2011XH17006；2. 基于 GEP 算法的山东省农村汽车保有量预测模型研究，山东省科技厅，项目编号为 2012G0022207），在 GEP 算法的应用方面积累了一定的经验，本书即是这些经验的总结。

本书共分为九章。

第一章介绍进化计算的生物学基础，并介绍了 GEP 算法的产生与发展，以及与 GA、GP 相比的特点与优势。第二章介绍 GEP 算法的基本概念，包括染色体、基因、各种遗传算子、适应度函数、选择策略等，并介绍了 GEP 算法的一般框架。最后以一元函数挖掘为例，说明了 GEP 算法在函数建模方

面的应用。第三章借鉴遗传算法的模式定理，得出了GEP算法的模式定理，从理论上解决了GEP算法得到最优解的可能性问题。第四章介绍了GEP算法在基因结构、遗传算子、种群多样性、适应度值、解码方法等方面的改进。另外，本章还介绍了GEP算法与小生境、免疫算法、信息熵等技术的结合，并将SA、PSO等方法引入常量的优化中。第五章介绍了函数挖掘的特点、现状和传统方法，并介绍了GEP算法在山东省汽车保有量预测、煤矿瓦斯喷涌量预测、地震等级预测等领域的应用。实践表明，在复杂函数挖掘中，GEP算法比人工神经网络等方法具有更高的精度和稳定性。第六章介绍了时间序列分析中嵌入维的计算方法：自相关系数法和关联积分法，并以股票数据和全国汽车保有量数据验证了GEP算法在时间序列预测中的可行性。另外，本章还介绍了基于GEP算法的常微分方程建模方法。第七章介绍了旅行商问题（TSP）的遗传算子，并介绍了形成初始解、提高搜索效率的方法，以及局部优化算子、优秀基因片段保护等技术。第八章介绍了GEP算法在分类、聚类问题中的应用，包括最大隶属度原则及K均值聚类与GEP算法的结合。第九章介绍了GEP算法在电路演化、作业车间调度、神经网络优化、支持向量机参数优化、图像恢复等领域的应用，并介绍了GEP算法的并行化方法。附录给出了全书中出现的缩写词汇，以及GEP算法的部分源代码，便于读者学习参考。

本书由山东管理学院陈青华、李亚梅、阮梦黎著，其中陈青华撰写第三、第四、第五、第六、第七、第八章，李亚梅撰写第一、第九章，阮梦黎撰写第二、第八章。陈青华对全书进行了统稿。

感谢山东管理学院信息学院袁锋教授以及科研处的领导为本书的出版做出的贡献。本书引用了许多学者的研究成果，在此表示衷心的感谢。

由于作者水平有限，加之GEP算法的应用不断发展，因此本书有不当之处在所难免，恳请读者批评指正。

作　者

2015年2月

目 录

第一章 绪 论

1.1 进化计算的生物学基础

1.1.1 遗传和基因表达

世界上的生物从其亲代继承特性或性状，这种现象称为遗传。由于遗传的作用，人们可以“种瓜得瓜，种豆得豆”。构成生物的基本功能单位是细胞，细胞中含有一种微小的丝状化合物称为染色体，生物的所有遗传信息都包含在染色体中。基因是遗传的生物学单位，基因决定了生物的显性特征，其中非常重要的部分之一就是指导合成包含生物酶在内的各种蛋白质。基因如何指导蛋白质的合成是我们所关心的。简单来说，在基因表达过程中有三种重要的物质：DNA，RNA 以及蛋白质。基因存在于 DNA 中，RNA 是中间媒介，而蛋白质是最终目的。

脱氧核糖核酸（DNA），又称去氧核糖核酸，是染色体的主要化学成分，同时也是组成基因的材料。有时被称为“遗传微粒”，因为在繁殖过程中，父代把自己 DNA 的一部分复制传递到子代中，从而完成性状的传播。DNA 是遗传信息的载体，遗传密码就由 DNA 中的四种不同的核苷酸结构表示。若干个碱基构成一个基因，基因主要包括结构基因和调节基因。其中，结构基因能够决定某一种蛋白质的氨基酸组成及排列顺序；而调节基因则对结构基因的表达有调节控制的作用。

一个基因是以 ORF（Open Reading Frame）的形式存在。ORF 是以一个

起始密码子开始，后面跟一系列的氨基酸密码子，最后以结束密码子结束。基因则除了包含 ORF 以外，还包括起始密码子的前序列，以及结束密码子的后序列。于是怎样确定一个基因的开始处，也就是找到 ORF 的起始密码子，这是一个很复杂的问题。基因中也可能存在未编码的部分，这些部分称为基因内区。

核糖核酸（RNA）是存在于生物细胞以及部分病毒、类病毒之间的遗传信息载体。蛋白质是一种复杂的有机大分子的组合，是生命最基本的组成部分之一，绝大多数具有生物活性物质的基本成分都是蛋白质。蛋白质的种类多种多样，每一种不同种类的蛋白质都具有特定的功能。蛋白质的基本组成单元是氨基酸。然而，自然界只有种类极其有限的氨基酸。将不同的顺序、不同的结构、不同种类的氨基酸组合在一起，从而形成各种各样的蛋白质。DNA 中的遗传信息指导蛋白质合成的过程称为“基因表达”。

生物按照如下的过程进行基因表达，根据承载于 DNA 上的基因制造出特定的蛋白质：

① 双螺旋结构的 DNA 在细胞核内解开双螺旋结构，形成两条单链。

② 位于细胞核内的 DNA 经过转录调控和加工调控传递给 mRNA。该过程中，每个 DNA 分子均以其自身作为模板，根据碱基配对的原则，即 A－U，T－A，G－C，C－G 来转录成为互补的 mRNA。

③ 细胞质的核糖体上进行蛋白质的合成。每一种蛋白质均是由直线序列的氨基酸组成。而这些氨基酸的种类和位置则是由 mRNA 分子上的连续三个碱基，即遗传密码子所决定。这些氨基酸通过肽键排列成为一定的顺序和结构，形成特定功能的蛋白质。

1.1.2 生物进化的系统观

生物在长期的生存过程中，逐渐适应其生存环境，品质不断得到改良，这种现象称为进化。生物的进化是以集团的形式共同进行的，这样的集团称为种群，组成种群的单个生物称为个体。不同的个体对生存环境有不同的适应能力，称为个体的适应度。

现代生物进化论主要依据达尔文的自然选择学说。达尔文认为，通过不同生物的交配以及一些其他原因（如环境的变化），生物的基因有可能发生变异而形成新的基因，这部分变异的基因也将遗传到下一代。生物基因变化的概率很小，具体哪一个个体发生变异完全是偶然的。新基因与生存环境的适应程度决定了其增殖能力，适应环境的个体将逐渐增多，不适应的将逐渐减少。通过这种自然的选择，物种将逐渐向适应生存环境的方向进化，从而产生优良物种。

生物界尚未完全揭开遗传和进化的奥秘，也不完全清楚染色体编码和译码的细节，但遗传和进化的以下几个特点却为人们所共识：

① 生物的所有信息都包含在染色体中，染色体决定了生物的特性。

② 染色体是由基因及其有规律的排列所构成的，遗传和进化过程是发生在染色体上。

③ 生物的繁殖过程由其基因的复制过程来完成。

④ 通过同源染色体之间的交叉或染色体的变异会产生新的物种，使生物呈现新的性状。

⑤ 对环境适应性好的基因（或染色体）比适应性差的基因（或染色体）有更多的机会遗传到下一代。

1.2 基因表达式编程算法（GEP）概述

受生物界遗传和进化思想的启发，研究者们陆续提出了很多进化算法，其中以遗传算法 GA 和遗传编程 GP 最受关注。基因表达式编程的思想也是起源于 GA 和 GP。本节首先介绍 GA 和 GP 的基本思想。

1.2.1 遗传算法 GA 概述

遗传算法 GA [1] 是 1962 年 Holland 教授首次提出的，它的基本思想是基于达尔文的进化论和孟德尔的遗传学说。遗传算法正是借用了仿真生物遗传学和自然选择机理，通过自然选择、遗传、变异等作用机制，实现各个个体的适应性的提高。从某种程度上说遗传算法是对生物进化过程进行的数

学方式仿真。这一点体现了自然界中“物竞天择、适者生存”的进化过程。

与自然界相似，遗传算法对求解问题的本身一无所知，从代表问题可能潜在解集的一个种群开始，每一个种群则由经过基因编码的一定数目的个体构成。每个个体实际上是染色体带有特征的实体。把问题的解表示成染色体，并基于适应值来选择染色体，遗传算法所需要的仅是对算法所产生的每个染色体进行评价，使适应性好的染色体有更多的繁殖机会，在算法中也就是以二进制编码的串。并且在执行遗传算法之前，给出一群染色体，也就是假设解。然后，把这些假设解置于问题的“环境”中来评价，并按适者生存的原则，从中选择出较适应环境的染色体进行复制，淘汰低适应度的个体，再通过交叉、变异过程产生更适应环境的新一代染色体群。对这个新种群进行下一轮进化，直到最适合环境的值出现。

遗传算法具有以下特点：

① 遗传算法从问题解的串集开始嫂索，而不是从单个解开始，这是遗传算法与传统优化算法的极大区别。传统优化算法是从单个初始值迭代求最优解，容易误入局部最优解。遗传算法从串集开始搜索，覆盖面大，利于全局择优。

② 许多传统搜索算法都是单点搜索算法，容易陷入局部的最优解。遗传算法同时处理群体中的多个个体，即对搜索空间中的多个解进行评估，减少了陷入局部最优解的风险，同时算法本身易于实现并行化。

③ 遗传算法基本上不用搜索空间的知识或其他辅助信息，而仅用适应度函数值来评估个体，在此基础上进行遗传操作。适应度函数不仅不受连续可微的约束，而且其定义域可以任意设定。这一特点使得遗传算法的应用范围大大扩展。

④ 遗传算法不是采用确定性规则，而是采用概率的变迁规则来指导其搜索方向。

⑤ 具有自组织、自适应和自学习性。遗传算法利用进化过程获得的信息自行组织搜索时，适应度大的个体具有较高的生存概率，并获得更适应环境的基因结构。

由于遗传算法的整体搜索策略和优化搜索方法是不依赖于梯度信息或其他辅助知识，而只需要影响搜索方向的目标函数和相应的适应度函数，所以遗传算法提供了一种求解复杂系统问题的通用框架。它不依赖于问题的具体领域，所以广泛应用于许多科学领域。下面将介绍遗传算法的一些主要应用领域：

① 函数优化

函数优化是遗传算法的经典应用领域，也是遗传算法进行性能评价的常用算例，许多人构造出了各种各样复杂形式的测试函数：连续函数和离散函数、凸函数和凹函数、低维函数和高维函数、单峰函数和多峰函数等。对于一些非线性、多模型、多目标的函数优化问题，用其他优化方法较难求解，而遗传算法可以方便地得到较好的结果。

② 组合优化

随着问题规模的增大，组合优化问题的搜索空间也急剧增大，有时在目前的计算上用枚举法很难求出最优解。对这类复杂的问题，人们已经意识到应把主要精力放在寻求满意解上，而遗传算法是寻求这种满意解的最佳工具之一。实践证明，遗传算法对于组合优化中的 NP 问题非常有效。例如遗传算法已经在求解旅行商问题、背包问题、装箱问题、图形划分问题等方面得到成功的应用。

此外，GA 也在生产调度问题、自动控制、机器人学、图像处理、人工生命、遗传编码和机器学习等方面获得了广泛的应用。

1.2.2 遗传编程 GP 概述

自计算机出现以来，计算机科学的一个重要目标就是让计算机自动进行程序设计，即只要明确地告诉计算机要解决的问题，而不需要告诉它如何去做。ArthurSamuel 早在 20 世纪 50 年代作为计算机科学的核心问题而提出的“遗传程序设计”（Genetic Programming，GP，又称遗传编程 [2]）便是在该领域的一种尝试，它借鉴生物界的自然选择和遗传机制，采用遗传算法（GA）的基本思想，但使用一种更为灵活的表示方式（分层树结构）来表示

解空间。分层树结构的叶结点是问题的原始变量，中间结点则是组合这些原始变量的函数。它们很类似于 LISP 语言中的 S－表达式。这样，每一个分层树原始结构对应问题的一个解，也可以理解为求解该问题的一个计算机程序。遗传程序设计就是采用遗传操作，动态地改变这些结构，以获得解决该问题的可行的计算机程序。

与传统的处理方法（如机器学习、人工智能）不同，GP 注重解的适应性，而不太强调传统求解过程所遵循的原则，这反映了 GP 的主要特点。

① 正确性：GP 处理的对象是不确切的解，它以误差作为进化的驱动，它所得到的解可能按传统的观点是不精确的。

② 一致性：传统方法在求解过程中是保持一致性的，即所得到的解是无矛盾、不冲突的，而 GP 则同时处理很多不一致、有冲突的解。并且这种不一致性能增加群体的多样性，从而能帮助遗传程序设计算法较快地解决问题。

③ 不推理性：传统的方法根据假设和已知条件以及逻辑规则，通过推理来获取结论。而遗传程序设计则不以逻辑推理为依据，它通过不断尝试和反复实验来求解问题。

④ 确定性：传统的方法以确定性的转移规则来求解具有严格数学描述的问题。而 GP 则以概率性的转移规则为基础。因此，其结果具有不确定性。

⑤ 秩序性：大多数传统的求解方法与算法不仅是确定的，而且其处理过程与步骤都是紧凑有序和同步的。而 GP 则是采用一种无序、非对等、独立、分布的异步并行处理方式，而且不受某中心处理过程的控制。

⑥ 决断性：传统算法大都采用确定的终止准则，此时可在一定程度上判断解所处的位置。而进化是一个连续的过程，因此 GP 无法定义明确的终止点，它通常需要通过人为干预终止其进化过程。

GP 是在遗传算法基础上发展起来的全局搜索算法，它与遗传算法有一定的区别，主要表现在：

① 由于遗传算法直接对定长字符串进行操作，所以不能描述层次化的问题，而遗传程序设计个体的树形表达方式则弥补了这一点。

② 遗传算法定长的字符串描述方法不具备动态可变性，每一种结构仅适

用于某类问题的求解，而遗传程序设计的程序结构不再考虑等位基因的位置，这带来了极大的灵活性，具有动态改变大小、形状的能力。

③ 遗传编程涉及的交叉算子更能体现语法结构的合理性，因为子树的交换受上下文语义环境的限制。

至目前为止，GP 已应用于许多领域，如电子工程、化学、经济、生命科学、艺术等，具体如下：

① 软件复用和自动程序设计：基于 GP 组件的软件复用方法为组件复用的自动化和工程化提供了一种可行的新途径。

② 多目标决策优化与分析：通过 GP 方法产生的决策函数比通过经典的传统 AHP 方法产生的决策函数更稳定。

③ 预测、分类和符号回归：使用历史数据库来预测新事例，发现各变量间的隐含关系。

④ 人工生命和机器人：控制机器人行为，使其对环境做出反应。用计算机模拟生物的自然进化或发现规律。

⑤ 神经网络设计：设计神经网络结构，发现学习规则和相关权值。

⑥ 图像和信号处理：图像识别、恢复及图像和声音的压缩等。

1.2.3 基因表达式编程算法 GEP 的产生与发展

基因表达式编程的产生来源于遗传算法（GA）和遗传编程（GP），它是借鉴生物遗传的基因表达规律提出的知识发现新技术。葡萄牙进化生物学家 Ferreira 博士早在 1999 年就有了思想萌芽；2000 年 Ferreira 的初期研究报告开始在网上陆续有报道，他提出基因表达式编程的效率比传统遗传编程系统高出 $10^2 \sim 6\times10^6$ 倍，这时学术界还对他的观点持怀疑态度，但他依然创办了基因表达式编程公司，开发了基于这一技术的软件，在理论、工程和系统三个方面向传统方法提出了挑战；2001 年，Ferreira 在 Complex Systems 杂志发表了第一篇原创性论文《Gene Expression Programming：A New Adaptive Algorithm for Solving Problems》，同时用副标题 Complete Reference for the First GEP Paper 说明了他的原创性；2002 年，Ferreira 出版了有关基因

表达式编程的第一本专著《Gene Expression Programming Ⅰ：Mathematical Modeling by An Artificial Intelligence》，书中详细介绍了 GEP 的基本方法，以及将 GEP 应用到诸如神经网络、数据挖掘、符号回归、时间序列预测、进化动力学、组合优化等实际问题中，至此正式开辟了 GEP 课题的研究和应用。随后 Ferreira 又发表了十余篇相关的学术论文，并于 2006 年出版了专著的第二版《Gene Expression Programming Ⅱ：Mathematical Modeling by An Artificial Intelligence》，再次注入了更新的思想，获得了两项专利。在 Ferreira 的各项研究成果中，不仅详细地阐述了 GEP 的各个关键技术 [3]，还分析了 GEP 的编码特点和重要作用，同时也将 GEP 应用到了诸如神经网络等其他交叉领域中。国际上对于 GEP 的研究也随之如火如荼地开展起来，重点主要在 GEP 的应用领域上，比如将 GEP 应用于分类规则挖掘和时间序列预测等。

国内对于 GEP [4] 的研究几乎与国际同步，四川大学数据库与知识工程研究所以唐常杰教授和左劼博士为代表的研究团队，于 2001 年第一时间将 GEP 的研究引入国内并开始深入研究。在国际会议 WAIM 2002 中，该团队发表了国内第一篇关于 GEP 的研究论文《Mining Predicate Association Rule by Gene Expression Programming》，文中用数学归纳法严格证明了被 Ferreira 直接采用而未加证明的收敛定理，“对任意良好定义的基因，Ferreira 的解码过程（算法）总能成功返回到对应的表达式树的根节点 [5]”，这一定理在理论上扫除了基因表达式编程问题研究的障碍，也由此开辟了该项课题在国内的研究之路。目前国内从事基因表达式编程研究的学术队伍日益壮大，包括四川大学、中国地质大学、武汉大学、华中科技大学、华南理工大学、江西理工大学、广西师范大学、华北电力大学等数十所高校都加入到研究队伍之列，同时渗透到的其他专业领域也在飞速发展，取得了很多宝贵的研究成果，如将生物中的转基因与重叠基因表达技术与 GEP 融合 [6]，将 GEP 应用于股票时间序列数据的挖掘 [7]、中医方证关系的挖掘 [8]、灾情的分析与预测 [9]、关联规则的挖掘 [10]、硬件进化和优化 [11] 等应用领域，这些都反映出 GEP 的应用前景非常广阔。