# 图的顶点着色问题的 DNA 算法

高 琳[1],许 进[2]
(1. 西安电子科技大学雷达信号处理国家重点实验室,陕西西安 710071;2. 华中科技大学系统科学研究所,湖北武汉 430074)

**摘 要**: 图的顶点着色问题是指无向图中任意两个相邻顶点都分配到不同的颜色,这个问题是著名的 NP-完全问题,没有非常有效的算法. 但在 1994 年 Adleman[1]首次提出用 DNA 计算解决 NP-完全问题,设计出一种全新的计算模式 —模拟生物分子 DNA 的结构并借助于分子生物技术进行计算,使得 NP-完全问题的求解可能得到解决. 本文首先提出了基于分子生物技术的图的顶点着色问题的 DNA 算法,算法的关键是对图中的顶点和顶点的颜色进行恰当的编码,以便于使用常规的生物操作及生物酶完成解的产生及最终解的分离,依据分子生物学的实验方法,本文提出的算法是有效和可行的;其次指出了该算法的优点、存在的问题及将来进一步的研究方向.

**关键词**: DNA 计算; NP-完全问题; 顶点着色问题; 限制酶
**中图分类号**: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2003) 04-0494-04

# A DNA Algorithm for Graph Vertex Coloring Problem

GAO Lin[1], XU Jin[2]
(1. *National Key Lab. of Radar Signal Processing, Xidian Univ., Xi'an, Shanxi* 710071, *China*;
2. *System Science Research Institute, Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*)

Abstract: Given an undirected graph, the vertex coloring problem is to assign a different color for vertex mutually adjacent. This problem is an NP-complete one and has no effective solving method. But Adleman[1] introduced firstly the DNA computing in 1994, with which the NP-complete problems are likely to be solved. DNA-based algorithm simulates molecular biology structure of DNA by means of molecular biology technological computation. This paper first introduces the DNA algorithm for the vertex coloring problem based on bio-molecular technology. The key of the algorithm is coding for the vertex and the color of the vertex The problem is solved by tube operation that performs the basic core processing and extraction that makes the results visible. On the basis of the experimental bio-molecular method, the algorithm is an effective method. Finally, the advantage and disadvantage are discussed, and the future research directions are pointed out.

Key words: DNA computation; NP-complete problem; vertex coloring problem; endonucleases

## 1 引言

图的着色问题是一个著名的组合优化问题,是现代图论中的一个主要的研究课题之一,它无论在理论上还是工程应用上均有良好的应用背景,诸如电路布局问题,工序问题,排课表问题以及存储问题[2]等均有直接的应用. 然而,无论是图 $G$ 的色数 $x(G)$ 还是对一个图 $G$ 进行正常 $k$-顶点着色, $k$-边着色以及 $k$-全着色问题都是 NP-完全问题[3]. 因此,不管是从事数学研究的图论学者们,还是从事电路与系统等工程技术方面研究的图论学家们,或者是其它领域的科学家们,都对图的着色问题很感兴趣. 由于用 $k \geq x(G)$ 种颜色对图 $G$ 进行正常 $k$-顶点着色算法是一个 NP-完全问题,NP-完全问题的求解一直困扰着人们. 近些年,人们用神经计算,进化计算等方法来求解 NP-完全问题也取得了一些进展. 基于分子生物技术的 DNA 计算是一种模拟生物分子 DNA 的结构并借助于生化反应作为计算工具的超大规模并行计算,而且 DNA 的双螺旋结构具有巨大的信息存储容量. 1994 年 Adleman[1]开拓性地采用现代分子生物技术,在试管中进行了 DNA 的实验,解决了有向图的哈密尔顿路问题 (Hamiltonian Path Problem,简记为 HPP). 虽然在实验室进行了 7 天的实验,才使一个只有 7 个顶点的有向图的哈密尔顿路问题得到解决. 但是由于他首先提出 DNA 计算的方法来解决 NP-完全问题,这一结果不但激发了人们对分子生物计算进一步研究的兴趣,对 NP-完全问题产生了新的希望与信心,而且开创了用分子生物技术研究组合优化问题的新途径,因而在国际上引起了巨大的轰动. 其后有许多学者沿此"路"而行,用 DNA 计算求解 SAT 问题[4,5]、图的最大团问题[6]、图的最大独立集问题[7]、最小集覆盖[8]问题等.

收稿日期:2001-10-26;修回日期:2002-07-31
基金项目:国家自然科学基金(No. 69971018, 60071026);陕西省自然科学基金(2001X05)

本文首先讨论了图的 3-顶点着色的 DNA 算法,对图的每个顶点用任意碱基排列的寡聚核苷酸片段编码,对顶点的颜色用具有特殊酶切位点的片段进行编码,通过并行重叠放大技术 POA(Parallel Overlap Assembly)建立数据池,然后运用分子生物操作如连接反应,聚合酶链式反应 PCR(Polymerase China Reaction),酶切反应,凝胶电泳对数据池进行运算,最终通过分子检测得到问题的解;其次分析了算法的优、缺点,并指出进一步的研究方向.

## 2　DNA 分子的计算特性

生物的各种生命活动都有它的物质基础,生物的遗传和变异也是这样.根据现代细胞学和遗传学的研究得知[9],控制生物性状遗传的主要物质是脱氧核糖核酸 DNA(deoxyribonucleic acid).DNA 是一种高分子化合物,组成它的基本单位是脱氧核苷酸.每个脱氧核苷酸是由一分子磷酸、一分子脱氧核糖和一分子含氮碱基组成的.含氮碱基有四种 A(Adenine,腺嘌呤)、G(Guanine,鸟嘌呤)、C(Cytosine 胞嘧啶)和 T(Thymine,胸腺嘧啶).DNA 不仅具有一定的化学组成,还具有规则的双螺旋结构.这一结构的主要特点是:(1)DNA 分子是由两条平行的脱氧核苷酸长链盘旋而成的;(2)两条链上的碱基通过氢键连接起来,形成碱基对,碱基对的组成有一定的规律,这就是嘌呤与嘧啶配对,而且腺嘌呤(A)一定与胸腺嘧啶(T)配对,鸟嘌呤(G)一定与胞嘧啶(C)配对.组成 DNA 的碱基虽然只有四种,而且这四种碱基的配对方式只有两种,但由于碱基对具有多种不同的排列顺序,因而就构成了 DNA 分子的多样性.在分子生物计算中,通常都采用 DNA 这种高分子化合物,这不仅因为 DNA 是生命信息的载体,而且在遗传工程实验中 DNA 易于操作.

DNA 算法解决计算问题的基本思想是:利用 DNA 特殊的双螺旋结构和碱基互补配对原则对问题进行编码,把要运算的对象映射成 DNA 分子链,在 DNA 溶液的试管里,在生物酶的作用下,生成各种数据池(Data Pool),然后按照一定的规则将原始问题的数据运算高度并行地映射成 DNA 分子链的可控的生化过程.最后,利用分子生物技术获得运算结果.

从 DNA 的原理来看,它与数学操作非常类似.DNA 的单链可看作由四个不同符号 A、G、C 和 T 组成的串.它在数学上就像计算机中的编码"0"和"1"一样,可表示成四个字母的集合 $\sum$ = {A,G,C,T}来译码信息.DNA 串可作为译码信息.酶可看作模拟在 DNA 序列上简单的计算.不同的酶相当于作用在 DNA 串上的不同的算子[10].

## 3　DNA 算法

### 3.1　问题描述

本文所言之图皆指无环、无重边的无向简单图,通常用 $V(G)$ 和 $E(G)$ 分别表示图 $G$ 的顶点集和边集,图 $G$ 的一个正常 $k$-顶点着色,简称为图的 $k$-点着色[2],是指用 $k$ 种颜色 1,2,…,$k$ 对 $G$ 的各顶点都分配(或称着)不同的颜色.换句话讲,简单图 $G$ 的一个正常 $k$-点着色,就是把 $V(G)$ 划分成 $k$ 个独立集的一个分类$\{V_1, V_2, \cdots V_k\}$,其中 $V_i(i = 1, \cdots, k)$ 是 $G$ 的独立集.我们用 $C(k)$ 表示 $k$ 种颜色,即 $C(k) = \{1,2, \cdots, k\}$.现在我们可以更确切地给出图 $G$ 的 $k$-点着色的定义:图 $G$ 的 $k$-点着色,是从 $V(G)$ 到 $C(k)$ 的一个映射 $\sigma$,当且仅当 $u, v \in V(G)$ 且 $(u, v) \in E(G)$ 时,$\sigma(u) \neq \sigma(v)$,全体 $G$ 的 $k$-点着色 $\sigma$ 构成的集合通常记作 $C_{\sigma k}(G)$,简记为 $C_k(G)$.若 $C_k(G) \neq \phi$(空集),既 $G$ 至少有一个正常 $k$-点着色,就称 $G$ 是 $k$-点可着色的.如图 1 所示为一个 6 顶点的图及其两种着色模式.

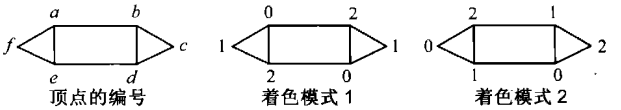


图 1　顶点的编号及两种着色模式

### 3.2　图的着色问题的 DNA 算法

对于具有 $n$ 个顶点的图,图的每一种可能的 3-顶点着色方案都可以表示为由 0、1 和 2 组成的 $n$ 位数字串,其中 0、1 和 2 分别表示三种颜色,如图 1 所示的一种着色方案表示为 021021.用这样的方法,我们可以把具有 $n$ 个顶点的图可能的各种着色方案转化为由 0、1 和 2 组成的 $n$ 位数字串的集合,称其为完全数据池.依照上述思想,为了求解图的着色问题,首先对图中的每个顶点及顶点的颜色用寡聚核苷酸片段进行编码,然后将这些寡聚核苷酸片段放在溶液中进行生化反应,生成问题的解,最终通过凝胶电泳分离问题的解.

具体的计算步骤如下:

Step 1　:对运算对象编码,建立完全数据池,将其作为 DNA 分子计算的输入数据.

对图 $G$ 的 $n$ 个顶点进行编码,每个顶点的编码由三部分构成,如图 2 所示,第一段的 $P_i$ 和第三段的 $P_{i+1}$ 表示位置,目的是为了在生化反应中各顶点的 DNA 片段通过并行重叠装配技术形成长的 DNA 链,中间部分 $V_i$ 表示各顶点颜色的编码.对每个顶点而言,有三种不同的编码,用来表示每个顶点的三种颜色,因此这三种编码除了中间的颜色链不同外,两边表示位置的链是完全相同的,如图 1 所示的顶点 1 的三种编码分别为:

CCCTGGGTACTGGATGCtcgaattcatAATGCTGAATGGCCCTT

CCCTGGGTACTGGATGCtctgacgaAATGCTGAATGGCCCTT

CCCTGGGTACTGGATGCggatccAATGCTGAATGGCCCTT


图 2　顶点编码示意图

为了区分每个顶点的颜色,中间颜色段的编码采用了具有特殊酶切位点的寡聚核苷酸序列,为了解的分离,序列的长度不同,采用这种方法建立的完全数据池有 $3^n$ 个元素.

Step 2　搜索满足着色条件的数据集.

对数据池中所有的串进行筛选,相邻顶点不能着同样的颜色,即在相应的数字串中对应位不能同时为 0、1 或 2,从数据池中删除(去掉)顶点相邻且相应的串值同时为 0、1 或 2 的所有数字串,如图 1 所示的顶点 $a$ 和 $b$ 相邻,不能着同样的颜色,因此 $a$ 和 $b$ 对应的串值不能相同,必须从数据池中删除这样的串,如图 3 所示.

Step 3　输出运算结果.

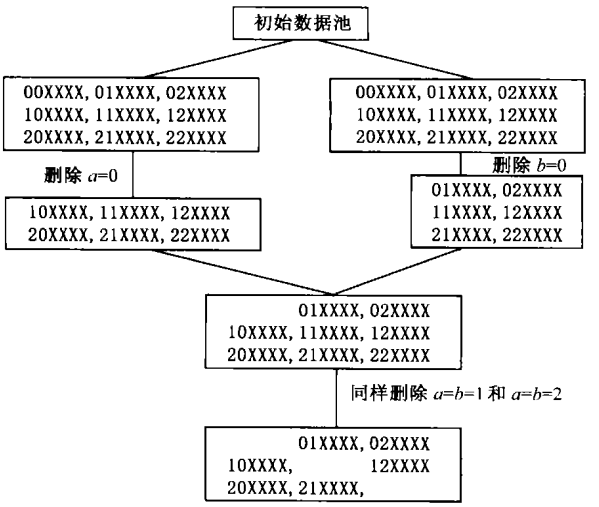


图 3 相邻顶点着色示意图

## 4 算法的生物实现

Step 1 建立数据池,以双链 DNA 表示数据结构,数字串中的每一位在 DNA 链上由三段组成,用 $i$ 表示所处的位置,当 $i$ 为奇数时表示为 $P_iV_i^mP_{i+1}$ ($m=0,1,2$),当 $i$ 为偶数时表示为 $\overline{P_{i+1}V_i^mP_i}$[6],上划线表示补序列,例对图 1 所示的例子,数字串长度为 6 位,在相应的 DNA 链上,有 6 个位的值序列 $V_1$ 到 $V_6$,依次间隔插入 7 个位置序列 $P_1$ 到 $P_7$,其中 $P_i$ ($i=1,\cdots,6$) 表示 $V_1$ 的位置,$P_7$ 用于生化反应过程中聚合酶链反应的引物,如图 4(a) 所示. 初始序列随机产生,但为了避免反应过程中错配现象的发生,不同序列具有相同碱基的长度不能超过 4bp,其次为了有效地识别顶点的不同颜色,必须加入

表 1 顶点的编码序列

| DNA 片段 | 序列(方向从 5 到 3′) |
|---|---|
| $P_1V_1^0P_2$ | CCCTGGGTACTGGATGCtcgaattcatAATGCTGAATGGCCCTT |
| $P_1V_1^1P_2$ | CCCTGGGTACTGGATGCtctgacgaAATGCTGAATGGCCCTT |
| $P_1V_1^2P_2$ | CCCTGGGTACTGGATGCggatccAATGCTGAATGGCCCTT |
| $P_3V_2^0P_2$ | AAAGGTCGGCTTGAATTaaggtaccggAAGGGCCATTCAGCATT |
| $P_3V_2^1P_2$ | AAAGGTCGGCTTGAATTgtcgaaAAGGGCCATTCAGCATT |
| $P_3V_2^2P_2$ | AAAGGTCGGCTTGAATTagcatgcgAAGGGCCATTCAGCATT |
| $P_3V_3^0P_4$ | AATTCAAGCCGACCTTTgagctcgcAAGGTTGGGTAACCCGT |
| $P_3V_3^1P_4$ | AATTCAAGCCGACCTTTataagcttgcAAGGTTGGGTAACCCGT |
| $P_3V_3^2P_4$ | AATTCAAGCCGACCTTTcccgggAAGGTTGGGTAACCCGT |
| $P_5V_4^0P_4$ | GTGTCGTCACGTGGTCCacctctatACGGGTTACCCAACCTT |
| $P_5V_4^1P_4$ | GTGTCGTCACGTGGTCCgtcgacACGGGTTACCCAACCTT |
| $P_5V_4^2P_4$ | GTGTCGTCACGTGGTCCattgcgcgccACGGGTTACCCAACCTT |
| $P_5V_5^0P_6$ | GGACCACGTGACGACACgtcaacAATTCCGGGCCCAATTG |
| $P_5V_5^1P_6$ | GGACCACGTGACGACACatcggccggcAATTCCGGGCCCAATTG |
| $P_5V_5^2P_6$ | GGACCACGTGACGACACatttaaagAATTCCGGGCCCAATTG |
| $P_7V_6^0P_6$ | CCTTGGGCCAATGGTGGagatctCAATTGGGCCCGGAATT |
| $P_7V_6^1P_6$ | CCTTGGGCCAATGGTGGttgatcatCAATTGGGCCCGGAATT |
| $P_7V_6^2P_6$ | CCTTGGGCCAATGGTGGatgcgcgcatCAATTGGGCCCGGAATT |

具有特殊酶切位点的限制性序列,这样就可以区分顶点的颜色. 采用并行重叠技术建立数据池[11],开始时用 18 个寡聚核苷酸片段,每个寡聚核苷酸片段的编码如表 1 所示. 18 个寡聚核苷酸片段放在一起进行热循环,在热循环过程中,一个寡聚核苷酸片段的位置串与另一个具有互补位置串的寡聚核苷酸片段退火,在聚合酶的作用下沿 3′方向延伸形成长的双链,最终形成了 $V_1V_2V_3V_4V_5V_6$ 的各种组合的数据池,如图 4(b) 所示. 随后以 $P_1$ 和 $\overline{P_7}$ 作为引物,利用多聚酶链式反应技术 PCR 进行扩增,就可以选择性地扩增那些以 $P_1$ 开始,$P_7$ 结束的 DNA 链.

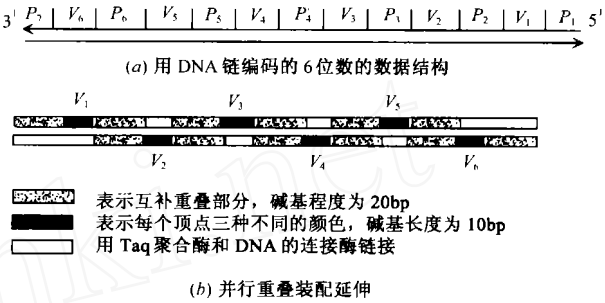


图 4 DNA 链编码的数据

Step 2 对数据池中的串用限制性内切酶进行筛选. 根据图的着色的定义,有边相连的顶点不能着同样的颜色,为了满足这个要求,用限制性酶的特殊酶切位点来完成. 限制性内切酶是一类能识别双链 DNA 分子中特异核苷酸序列的水解酶,如果某一核苷酸序列含有限制酶的识别序列,当加入这种酶后,就会将双链的 DNA 分子在酶切点切开,如 EcoRII[11]的识别序列为 GAATTC,酶切点在 G 与 A 之间,不同的酶具有不同的识别序列及相应的切割点,因此在图的着色中,为了区分不同的颜色,分别用含有不同酶的序列来表示这些颜色的特征. 如果在某条链上,两条相邻的顶点有相同的颜色,用限制内切酶将其对应的 DNA 链切开,在引物 $P_1$ 和 $\overline{P_7}$ 的作用下进行 PCR 扩增,切断的链将不会被扩增. 如图 1 所示 $a$ 和 $b$ 相邻,若同为 0(表示红色),将试管中的液体分为两个试管 $t_1$ 和 $t_2$,在 $t_1$ 中用 ECORI 切断含 $V_1^0$ 的串,在 $t_2$ 中用 KpnI 切断含的 $V_2^0$ 串,然后将两个试管中的液体进行合并,得到的数据池不含 00××××,若同为 1(表示蓝色),将试管中的液体重新分为两个试管 $t_1$ 和 $t_2$,在 $t_1$ 中用 PstI 切断含 $V_2^1$ 的串,在 $t_2$ 中用 SalI 切断含 $V_2^1$ 的串,然后将两个试管中的液体进行合并,得到的数据池不含 11××××,若同为 2(表示绿色),类似的操作加入酶 BamHI 和 SphI,得到的数据池将不含 22××××,最终得到的数据池如图 3 所示. 对于其它相邻的顶点,如果具有相同的颜色,其操作和上面的描述完全相同,只不过每次针对某个顶点加入不同的限制性内切酶,每个顶点不同颜色所对应的酶如表 1 中小写字母所示.

Step 3 输出运算结果. 上述运算结束后,剩余的 DNA 分子链对应的编码就是图的正常着色. 为了将这些 DNA 链分离开,采用聚丙烯酰胺凝胶电泳鉴定酶解产物[11],聚丙烯酰胺分离小片段(5～500bp)的效果较好,甚至可以分辨相差 1bp 的 DNA 片段. 根据我们的编码设计,图 1 中第一种着色方案

*021021* 的编码链长为 *171*bp,第二种着色方案 212010 的编码链长为 161bp,通过凝胶电泳将其分离开. 为了进一步知道每个顶点的着色情况,必须对 DNA 链进行序列测定,采用基因工程的方法进行,将目的 DNA 片段克隆于适当的载体(如 M13 噬菌体),产生一个能方便地进行测序的重组 DNA 分子,然后将重组子导入 E. Coli(大肠杆菌),进行克隆和测序,根据基因的测序结果就可知道每个顶点的着色情况.

## 5　结束语

本文讨论了图的 3-顶点着色的 DNA 算法,对图的每个顶点用任意碱基排列的寡聚核苷酸片段编码,顶点的颜色用具有特殊酶切位点的片段进行编码,通过并行重叠放大技术建立数据池,然后运用分子生物操作如连接反应,聚合酶链式反应 PCR,酶切反应,凝胶电泳对数据池进行运算,最终通过 DNA 测序得到问题的解.

本文的算法编码思想充分利用 DNA 分子结构的特征及常规的生物操作,将数学问题的求解同生物技术密切结合起来,但在具体的生物实现过程中仍有不足之处,具体表现为:(1)在 PCR 过程中,有可能产生单链 DNA,而限制性内切酶不能作用于单链 DNA;(2)内切酶的切除不完全. 由以上两个因素可能导致非法解的出现,但随着现代分子生物学技术的发展,这个问题会逐步得到克服,使得差错率控制在一定的范围内. (3)DNA 计算的最大特点是大规模的并行运算及巨大的信息存储能力,但编码过程中 DNA 链的数目随顶点数呈指数形式增长($3^n$),这是目前困扰 DNA 计算的一个障碍,也是 DNA 计算理论研究工作的一个研究问题[10]. (4)酶的种类为顶点个数的 3 倍,随着顶点数的增加,需要种类更多的酶,计算代价比较大,这个问题可以通过不同的生物操作过程得到改善,我们在这方面将做进一步的研究工作.

目前,DNA 计算的研究涉及到:(1)DNA 计算的能力及数学基础[13];(2)对于各种计算问题,怎样寻找一种直接的翻译方式,变换成 DNA 计算系统,也即 DNA 生物化学反应的运算途径,以至鉴别和输出最优解技术路线,使得 DNA 计算适应广阔的问题面,并具有实用性[14];(3)DNA 计算的复杂性、生物复杂性和可实现的衡量尺度[15];(4)基于 DNA 计算求解问题的装置并使之自动化,研究未来 DNA 计算机的可行性;(5)将 DNA 计算与遗传算法、神经网络等智能计算方法相结合[16,17].

## 参考文献:


[1] L Adleman. Molecular Computation of Solution to Combinatorial problems [J]. Science,1994,266 (11):1021 - 1024.

[2] J A Bondy,U S R Murty. Graph theory with application,the Macmillan press LTD [M]. London:Basingtoke and New York,1976.

[3] A Gibbons. Algorithmic graph Theory,Cambridge University dress [M]. London:Cambridge,1985.

[4] Richard J Lipton. DNA Solution of Computation Problems [J]. Science,1995,268(4):542 - 545.

[5] Qinghua Liu,et al. DNA Computing on Surface [J]. Nature,. 2000,403(13):175 - 179.

[6] Q Ouyang,et al. Solution of the Maximal Clique Problem [J]. Science,1997,278(17):446 - 449.

[7] T Head,et al. Computing with DNA by Operating on Plasmids [J]. Biosystem,2000,57:87 - 93.

[8] S Rowise,et al. A sticker Based models for DNA computation [EB/OL]. http://www.corninfo.chem.wisc.edu

[9] 黄翠芬,主编. 遗传工程理论与方法 [M]. 北京:科学出版社,1987.

[10] 高琳,许进,张军英. DNA 计算的研究进展与展望 [J]. 电子学报,2001,29(7):945 - 949.

[11] P C Turner,A G McLennan,A D Bates,M R H White. Molecular Biology [M]. Bios Scientific Publishers Limited. 2001.

[12] 姜泊,张亚历,周殿元,主编. 分子生物学常用实验方法 [M]. 北京,人民军医出版社,2000.

[13] D Boneh,et al. On the Computation Power of DNA [R]. USA:Prinecton University,1995.

[14] M H Garzon,et al. Biomolecular Computing and Programming [J]. IEEE Trans. On Evolutionary Computation,1999,3(3):236 - 250.

[15] H Garzon. The Bounded Complexity of DNA Computing [J]. Biosystems,1999,52:63 - 72.

[16] Deaton R,et al. A DNA Based Implementation of an Evolutionary Search for Good Encodings for DNA Computation [A]. Proceeding of 1997 IEEE International Conference on Evolutionary Computation,Indianapolis [C]. USA:IEEE,1997

[17] Russo M F,et al. An Improved DNA Encoding Scheme for Neural Network Modeling,World Congress on Neural networks San diego [A]. 1994 International Neural networks Society Annual Meeting [C]. USA:CA,1994. 354 - 359.


## 作者简介:


**高　琳**　女,1964 年生于陕西,1987 获西安交通大学数学系理学士学位,1990 年获西安电子科技大学数学系硕士学位,现为西安电子科技大学计算机学院副教授,该校雷达信号处理国家重点实验室博士生,感兴趣的研究领域为 DNA 分子生物计算、神经网络、遗传算法及其在组合优化问题中的应用.

**许　进**　男,1958 年生于陕西,教授,博士生导师,西安交通大学管理学院管理工程专业工学博士,北京理工大学应用数学系应用数学专业理学博士,西安电子科技大学电路与系统专业博士后,现任华中理工大学系统工程专业特聘教授,感兴趣的研究领域为神经网络、遗传算法、图论以及 DNA 计算等.