最小顶点覆盖问题的改进粘贴模型

董亚非* 张家秀* 殷志祥*** 许 进*
*(华中科技大学控制科学与工程系 武汉 430074)
**(安徽理工大学数理系 淮南 232001)

摘 要：DNA 计算是一种模拟生物分子 DNA 的结构并借助于分子生物技术进行计算的新方法。它开创了以化学反应作为计算工具的先例，具有广阔的应用前景。本文简单回顾了 DNA 计算的发展，并简要介绍了分子计算的一种模型——粘贴模型。最后我们利用粘贴模型的基本原理，运用荧光标记技术，提出了最小顶点覆盖问题的表面技术解决方案。

关键词：DNA 计算，粘贴模型，荧光标记技术，最小顶点覆盖问题
中图分类号：TP18 **文献标识码**：A **文章编号**：1009-5896(2005)04-0556-05

An Improved Sticker Model of the Minimal Covering Problem

Dong Ya-fei* Zhang Jia-xiu** Yin Zhi-xiang*** Xu Jin*
*(*Department of Control Science and Engineering, HUST, Wuhan* 430074, *China*)
**(*Department of Mathematics and physics, AUST, Huainan* 232001, *China*)

Abstract DNA computing is a new computation method with simulating molecular biology structure of DNA and by means of molecular biology technology. This method has been widely used in many respects. Simply reviewed the progress of DNA computing, the paper introduces a new model of molecular computation that is called the sticker model. Finally, the solution of the minimal covering problem on surface using fluorescence marking technology is proposed based on the principle of sticker model.

Key words DNA computing, Stickers model, Fluorescence marker, Minimal covering problem

1 引言

自从 Adleman 教授[1]于 1994 年开创性地给出了基于 DNA 分子与 DNA 连接酶等、通过可控的生化反应解决了 7 个顶点的有向 Hamilton 路问题之后，关于 DNA 计算模型、DNA 计算的实验方法与技术、以及 DNA 计算模型机理等方面的研究越来越多，特别是 DNA 计算机的研究倍受学者们的关注。DNA 计算机的研究已经在国际上形成了科学研究领域的一个新的“热点”、新的生长点。1995 年，Lipton[2]仿效 Adleman 的方法对另一个经典的 NP-完全问题——可满足性问题(SAT 问题) ——给出了求解方案。Lipton 的主要贡献是把 DNA 分子上的碱基对翻译成一串 1 和 0 的编码过程，然后把已经测定了序列的 DNA 分子从不同的试管里取出并混合在一起，从而允许 DNA 分子模仿电子门做出“是”与“非”的判断，也就是说他让 DNA 具有了“思维”，有了逻辑判断能力。2000 年，Faulhammer 等人[3]在实验里用生物技术实现了 Lipton 的上述设想。此后，有诸多学者给出了不同类型的图和组合优化问题的 DNA 计算方法和结果，如 Adleman 提出了的 3 可着色图的线性算法[4]；Lipton 给出的连接网络的可满足性问题[5]、电路、最大电路以及正规电路等的可满足性问题[6]；2002 年，Barich 等人[7] 成功地应用粘贴 DNA 计算模型对具有 20 个变量的可满足性问题进行了求解，这是目前从规模上应用 DNA 计算模型所得到的最好的结果。不同的分子材料也被引入到 DNA 计算中来，2000 年，Head 提出了用质粒 DNA 分子来进行可满足性问题[8]。同年，Dirk 等人[9]用 RNA 分子代替 DNA 分子给出了可满足性问题的一种实验性的计算模型，并讨论了国际象棋问题的 RNA 计算模型。2000 年，Liu 等人[10]成功在镀金表面对可满足性问题进行了 DNA 计算实验。

从 DNA 计算的发展现状来看，目前关于 DNA 计算机的研究仍然处于理论探索阶段，其中模型研究是其核心内容。在模型方面的研究上，已经有不少的 DNA 计算模型被

2003-12-04 收到，2004-04-05 改回
国家自然科学基金(30370356, 60274026)和国家博士点基金资助课题

提出，其中近几年来比较流行的模型是剪接系统模型(splicing system model)[11-13]和粘贴模型(sticker model)[13-17]。剪接系统模型是基于剪接操作(分子生物操作的抽象，每个操作是切割双链 DNA 分子和再重新连接切割部分以获得新的 DNA 分子)，剪接系统是计算完备的，即每个 PASCAL 程序可通过一个适当的剪接系统来模拟,反之亦然。2001 年 11 月，一个可编程的有限自动机已经通过剪接系统模型实现[18]。粘贴模型不仅与剪接系统模型一样的计算能力，且具有操作简单明了，理论性强，特别是易于硬件实现等优点。

在本文中，我们利用粘贴模型的基本原理，运用荧光标记技术，提出了最小顶点覆盖问题的表面技术解决方案。

2　粘贴模型

粘贴模型在对二进制串进行表述时，通常使用两组基本的单链 DNA 分子。将长度为 N 个碱基的存储链细分为含有 K 个长度为 M 个碱基的无悬垂区段（因此 $N \geqslant MK$）。我们称 M 个碱基长度的序列为一个位点，在计算的过程中，每个位点实际定义为一个二进制位点（或等价于一个布尔变量）。同时设计 K 个不同的寡聚核苷酸，我们称之为粘贴串，每个粘贴串长度为 M 个碱基，并且与 K 个存储位点中的一个且只有一个位点成为互补关系。

粘贴模型在位串集合上是通过四种基本的操作完成计算的：合并、分离、设置和清除。

合并(merge)即将两个位串的集合合并在一起生成一个新的集合，该集合包含了两个初始集合中全部串的合成物。这就相当于产生了一个新的试管，该试管含有两个初始试管混合后所生成的全部存储合成物。

分离(separate)是将一个集合分成两个新的集合，操作的结果是其中一个集合包含了特殊位元全部为开(on)的初始链，而另一个集合则包含了特殊位元全部为关(off)的初始链。这就相当于用粘贴串退火到给定位元的区，从而从集合试管中实际隔离那些合成物。初始输入集合(试管)按照给定的限制条件而被分开。

设置(set)是在集合中，对所有串的任意位元进行设置，当集合中的某一特殊位元设置为开时，即意味着代表那个位元的粘贴串在该集合试管里所有合成物上发生退火反应。

清除(Clear) 是对一个集合的所有串的第 i 个位元进行清除(关)，即那个位元的粘贴串必须从集合试管里的所有存储合成物中除去。

在粘贴模式里，计算由一系列的合并、分离和设置操作按顺序组成。这个顺序起始于某些初始位串集合，终止于产生一个“答案串”。将输出结果读出，就是将一个存储合成物从输出试管中分离出来，并确定其已复性的粘贴串；否则就显示输出试管中没有存储合成物。

粘贴模型的计算模式就是在长度为 L 的输入上进行穷举组合搜索来处理一些难题的。所有 2^L 个输入都以并行方式处理。可以认为这就是 DNA 计算的最一般的本质。

3　最小顶点覆盖问题

图的顶点覆盖问题是指找出给定图中顶点的的一个最小子集，图中的任意一条边的两个端点都至少有一个属于该子集。它在分子生物学，调度问题，信息检索，错误诊断和恢复，集装线平衡，油轮行程安排，及开关理论有着广泛的应用。

设简单图 $G=(V, E)$是一无向图，其中 V 是图中顶点集合，E 是边的集合。p 表示图 G 中顶点的个数，q 表示图 G 中的边数。若图 G 中的一个顶点和一条边相互关联，则称它们相互覆盖。覆盖图 G 的所有边的一个顶点子集称为图 G 的一个顶点覆盖。类似地，覆盖图 G 的所有顶点的一个边子集称为图 G 的一个边覆盖。图 G 的所有顶点覆盖中顶点最少的数目称为图 G 的顶点覆盖数，或者简称为点覆盖数，记为 $\alpha_0(G)$，或简记为 α_0。图的顶点覆盖问题是找一 V 的子集 $S \subseteq V$，使得 G的每条边至少有一个顶点在 S中，即 对于$(i, j) \in E$, i 和 j 至少有一个属于 S。而图的最小覆盖问题是找出 S中的元素最少的一个子集。

对于上述问题，Roweis 运用粘贴模型在试管中给出了 DNA 计算的解决方案[16]。设存储链有 $k=A+B$ 个子串。初始试管 N_0 是一个（$A+B$，B）的库。初始试管 N_0 中的存储合成物代表$\{1, 2, \cdots, B\}$的所有可能的子集 L。每一个存储合成物的前 B 个子集的开或关的状态表示了 $1, 2, \cdots, B$ 中的一个数，它属于该存储合成物所代表的特定的子集 L。开始时，每一个存储合成物 M 的后 A 个子串是关闭的。前 $B+J$ 个子串最终会打开，其中 $1 \leqslant J \leqslant A$，$J$ 为一些集合 C_i 的数目，而 i 就属于 M 所代表的指示集合 L。因此，给定 M，进行如下过程：首先检查 M 的前 B 个子串；只要前 B 个子串为开时（假定为第 i 位，可以用分离操作区分其“开”或“关”），就使用设置操作将 M 的后 A 个子串之中代表集合 C_i 中元素个数的子串置“开”。按照这种方式将 M 的前 B 个子串检查完后，再来看每存储合成物的后 A 个子串是否已置“开”。这个步骤同样可以通过分离来完成。后 A 个子串已置“开”就意味着 M 所代表的指示集合 L 的确覆盖了集合 $S=\{1, 2, \cdots, A\}$。因此，去掉不满足这种条件的存储合成物，然后在其中寻找满足此条件的最小的指示集合 L。

高琳[16]基于粘贴模型设计了顶点覆盖问题的 DNA 算法，对图 G 有 n 个顶点，m 条边，DNA 链有 $n+m$ 位，前 n 位表示顶点，后 m 位表示边，该算法设计思路精巧，通过反

复进行合并、设置、清除、分离四个基本操作来完成复杂的计算任务。这些操作需要通过杂交、退火和探针的设计完成，步骤很多，另外，在操作的过程中，由于杂交及退火的不完整性，会导致大量的伪解出现。

我们曾运用并行重叠放大技术(Parallel Overlap Assembly, POA)[19]建立初始数据池，通过运用分子生物操作如连接反应，聚合酶链式反应(Polymerase Chain Reaction, PCR)，酶切反应，凝胶电泳对数据池进行运算，对图 $G=\{V,E:V=6,E=11\}$ 的最小顶点覆盖做了讨论[20]。

4 最小顶点覆盖问题的表面粘贴模型算法设计

最小顶点覆盖问题的算法可以由以下几步完成：

步骤 1 生成给定集合覆盖的所有可能组合；

步骤 2 利用每一约束条件剔除非可行解(保留可行解)；

步骤 3 生成剩余的解；

步骤 4 重复进行步骤 2, 3，我们可以排除所有的非解，从而得到问题的所有可行解；

步骤 5 比较各可行解对应的覆盖值，进而得到最优解。

对应于该算法的生物步骤简单的描述如下：

步骤 1 制作代表一个给定集合的覆盖的计算模板以及集合中所有顶点和边的寡聚核苷酸片断，并加上荧光猝灭分子。然后将其标号后分别固定在表面上；

步骤 2 通过寡聚核苷酸片断在被其互补链杂交时的荧光猝灭，判断不满足顶点覆盖的组合；

步骤 3 加热解链（对于步骤 2 已经判断为不满足的不再考虑），清洗掉所有杂交的补；

步骤 4 重复进行步骤 2，3，可以排除所有的不满足条件的组合，从而也就判断了给定问题的可行解；

步骤 5 对于步骤 4 所得到的所有的可行解，通过观察荧光判断出最优解。

5 最小顶点覆盖问题的模型系统

对于一个含有 n 个顶点 $v_1,v_2,\cdots,v_n$ 和 m 条边的图。对应算法的步骤 1，我们分为两步，首先合成计算模板的寡聚核苷酸片断。我们把它分为两部分，前 n 个寡聚核苷酸片断分别表示顶点 $v_1,v_2,\cdots,v_n$；后 m 个寡聚核苷酸片断分别表示边 $e_1,e_2,\cdots,e_m$；其次合成 n 个寡聚核苷酸片断代表顶点对应的补链，并分别记为 $\overline{v}_1,\overline{v}_2,\cdots,\overline{v}_n$； m 种寡聚核苷酸片断分别表示边对应的补链，并分别表示为 $\overline{e}_1,\overline{e}_2,\cdots,\overline{e}_m$。为了避免它们之间的错误杂交，我们选择的每组寡聚核苷酸应具有较大差异，至少有 4 个以上是不同的。然后我们利用生成的寡聚核苷酸片断构造 DNA 计算初始数据库，这里分为 2 步：(1) 在计算模板 DNA 片断上连接上荧光素，其中代表顶点的寡聚核苷酸片断用一种荧光素，代表边的寡聚核苷酸用另一种荧光素；(2) 将构造好的 DNA 片断分别点样到固体表面上。对应算法的步骤 2，根据每组顶点集合所关联的各个边，把他们的补链 $\overline{v}_i,\overline{e}_j$ 加到表面上（在补链上连接上荧光猝灭剂）。通过杂交使表面上的 DNA 链荧光猝灭程度，利用激光共聚焦显微镜观察 DNA 链是否完全覆盖。对应算法的步骤 3，我们对步骤 2 的产物进行加热解开双链，冲洗掉与 DNA 链杂交的所有补链。对应算法步骤 4，我们重复步骤 2,3 的操作，我们就可以得到满足约束方程组的可行解。对应算法的步骤 5，我们对得到的所有覆盖集合加以比较后，就可以找到问题的最优解。下面我们来具体讨论一个简单顶点覆盖问题，见图 1。

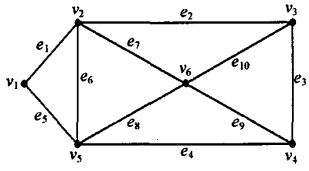
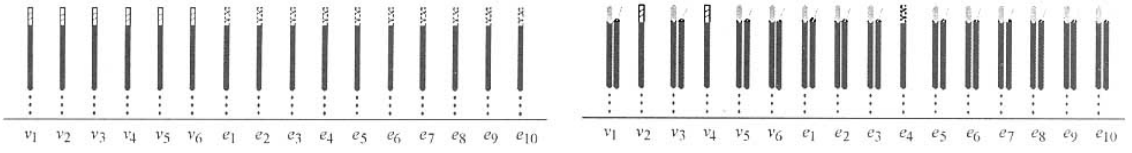


图 1 简单的覆盖图

要讨论这个顶点覆盖问题，首先我们构造 32 种短的寡聚核苷酸，它们中的 16 种分别代表 $v_1,v_2,\cdots,v_6$ 和 $e_1,e_2,\cdots,e_{10}$，剩下的 16 种表示其补链(注意到这里只需构造 16 个寡聚核苷酸，另外 16 个就已确定)，我们把它们分别记为 $\overline{v}_1,\overline{v}_2,\cdots,\overline{v}_6$ 和 $\overline{e}_1,\overline{e}_2,\cdots,\overline{e}_{10}$ (如图2)。现在我们将这 16 种寡聚核苷酸 $v_1,v_2,\cdots,v_6$ 和 $e_1,e_2,\cdots,e_{10}$ 按顺序用巯基修饰固定化核苷酸片断,将他们固定到表面上，同时前 6 个表示顶点的寡聚核苷酸片断连接上红(Cy5 ™，激发波长为 649nm, 发射波长为 670nm)，后 10 个表示边的寡聚核苷酸片断连接上蓝(Cy3 ™，激发波长为 550nm, 发射波长为 570nm) 两种颜色（如图 3 所示）。

v_1 : CAACCCAA;	$\overline{v}_1$: GTTGGGTT;	v_2 : AACCTGGT;	$\overline{v}_2$: TTGGACCA
v_3 : ACCAAACC;	$\overline{v}_3$: TGGTTTGG;	v_4 : AGAGTCTC;	$\overline{v}_4$: TCTCAGAG
v_5 : ATATCGCG;	$\overline{v}_5$: TATAGCGC;	v_6 : CCAAGTTG;	$\overline{v}_6$: GGTTCAAC
e_1 : GGTTCAAC;	$\overline{e}_1$: CCAAGTTG;	e_2 : GTTGGGTT;	$\overline{e}_2$: CAACCCAA
e_3 : TATAGCGC;	$\overline{e}_3$: ATATCGCG;	e_4 : TCTCAGAG;	$\overline{e}_4$: AGAGTCTC
e_5 : TGGTTTGG;	$\overline{e}_5$: ACCAAACC;	e_6 : TTGGACCA;	$\overline{e}_6$: AACCTGGT
e_7 : ACTGGTCA;	$\overline{e}_7$: TGACCAGT;	e_8 : CAGTTGAC;	$\overline{e}_8$: GTCAACTG
e_9 : ATGCAGGA;	$\overline{e}_9$: TACGTCCT;	e_{10} : ATCGAGCT;	$\overline{e}_{10}$: TAGCTCGA

图 2 编码示意图



(a) 初始表面 DNA 模板示意图　　(b) 表面 DNA 杂交后的模板示意图

图 3

在图 3(a)中，黑色柱状物表示顶点和边的编码序列，红色柱状物（▮）表示红色荧光剂，蓝色柱状物（▮）表示蓝色荧光剂，虚线表示将寡核苷酸连接到玻璃镀膜表面上的联接臂。在图 3(b)中，粘贴串用较短的黑色柱状物表示，其顶端的细线状物用于表示荧光猝灭剂，猝灭效果用灰色表示。例如，在图 3(b)中，顶点 v_2,v_4 和边 e_4 没有互补链杂交，则其顶端的荧光剂仍在发不同颜色的荧光；而其余各条 DNA 链分别有互补链杂交其上，在荧光猝灭剂的作用下，没有荧光出现。

现在我们用 6 – 9 个原子的联接臂将 $\overline{v}_1,\overline{v}_2,\cdots,\overline{v}_6$ 和 $\overline{e}_1,\overline{e}_2,\cdots,\overline{e}_{10}$ 分别连接上红蓝两种颜色对应的荧光猝灭剂,荧光猝灭剂是二甲氨基偶氮苯甲酰(DABSYL)基,对多种荧光素都有很强荧光猝灭效率.对于任意一组解空间 $\{v_i\}$，我们在表面上加入核苷酸 $\{v_i\}$ 对应的补链 $\{\overline{v}_i\}$ 及其所关联的边 $\{e_j\}$ 的补链 $\{\overline{e}_j\}$，利用激光共聚焦显微镜观察对应于边的寡聚核苷酸的颜色变化情况，有蓝色亮点的寡聚核苷酸出现，说明未完全覆盖，加热表面解开双链，并冲洗；而蓝色全部猝灭的寡聚核苷酸对应的编码变量满足覆盖条件（对于本例，它们是“$\{v_1,v_2,v_3,v_4,v_5,v_6\}$，$\{v_2,v_3,v_4,v_5,v_6\}$，$\{v_1,v_3,v_4,v_5,v_6\}$，$\{v_1,v_2,v_4,v_5,v_6\}$，$\{v_1,v_2,v_3,v_5,v_6\}$，$\{v_1,v_2,v_3,v_4,v_6\}$，$\{v_1,v_2,v_3,v_4,v_5\}$，$\{v_1,v_2,v_4,v_6\}$，$\{v_1,v_3,v_5,v_6\}$，$\{v_2,v_3,v_4,v_5\}$,$\{v_2,v_3,v_5,v_6\}$，$\{v_2,v_4,v_5,v_6\}$”，见图 4。拍照其图片并保留，加热表面解开双链，并冲洗。对应于步骤 5，我们将所得到的所有满足覆盖条件的解进行比较，观察荧光数可得到最小顶点覆盖，红色荧光最多的为最小覆盖集合，覆盖点为红色荧光猝灭的顶点（本例为 $\{v_1,v_2,v_4,v_6\}$，$\{v_1,v_3,v_5,v_6\}$，$\{v_2,v_3,v_4,v_5\}$,$\{v_2,v_3,v_5,v_6\}$，$\{v_2,v_4,v_5,v_6\}$）。

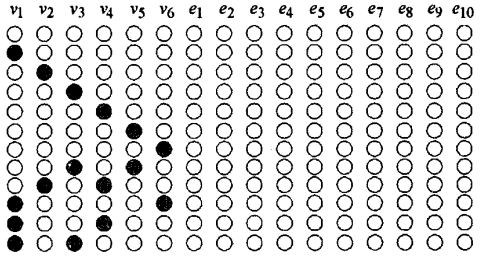


图 4　最优解检测图

6　结论分析

我们在算法的设计中采用了荧光猝灭的有关技术，利用观察荧光来排除非解，这种读解方法简单而有效且错误率低。由于检测方法的改变，省略了在试管计算中常用的的酶切、磁珠分离、PCR 扩增、凝胶电泳等步骤，避免了在这些步骤中可能出现的计算误差和数据丢失；并且我们使用的材料可重复使用，与在试管溶液中的 DNA 计算相比，更接近于工程实践。但是，自然界可供我们使用的荧光素种类有限，如果 DNA 计算的规模变得巨大，有可能造成计算材料上的困难，这一点与在试管溶液中进行 DNA 计算，有时受到各种酶的限制一样，这一点也是目前 DNA 计算的瓶颈。然而，DNA 计算要实现工程化，这一点是必须逾越的，这将有待于分子生物学及生物工程技术的进一步发展。我们的方法可以非常方便地解决最小覆盖问题，该问题是一个 NP 完全问题。我们的方法是DNA计算粘贴模型利用表面技术的一种尝试，随着分子生物学及生物工程技术的进一步发展，也许，在不远的将来，利用 DNA 计算粘贴模型求解更为复杂的问题将成为现实。

参 考 文 献

[1] Adleman L, Molecular computation of solutions to combinatorial problems. *Science*, 1994, 266(11): 1021 – 1024.

[2] Lipton R. DNA solution of hard computation problems. *Science*, 1995, 268 (4): 542 – 54s5.

[3] Faulhammer D, Cukras A R, Lipton R J, *et al*.. Molecular computation: RNA solution to chess problem. *Biochemsty*, 2000, 97: 1385 – 1389

[4] Adleman L. On constructing a molecular computer. Technical Report TR. 79-387, Computer Science Department, University of Southern California, USA, January, 1995.

[5] Lipton R. Using DNA to solve SAT, 1995.http://www.cs.princeton.edu/rj1/bio.ps, December 1994.

[6] Boneh D, Dunworth C, Lipton R, *et al*.. On the computational power of DNA, Technical Report TR-499-95, Princeton University, USA, October 1995.

[7] Braich R S, Chelyapov N, Johnson C, *et al*.. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*, 2002, 296(19): 499 – 502.

[8] Head T, Rozenberg G , Bladergroen R B, *et al*.. Computing with DNA by operating on plasmids. *BioSystems*, 2000, 57: 87 – 93.

[9] Dirk F, Cukras A R, Lipton R J, *et al*.. Molecular computation: RNA solutions to chess problem. *Biochemsty*, 2000, 97: 1385 – 1389.

[10] Liu Q H, Wang L M, Frutos A G , *et al*.. DNA computing on surfaces. *Nature*, 2000, 403(13): 175 – 178.

[11] Head T. Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviors, *Bull. Math. Biology*, 1987, 49: 737 – 759.

[12] Kari L. DNA computing: arrival of biological mathematics. *Math. Intelligencer*, 1997, 19(2): 9 – 22.

[13] Praun G, Rozenberg G, Salomaa A. DNA Computing—New Computing Paradigms. Berlin: Springer, 1998: 32 – 63.

[14] Roweis S, Winfree E, Burgoyne R, *et al*.. A sticker based archtecture for DNA computation, in: Baum E B *et al*.(Eds), DNA Based Computers, Proc. 2nd Annual Meeting, Princeton, 1999: 1 – 27.

[15] Paun G , Rozenberg G . Sticker systems. *Theoretical Computer Science*, 1998, 204: 183: 203.

[16] Gao L, Xu J, DNA Solution of Vertex Cover Problem Based on Sticker Model. *Chinese Journal of Electronics*, 2002, 11(2): 280 – 284.

[17] Zimmermann K H. Efficient DNA sticker algorithms for NP-complete graph problems. *Computer Physics Communications*, 2002, 144: 297 – 309.

[18] Benenson Y, Tamar P E, Rivka A, *et al*.. Programmable and autonomous computing machine made of biomolecules, *Nature*, 2001, (414): 430 – 434.

[19] Ouyang Q, *et al*.. DNA solution of the maximal clique problem, *Science*, 1997, 278(17): 446 – 449.

[20] Dong Ya-fei, Wang Shu-dong, Yin Zhi-xiang, *et al*.. DNA solution of the minimal covering problem. *Advances in Systems Science and Applications*, 2003, 3(2): 152 – 156.

董亚非： 男，1963年生，副教授，主要研究方向为图与组合优化、DNA计算.

殷志祥： 男，1966年生，教授，主要研究方向为图与组合优化、DNA计算、蛋白质结构预测.

许 进： 男，1959年生，教授，博士生导师，主要研究方向为图与组合优化、神经网络、DNA计算、蛋白质结构预测.