最小自由能约束的 DNA 编码设计研究

殷 脂[1,2],叶春明[1],温 蜜[2]
YIN Zhi[1,2],YE Chun-ming[1],WEN Mi[2]

1.上海理工大学 管理学院,上海 200093
2.上海电力学院 计算机信息工程学院,上海 200090
1.Business School,University of Shanghai for Science and Technology,Shanghai 200093,China
2.Department of Computer and Information Engineering,Shanghai University of Electric Power,Shanghai 200090,China
E-mail:yzzhizhi@163.com

YIN Zhi,YE Chun-ming,WEN Mi.Research on DNA encoding design constraint by minimal free energy.Computer Engineering and Applications,2010,46(12):25-27.

Abstract: This paper firstly introduces the importance of free energy based constraint in DNA sequence design and the formulas for calculating free energy,then adopts an improved Ant Colony Optimization(ACO) algorithm to solve sequence encoding problem. Emulation exercise shows that this method can generate a group of DNA sequences which satisfies free energy constraint and certain range of melt temperature.This method improves efficiency of DNA sequence design by leveraging concurrent executing of ACO,and generates more stable DNA sequence.

Key words: DNA computing;DNA encoding;free energy;Ant Colony Optimization(ACO)

摘 要:首先介绍了 DNA 编码设计中自由能约束的重要性,以及自由能约束的计算公式,进而采用一种改进的蚁群优化算法来求解。仿真实验表明此算法产生一组能满足特定自由能约束和统一的解链温度约束的 DNA 序列,算法利用蚁群算法的并行性提高了编码设计算法的效率,利用最小自由能约束产生更稳定的 DNA 序列。

关键词:DNA 计算;DNA 编码;自由能;蚁群优化
DOI:10.3778/j.issn.1002-8331.2010.12.007 **文章编号**:1002-8331(2010)12-0025-03 **文献标识码**:A **中图分类号**:TP301

1 引言

20 世纪 50 年代,Richard Feynman 首次提出在分子水平上计算的思想,阐述分子计算的可行性。1994 年,Adleman 首次将 DNA 计算用于解决 7 个顶点有向 Hamilton 路有向图问题[1]。DNA 计算过程分为 3 个阶段:(1)编码阶段:将待解决的问题模型通过编码,映射为 DNA 分子的集合,作为输入;(2)计算过程:即生化反应过程,在编码的 DNA 分子间进行各种可能的生化反应,也是产生解空间的过程;(3)提取解的过程:即标示目标解的过程,在计算结果的解空间中,运用分子生物技术,例如 PCR(Polymerize Chain Reaction)技术、分子纯化、电泳、磁珠分离等,将满足特定条件的解提取。由此可见,生化反应的精确度对于 DNA 计算是非常重要的,而 DNA 序列的编码又是影响生化反应的关键因素,学者们进行了深入研究。文献[2]重点分析了汉明距离约束的特点及编码方法;文献[3]对传统的基于汉明距离和基于自由能的 DNA 序列编码方法进行了分析,强调自由能的重要性;文献[4]提出了基于部分字的编码设计方法。以上编码设计方法均采用汉明距离近似代替最小自由能来评价 DNA 序列的稳定性。

该文将自由能评价公式的特点与蚁群算法相结合,充分考虑 GC 含量约束并根据给定的最小自由能值产生一组 DNA 编码序列,由此产生的 DNA 序列的碱基分布更均匀,具有更高的热力学稳定性,有助于避免生化反应中的假阳性现象,使用者也可以在此基础上做更进一步的编码过滤。

2 DNA 计算中的编码问题以及编码质量评价方式

Garzon 将 DNA 计算中的编码问题定义为[5]:以构成 DNA分子的 4 个碱基(A,C,G,T)为字母表,在一个长度为 N 的 DNA 分子的编码集合 S 中,求 S 的一个子集 $C\subseteq S$,使得 $\forall s_i,s_j\in C$ 满足:$\tau(s_i,s_j)\geqslant k$,其中 k 为正整数,τ 是评价编码的期望准则,即编码应满足的约束条件。例如汉明距离、GC 含量、解链温度 T_m、自由能 ΔG 等。

2.1 解链温度 T_m 约束

解链温度是决定 DNA 反应效率的重要参数,为有效地降低不完全匹配双链产生的概率,提出了解链温度约束,其主要

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60903188);上海市高校选拔培养优秀青年教师科研专项基金资助项目(the Supporting Excellent Young Teacher Foundation of Shanghai under Grant No.sdl-07013)。

作者简介:殷脂(1981-),女,博士研究生,讲师,CCF 会员,主要研究领域为智能优化、人工智能;叶春明(1964-),男,教授,博导,主要研究领域为智能优化、人工智能;温蜜(1980-),女,博士,讲师,主要研究领域为人工智能、网络安全。

收稿日期:2010-01-13 **修回日期**:2010-03-01

用来检查候选序列 x 及其补链 $\bar{x}$ 形成的双链的解链温度是否在给定的区间(Tm_{max},Tm_{min})。如果解链温度不属于以上区间,候选序列 x 被拒绝。目前 T_m 主要计算公式[6-7]有:

$$T_m(x)=\frac{\Delta H}{\Delta S+\Delta R\ln C_T}+16.6\log(Na^+) \quad (1)$$

式(1)中,ΔS 和 ΔH 分别表示以一系列人工合成的寡核苷酸为模型测算出的各碱基之间的熵变与焓变。

$$T_m(x)=81.5+16.6\times\log(\frac{[salt]}{1.0+0.7\times[salt]})+41\times GC(x)-\frac{500}{|x|}W \quad (2)$$

式(2)中,|x|表示序列长度,GC 含量是影响解链温度的唯一因素,该算法将采用公式(2)计算解链温度。

2.2 自由能约束

自由能约束主要用于检查杂交状态的稳定性,文献[6,8-9]研究了主要的计算方法,该算法采用文献[6]的计算公式:

$$\Delta G^{\circ}(x)=\sum_i n_i\Delta G^{\circ}(i)+\Delta G^{\circ}(Init.W/term.G.C)+\Delta G^{\circ}(Init.W/term.A.T)+\Delta G^{\circ}(sym) \quad (3)$$

式(3)中各近邻热力学参数 ΔG°_{37} 见表 1。

根据式(3)计算序列CAATCATGAA 的最小自由能:

表 1 近邻热力学参数

近邻碱基	ΔG°_{37}
AA/TT	−1.0
AT/TA	−0.88
TA/AT	−0.58
CA/GT	−1.45
GT/CA	−1.44
CT/GA	−1.28
GA/CT	−1.30
CG/GC	−2.17
GC/CG	−2.24
GG/CC	−1.84
Init.W/term.G.C	+0.98
Init.W/term.A.T	+1.03
$\Delta G^{\circ}(sym)$	若 $x=\bar{x}$,+0.43;若 $x\neq\bar{x}$,0

$$\Delta G^{\circ}_{37}(5'-CAATCATGAA-3')=\Delta G^{\circ}(CA)+\Delta G^{\circ}(AA)+\Delta G^{\circ}(AT)+\Delta G^{\circ}(TC)+\Delta G^{\circ}(CA)+\Delta G^{\circ}(AT)+\Delta G^{\circ}(TG)+\Delta G^{\circ}(GA)+\Delta G^{\circ}(AA)+\Delta G^{\circ}(Init)+\Delta G^{\circ}(sym) \quad (4)$$

因 CAATCATGAA 非自补序列,

$\Delta G^{\circ}(sym)=0$,$\Delta G^{\circ}(Init)=0.98+1.03=2.01$

$\Delta G^{\circ}_{37}(5'-CAATCATGAA-3')=-1.45-1.00-0.88-1.30-1.45-0.88-1.45-1.30-1.00+2.01+0=-8.7$

2.3 GC Content 约束

因为 G-C 之间有 3 个氢键,而 A-T 之间只有 2 个氢键,所以 GC 含量较高的 DNA 分子结构更稳定;相同的 GC 含量能保持序列具有相似的解链温度[10],因此一般要求编码集合中的序列 GC 含量在 50%左右。

$$f_{GC}(x)=-|GC(x)-GC_{defined}| \quad (5)$$

其中,$GC(x)$为序列 x 中字母 G、C 在序列中的比例;$GC_{defined}$ 表示所指定的 GC 含量。

3 蚁群优化算法(Ant Colony Optimization,ACO)

3.1 算法基本思想

蚁群算法[11-13]是一种源于大自然的新型仿生类进化算法,源于对蚂蚁觅食模型的研究。蚁群算法的基本思想是:模仿蚂蚁依赖信息素进行通信而显示出的社会性行为,在智能体定义的基础上,由一个贪心法指导下的自催化过程引导每个智能体的行动。它是一个随机的通用试探法,具有分布式的计算特性和很强的通用性,是基于总体优化的方法,是解决 NP 完全问题的有效工具。

3.2 蚂蚁路径选择与信息素的更新

设 $T_{ij}(t)$为 $t(t=0,1,\cdots)$时刻有向线段 $a[i,j]$上的信息素,初始时刻各有向线段上的信息素为 c。在时刻 t,蚂蚁 $k(k=1,2,\cdots,m)$从节点 $i(i=1,2,\cdots,n)$经由线段 $a[i,j]$转移到节点 $i+1$ 的转移概率为:

$$p_{ij}^{k}=\begin{cases}\dfrac{[T_{ij}(t)]^{\alpha}[\eta i_j]^{\beta}}{\sum[T_{ij}(t)]^{\alpha}[\eta i_j]^{\beta}} & s\in R\\ 0 \quad \text{Otherwise}\end{cases} \quad (6)$$

信息素的更新,基本公式:

$$T_{ij}^{new}=\rho\times T_{ij}^{old}+\sum_k\Delta T_{ij}^{k} \quad (7)$$

$$\Delta T_{ij}^{k}=\begin{cases}\dfrac{Q}{Z_k} & \text{若}(i,j)\text{在最优路径上,}Z_k\text{为目标函数值}\\ 0 \quad \text{Otherwise}\end{cases} \quad (8)$$

其中,α 为信息启发式因子,反映蚂蚁在运动过程中所积累的信息素在蚂蚁运动时所起的作用;β 为期望启发式因子,表示能见度的相对重要性,反映了蚂蚁在运动过程中启发信息在选择路径中受到重视的程度;R 是位于节点 i 上的蚂蚁可以选择的所有路径的集合;$\boldsymbol{\eta}$ 为能见度矩阵;$\rho(0\leqslant\rho<1)$为轨迹的持久性。

3.3 DNA 编码算法

DNA 编码算法设计了 2 个约束条件,分别为解链温度和最小自由能约束,假设 Tm_{max} 和 Tm_{min} 是 DNA 链 X 与补链 $\bar{X}$ 构成的双链的解链温度的上限和下限,ΔG_{min} 是实验要求的最小自由能的上限,即 $\Delta G\leqslant\Delta G_{min}$。

算法基本步骤如下:

步骤 1 根据 Tm_{max}、Tm_{min} 按公式(2)反向计算 $GC(x)$的取值范围 $GC(x)_{min}<GC(x)<GC(x)_{max}$;

步骤 2 从字母表(A,C,G,T)中选择 n 个元素作为蚁群算法的节点,每个节点代表一个碱基,n 个元素组成 DNA 序列,确保满足步骤 1 的取值范围,否则循环步骤 2;

步骤 3 初始化 ACO 算法的迭代步数以及相关参数;根据表 1 初始化节点间 $n\times n$ 距离矩阵;当 $n=8$ 时,矩阵如式(9):

$$\boldsymbol{D}=\begin{bmatrix} & A & T & G & C & A & T & G & C\\ A & -1.00 & -0.88 & -1.28 & -1.44 & -1.00 & -0.88 & -1.28 & -1.44\\ T & -0.58 & -1.00 & -1.45 & -1.30 & -0.58 & -1.00 & -1.45 & -1.30\\ G & -1.30 & -1.44 & -1.84 & -2.24 & -1.30 & -1.44 & -1.84 & -2.24\\ C & -1.45 & -1.28 & -2.17 & -1.84 & -1.45 & -1.28 & -2.17 & -1.84\\ A & -1.00 & -0.88 & -1.28 & -1.44 & -1.00 & -0.88 & -1.28 & -1.44\\ T & -0.58 & -1.00 & -1.45 & -1.30 & -0.58 & -1.00 & -1.45 & -1.30\\ G & -1.30 & -1.44 & -1.84 & -2.24 & -1.30 & -1.44 & -1.84 & -2.24\\ C & -1.45 & -1.28 & -2.17 & -1.84 & -1.45 & -1.28 & -2.17 & -1.84 \end{bmatrix} \quad (9)$$

步骤 4 将 m 个蚂蚁随机置于 n 个顶点上,并将各蚂蚁的初始出发点置于当前解集;对每个蚂蚁 k,按概率 p_{ij}^{k} 移动至下一

个顶点 j；将顶点 j 置于当前解集；

步骤 5　根据公式(3)以及对应矩阵 $\boldsymbol{D}$ 计算各蚂蚁的目标函数值 Z_k；记录当前最好解；

步骤 6　根据信息素的更新公式更新信息素矩阵；

步骤 7　对各路经，置信息素为 0；迭代步数增加 1；

步骤 8　若迭代步数等于预设值，且无退化行为(即所有找到的解均为同解)，则继续步骤 4；

步骤 9　输出所有满足 $Z_k<\Delta G_{\min}$ 的解。

4　实验结果与分析

应用以上算法，将 ACO 算法的参数设置为：进化代数500，蚂蚁数 20，信息启发式因子 1.2，期望启发式因子 2，轨迹信息素持久性 0.5，尝试构造长度为 20 的序列，在无其他约束条件过滤的情况下，得到一组 DNA 序列，如表 2 所示，各序列均满足 $\Delta G\leqslant -22$ kcal/mol，且形成的双链解链温度位于给定区间，即 69.58℃<T_m<72.58℃。

表 2　该文算法序列与 DNA-SDT 序列

该文算法	GC 含量/(%)	DNA-SDT[15]	GC 含量/(%)
TTGGCGCGCAAAACGTTTCA	50	CCGATGACTTAACTGGCGTC	55
AGCAAATTTGCGCCGTCGAT	50	GAAAAGGCCAAGATAGACCA	45
TAATTGCTGGCCAAACGCGT	50	TTAGAAGTCCCACTGCACTG	50
AAATATATTCGCGGCGCGCT	50	TCCAGCCATTAGACTGACGA	50
GGTAAATCGCGCGCATACTT	50	CACCCATTGCAAGGCACATA	50
AAATATGGCGTTACCCTGCG	50	CGCACCGGAAAACAACTTTA	45
TTAAATGATTCGCGCGCGCA	50	GCTGCTTTTACCAGCACGTA	50
TGCAAACGCTTTCGACGGAT	50	TCCCCAABACACTGTTTCAC	50

由于采用了 ACO 智能搜索算法，算法时间复杂度为 $O(n^3)$，编码算法的效率优于文献[14]的全局遍历算法，其时间复杂度为 $O(4^n)$(n 表示序列长度)。但也存在 ACO 算法的不足之处，如图 1 所示，算法容易早熟。同时将产生的序列集与 DNA-SDT[15]生成的序列进行比较，可以发现 GC 含量的稳定性更优(见表 2)。

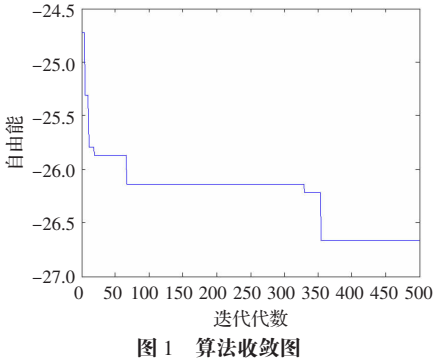

图 1　算法收敛图

5　结论

提出一种基于蚁群优化算法的 DNA 序列自由能约束过滤算法，该算法生成的 DNA 序列能够很好地满足用户给定的最小自由能的能量范围和实验需要的解链问题范围。根据热力学 Nearest-Neighbors 模型计算出 DNA 序列的最小自由能，可以极大地提高 DNA 序列自由能的计算精度。同时，该算法发挥蚁群算法并行运算的优势，提高了编码设计算法的效率。

参考文献：

[1] Adleman L.Molecular computation of solution to combinatorial problems[J].Science,1994,66(11):1201-1204.

[2] 李珍，王淑栋.DNA 编码限制条件与编码策略[J].计算机工程与应用，2009，45(5)：43-45.

[3] 张凯，耿修堂.DNA 计算中核苷酸序列设计方法比较研究[J].计算机学报，2008，31(12)：2149-2154

[4] 李珍，王淑栋，李二艳.基于部分字的 DNA 编码设计与分析[J].计算机应用研究，2010，27(1)：86-88

[5] Garzon M,Deaton R,Neathery P,et al.On the encoding problem for DNA computing[C]//Preliminary Proceedings 3rd DLMACS Workshop on DNA Based Computers,Philadelphi,University of Penns,1999,48:230-237.

[6] SantaLucia J Jr.A unified view of polymer,dumbbell,and oligonucleotide DNA nearest-neighbor thermodynamics[J].Proc Natl Acac,1998,95(4):1460-1465.

[7] Wetmur J G.DNA probes:Applications of the principles of nucleic acid hybridization[J].Biochemical Molecular Bio,1991,26(3):227-259.

[8] Tanaka F,Kameda A,Yamamoto M,et al.Thermodynamic parameter based on a nearest-neighbor model for DNA sequences with a single-bulge loop[J].Biochemistry,2004,43(22):7143-7150.

[9] Bommarito S,Peyret N,SantaLucia J Jr.Thermodynamic parameters for DNA sequences with dangling ends[J].Nucleic Acids Res,2000,28:1929-1934.

[10] Bloomfield V A,Crothersn D M,Timoco I.Physical chemistry of nucleic acids[M].New York:Harp and Row,1974.

[11] Dorigo M,Maniezzo V,Colorini A.Ant system:Optimization by a colony of cooperating agents[J].IEEE Trans on SMC,1996,26(1):29-41.

[12] 刘士新，宋健海，唐加福.蚁群最优化：模型、算法及应用综述[J].系统工程学报，2004，19(5)：496-502.

[13] 马良，朱刚.蚁群优化算法[M].北京：科学出版社，2008：20-27.

[14] 张凯，肖建华.基于最小自由能的 DNA 编码设计算法[J].武汉理工大学学报，2008，30(2)：176-179.

[15] Tanaka F,Kameda A,Yamamoto M,et al.Design of nucleic acid sequences for DNA computing based on a thermodynamic approach[J].Nucleic Acids Research,2005,33:903-911.

(上接 24 页)

参考文献：

[1] Fuhrmann P,Gulzow V.dCache,storage system for the future[C]//Euro-Par 2006 Parallel Processing,2006:1106-1113.

[2] Ranganathan K,Foster I.Identifying dynamic replication strategies for a high-performance data grid[C]//Lecture Notes in Computer Science,2001,2242:75-86.

[3] 亓雪冬，全兆岐，何潮观.快速瀑布模型动态副本创建策略研究[J].系统仿真学报，2008，20(15)：4054-4056.

[4] Tang M.Dynamic replication algorithms for the multi-tier data grid[J].Future Generation Computer Systems,2005,21(5):775-790.

[5] 钱晔蕾，董健全.基于非结构化 P2P 的副本技术的研究和应用[J].计算机工程与应用，2007，43(10)：148-152.

[6] 陈宇，董健全.非结构化 P2P 网络中的副本管理策略[J].计算机工程，2008，34(18)：108-110.

[7] USATLAS[EB/OL].http://www.usatlas.bnl.gov/.