一种基于编码等价变换和遗传算法的 DNA 序列优化设计

郑 学 东

(大连大学先进设计与智能计算省部共建教育部重点实验室 辽宁 大连 116622)

摘 要 针对 DNA 计算中的 DNA 序列设计问题,基于 6 个 DNA 序列设计约束条件,将 DNA 序列设计问题转化为多目标优化问题,提出小生境遗传算法进行求解。算法利用 DNA 序列设计中的相似性约束与 H-测度约束,在单链 DNA 序列集合上定义共享函数,利用两种类型的编码等价变换以及模 4 算术运算,构造了 5 个遗传算子,并给出具体的 DNA 序列设计结果。通过比较,算法可以得到质量更好的 DNA 序列,且在种群规模与进化代数方面具有更高的计算效率。

关键词 DNA 计算 DNA 序列设计 DNA 编码 小种群遗传算法 线性码

中图分类号 TP306.6 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.10.059

A DNA SEQUENCE OPTIMISATION DESIGN BASED ON CODE EQUIVALENT TRANSFORMATION AND GENETIC ALGORITHM

Zheng Xuedong

(*Key Laboratory of Advanced Design and Intelligent Computing* (*Dalian University*), *Ministry of Education*, *Dalian 116622*, *Liaoning*, *China*)

Abstract For the problem of DNA sequence design in DNA computing, we transform it to a multi-objective optimisation problem based on six constraints in DNA sequence design, and present the niche genetic algorithm to solve it. In the algorithm, a sharing function over the set of single DNA strand is defined using similarity constraint and H-measure constraint in DNA sequence design; five genetic operators are constructed based on two kinds of code equivalent transformations and modulo 4 arithmetic operation; and the specific results of DNA sequence design is presented as well. Compared with previous results, this algorithm can get DNA sequences with better quality and has higher computation efficiency in terms of population size and evolution algebra.

Keywords DNA computing DNA sequence design DNA encoding Small population genetic algorithm Linear code

0 引 言

DNA 序列设计是 DNA 计算[1,2]研究中的首要与核心问题[3]。在 DNA 序列设计中,首要的目标是提高 DNA 编码序列的特异性识别能力,即降低 DNA 序列的相似性。Baum 最早提出用最小相同子序列来定义不同 DNA 编码序列之间的相似程度[4]。由于 DNA 分子在溶液中可以自由扩散与相对移动,Garzon 等提出用 H-测度(h-measure)来刻画 DNA 编码序列之间的相似性[5]。2009 年 Phan 等基于 H-测度进一步定义了 h 距离(h-distance)[6],并对相应距离空间的几何性质进行了讨论,给出了 DNA 序列集合设计的相应结果。由于 DNA 分子具有特殊的理化性质,为了保证参与生化反应的 DNA 分子具有一致的理化性质,在 DNA 编码序列设计的过程中还需要考虑其他诸如自由能、解链温度、连续性和 GC 含量等 DNA 分子的热力学约束以及发卡等二级结构约束[7]。

对 DNA 编码序列设计问题,求解方法主要有随机搜索[8]、进化算法[9-13]、模板映射方法[14]、模计算方法[15]、多目标优化[16,17]和引力搜索算法[18]等。DNA 编码序列设计问题是一个难于求解的问题[19],目前还没有一个统一的规范方法来求解。

本文利用小生境技术,采用小种群遗传算法,定义了 DNA 编码序列间的共享函数,基于编码的两种等价变换及模 4 算术运算,设计了 5 个遗传算子,对 DNA 编码序列设计问题进行了求解,并给出具体的 DNA 编码序列。通过与已有结果的比较,说明了算法的有效性及计算效率。

1 约束条件与优化模型

1.1 约束条件

DNA 编码序列的相关约束大致可以分为两类,一类是以增加 DNA 分子特异性识别为目标的汉明距离类约束(或称为组合类约束);另一类是以保证 DNA 分子理化性质一致性为目标的热力学约束及二级结构类约束,本文采用的相关约束[20]具体如下。

(1) 相似性 Sm(Similarity):对长度为 n 的任意两个 DNA 序列 $D_j = N_{j1}\cdots N_{jn}, D_k = N_{k1}\cdots N_{kn} \in \{A,C,G,T\}^n$,$\{A,C,G,T\}$ 为碱基字母表,$\{A,C,G,T\}^n$ 表示字母表 $\{A,C,G,T\}$ 上所有长

收稿日期:2014-04-21。国家自然科学基金项目(61370005,31170797,61103057,31370778);长江学者和创新团队发展计划项目(IRT1109);教育部科学技术研究重点项目(211036);辽宁省教育厅科研项目(L2011218)。**郑学东**,讲师,主研领域:DNA 计算。

度为 n 的 DNA 序列集合，D_j 与 D_k 的相似性度量定义为：

$$S(D_j,D_k) = \min_{-n<l<n} H(D_j,D_k,l) \tag{1}$$

式中：

$$H(D_j,D_k,l) = \begin{cases} l + \sum_{i=1}^{n-l} h(N_{j(l+i)},N_{ki}) & l \in [0,n) \\ -l + \sum_{i=1}^{n+l} h(N_{ji},N_{k(-l+i)}) & l \in (-n,0) \end{cases}$$

$$h(N_{jp},N_{kp}) = \begin{cases} 0 \text{ if } N_{jp} = N_{kp} \\ 1 \text{ if } N_{jp} \neq N_{kp} \end{cases} \quad p = 1,2,\cdots,n$$

式(1)用于刻画在相对移动 l 位的情况下，DNA 编码序列 D_j 与 D_k 的相似程度，$S(D_j,D_k)$ 的值越大则编码序列 D_j 与 D_k 的特异性设别能力越高。

(2) H-测度 HM(h-measure)：在考虑相对移动的情况下，为避免两个 DNA 编码序列之间的杂交，Garzon 等人引入了 h-measure，其数学定义为：

$$H_m(D_j,D_k) = \min_{-n<l<n} H(D_j,\overline{D_k^R},l) \tag{2}$$

式中，$\overline{D_k^R}$ 表示 DNA 序列 D_k 的逆补序列。式(2)用于刻画在相对移动 l 位的情况下，DNA 编码序列 D_j 与 D_k 之间发生杂交的可能性，$H_m(D_j,D_k)$ 的值越大，则这种可能性越小。

(3) GC 含量 GC(GC-Content)：即碱基 G 与 C 在 DNA 序列中出现次数的百分比，GC 含量约束可表示为：

$$f_{GC}(D_j) = | GC(D_j) - GC_{goal} | \leqslant k_{GC} \tag{3}$$

式中，$GC(D_j)$ 表示序列 D_j 的 GC 含量，GC_{goal} 表示 GC 含量的目标值，k_{GC} 表示可接受范围。

(4) 连续性 Con(Continuity)：在 DNA 序列中，如果相同碱基连续重复出现，则生化反应将难以控制，在 DNA 序列设计中要尽量避免这种情况。连续性约束可表示为：

$$f_{Con}(D_j) = \sum_{N \in \{A,C,G,T\}} \sum_{i=1}^{n} (i-1) N_{D_j}^i \tag{4}$$

式中，$N_{D_j}^i$ 表示序列 D_j 中由相同碱基 N 构成的长度为 i 的子串出现的次数。

(5) 发卡 Hp(Hairpin)：DNA 序列可能发生自杂交，并形成环状结构，一般发卡结构是希望避免的。一个 DNA 分子可能出现的发卡数量可以通过以下公式进行估计：

$$f_{Hp}(D_j) = \sum_{s+1 \leqslant p \leqslant n-r-s} \sum_{5 \leqslant r \leqslant n-2\times s} Hp(p,r,s,D_j) \tag{5}$$

式中，$Hp(p,r,s,D_j)$ 表示在序列 D_j 中，开始于位置 p，环长为 r，茎长为 s 的发卡数。

(6) 解链温度 Tm(Melting Temperature)：即在 DNA 变性过程中，50% 的 DNA 双链变成单链时的温度，本文采用最近邻模型[21]。令 $T_m(D_j)$ 表示序列 D_j 的 Tm 值，$T_{m_{goal}}$ 表示 Tm 的目标值，k_{T_m} 表示可接受的 Tm 值范围，则 Tm 约束可表示为：

$$f_{T_m}(D_j) = | T_m(D_j) - T_{m_{goal}} | \leqslant k_{T_m} \tag{6}$$

在上述约束条件中，相似性约束与 H-测度约束反映了 DNA 序列及其逆补序列之间公共子序列的共享程度，是一类施加于编码序列集合的群体性约束；其他热力学与二级结构类约束则为施加于 DNA 序列自身的约束，是一种个体约束。另外，Gibbs 自由能可以通过其他约束的组合来替代，故这里不考虑 Gibbs 自由能约束[16]。

1.2 优化模型

由于 DNA 分子在溶液中可以相对移动与自由扩散，对于 DNA 编码序列集合 $C \subset \{A,C,G,T\}^n$，DNA 编码序列的特异性识别能力取决于彼此间的最小相似性，DNA 编码序列间发生错误杂交的可能性取决于彼此间的最小 H-测度，故这里分别取相似性与 H-测度的最小值作为目标函数，其定义为：

$$f_{Sm}(D_j) = \min_{D_k \in C-\{D_j\}} S(D_j,D_k) \tag{7}$$

$$f_{HM}(D_j) = \min_{D_k \in C} H_m(D_j,D_k) \tag{8}$$

对于连续性与发卡，将 $-f_{Con}(D_j)$ 与 $-f_{Hp}(D_j)$ 取为目标函数。将 GC 含量约束 $f_{GC}(D_j)$ 与 Tm 值约束 $f_{T_m}(D_j)$ 作为优化模型的约束条件。

于是，DNA 编码序列设计问题可以转化寻找 DNA 序列集合 $C \subset \{A,C,G,T\}^n$，使得对 $\forall D_j \in C$，满足如下的最大值多目标优化问题：

$$\max F(D_j) = [f_{Sm}(D_j), f_{HM}(D_j), -f_{Con}(D_j), -f_{Hp}(D_j)] \tag{9}$$

$$\text{s.t.} \quad \begin{cases} f_{GC}(D_j) \leqslant k_{GC} \\ f_{T_m}(D_j) \leqslant k_{T_m} \\ D_j \in C \end{cases}$$

2 算法设计

针对模型式(9)，构建基于共享函数的小种群遗传算法，将 DNA 编码序列作为种群个体，用 4 进制整数编码，种群编码为 4 进制整数矩阵 $P_{m\times n}$，其中 n 表示个体长度，m 表示种群规模，算法输出 $P_{m\times n}$ 的子式 $C_{M\times n}$ 作为 DNA 编码序列设计的结果，M 表示 DNA 编码序列的数量。适应度函数为模型式(9)中目标函数与约束条件的线性组合，将其转换为无约束问题，适应度函数定义为：

$$f(D_j) = \sum_{i=1}^{6} \alpha_i f_i(D_j) \tag{10}$$

式中，$f_i \in \{f_{Sm}, f_{HM}, -f_{GC}, -f_{Tm}, -f_{Hp}, -f_{Con}\}$，$\alpha_i$ 为权重系数。对种群 $P_{m\times n}$ 中的 DNA 序列 D_j 与 D_k，定义共享函数为：

$$sh(D_j,D_k) = 2n - \gamma_1 S(D_j,D_k) - \gamma_2 H_m(D_j,D_k) \tag{11}$$

式中，γ_1 与 γ_2 为控制参数，个体小生境数定义为：

$$f_{sh}(D_j) = \sum_{k=1}^{m} sh(D_j,D_k) \tag{12}$$

小生境淘汰时，序列 D_j 的适应度为 $f(D_j)/f_{sh}(D_j)$。

在编码理论[22]中，一个 q 元 (n,m,d) 码可以排列为 $m\times n$ 阶矩阵 P，对 P 可以进行两种编码等价变换：一是对 P 的列进行置换，二是对固定列上的字符进行置换。经过两种变换后的编码仍为 (n,m,d) 码，即等价变换不改变编码的最小汉明距离 d。由于式(1)与式(2)均基于汉明距离定义，故考察目标函数(式(7)，式(8))与编码序列集合最小汉明距离的相关性。对含有 m 个长为 n 的 DNA 序列构成的集合 C，在搜索空间 $\{A,C,G,T\}^n$ 中随机选择 m 个 DNA 序列构成一个样本，计算 m 个 DNA 序列之间的最小汉明距离(MH)，最小相似性(MSm)与最小 H-测度(MHM)，考察三者之间的相关性，这里样本数量为 500，分别计算 9 次，结果如表 1 与表 2 所示，其中 ρ 为相关系数，P 为相关系数显著性检验的概率值，显著性水平为 0.05，数值计算保留小数点后 4 位。

表 1 m=7,n=20 情形的相关性

(MH, MSm)		(MH, MHM)		(MSm, MHM)	
ρ	P	ρ	P	ρ	P
0.4470	0	-0.0119	0.7905	-0.0138	0.7578
0.4268	0	-0.0910	0.0420	-0.0837	0.0614
0.4245	0	0.0210	0.6397	-0.0048	0.9154
0.4420	0	0.0089	0.8425	0.0521	0.2451
0.4683	0	0.0281	0.5304	0.0074	0.8691
0.5236	0	-0.0089	0.8419	0.0168	0.7079
0.3903	0	0.0333	0.4569	0.0272	0.5440
0.3831	0	0.0290	0.5170	0.0258	0.5644
0.4586	0	-0.0157	0.7260	-0.0441	0.3251

表 2 m=20,n=15 情形的相关性

(MH, MSm)		(MH, MHM)		(MSm, MHM)	
ρ	P	ρ	P	ρ	P
0.6011	0	0.0252	0.5745	0.0031	0.9443
0.5314	0	-0.0363	0.4182	-0.0125	0.7799
0.6192	0	-0.1034	0.0208	-0.0236	0.5979
0.5698	0	0.0332	0.4584	0.0308	0.4916
0.5661	0	0.0098	0.8271	-0.0406	0.3655
0.5009	0	0.0227	0.6123	0.0286	0.5235
0.5560	0	-0.0009	0.9831	0.0595	0.1844
0.5747	0	-0.0337	0.4526	-0.0552	0.2176
0.5542	0	-0.0423	0.3455	-0.0227	0.6132

由表 1 和表 2,对于随机选择的 DNA 序列集合 C, C 的最小汉明距离与最小相似性具有中度相关性,且由 P 值说明这种相关性是显著的,而最小汉明距离与最小 H-测度以及最小相似性与最小 H-测度则几乎没有相关性。式(4)-式(6)的值与 DNA 序列中不同碱基的排列顺序显然是密切相关的。基于上述考虑,在遗传算法中定义如下的遗传算子[23]:

(1) 模交叉 MAXO(Modular Arithmetic Crossover Operator):以概率 p_{MAXO},随机选择 $P_{m\times n}$ 的两行作为父代个体,通过模 4 的加法与减法得到两个子代个体。

(2) 多点交叉 MXO(Multipoint Crossover Operator):以概率 p_{MXO},随机选择 $P_{m\times n}$ 的两行,随机选择 r 个位点进行交叉。

(3) 列置换变异 CPMO(Column Permutation Mutation Operator):以概率 p_{CPMO},随机选择 $P_{m\times n}$ 的 k 列,进行随机置换。

(4) 字符置换变异 LPMO(Letter Permutation Mutation Operator):以概率 p_{LPMO},随机选择 $P_{m\times n}$ 的 l 列,对列上的字符进行随机置换。

(5) 子式变异 MMO(Minor Mutation Operator):以概率 p_{MMO},随机选择 $P_{m\times n}$ 的子式,将其上整数值随机替换为其余碱基对应的整数值。

(6) 选择 SO(Selection Operator):采用锦标赛选择。

算法连续进化 500 代终止。算法流程如图 1 所示。

在上述算法中,通过列置换变异与字符置换变异进行微调,微调条件设定为连续 10 代种群平均适应度值没有得到改善。

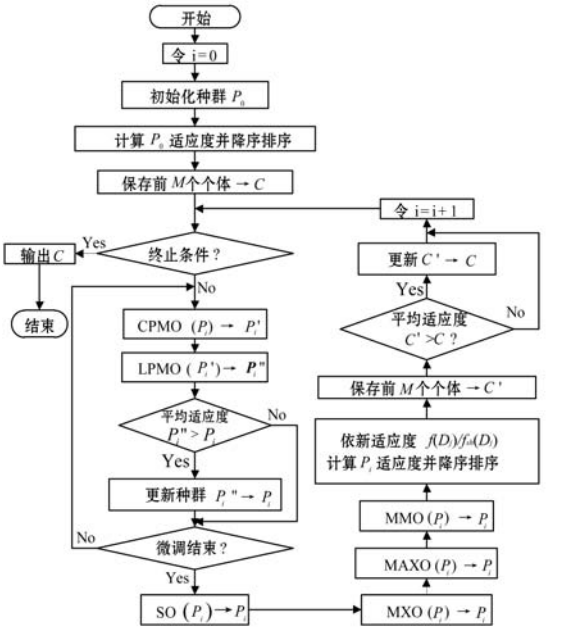

图 1 算法流程图

3 实验结果与比较

本文算法用 Matlab 语言实现,算法的参数设置如表 3 所示,DNA 序列的 Tm 值计算采用 Matlab 中 Bioinformatics Toolbox 中的相应函数。

表 3 算法参数设置

参数项	7 个 DNA 编码序列	20 个 DNA 编码序列
n	20	15
m	30	60
M	7	20
$\alpha_1,\cdots,\alpha_6$	1.2, 1.2, 1, 1, 1, 1	1.2, 1.2, 1, 1, 1, 1
$\gamma 1$, γ_2	1, 1	1, 1
p_{MAXO}	0.5	0.6
p_{MXO}	0.6	0.6
r	3	3
p_{CPMO}	0.45	0.6
k	10	7
p_{LPMO}	0.42	0.6
l	8	6
p_{MMO}	0.2	0.3
锦标赛选择规模	3	3

理论上,本文定义的相似性与 H-测度的值越大,则 DNA 序列之间发生错误杂交的可能性越低;GC 含量与 Tm 值的分布越集中、连续性值越低,则 DNA 序列理化性质的一致性越好;发卡的值越低则 DNA 序列出现二级结构的可能性越低。基于上述考虑,提出以下的 DNA 编码序列设计的评估标准:选择相似性与 H-测度的最小值作为评估项;选择 GC 含量与 Tm 值的标准差作为评估项,选择发卡与连续性的和作为评估项。为说明算法的有效性,将实验结果与文献[17]中的结果进行比较,比较结果如表

4 – 表 7 所示。同时在表中给出依文献[17]中约束条件定义计算的相似性(Sm′)、最大相似性(MSm′)、H-测度(HM′)、最大 H-测度(MHM′)、连续性(Con′)的评估值,由于 Hp 项的值均为 0,故未单独列出,Tm 值计算均采用 Matlab Bioinformatics 工具箱的默认条件。编码评估项的比较如表 8 至表 11 所示。

表 4　文献[17]7 个 DNA 编码序列结果

DNA 序列	Sm	Sm′	MSm′	HM	HM′	MHM′	GC	Con	Con′	Tm
TCCTTCCTCCTTACTCCTTC	7	49	23	7	31	23	50	7	0	53.94
CTTCTGCTCTCTTCTCTTCC	7	53	23	10	25	12	50	4	0	52.89
CCTCTCGTTCTTGTCTCTCT	9	45	18	10	28	12	50	3	0	54.19
TCCTTCTCTGTTCCTCTTCC	7	52	22	9	30	19	50	6	0	54.42
GAGGAAGGAAGGAAGTAAGG	10	13	13	7	65	23	50	8	0	52.77
AGGTAAGAGGAGGCGAAGAA	10	13	13	9	69	19	50	6	0	56.48
TCTCTTCTTCTCTCCGCTTC	7	53	22	10	25	13	50	4	0	54.52

表 5　本文 7 个 DNA 编码序列结果

DNA	Sm	Sm′	MSm′	HM	HM′	MHM′	GC	Con	Con′	Tm
TACATCAGCTCTGCGACAGT	10	58	11	11	64	10	50	0	0	57.1
ATGATGCTGCGTCGTACTCT	12	57	11	11	65	12	50	0	0	57.39
TACACGCTATCAGACGCTGT	10	55	10	11	68	12	50	0	0	56.96
GCGTTCGATTCAGACAGCAT	12	53	10	11	65	10	50	2	0	56.73
TTCGTCGTGTCTGTCTTGCT	11	54	11	11	58	10	50	2	0	57.65
TGCTGTGAGCGCATAGACAT	11	56	11	11	67	12	50	0	0	58.07
TACGTAGCGAGACTGTGCAT	11	57	11	11	65	10	50	0	0	56.96

表 6　文献[17] 20 个 DNA 编码序列结果

DNA 序列	Sm	Sm′	MSm′	HM	HM′	MHM′	GC	Con	Con′	Tm
AGGAGTGGAGAAGAG	4	128	20	5	84	20	53.33	3	0	47.45
GGAGACAGAGAGAAG	4	121	19	4	79	17	53.33	2	0	45.19
AAGGAGAGAATGAGA	4	128	19	5	82	19	40	3	0	43.5
CTCTTATCCTTCTTC	4	64	17	4	144	20	40	4	0	39.26
GGATGAAGAGAAGAG	5	128	20	5	79	21	46.67	3	0	42.69
AGTGGAAGGAAGGAA	5	119	19	3	87	25	46.67	6	0	47.85
CTCCTCTCACTCTCT	4	64	12	4	148	17	53.33	1	0	46.6
AGAAGAATAAGAAGA	5	118	11	4	81	13	26.67	4	0	37.18
AAGGAACAGGAGAGG	5	126	17	6	83	17	53.33	5	0	48.31
AAGAGGTGAGGAGAG	3	130	18	6	76	18	53.33	3	0	47.45
TCCTCCTTCTCCATT	6	66	17	6	144	18	46.67	5	0	47.17
AGGAGGACGGAAGAA	5	122	11	5	78	17	53.33	5	0	50.54
TGCGCCTCTCTATTC	7	105	12	6	152	17	53.33	2	0	49.15
GAGAAGAGGATGAGG	4	129	20	5	84	19	53.33	3	0	45.36
TCTCTTCTCCGTCTC	4	70	17	4	158	21	53.33	2	0	47.34
TTCTCTTCCTTCCAC	5	67	15	3	157	24	46.67	5	0	45.54
CAACCACCAACTCCT	7	105	9	7	138	9	53.33	5	0	49.84
TAGAGAGAGAGAGAG	4	119	18	4	80	21	46.67	0	0	41.86
AAGAGTAGAAGAGAG	3	129	21	5	81	20	40	2	0	40.4
TTCCGATTGGTTGGT	9	110	7	7	130	9	46.67	6	0	49.5

表 7　本文 20 个 DNA 编码序列结果

DNA	Sm	Sm′	MSm′	HM	HM′	MHM′	GC	Con	Con′	Tm
TCTGATGACTCACGT	6	147	10	8	152	11	46.67	0	0	47.7
GCGTGTATGTGTCTG	6	146	18	3	174	24	53.33	0	0	47.93
TGTCACTCACTACGA	6	148	9	7	166	16	46.67	0	0	47.16
ATACTGTCTCACGCT	6	160	15	7	166	10	46.67	0	0	47.81
AGACATCATCGACGT	7	141	9	8	169	15	46.67	0	0	48.19
AGTGTATCACAGCGA	6	156	11	7	161	10	46.67	0	0	48.17
ACTCATACGACACGT	6	146	9	8	162	15	46.67	0	0	48.06
TGAGACGCGATGATA	6	152	10	6	172	15	46.67	0	0	47.87
AGAGCGTAGTGTCAT	8	142	11	8	155	16	46.67	0	0	47.81
ACACAGAGATAGCGT	6	172	18	6	156	14	46.67	0	0	47.81
ACACATACACGCTGA	5	140	10	3	175	24	46.67	0	0	48.92
TGTCTACGCGTGATA	6	152	10	6	171	17	46.67	0	0	47.73
TCTCACACACGATCT	6	149	10	7	166	20	46.67	0	0	47.7
TGTGACAGAGATCGT	6	161	18	8	171	20	46.67	0	0	47.7
GCATGTGTCTCAGTC	6	165	18	7	159	11	53.33	0	0	47.59
CGACAGACACGAAGA	5	145	10	6	154	18	53.33	1	0	49.1
ATGCATAGTCGCGAT	6	156	18	6	160	15	46.67	0	0	49.26
ATAGCGTACGTCTCA	6	164	18	7	160	16	46.67	0	0	47.38
GCGTATCTATGCGTC	6	146	10	6	160	10	53.33	0	0	47.38
GCGATATAGCGTGTC	6	174	17	7	175	19	53.33	0	0	47.38

表 8　7 个 DNA 序列评估项比较——依本文约束条件定义

编码方案	相似性最小值	H-测度最小值	GC 含量标准差	连续性求和	发卡求和	Tm 值标准差
文献[17]	7	7	0	38	0	1.24
本文	10	11	0	4	0	0.47

表 9　7 个 DNA 序列评估项比较——依文献[17]约束条件定义

编码方案	相似性求和	最大相似性	H-测度求和	最大 H-测度	连续性求和
文献[17]	278	23	273	23	0
本文	390	11	452	12	0

表 10　20 个 DNA 序列评估项比较——依本文约束条件定义

编码方案	相似性最小值	H-测度最小值	GC 含量标准差	连续性求和	发卡求和	Tm 值标准差
文献[17]	3	3	7.04	69	0	3.71
本文	5	3	2.96	1	0	0.57

表 11　20 个 DNA 序列评估项比较——依文献[17]约束条件定义

编码方案	相似性求和	最大相似性	H-测度求和	最大 H-测度	连续性求和
文献[17]	2148	21	2145	25	0
本文	3062	18	3284	24	0

由上述比较结果(表 8 – 表 11),依据两种 DNA 序列设计约束条件的定义,在所有的评估项上,本文算法均可以获得较好的解,尤其是在连续性方面,可以获得极大的改善(依文献[17]中连续性约束的定义,由两个相同碱基构成的子串的连续性为 0,且连续碱基序列中子串的连续性不重复计数,如 DNA 序列 AAAA 依文献[17]中连续性的值为 4,在本文中连续性的值为 10,故本文采用的连续性约束较文献[17]中的定义更为严格),序列 Tm 值与 GC 含量分布更为集中,说明 DNA 序列理化性质

具有更好的一致性。在计算效率方面,与文献[17]相比,本文算法的种群规模分别为文献[13]的 1/10(7 个 DNA 编码序列情形)与 1/5(20 个 DNA 编码序列情形),迭代次数至多仅为文献[17]的 1/3,计算量要小得多。

4 结 语

DNA 序列设计是 DNA 计算研究中的重要问题,还没有一个规范的统一求解方法。本文选取了 DNA 序列设计中的几个约束条件,将 DNA 序列设计问题转换为多目标优化问题,通过在 DNA 序列间定义共享函数,给出了基于小生境技术的遗传算法。对 DNA 序列设计问题进行了求解,并给出具体的 DNA 序列设计结果,通过与已有结果的比较,验证了算法的可行性与有效性。同时从种群规模与进化迭代次数两方面,说明本文的算法具有更高的计算效率。本文中共享函数的应用,说明 DNA 编码序列设计问题是一个多模函数的优化问题,也说明了 DNA 编码序列设计问题是难于求解的。未来自适应方法的应用,可以期望获得更好的结果与计算效率的进一步提高。

参 考 文 献

[1] Adleman L. Molecular computation of solution to combinatorial problems[J]. Science, 1994, 266(5187):1021 - 1023.

[2] Ignatova Z, Martinez-perez I, Zimmermann K H. DNA 计算模型[M]. 郁方, 王淑栋, 强小利, 译. 北京:清华大学出版社, 2010.

[3] Garzon M H, Deaton R J. Codeword design and information encoding in DNA ensembles[J]. Natural Computing, 2004, 3(3):253 - 292.

[4] Baum E B. DNA Sequences useful for computation[C]//Proceedings of 2nd DIMACS Workshop on DNA Based Computers, Princeton University:American Mathematical Society, 1996:122 - 127.

[5] Garxon M, Neathery P, Deaton R J, et al. A new metric for DNA computing[C]//Proceedings of 2nd Annual Genetic Programming Conference, San Francisco:Morgan Kaufmann, 1997:472 - 487.

[6] PHan V, Garzon M. On codeword design in metric DNA spaces[J]. Natural Computing, 2009, 8(3):571 - 588.

[7] Sager J, Stefanovic D. Designing Nucleotide Sequences for Computation:A survey of constraints[C]//Proceedings of 11th International Meeting on DNA Computing, Lecture Notes in Computer Science Vol. 3892, London:Springer-Verlag, 2006:275 - 289.

[8] Penchovsky R, Ackermann J. DNA library design for molecular computation[J]. Journal of Computational Biology, 2003, 10(2):215 - 229.

[9] 殷脂, 叶春明, 马慧民. 基于文化进化粒子群算法的 DNA 序列设计[J]. 计算机工程与应用, 2011, 47(1):40 - 42.

[10] Cui G, Li X. The optimization of DNA encodings based on modified PSO/GA algorithm[C]//Proceedings of International Conference on Computer Design and Applications, Qinhuangdao, 2010:609 - 614.

[11] Ibrahim Z, Khalid N K, Buyamin S, et al. DNA sequence design for DNA computation based on binary particle swarm optimization[J]. International Journal of Innovative Computing, Information and Control, 2012, 8(5B):3441 - 3450.

[12] Mustaza S M, Abidin A F Z, Ibrahim Z, et al. A modified computational model of ant colony system in DNA sequence design[C]//Proceedings of IEEE Student Conference on Research and Development, 2011:169 - 173.

[13] Chaves-Gonzalez J M, Vega-Rodriguez M A, Granado-Criado J M. A multiobjective swarm intelligence approach based on artificial bee colony for reliable DNA sequence design[J]. Engineering Applications of Artificial Intelligence, 2013, 26:2045 - 2057.

[14] 王向红,刘文斌,朱翔鸥,等. DNA 计算中的单模板编码方法改进研究[J]. 电子学报,2009,37(12):2720 - 2724.

[15] Xiao J H, Zhang X Y, Xu J. A membrane evolutionary algorithm for DNA sequence design in DNA computing[J]. Chinese Science Bulletin, 2012, 57(6):698 - 706.

[16] Shin S Y, Lee I H, Kim D, et al. Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing[J]. IEEE Trans on Evolutionary Computation, 2005, 9(2):143 - 158.

[17] Cervantes-salido V M, Jaime O, Brizuela C A, et al. Improving the design of sequences for DNA computing:a multi-objective evolutionary approach[J]. Applied Soft Computing, 2013,13(12):4594 - 4607.

[18] Xiao J H, Cheng Z. DNA sequences optimization based on gravitational search algorithm for reliable DNA computing[C]//Proceedings of 6th International Conference on Bio-Inspired Computing, 2011:103 - 107.

[19] 张凯, 耿修堂, 肖建华, 等. DNA 编码问题及其复杂性研究[J]. 计算机应用研究, 2008, 25(11):3264 - 3267.

[20] Tanaka F, Nakatsugawa M, Yamamoto M, et al. Developing support system for sequence design in DNA computing[C]//Proceedings of 7th International Workshop DNA Based Computers, Lecture Notes in Computer Science Vol. 2340, 2002:129 - 137.

[21] Santalucla J JR, Hicks D. The thermodynamics of DNA structural motifs[J]. Annual Review of Biophysics and Biomolecular Structure, 2004, 33:415 - 440.

[22] 王育民, 李晖. 信息论与编码理论[M]. 北京:高等教育出版社, 2013.

[23] 陈国良,王煦法,庄镇泉,等. 遗传算法及其应用[M]. 北京:人民邮电出版社,1996.

(上接第 226 页)

[3] Yang Xing, Huang Chaochao, Qin Min. Research on Acquiring Binarized Vertical Edge Image for Texture-Based Adapting License Plate Location[J]. International Journal of Systems and Control, 2008, 3:192 - 198.

[4] Yingjun Wu, Shouxun Liu, Xuan Wang. License plate location method based on texture and color[C]//Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on. IEEE, 2013:361 - 364.

[5] Peter Trebuňa, Jana Halcnova. Mathematical Tools of Cluster Analysis[J]. Applied Mathematics, 2013, 4(5):814 - 816.

[6] Rozhentsov A A, Morozov K V, Baev A A. Modified generalized Hough transform for 3D image processing with unknown rotation and scaling parameters[J]. Optoelectronics, Instrumentation and Data Processing, 2013, 49(2):131 - 141.

[7] Longqin Xu, Shuangyin Liu. Study of short-term water quality prediction model based on wavelet neural network[J]. Mathematical and Computer Modelling, 2013, 58(3):807 - 813.

[8] Bilal Bataineh, Siti Norul Huda Sheikh Abdullah, Khairuddin Omar. An adaptive local binarization method for document images based on a novel thresholding method and dynamic windows[J]. Pattern Recognition Letters, 2011, 32(14):1805 - 1813.

[9] 平源,李慧娜. 快速精确识别车牌字符的方法[J]. 计算机工程与设计,2008,29(9):2410 - 2412.

[10] 黄凡,李志敏,张晶,等. 基于 K-L 变换和 LS-SVM 的车牌字符识别新方法[J]. 微计算机信息,2008(24):127 - 129.