基于多目标进化策略算法的DNA核酸编码设计

张 凯[①②] 陈 彬[①] 许志伟*[①]

[①](武汉科技大学计算机科学与技术学院 武汉 430065)

[②](智能信息处理和实时工业系统湖北省重点实验室 武汉 430065)

摘 要：设计高质量的核酸分子集合能有效提高DNA计算的可靠性、有效性和可求解问题的规模。DNA分子需要满足热力学约束、相似度约束、GC含量约束等多个相互冲突的目标函数，是典型的多目标优化问题。该文提出一种多目标进化策略(MOES)算法求解DNA分子序列设计问题，算法设计了随机碱基变异算子实现高效的局部搜索和全局搜索。改进的评价函数综合考虑了候选解的支配关系和冲突目标的平衡程度，选取符合DNA编码约束的核酸序列。实验结果证明，该文提出的算法具有高效的搜索效率和快速收敛能力，可以产生高质量的DNA序列集合，优于其他对比算法产生的DNA分子序列集合。

关键词：DNA计算；DNA序列设计；多目标进化算法；进化策略

中图分类号：TP301 **文献标识码**：A **文章编号**：1009-5896(2020)06-1365-09

DOI: 10.11999/JEIT190869

A Multiobjective Evolution Strategy Algorithm for DNA Sequence Design

ZHANG Kai[①②] Chen Bin[①] Xu Zhiwei[①]

[①](*School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan* 430065, *China*)

[②](*Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan* 430065, *China*))

Abstract: It is important to design high-quality DNA sequences set, which can improve the reliability and efficiency of DNA computing. DNA sequence design problem is an multiobjective optimization problem that needs to satisfy multiple conflict objectives which are thermodynamic constraint, similarity constraint and GC content constraint simultaneously. A MultiObjective Evolutionary Strategy (MOES) is proposed to solve the DNA sequence design problem. The random base mutation operator is designed for exploration and exploitation the search space. The fitness function is improved for obtaining balanced similarity and H-measure objective functions. Some state-of-the-art approaches are chosen to evaluate the effectivity of proposed algorithm. The experiment results show that the proposed multiobjective evolution strategy algorithm obtains very promising DNA sequences and outperforms previous approaches.

Key words: DNA Computing; DNA sequence design; Multiobjective evolutionary algorithm; Evolution strategy

1 引言

DNA计算是一种基于DNA分子的新型计算模型，由于具有大规模并行性，低能耗，海量存储等能力，DNA计算技术对求解NP完全问题的展现出极大潜力。1994年，Adleman[1]通过DNA分子成功求解7个顶点的哈密尔顿路问题。随后，DNA计算被用于求解各种NP完全问题[2]，如满意度问题(SATisfaction problem, SAT)[3]，旅行商问题(Traveling Salesman Problem, TSP)[4]和图着色问题[5]。在DNA计算过程中，待求解的问题被编码在DNA序列中，利用DNA分子的特异性杂交生成代表问题解的DNA分子。然而，低质量的DNA序列设计会导致非特异性杂交、不一致的解链温度，甚至DNA计算的失败。高可靠的DNA分子集合设计是提高DNA计算效率的关键。DNA分子设计需要从4^n的海量解空间中挑选出满足热力学约束、相似度约束、GC

收稿日期：2019-11-01；改回日期：2020-03-01；网络出版：2020-04-09

*通信作者： 许志伟 xuzhiwei@wust.edu.cn

基金项目：国家自然科学基金(61472293, 61702383, 61602328)

Foundation Items: The National Natural Science Foundation of China (61472293, 61702383, 61602328)

含量约束等条件的DNA分子集合，DNA序列设计问题也被证明是一个NP完全问题[6]。

近年来，进化计算被广泛的应用于求解DNA序列设计问题，通过模仿自然界各种生物的进化过程，高效的搜索4^n的指数级解空间。国内外研究人员已提出多种进化算法求解DNA序列设计问题，如微遗传算法(Micro-Genetic Algorithm, MGA)[7]、杂草入侵优化(Invasive Weed Optimization, IWO)算法[8]、改进的非支配排序遗传算法(Improved the Nondominated Sorting Genetic Algorithm, INSGA-II)[9]、并行多目标进化算法(parallel Multi-Objective evolutionary algorithm, pMO-ABC)[10]等。然而，目前的群智能算法需要从整个种群中选择最优个体进行下一次迭代，虽然扩大了搜索空间，但也极大地增加了时间复杂度。当目标函数较多时，种群的候选解相互非支配，也会因此缺少选择压力导致算法难以收敛。进化策略(Evolutionary Strategy, ES)也是一种进化算法，种群中的个体不需要像遗传算法进行交叉，只通过变异实现解空间的探索和开发，利用较少的计算资源找到全局最优解。ES在多目标优化问题[11]、带噪声的优化问题[12]、离散优化问题[13]和约束优化问题[14]中均表现出良好的性能。

本文提出一种多目标进化策略(MOES)算法求解DNA序列设计问题，设计的随机碱基变异算子可以兼顾局部搜索和全局搜索能力。此外，改进的评价函数综合考虑冲突指标相似度和H-measure的平衡性，可以有效地减少DNA分子集合中的非特异性杂交，可有效地提高DNA计算的可靠性。最后，通过实验跟几种最先进的DNA核酸编码算法进行比较，结果表明本文算法设计的DNA序列分子集合具有更高的质量。

2 DNA编码设计问题

DNA编码问题可表示为：设X=5′–$x_1x_2\cdots x_n$–3′为一个长度为n的单链DNA序列，其中x_i代表碱基，$x_i\in${A, C, G, T}。令S是长度为n的DNA序列集合，显然集合S的解空间大小为$|S|=4^n$。求S的一个子集$C\subseteq S$，使得C中的任意两条序列X, Y满足给定的多个编码约束准则，由于多个目标函数间相互冲突，DNA编码问题可看作多目标优化问题，如式(1)所示。

$$\begin{aligned}&\min f(X)=\min\left[f_1(X),f_2(X),\cdots,f_M(X)\right]^{\mathrm{T}},\\&f(X)\in R^M\end{aligned}\tag{1}$$

其中，$f(X)$由M个目标函数 $f_m(x)$组成, m =1, 2, …, M。通常DNA序列设计需要满足6个编码目标函数，如相似度、H-measure、解链温度、连续性、GC含量、发卡结构。

2.1 相似度约束

相似度约束(similarity)可以描述两个DNA序列X_i和X_j碱基组成的相似程度[15]。满足相似度约束的两条DNA链的同向序列尽可能唯一，且序列滑动后也尽量不会重复[16]。相似度可以通过计算序列X_i和X_j之间移动后取最大相似距离得到，其计算公式如式(2)所示

$$\begin{aligned}f_{\mathrm{Si}}(X)&=\sum_{i=1}^{n}\sum_{j=1}^{n}Si(X_i,X_j)\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\max_{-n<k<n}E(X_i,\sigma^k(X_j))\end{aligned}\tag{2}$$

其中，E(*,*)表示相似距离，即当对应位点碱基相等时结果为1，当k>0时，σ^k表示右移；当k<0时，σ^k表示左移，k表示移动的位数。如果X_i和X_j经过距离为k的移动后相似距离减小，那么相似度约束的值也随之减小。相似度约束值较大时序列X_i和X_j就非常相似，序列X_i和X_j的补链${X_j}^C$之间互补的碱基就多，容易出现非特异性杂交；当相似度约束值较小时序列X_i和X_j相同的碱基则很少，X_i和${X_j}^C$之间互补碱基就很少，从而X_i和${X_j}^C$之间不会出现非特异性杂交[17]。

2.2 H-measure约束

文献[18]将核酸碱基互补的信息扩充到汉明距离中，提出了H-measure约束以限制两条DNA序列之间的碱基互补。给定两条DNA序列X_i和X_j，通过计算X_i和X_j互补碱基的个数，H-measure可以防止X_i和X_j之间的交叉杂交[19]。H-measure可以通过计算序列X_i和X_j之间的最大滑动汉明距离得到，其公式如式(3)所示

$$\left.\begin{aligned}f_{\mathrm{Hm}}(X)&=\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Hm}(X_i,X_j)\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\max_{-n<k<n}H(X,\sigma^k(X_j))\\H(X,Y)&=\sum_{i=1}^{n}\mathrm{bp}(x_i,y_i)\\\mathrm{bp}(x_i,y_i)&=\begin{cases}1,&x_i=\overline{y_i}\\0,&x_i\neq\overline{y_i}\end{cases},x_i,y_i\in\{\mathrm{A,C,G,T}\}\end{aligned}\right\}\tag{3}$$

2.3 连续性约束

如果一个DNA序列中出现连续相同的碱基，则在碱基分子氢键力作用下出现不期望的2级结构，碱基连续性的评价函数，其公式如式(4)所示

$$f_{\mathrm{con}}(X)=\sum_{i=1}^{n}\mathrm{Continuity}(X_i)=\sum_{i=1}^{n}\sum_{i=1}^{l-t+1}(j-1)C_j^{(i)}\tag{4}$$

其中，$C_j^{(i)}$表示在DNA序列X_i中，相同碱基连续出现j次的数目。

2.4 GC含量约束

DNA序列的GC含量影响该序列的化学性质，例如解链温度可以由GC含量计算得到。GC含量即为DNA序列中碱基G和碱基C的个数或者百分比，其公式如式(5)所示

$$\left.\begin{aligned} &f_{\mathrm{GC}}(X)=\max_i\{\mathrm{GC}(X_i)\}-\min_j\{\mathrm{GC}(X_j)\}\\ &\mathrm{GC}=\sum_{i=1}^{n}\sum_{i=1}^{l}\mathrm{gc}(x_i)\\ &\mathrm{gc}(x_i)=\begin{cases}1, x_i=G\text{或 } x_i=C\\ 0, x_i=A\text{或}x_i=T\end{cases}\end{aligned}\right\} \tag{5}$$

2.5 发卡结构约束

单链DNA分子产生2级结构通常由自身反向折叠而形成，发卡结构(Hairpin, Hp)为典型的自身折叠结构，许多以特异性杂交反应为基础的DNA计算模型，都要求避免单链DNA形成2级结构，这样单链DNA分子才能和自身的补链充分有效地发生特异性杂交，其公式如式(6)所示

$$\begin{aligned} f_{Hp}(X)&=\sum_{i=1}^{n}\mathrm{Hp}(X_i)\\ &=\sum_{i=1}^{n}\sum_{s=S_{\min}}^{(l/R_{\min})/2}\sum_{r=R_{\min}}^{l-2s}\sum_{i=1}^{l-2s-r}T\\ &\quad\cdot\left(\sum_{j=1}^{s}\mathrm{bp}(x_{s+i-j},x_{s+i+r+j}),\frac{s}{2}\right)\end{aligned} \tag{6}$$

其中，X_i表示DNA序列X中的第i个碱基，s为发卡结构茎长，$S_{\min}$为设定的最小茎长，r表示环的长度，$R_{\min}$为设定的最小环长，本文中$S_{\min}$和$R_{\min}$均被设置为6。

2.6 解链温度约束

解链温度(melting Temperature, Tm)是双链DNA分子在加温变性过程中，有50%的DNA分子打开双链变成单链时的温度。解链温度它是评价DNA分子化学热力学稳定性的一个重要参数。DNA计算要求DNA分子具有一致的解链温度。影响解链温度Tm的因素为：DNA分子组成、DNA分子浓度、溶液PH值等。根据Nearest-Neighbors热力学模型，其公式如式(7)所示

$$\left.\begin{aligned} &f_{\mathrm{Tm}}(X)=\max_i\{\mathrm{Tm}(X_i)\}-\min_j\{\mathrm{Tm}(X_j)\}\\ &\mathrm{Tm}(X_i)=\sum_{i=1}^{n}\frac{\Delta H^\circ}{S^\circ+R\ln(|C|/4)}\end{aligned}\right\} \tag{7}$$

其中，H°是相邻碱基的总焓，S°是相邻碱基的总熵，R为气体常数(1.987 cal/kmol)，C为DNA分子浓度。根据上面3个Tm值计算式可以看出，GC含量高，Tm值大；DNA分子浓度大，Tm值大；溶液PH值大，Tm值大。

3 多目标进化策略DNA编码算法

Shin等人[20]通过仿真实验证明了DNA编码约束函数中存在相互冲突的目标函数，并且它们都是具有很多局部最优解的不连续的函数，而且没有全局最优解的梯度信息。传统的基于遗传算法的多目标优化算法NSGA-II和MOEA/D，通常只能处理2～3个目标函数的优化问题，产生的新种群由于高维目标函数无法比较支配关系，缺少选择压力推进种群收敛。此外，由于交叉算子实现近邻局部搜索的过程中会产生大量跟父代相似的DNA分子候选解，跟DNA编码需要降低相似度的要求相互矛盾。本文采用多目标进化策略可有效降低种群大小，提高局部和全局搜索效率。

3.1 DNA分子编码及变异

进化策略中的核心运算是变异，对于一个长度为n的单链DNA序列X=5'–$x_1x_2\cdots x_n$–3' 来说，可以变异的有两个变化的方面，一个是每一位碱基，另一个是同时变化碱基的个数。每一位碱基$x_i\in${A, C, G, T}, 本文采用四进制整数进行编码{A, C, G, T}→{0, 1, 2, 3}，因此单个碱基x_i可以采用式(8)进行变异。

$$x_i=(x_i+\mathrm{random}(3)+1)\bmod 4 \tag{8}$$

例如，x_i为碱基C=1, random(a)随机产生[0, a-1]间的任意整数，则式(8)将x_i随机变换为{0, 2, 3}→{A, G,T}。

此外，对于DNA序列X=5′–$x_1x_2\cdots x_n$–3′，共有n个碱基x_i可以发生变异。变异的碱基的个数将对算法的局部搜索能力和全局搜索能力具有重要的影响。令变异碱基的个数为k，如果k=1则每次只有1个碱基发生变化，新产生的候选序列跟原序列X非常接近，算法就产生了很强的局部搜索能力，但是容易陷入局部最优。当k=n，则每次所有的碱基都随机发生变化，新产生的候选序列跟原序列每一位碱基都不同，极大地降低了序列的相似性，但是容易导致算法不收敛。我们通常期望算法前期先尽量探索全局，后期增强局部搜索能力。然而，在算法执行的过程中，没有明显的线索判断算法处于前期还是后期阶段。

本文采用的进化策略同时考虑局部搜索和全局搜索能力，将参数k取为单链DNA序列长度的随机数，随机确定变异碱基的总个数。进而，随机挑选k个不同的碱基位置对DNA序列进行变异，算法流程如表1所示。

3.2 评价函数

由于DNA序列设计问题中编码约束函数相互冲突，并且每个指标的值域不相同，进化策略的个体在变异中会存在相互非支配的解，通常采用式(9)对个体进行评价，选取适应度函数较优的个体。

$$\text{Fitness}(x)=\sum_{i=1}^{m}\frac{f_i(x)-f_i^{\min}}{f_i^{\max}-f_i^{\min}} \tag{9}$$

当整个种群适应度值越来越小的时候，算法就收敛的越来越好。传统的多目标进化算法需要找到分布均匀，且相互非支配的多组候选解。利用各种有效的小生境技术可以帮助让算法在收敛的过程中，保持较好的多样性。经过分析可以发现，DNA编码问题中的相似度约束(f_{Si})已经对DNA分子的多样性进行了限制，将多样性的保持转换为了目标函数的优化。此外，虽然从多目标优化的角度，相互非支配的候选解无法做出取舍。但是从DNA计算的实验需求上，DNA计算更倾向于选择各目标函数值较为平衡的解。以相似度和H-measure为例，高相似度的DNA分子集合，容易导致DNA分子X和Y的补链发生非特异性杂交。高H-measure的DNA分子集合，容易导致DNA分子X和Y的反向序列直接发生互补，导致非特异性杂交错配，如图1所示。

因此，在所有的相互非支配的候选解集中，理想的选择是相似度和H-measure两个指标较为平衡的解，这种类型的DNA分子集合，将更有效的减少DNA分子间非特异性杂交，保证DNA计算的可靠性。因此，我们引入相似度和H-measure目标函数值之差的平方项，引导算法选择冲突目标较均衡的解，如式(10)所示

$$\text{Fitness}(x)=(f_{\text{Si}}-f_{\text{Hm}})^2+\sum_{i=1}^{m}\frac{f_i(x)-f_i^{\min}}{f_i^{\max}-f_i^{\min}} \tag{10}$$

表 1 个体变异算子伪代码

输入: X=5′–$x_1x_2\ldots x_n$–3′
输出: Y=5′–$y_1y_2\ldots y_n$–3′
1: for j=1 to n
2: List.add(j)
3: end for
4: k=random(n)
5: for j=1 to k
6: i = random(List.count)
7: y_i=(x_i+random(3)+1) mod 4
8: List.delete(i)
9: end for

3.3 算法流程

算法首先生成初始化种群P_t，然后依次计算个体p的目标函数值。然后，种群个体变异产生新的候选个体q，如果新的候选解q能支配原个体p，则用新的候选解q替换掉原个体p，表示为$q \prec p$。其中，个体q的目标函数值都小于等于个体p，且至少有1个目标函数值小于p。如果新的候选解q被原个体p支配即$p \prec q$，则放弃新候选解。如果p和q之间的关系不满足上述两种情况，则定义p和q两个个体相互非支配，即$(q \prec p)$ and $(p \prec q)$，则比较式(10)的适应度函数值，如果新的候选解综合目标函数较小，且两个指标更平均。则新候选解替代原个体，算法流程如表2所示。

4 实验结果

为了验证算法的有效性，本文将算法和MGA[7], IWO[8], pMO-ABC[10]算法进行对比，根据Adelman的经典实验，产生7条长度为20的DNA序列集合进行比较，实验结果如表3所示。

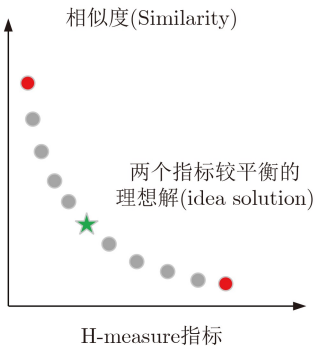


图 1 非支配解集中的边界点和理想解

表 2 算法总体流程伪代码

1: Initialization
2: while (t < max iteration)
3: for i=1 to P
4: $p = P_t(i)$
5: q = Individual Mutation(p)
6: if $q \prec p$ then
7: $P_t(i)$=q
8: else if $(q \prec p)$ and $(p \prec q)$ then
9: if Fitness(q) > Fitness(p) then
10: $P_t(i)$=q
11: end if
12: end if
13: end for
14: t=t+1
15: end while

表 3　7条长度为20的DNA序列的结果比较

方法(序列)	连续性	发卡结构	H-measure	相似度	Tm	GC(%)
MGA[7]						
TAGACCACTGTTGCACATGG	0	0	58	52	50.2794	50
ATTCGGTCAGACTTGCTGTG	0	0	64	52	48.6650	50
ATAGTGCGGACAGTAGTTCC	0	0	66	59	50.1634	50
AATACGCGGAACGTAACCTC	0	0	61	85	50.4158	50
AATACGCGGAACGTAACCTC	0	0	61	85	50.4158	50
ACAGCCTTAAGCCTAACTCC	0	3	65	54	49.0641	50
ATGCTTCCGACATGGAATGG	0	3	63	57	49.8160	50
$f(x)$	0	6	438	444	1.7508	0
IWO[8]						
ACACCAGCACACCAGAAACA	9	0	55	55	48.4670	50
GTTCAATCGCCTCTCGGTAT	0	0	57	57	49.3935	50
GCTACCTCTTCCACCATTCT	0	0	55	55	49.2453	50
GAATCAATGGCGGTCAGAAG	0	0	66	66	49.9440	50
TTGGTCCGGTTATTCCTTCG	0	0	65	65	50.6418	50
CCATCTTCCGTACTTCACTG	0	0	56	56	51.0993	50
TTCGACTCGGTTCCTTGCTA	0	0	58	58	47.6049	50
$f(x)$	9	0	412	412	3.4944	0
pMO-ABC[10]						
TGTGGTTGGTTAGTCGGTTG	0	0	46	49	51.0421	50
GGTGGTATTGGTGGTATTGG	0	0	47	47	53.8027	50
CTTCTCTTCTCTTGCCGCTT	0	0	39	56	46.4112	50
AACAACCTCCACACCGAACA	0	0	62	32	49.1737	50
CTCTCTCTCTCACTCTCTCA	0	0	41	48	46.5220	50
CTCTCATTCCTTCTTACCCC	16	0	43	51	50.8283	50
TGGTGTTGCTGGTGTAGGTT	0	0	48	51	49.3985	50
$f(x)$	16	0	326	334	7.3915	0
MOES						
GGAGAGGAGAAGAAGAAGAG	0	0	52	25	48.1727	50
CCTCCACATCACCATTAACC	0	0	56	31	52.3807	50
CTCTCTCTCTCTCTCTCTCT	0	0	34	37	45.6658	50
TTCCTTCCTTCCTTCCTTCC	0	0	36	39	48.7500	50
TTGGTTGGTTGGTTGGTTGG	0	0	30	46	51.3054	50
TTGTTGTTGGTGGTGGTGGT	0	0	30	48	50.2236	50
TGTGTGTGGTGTGTGTTGTG	0	0	30	46	51.0025	50
$f(x)$	0	0	268	272	6.7149	0

从表3中可以看出4种算法在GC含量约束上的表现都很一致。在温度约束方面，4种算法都满足比较统一的解链温度。本文算法MOES产生DNA序列的连续性和发卡结构均为0，因此不会产生不期望的二级结构。此外，相似度和H-measure的数值最低，表明DNA序列集合具有更低的概率发生相互的非特异性杂交和错配。因此，从实验结果可以看出本文所提算法在解决DNA序列设计问题上比对比算法质量更高。

由于更大规模的DNA分子集合，可以用于求解更大规模的计算问题。本文生成14条长度为20的DNA序列集合，和MGA[7], INSGA-II[9], pMO-ABC[10]进行比较，实验结果如表4所示。可以看出，所有算法还是较好地满足了解链温度和GC含量的要

表 4 14条长度为20的DNA序列的结果比较

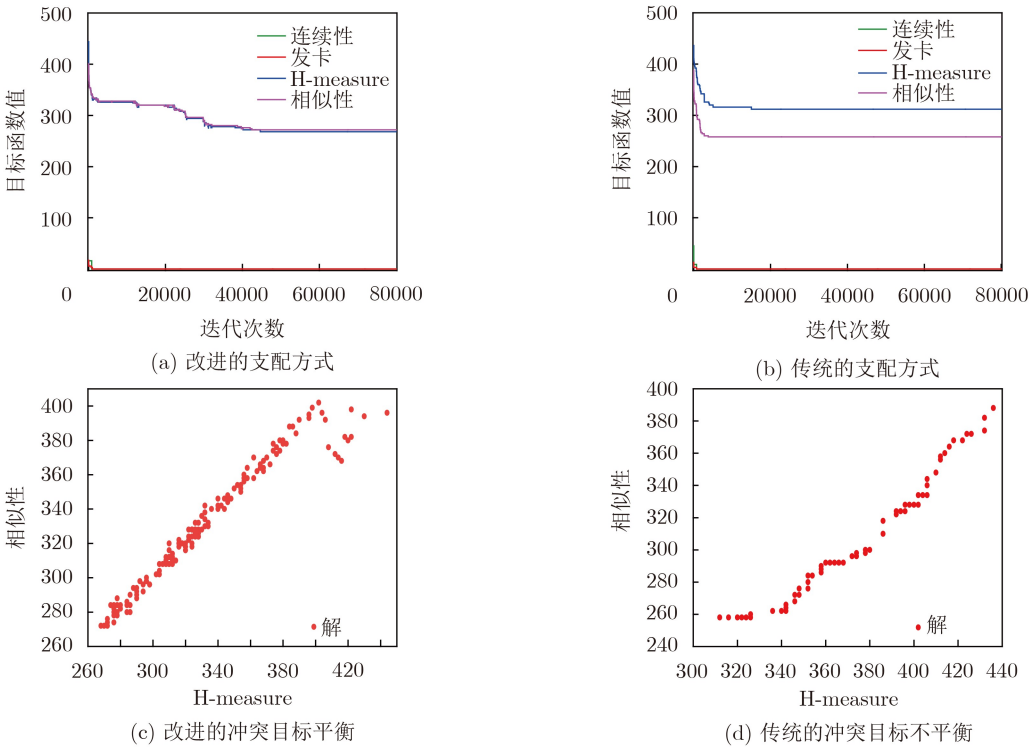
方法(序列)	连续性	发卡结构	H-measure	相似度	Tm	GC(%)
MGA[7]						
CTCATCTAATCAGCCTCGCA	0	0	135	114	48.1554	50
CTAATAGTGACAGCTGCGTG	0	3	131	119	50.2421	50
GCATCGTTAGAGACACCTAC	0	3	134	124	50.7932	50
GCATCAATATGCGCGACTAC	0	0	131	125	50.2815	50
CATTAAGTAGACGCTGTCGG	0	3	132	114	50.9507	50
TATGGATGAGGAGGACCTAG	0	3	133	117	50.6387	50
CAGAGATGTTCTGTACCACC	0	3	128	117	51.2232	50
CGTCGAGAATTCGTAGCTCA	0	0	137	119	48.3224	50
TCTGTTACCGTATCGGATCG	0	3	129	115	50.8791	50
AGAAGAGTTCGACTTGCTGG	0	3	134	121	47.5507	50
GCAAGGAATTCACCGTCTGT	0	3	133	129	48.9881	50
CGTGTGAAGAGAGTGGTTCA	0	0	127	123	48.9355	50
CGACTGAATCATGGACCTGT	0	3	134	126	49.7624	50
TACCGAGAAGTAGGACTGCA	0	3	134	124	48.3847	50
$f(x)$	0	30	1852	1687	3.6725	0
INSGA-II[9]						
CGAGACATCGTGCATATCGT	0	4	143	124	49.6393	50
TATAGCACGAGTGCGCGTAT	0	3	137	130	48.5659	50
GATCTACGATCATGAGAGCG	0	4	135	126	49.6673	50
TCTGTACTGCTGACTCGAGT	0	3	163	124	47.1312	50
CGAGTAGTCACACGATGAGA	0	0	152	132	49.2836	50
AGATGATCAGCAGCGACACT	0	3	133	133	46.5546	50
TGTGCTCGTCTCTGCATACT	0	4	159	130	47.1507	50
AGACGAGTCGTACAGTACAG	0	0	152	134	49.9091	50
ATGTACGTGAGATGCAGCAG	0	0	139	121	48.9270	50
ATCACTACTCGCTCGTCACT	0	3	141	132	47.5190	50
TCAGAGATACTCACGTCACG	0	3	142	123	49.2836	50
GACAGAGCTATCAGCTACTG	0	3	129	124	49.2927	50
GCTGACATAGAGTGCGATAC	0	0	130	133	50.1725	50
ACATCGACACTACTACGCAC	0	3	133	144	50.1554	50
$f(x)$	0	33	1988	1810	3.6179	0
pMO-ABC[10]						
GTTATTGGTGGTGTGCGTTG	0	0	143	82	51.9305	50
ACGGAAGTAGGAAGGAGAGA	0	0	137	106	47.8089	50
GGAAGACGCAGAAGAGAAAG	9	0	121	110	48.2609	50
CCTCCTTATTGCCTTCCTTC	0	0	114	102	50.3081	50
AACTAACCACCGACCAACCA	0	0	95	118	50.1102	50
ACACACAACACACACACTCC	0	0	88	119	50.4577	50
ACACCACCACATTACCACAC	0	0	97	119	51.9161	50
CTTCCGTCTCTTCTCTCTCT	0	0	134	105	46.9561	50
AAGGAGCGAGGAAGCGAAAA	16	0	107	95	45.8306	50
AACACCAGAACATCCACACC	0	0	90	131	50.5474	50
CCAACACCATACAACAGACC	0	0	95	130	52.3720	50

续表 4

方法(序列)	连续性	发卡结构	H-measure	相似度	Tm	GC(%)
AAGGCGGAAGGATAGAAGAG	0	0	128	115	48.5370	50
TCTGCCGCTTCTTCTTCTTC	0	0	118	95	46.4000	50
TCCTCTCGTCTATTCTCCTC	0	0	111	98	48.4427	50
$f(x)$	25	0	1578	1525	6.5414	0
MOES						
CATACACACTCACACTCACC	0	0	112	89	51.6792	50
TTGTTGTGGGTTGTCCGGTT	9	0	105	90	49.7949	50
ACACACACACACACACACAC	0	0	93	78	50.9244	50
TTGTGGTCCTGGTGTTCTCT	0	0	112	90	48.4957	50
GAGAGAGAGAGAGAGAGAGA	0	0	72	100	45.6568	50
TGGTGTGGTGTGGTTAGGTT	0	0	96	93	50.5325	50
TTGGTGGTGGTGGTTGTAGT	0	0	96	95	50.5325	50
CCAACCAACCAACCAACCAA	0	0	95	78	51.3054	50
AACAAGCCAGAAGCCAGAAG	0	0	94	102	47.5066	50
GTTGGTGCTGTTGTTGAGGT	0	0	101	99	49.4550	50
GAAGAAGGGAGGAGAGAAGA	9	0	77	108	47.4961	50
AATGGAAGCGAAGCGAAGTG	0	0	93	104	47.6766	50
AACCATCAACCGCCGAAGAA	0	3	104	95	48.1694	50
AAGGTGGAGAGGAAGGAGAA	0	0	82	111	47.4098	50
$f(x)$	18	3	1332	1332	6.0224	0

求。本文算法显著地降低了相似度和H-measure的目标函数值，可以极大地避免DNA分子间的相互干扰，反映出算法具有最好的收敛性能。

为了进一步证明本文所提的改进评价函数在解决算法早熟和目标不平衡问题上的有效性，本文和传统的评价函数进行了比较。图2(a)和图2(b)分别



(a) 改进的支配方式　(b) 传统的支配方式

(c) 改进的冲突目标平衡　(d) 传统的冲突目标不平衡

图 2 目标函数收敛过程

显示了使用改进的评价函数和传统的评价函数时目标函数值的收敛过程。首先，连续性和发卡结构都能在早期就收敛到0，这表明基于本文算法的具有很强的局部和全局搜索能力。但是对于相似度和H-measure，传统的评价函数在大约6000代时就停止继续下降并且H-measure一直比相似度大很多，陷入了局部最优。可以认为传统的方式，由于相似度目标上的值优化变小，而舍弃H-measure目标上的优化。而改进的支配方式中，相似度和H-measure能够一直一起进化，直到大约45000代才收敛。在改进的算法中，需要优化的目标是相似度和H-measure中较大的值，因此即使一个目标收敛，程序仍然乐于接受能够优化较差目标的变异。使得程序不会过早收敛并且维持目标函数值之间的平衡。图2(c))和图2(d)显示的是整个迭代过程中相似度和H-measure的分布。可以看出改进的方式中算法会优先优化相似度和H-measure中较劣的解决方案，两个目标间差距不大，证明了本文算法的有效性。

5 结束语

本文提出一种基于进化策略的多目标优化算法求解DNA序列设计问题，设计的随机碱基变异算子可以兼顾局部搜索和全局搜索能力，跟传统的优化算法相比，没有需要设置的敏感参数。此外，改进的评价函数可以在优化目标函数的同时，综合考虑冲突指标相似度和H-measure的平衡性，可以有效地减少DNA分子集合中的非特异性杂交。由于算法运行在较小的种群上，可以极大地减少不必要的时间复杂度。通过实验跟几种最新的DNA编码设计算法进行比对，结果表明本文算法可以设计出高质量的DNA分子集合，可有效提高DNA计算的有效性和可靠性。

参 考 文 献

[1] ADLEMAN L M. Molecular computation of solutions to combinatorial problems[J]. *Science*, 1994, 266(5187): 1021–1024. doi: 10.1126/science.7973651.

[2] DE SILVA P Y and GANEGODA G U. New trends of digital data storage in DNA[J]. *BioMed Research International*, 2016: 8072463.

[3] BRAICH R S, CHELYAPOV N, JOHNSON C, *et al*. Solution of a 20-variable 3-SAT problem on a DNA computer[J]. *Science*, 2002, 296(5567): 499–502. doi: 10.1126/science.1069528.

[4] ZIMMERMANN K H. Efficient DNA sticker algorithms for NP-complete graph problems[J]. *Computer Physics Communications*, 2002, 144(3): 297–309. doi: 10.1016/S0010-4655(02)00270-9.

[5] XU Jin, QIANG Xiaoli, ZHANG Kai, *et al*. A DNA computing model for the graph vertex coloring problem based on a probe graph[J]. *Engineering*, 2018, 4(1): 61–77. doi: 10.1016/j.eng.2018.02.011.

[6] CHAVES-GONZÁLEZ J M and VEGA-RODRÍGUEZ M A. A multiobjective approach based on the behavior of fireflies to generate reliable DNA sequences for molecular computing[J]. *Applied Mathematics and Computation*, 2014, 227: 291–308. doi: 10.1016/j.amc.2013.11.032.

[7] PENG Ximei, ZHENG Xuedong, WANG Bin, *et al*. A micro-genetic algorithm for DNA encoding sequences design[C]. The 2nd International Conference on Control Science and Systems Engineering, Singapore, 2016: 10–14.

[8] YANG Gaijing, WANG Bin, ZHENG Xuedong, *et al*. IWO algorithm based on niche crowding for DNA sequence design[J]. *Interdisciplinary Sciences*: *Computational Life Sciences*, 2017, 9(3): 341–349. doi: 10.1007/s12539-016-0160-0.

[9] WANG Yanfeng, SHEN Yongpeng, ZHANG Xuncai, *et al*. An improved non-dominated sorting genetic algorithm-Ⅱ (INSGA-Ⅱ) applied to the design of DNA codewords[J]. *Mathematics and Computers in Simulation*, 2018, 151: 131–139. doi: 10.1016/j.matcom.2018.03.011.

[10] CHAVES-GONZÁLEZ J M and MARTÍNEZ-GIL J. An efficient design for a multi-objective evolutionary algorithm to generate DNA libraries suitable for computation[J]. *Interdisciplinary Sciences*: *Computational Life Sciences*, 2018, 11(3): 542–558.

[11] YANG Shuming, SHAO Dongguo, and LUO Yangjie. A novel evolution strategy for multiobjective optimization problem[J]. *Applied Mathematics and Computation*, 2005, 170(2): 850–873. doi: 10.1016/j.amc.2004.12.025.

[12] ARNOLD D V and BEYER H G. Investigation of the (μ, λ)-ES in the presence of noise[C]. The 2001 Congress on Evolutionary Computation, Seoul, South Korea, 2001, 1: 332–339.

[13] EBENAU C, ROTTSCHÄFER J, and THIERAUF G. An advanced evolutionary strategy with an adaptive penalty function for mixed-discrete structural optimisation[J]. *Advances in Engineering Software*, 2005, 36(1): 29–38. doi: 10.1016/j.advengsoft.2003.10.008.

[14] MEZURA-MONTES E and COELLO C A C. A simple multimembered evolution strategy to solve constrained optimization problems[J]. *IEEE Transactions on Evolutionary Computation*, 2005, 9(1): 1–17. doi: 10.1109/TEVC.2004.836819.

[15] XIAO J, XU Jin, CHEN Zhihua, *et al*. A hybrid quantum chaotic swarm evolutionary algorithm for DNA encoding[J]. *Computers & Mathematics with Applications*, 2009, 57(11/12): 1949–1958.

[16] CHAVES-GONZÁLEZ J M, VEGA-RODRÍGUEZ M A, and 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(9): 2045–2057. doi: 10.1016/j.engappai.2013.04.011.

[17] MUHAMMAD M S, SELVAN K V, MASRA S M W, *et al*. An improved binary particle swarm optimization algorithm for DNA encoding enhancement[C]. 2011 IEEE Symposium on Swarm Intelligence, Paris, France, 2011: 1–8.

[18] CHAVES-GONZÁLEZ J M and VEGA-RODRÍGUEZ M A. DNA strand generation for DNA computing by using a multi-objective differential evolution algorithm[J]. *Biosystems*, 2014, 116: 49–64. doi: 10.1016/j.biosystems.2013.12.005.

[19] BUI L T and ALAM S. Multi-Objective Optimization in Computational Intelligence: Theory and Practice[M]. Hershey: IGI Global, 2008.

[20] SHIN S Y, LEE I H, KIM D, *et al*. Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing[J]. *IEEE Transactions on Evolutionary Computation*, 2005, 9(2): 143–158. doi: 10.1109/TEVC.2005.844166.

张　凯：男，1979年生，教授，研究方向为DNA计算、多目标进化算法.

陈　彬：男，1982年生，硕士生，研究方向为智能优化算法.

许志伟：男，1995年生，博士生，研究方向为DNA编码、演化计算.