第 28 卷第 3 期　　四川理工学院学报(自然科学版)　　Vol. 28　No. 3
2015 年 6 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Jun. 2015
文章编号:1673-1549(2015)03-0046-05　　DOI:10.11863/j.suse.2015.03.10

MO_DE:一种结合多目标优化机制的 DNA 编码序列算法

岑　巍

(上海浦东发展银行, 上海　200042)

摘　要:针对现有 DNA 计算中存在的编码序列设计稳定性不足、可靠性不完善等问题,充分考虑基本编码问题,设计出一种基于多目标优化机制的 DNA 编码序列设计算法(MO_DE:multiobjective design algorithm)。在一定的约束条件下,该算法利用了多目标优化机制以及采取小种蚁群算法,将 h-distance 因子添加到单链 DNA 架构中,建立一种 DNA 序列公用方法。通过模拟实验表明,该算法与同类型算法相比,在计算效率、优化性方面具有一定优势。

关键词:DNA 计算;多目标优化;小种蚁群;编码序列;MO_DE
中图分类号:TP301　　**文献标志码**:A

引　言

近年来,智能化生物计算机作为一个新型的热点研究领域,得到了长足发展与广泛关注。DNA 计算则是其中重要一环。DNA 计算是以 DNA 链条作为操作处理手段,使用 DNA 链中多行反应计算能力的一种智能化生物计算方式,现已成功解决了众多世界性难题,如 Hamilton 路径、NP 完全难题等[1]。

在 DNA 计算中,DNA 编码序列算法是影响 DNA 计算的核心与关键问题。Cui G Z 等[2]提出,如何加强 DNA 编码序列的特异性设计与计算区别功能是 DNA 编码序列问题的关键研究方向。在进行 DNA 编码序列设计之前,首先需要关联 DNA 组合限制约束条件、热力学问题(如:汉明距离限制、二级结构限制、连续性限制、解链温度限制、GC 含量限制等),所以,张强等[3]基于上述相关约束条件,提出了一种 DNA 模板编码方法。Zhang X 等[4]在文献[3]的基础之上,提出了一种基于 DNA 编码序列的编译算法。Shin S Y 等[5]基于遗传算法设计了一种新型的 DNA 序列类型;Wang W 等[6]将 DNA 编码序列问题成功转化成一个附有约束限制的多目标优化问题。殷脂等[7]分析了 DNA 编码序列对 DNA 计算的稳定性和可靠性的影响,文中指出,DNA 编码的多约束条件是影响 DNA 计算稳定性和可靠性的主要因素。目前,DNA 编码序列设计问题解决手段一般包括随机查询、种族进化、模板编码以及多目标优化等。因此,如何找到一种标准条件下的 DNA 编码序列问题求解方法,成为首要解决的目标。

针对上述描述,本文充分考虑到 DNA 基本编码问题[8],结合多目标优化机制,设计出一种新型的 DNA 编码序列设计算法,即 MO_DE 算法(MO_DE:multiobjective design algorithm)。在一定的约束条件下,并采取小种蚁群算法,将 h - distance 因子添加到单链 DNA 架构中,构造一种 DNA 序列公用方法。通过与相关文献作比较分析,说明了该算法在计算效率、优化性等方面具有一定优势。

收稿日期:2015-04-23
作者简介:岑　巍(1974-),男,浙江余姚人,工程师,硕士,主要从事商业银行中间业务产品研发和数据库优化、云计算等方面的研究,(E-mail)2656543164@qq.com

1　约束限制与优化结构

1.1　约束限制

具体包括以下几种约束条件与限制因素,即:

(1) 相似性

相似性主要应用于调整 DNA 编码序列设计中的移位相似度,对于 DNA 单序列链条 D_j 与 D_k, 由于存在相关关联公式:

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

$$h(N_{ji},N_{ki}) = \begin{cases} 0, N_{ji} = N_{ki} \\ 1, N_{ji} \neq N_{ki} \end{cases} \tag{2}$$

由此可知:

$$S(D_j,D_k) = \min_{-n<l<n} H(D_j,\sigma^l(D_k)) \tag{3}$$

依据 DNA 编码序列的无限制运动与全面扩展[9], 利用以下最小相似性作为目标方法:

$$f_H(D_j) = \min_{D_k \in C-\{D_j\}} S(D_j,D_k) \tag{4}$$

(2) 移动测度

移动测度主要是使用于标记 DNA 编码序列中各个移位运动概率值,设定:

$$H_m(D_j,D_k) = \min_{-n<l<n} H(D_j,\sigma^l(\overline{D_k^R})) \tag{5}$$

可知, $\overline{D_k^R}$ 说明 DNA 编码序列 D_k 的逆序队列。依据 DNA 编码序列的无限制运动与全面扩展,利用其最小化移动测度作为目标方法:

$$f_{HR}(D_j) = \min_{D_k \in C} \min_{-n<l<n} H(D_j,\sigma^l(\overline{D_k^R})) \tag{6}$$

(3) h - 距离

h - 距离(汉明)说明了 DNA 编码序列在相似性以及移动测度相关特征,表达了 DNA 编码序列中公用分享度。

$$h_d(D_j,D_k) = min[S(D_j,D_k),H_m(D_j,D_k), S(\overline{D_j^R},\overline{D_k^R}),H_m(\overline{D_j^R},\overline{D_k^R})] \tag{7}$$

(4) GC 含量

GC 含量是碱基 G 和 C 量之间在 DNA 编码序列中的存在占有比重。设定 $GC(D_j)$ 是 DNA 编码序列 D_j 的 GC 含量, GC_{goal} 表示目标参数, k_{GC} 表示 GC 含量能够接受上下限区域,可知 GC 含量的约束限制:

$$f_{GC}(D_j) = |GC(D_j) - GC_{goal}| \leqslant k_{GC} \tag{8}$$

(5) T_m 温度

在 DNA 编码设计过程中,一半左右的 DNA 链条所需的温度参数值。设定 $T_m(D_j)$ 是 DNA 编码序列 D_j 的 T_m 温度参数值, $T_{m_{goal}}$ 表示目标参数, k_{T_m} 表示 T_m 温度能够接受上下限区域,可知 T_m 温度约束限制:

$$f_{T_m}(D_j) = |T_m(D_j) - T_{m_{goal}}| \leqslant k_{T_m} \tag{9}$$

(6) HP 环形模型

HP 环形模型是 DNA 编码序列中杂交而成,设定:

$$f_{HP}(D_j) = \sum_{s+l \leqslant p \leqslant n-r-s} \sum_{5 \leqslant r \leqslant n-2s} Haripin(p,r,s,D_j) \tag{10}$$

其中, $Haripin(p,r,s,D_j)$ 说明 DNA 编码序列 D_j 中起始于点 p、DNA 环形周长是 r、DNA 环形杆长是 s 的发卡参数。设定 k_{HP} 表示 HP 能够接受上下限区域,可知 HP 约束限制:

$$f_{HP}(D_j) \leqslant k_{HP} \tag{11}$$

(7) Con 函数因子

基于 DNA 编码序列中 Con 函数知识,设定:

$$f_{Con}(D_j,N) = \sum_{i=1}^{n} (i-1) N_{D_j}^i \tag{12}$$

式中, $N_{D_j}^i$ 说明 DNA 编码序列中碱基 N 元素组合的长度数值 i 所具有子串出现的次数值。设定 k_{Con} 表示 Con 函数因子能够接受上下限区域。可知 Con 函数因子的约束限制:

$$f_{Con}(D_j,N) \leqslant k_{Con} \tag{13}$$

1.2　优化结构

依据上述提供的多种目标方法与约束条件的局限限制,DNA 编码序列设计思路能够构成一种标准的优化结构,即:

$$\begin{aligned} &\max F(D_j) = [f_H(D_j),f_{HR}(D_j)] \\ &s.t. f_{GC}(D_j) \leqslant k_{GC} \\ &\quad f_{T_m}(D_j) \leqslant k_{T_m} \\ &\quad f_{HP}(D_j) \leqslant k_{HP} \\ &\quad f_{Con}(D_j,N) \leqslant k_{Con} \end{aligned} \tag{14}$$

2　MO_DE 算法

按照上面 DNA 编码序列设计过程中的约束限制与优化结构,设计出一种结合多目标优化机制的小种蚁群算法,将其应用与 DNA 编码序列中,从而产生出 MO_DE 算法。

在 MO_DE 算法中,对于小种蚁群个体是 DNA 编码序列,其碱基参数变量 $\{a,g,c,t\}$ 编码设定是 $\{0,1,2,$

3},种群结构使用矩阵表示,即:四进制整数矩阵 $P_{m\times n}$,其中,n 表示 DNA 编码序列长度参数,m 表示矩阵大小,输出值是 $P_{m\times n}$ 的子公式 $C_{m\times n}$,m'表示 DNA 编码序列集合的大小。DNA 编码序列适应度表示为:$F(D_j)=\sum_{i=1}^{6}\alpha_i f_i(D_j)$,其中 α_i 表示权重参数,$f_i\in\{f_H, f_{HR}, -f_{GC}, -f_{T_m}, -f_{HP}, -f_{Con}\}$。

针对 DNA 编码序列 D_j 以及 D_k,可知共同分享方法设置:

$$sh(D_j,D_k)=\begin{cases}1-(h_d(D_j,D_k)/\sigma)^{\alpha}, d_{ij}<\sigma\\ 0, other\end{cases} \quad (15)$$

其中,σ 表示峰体半径参数,α 表示调控参数,DNA 编码序列小生境参数是:$f_{share}(D_j)=\sum_{k=1}^{n}sh(D_j,D_k)$,其共同分享适应度参数表示为:$F(D_j)/f_{share}(D_j)$。

基于 DNA 编码序列种群四进制整数矩阵 $P_{m\times n}$,添加一些算子元素:

(1) SX,即洗牌交叉处理。通过概率参数 P_{SXO},随机地对四进制整数矩阵 $P_{m\times n}$ 进行排序,其行表示 $P'_{m\times n}$,对于 $P'_{m\times n}$ 洗牌操作获取其行向量 P,对于行向量 P 分组操作获取 $P''_{m\times n}$,最终随机交叉配对。

(2) RX,即随机交叉处理。通过概率参数 P_{RXO},对四进制整数矩阵 $P_{m\times n}$ 的行参数进行随机选择并配对,其次随机进行交叉位置选取,进行随机交叉操作。

(3) RM,即逆向密码因子变异操作。设定 r 表示密码因子长度参数,通过概率参数 P_{RMO},对四进制整数矩阵 $P_{m\times n}$ 的 r 层子表达式进行随机选取,然后按行向量获取逆向序列。

(4) RCM(RM 的补集操作),即逆向补密码因子变异操作。设定 r 表示密码因子长度参数,通过概率参数 P_{RCMO},对四进制整数矩阵 $P_{m\times n}$ 的 r 层子表达式进行随机选取,然后按行向量获取逆向补集序列。

(5) M,即变异操作。通过概率参数 P_{MO},对四进制整数矩阵 $P_{m\times n}$ 的 L 层子表达式进行随机选取,将其碱基因子对应的整数参数随机代替上限整数参数。

(6) S,即选择操作。采取竞赛选取机制,大小为 2。

(7) 终止操作。持续进化时间长度超过几百年且达到历史最优化解时终止,不刷新操作。

通过上述算子元素,其 MO_DE 算法的具体流程如图 1 所示。

具体算法如下:

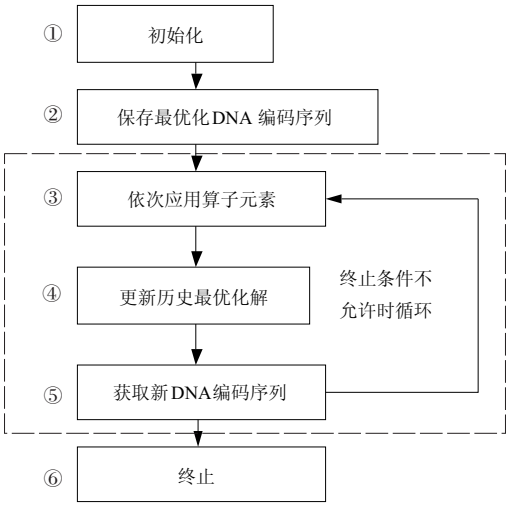

图 1 MO_DE 算法流程图

$P0, i=0$;//第一步;

C < = BestSave($P0, m$');//第二步

while 不符合终止操作 do//第三步

Pi < = $S(Pi)$,

$Pi\ 0$ < = $SX(Pi)$,

$Pi\ 1$ < = $RX(Pi\ 0)$,

$Pi\ 2$ < = $RM(Pi\ 0)$,

$Pi\ 3$ < = $RCM(Pi\ 0)$,

$Pi\ 4$ < = $M(Pi\ 0)$;

Pi' < = Union($Pi\ 1, Pi\ 2, Pi\ 3, Pi\ 4, C$);

C' < = BestSave(Pi', m');

If MeanFitness(C') > MeanFitness(C)

C < = C';//第四步

$Pi+1$ < = BestSave(Pi', m);//第五步

end while//第六步

在 MO_DE 算法流程中会出现对应的种群分散与扩展现象,其大小分散与扩展的上限大致是初始化状态的 5 倍至 7 倍左右。

3 模拟实验与结果分析

针对相关约束限制与优化模型,在 Matlab 7.0 的平台环境下,对设计的 MO_DE 算法进行模拟实验,操作环境为:Pentium Dual E2104, 2.5 GHz, 512 MB, Windows2008。设置参数因子:(1)DNA 编码序列发卡数值是 0;(2)GC 含量是 50%;(3)对应于 MO_DE 算法的各个参数($n, m', m, i, r, GC_{goal}, Tm_{goal}, a_1, a_2, a_3, a_4, a_5, a_6, \sigma, a, P_{SXO}, P_{RXO}, P_{RMO}, P_{RCMO}, P_{MO}$)分别是:(40, 6, 20,

400,3,40,47,1,1,1,1,1,1,8,1,0.8,0.2,0.2,0.2,0.4),以七个 DNA 编码序列为示例,实验结果见表 1。

依据 DNA 编码序列的无限制运动与全面扩展,对模拟实验的数据信息进行评估处理,特定在 DNA 编码序列中选取两个目标方法参数的最小数(min(MS),min(MH));选取 GC 含量与 T_m 温度的相对标准差参数,即(σ_{GC}, σ_{T_m}),从而使得 DNA 编码序列的理化性质保持相对统一;选取 S_{Con} 保证 DNA 编码序列的一定连续性。以七个 DNA 编码序列为示例,评估结果见表 2。

表 1　MO_DE 与同类型算法之间的 DNA 编码序列结果对比表

算　法	DNA 编码序列	MS	MH	Con	T_m
蚁群算法[10]	GAGCAAACAGCCAACCAAGC	13	8	17	56.28
	GCATACTTCACTACTCTCCC	14	7	9	55.38
	GGAGACAAGGAACAGAAGTA	10	9	6	54.32
	GTGGAAGGGACGGAAGAAGA	8	11	8	57.23
	AGTGGCTTGAAGCTTCCAA	14	10	7	51.87
	AGGAAGAGAGAAAGAAAAGG	11	8	13	56.54
	AAATAGGAAGGAAGGAAAGG	11	10	12	53.39
MO_EA 算法[11]	GCGCGGACAGCCAACCGAGA	7	8	8	52.44
	GCATACCTCACTTCTCTCTT	9	10	4	51.37
	GGAGACGGGGAACAGAAGTA	7	10	5	54.29
	GTGGCCGGGACGGAAGCCGA	10	8	4	55.47
	AGTGGCGGGTTGCTAACAA	11	7	3	51.28
	AGGAAGAGAGAAAGAGGAGG	8	9	2	56.83
	AAATATTAAGGAGGGTTAGG	7	9	3	53.27
MO_DE 算法	GAGCGCACAGCCAACCAAGC	11	13	0	57.23
	GCATACTTCACTACTCTCTT	10	12	0	57.87
	CCAGACAAGGAACAGAAGTA	11	13	0	57.18
	GTGGAAGGGACGGAAGAAGA	11	13	0	56.92
	AGTGGCTTGAAGCTTCCTT	10	13	0	57.77
	AGGAAGAGAGAAAGAAAAGG	11	13	0	57.55
	ATTTAGGGAGGTTGGTTAGG	11	12	0	56.29

注:MO_EA – Multiobjecticve evolutionary approach

表 2　DNA 编码序列结果评估项对比表

算法	min(MS)	min(MH)	S_{Con}	σ_{T_m}
蚁群算法[9]	8	7	63	1.94
MO_EA 算法[10]	9	8	48	1.38
MO_DE 算法	11	12	0	0.14

从表 1 与表 2 中可知,在各个评估项中,MO_DE 算法具有最优化结果,特别是在 DNA 编码序列的连续性上具有很大改进,同时 T_m 温度以及 GC 含量都足够统一集中,表示 DNA 编码序列的理化性质保持了相对统一,比同类型的两种算法都有明显的优势。除此之外,在种群分散膨胀方面,MO_DE 算法均比同类型两种算法更加完善,其迭代次数最少,计算操作量也最优。

在适应度函数方面,MO_DE 算法同样进行了改进设计,MO_DE 算法的进化收敛过程如图 2 所示。表明 MO_DE 算法在适应度函数方面具有很好的收敛性。

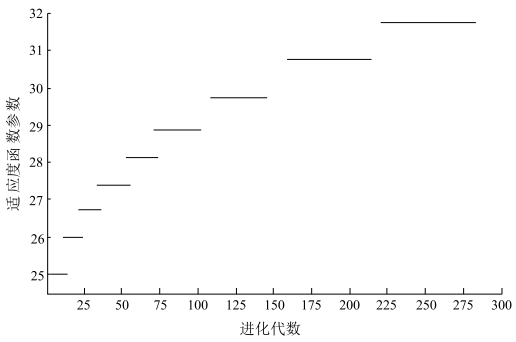

图 2　MO_DE 算法的进化收敛图

4　结束语

针对现有 DNA 计算中存在的编码序列设计稳定性、可靠性不完善等问题,从多层面分析,DNA 编码序列设计问题是一种多目标函数的优化问题。因此,本文充分考虑基本编码问题,结合多目标优化机制,设计出一种新型的、改进的 DNA 编码序列设计算法(MO_DE)。实验与分析表明,MO_DE 算法与同类型算法相比,在计算效率、优化性方面均具有一定优势。在将来,自适应算法[12]的应用分析,能够使其 DNA 编码序列设计问题得到更好的提高与解决,这将是下一步的研究方向。

参 考 文 献:

[1] Chang Wenglong, Vasilakos A V. DNA algorithms of implementing biomolecular databases on a biological computer[J]. IEEE Transactions on NanoBioscience, 2015,14(2):104-111.

[2] Cui G Z,Qin L M. Modified PSO algorithm for solving planar graph coloring problem[J]. Progress in Natural Science,2008,18(3):353-357.

[3] 张强,王宾,张锐,等.基于动态遗传算法的 DNA 序列集合设计[J].计算机学报,2008,31(12):2193-2199.

[4] Zhang X,Wang Y,Cui G,et al. Application of a novel IWO to the design of encoding sequences for DNA computing[J].Computers and Mathematics with Applications,2009,57(11/12):2001-2008.

[5] 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.

[6] Wang W,Zheng X,Zhang Q,et al. The optimization of DNA encodings based on GA/SA algorithm[J].Progress in Natural Science,2007,17(6):739-744.

[7] 殷脂,叶春明,马慧民,等.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.

[8] 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.

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

[10] 段海滨,王道波,朱家强,等.蚁群算法理论及应用研究的进展[J].控制与决策,2004,19(12):1321-1326.

[11] Chaves-Gonzaiez J M,Vega-Rodriguez M A,Granado-riado 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.

[12] Sharma S,Saxen R,Sharma S. Identification of mircrosatellites in DNA using adaptive S-transform[J]. IEEE Journal of Biomedical and Health Informatics,2014,19(3):1097-1105.

MO_DE: A DNA Coding Sequence Algorithm Based on Multi-objective Optimization Mechanisms

CEN Wei

(Shanghai Pudong Development Bank, Shanghai 200042, China)

Abstract: Aiming at the poor stability and reliability problems of sequence design existed in DNA computing, a DNA coding sequence design algorithm based on multi-objective optimization mechanism (MO_DE: multi-objective design algorithm) was designed with a full consideration of basic coding issues. Under certain constraints, MO_DE algorithm established a DNA sequence shared function by using multi-objective optimization mechanism and small populations ant colony algorithm, and adding the h-distance factor to the single stranded DNA architecture. The simulation experiments show that the MO_DE algorithm has certain advantages in computing efficiency and optimization compared with same type algorithms.

Key words: DNA computing; multi-objective optimization; small populations ant colony; coding sequence; MO_DE