一种基于关键路径法的DNA计算用寡核苷酸序列设计算法

吕鸣[1],张晓东[1],潘家奎[1],张治洲[3],胡钧[1,2]
(1.上海交通大学 Bio-X中心 DNA计算交叉团队、生命科学技术学院,上海,200240;
2.中国科学院上海应用物理研究所,上海,201800;3.天津科技大学,天津,300222)

摘要:在原有的生物大分子序列比对算法的基础上,结合图论中的关键路径法,提出了一种新的计算两寡核苷酸序列间最大配对程度的算法。采用此算法结合生成并测试的方法,能够寻找给定长度的一组适用于DNA计算的寡核苷酸序列。同时采用DNA芯片杂交方法验证了用该算法设计的一组序列的杂交特异性。

关键词:DNA计算;序列设计;序列比对;关键路径法
中图分类号:Q523+.8; O157.6 **文献标识码**:B **文章编号**:1672-5565(2007)-02-62-05

A sequences designing algorithm for DNA computation based on critical path method

LU Ming[1], ZHANG Xiao-dong[1], PAN Jia-kui[1], ZHANG Zhi-zhou[3], HU Jun[1,2]
(*1. Bio-X DNA Computer Consortium, College of Life Science & Biotechnology, Shanghai Jiao Tong University, Shanghai 200240;*
2. Shanghai Institute of Apply Physics, Shanghai, 201800; 3. Tianjin University of Science and Technology, Tianjin 300222)

Abstract: Based on the original biological macro molecule alignment algorithm and combined with critical path method in graph theory, an algorithm to calculate the maximal matches between two oligonucleotides was proposed. By using this algorithm and generate-and-test method, a group of oligonucleotides of a given length for DNA computation can be searched. A group of sequences designed by this algorithm were tested by hybridization on DNA chips.

Key Words: *DNA computation; sequences design; sequence alignment; critical path method*

在DNA计算[1-5]中,一个重要的问题就是如何设计一组满足一定约束条件的寡核苷酸序列,使得该组核酸的不同序列之间不存在非特异性的杂交。在早期研究中,Frutos[6]等人采用启发式算法寻找了总数为108个、长度为8bp的适用于DNA计算的寡核苷酸序列集合。但是此算法只能用于设计长度为8的序列,更多情况下,由于长度为8bp的序列之间杂交稳定性较差,往往需要用更长的序列。目前,用于考察两条寡核苷酸杂交能力的方法主要有两种:一种是基于热动力学的方法[7],另一种是基于序列比对的算法[8]。由于热动力学的方法需要首先获得很多相关的热动力学参数,而且这些参数往往与温度、浓度和离子强度有关,预知或调整热力学参数比较困难,通常只是采用估计值。相比之下,序列比对的方法相对简单,也足以区分完全杂交和非特异性杂合。但已有的基于动态规划的序列比对算法主要是针对大分子序列的同源性比对而设计的,这与考察寡核苷酸杂交能力存在较大区别。本文在序列比对的算法基础上,结合图论中的关键路径法,提出了一种新的寻找两条序列之间的最大碱基配对的算法,从而用于判断两寡核苷酸链间的配对程度。

收稿日期:2006-05-12;修回日期:2006-09-22
基金项目:上海市科委项目(NO.03D214025,045207);国家自主科学基金项目(NO.10335070)
作者简介:吕鸣(1981-),在读博士研究生,主要研究方向:纳米生物技术、生物信息学.E-mail:minglvxyz@gmail.com

2　算法

2.1　定义

一条长度为 L 的寡核苷酸序列表示为字符串 S，用 S^R 表示字符串 S 的反序列，S^C 表示 S 的补序列，S^{RC} 则表示 S 的反向互补序列。比如 S = AAGTCAGAAAGC，则 S^R = CGAAAGACTGAA，S^C = TTCAGTCTTTCG，S^{RC} = GCTTTCTGACTT。

给定两条序列 A、B，如果序列 A 表示的寡核苷酸和序列 B 表示的寡核苷酸有超过某一阈值的非特异性杂交，我们就说是不相容的，否则就说序列 A 和序列 B 是相容的，记为 A ~ B。

寻找一组不同序列之间没有非特异性杂交的寡核苷酸序列，实际上就是寻找一个这样的字符串集合 $P = \{O_i\}, i = 1,2,L,N$，N 为所需要的寡核苷酸序列的个数，通常让每个序列的长度相等，即 $|O_i| = L$。该集合需满足以下三个条件：(1)每条序列的 GC 含量都处在一个给定的范围，以保证各条序列与其各自的互补序列的结合强度基本一致。(2)集合中任意两条序列(包括序列与其自身)是相容的，即 $\forall(i,j) \in \{1,L,N\}^2, O_i \sim O_j$。(3)集合中任意一序列和任意另一序列的互补序列是相容的，即 $\forall(i,j) \in \{1,L,N\}^2, i \neq j, O_i \sim O_j^{RC}$。

为找到满足此要求的序列集合，首先用一种方法判断两条序列间是否相容。为找出一个满足上面提到的要求的一个子集，采取随机生成并测试的方法，每次随机生成一条序列，然后判断该序列是否与已有的集合中的序列相容，如果相容，则把它添加到集合中，然后寻找下一个满足要求的序列。

2.2　两条序列间的最大碱基配对

在序列比对方法的基础上，结合了关键路径法，提出一种新的算法解决寻找最大配对的问题。该算法主要分以下三步：(1)采用动态规划算法，找出两条序列间所有配对单元；(2)把每个配对单元作为节点，两个配对单元如果没有重叠部分，则添加一条前一配对单元到后一配对单元的有向边，后一配对单元的加权长度作为该有向边的权，从而构造出一个单起点到单终点的加权有向图(AOE 网)；(3)采用关键路径法[10]寻找出该加权有向图起点到终点的关键路径即最长路径，该路径上的节点就构成了该两条序列间的最大配对。下面以序列 A = AGTGATAACTAGCCG (5' - > 3')和序列 B = ACACCAGATTGGGCT (3' - > 5')为例详细说明此算法。

(1)给定一个阈值(此处取 6)，采用动态规划算法找出所有得分大于或等于该阈值的配对单元。考虑到 GC 配对含有 3 个氢键，AT 配对含有 2 个氢键，取 GC 配对得分为 3，AT 配对得分为 2。见(图 1)，共有五个满足要求的配对单元。

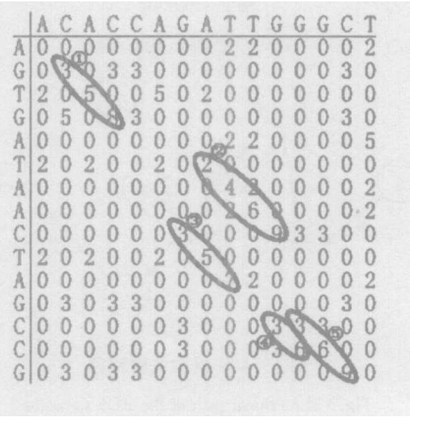

图 1　两条序列的得分矩阵

Fig. 1 The scoring matrix of the two sequences

(2)每个配对单元记为一个 4 元组(s1，e1，s2，e2)，作为图的节点集合，另加一个起点 S 和一个终点 E，其中 s1、e1 表示该配对单元在第一条序列上的起点和终点，s2、e2 表示该配对单元在第二条序列上的起点和终点。对于任两个配对单元 A($s1_A$，$e1_A$，$s2_A$，$e2_A$)和 B($s1_B$，$e1_B$，$s2_B$，$e2_B$)，如果 $s1_B > e1_A$ 且 $s2_B > e2_A$，则定义一条 A 到 B 的有向边，边的权值为 B 节点的分值；如果 $s1_A > e1_B$ 且 $s2_A > e2_B$，则定义一条 B 到 A 的有向边，边的权值为 A 节点的分值；否则，A 节点和 B 节点间就无边。另外，定义起点 S 到每个配对单元节点都有一条边，权值为该节点的得分，定义每个配对单元到终点 E 都有一条边，权值取 0。由此构造出一个加权有向图(AOE 网)，如图 2。

(3)采用关键路径算法求出图 2 中 AOE 网的关键路径，如图 3。关键路径经过的配对单元节点就构成了两条序列间的最大配对，如图 4，该两条序列的配对得分为 8 + 9 + 9 = 26。

2.3　寻找相容子集

采用随机生成并测试的方法寻找相容的寡核苷

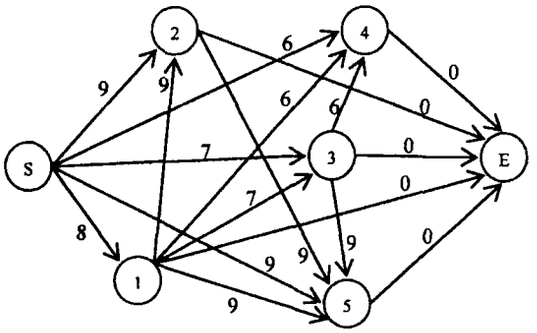

图2 AOE网

Fig.2 The AOE networks

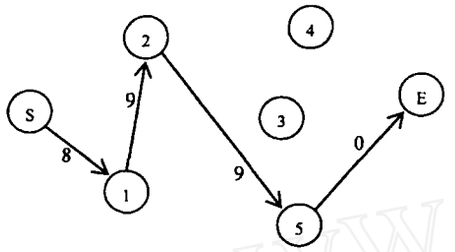

图3 AOE网的关键路径

Fig.3 The critical path of the AOE networks

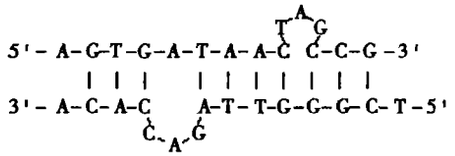

图4 序列A与序列B配对示意图

Fig.4 The matching schema of sequence A and sequence B

酸序列集合。方法如下：

(1) 初始化集合 P = Ø；

(2) 随机生成 GC 含量在指定范围内的指定长度的寡核苷酸序列；

(3) 判断该序列是否是简单重复序列，如果是，则排除简单重复序列，返回(2)

(4) 判断该序列是否与自身相容，如果否，则返回(2)

(5) 判断该序列是否与集合 P 中的所有序列相容。如果不相容，返回(2)；否则，将该序列添加到集合 P 中，并判断集合 P 中序列数目是否达到要求，是则退出，否则返回(2)。

3 实现

OligoPool 采用跨平台的 Python 语言实现。程序设计为图形用户界面，在序列设计窗口中，有 6 个参数，它们依次是：所设计的序列的长度(Length)；所期望得到的序列的个数(Number)；序列 GC 含量的下界(GC lower)和上界(GC upper)；判断序列是否相容的阈值(S_{max})，配对单元的最小得分(S_{unit})，默认值为 6，相当于连续 2 个 GC 对或连续 3 个 AT 对。考虑计算速度的原因，序列长度限制为 8 ~ 50，集合大小限制为 150 以下，原则上该算法并不受此限制。右边窗口为配对显示部分，对输入的两条序列进行配对，显示它们之间的最大配对情况。

4 实验测试结果

首先测试了在不同的参数下，在一定时间内所能生成寡核苷酸序列组中序列的个数，见表 1。由于采用的是生成并测试的方法，实质上不能得到序列个数的最大值，判断序列个数已接近最大值的方法是数小时后寻找的序列数仍不发生变化。从表中可以看出，对于长度为 50bp 的序列，如果配对单元的最小得分 S_{unit} 较小(如 $S_{unit} = 6$)，则能生成的序列个数为 47，也较小，而当 S_{unit} 调整为 8，则能生成的序列个数明显增加。

表 1 OligoPool 程序生成寡核苷酸序列个数测试

参数				序列数
长度	GC 含量%	S max	S unit	
12	45 ~ 55	15	6	54
20	45 ~ 55	25	6	75
50	45 ~ 55	62	6	47
50	45 ~ 55	62	8	150

为了检测由本文工具生成的寡核苷酸序列之间的特异性配对情况，采用了 DNA 芯片杂交的方式来进行观测。以参数 Length = 12，Number = 4，GC lower = GC upper = 50%，$S_{max} = 15$，$S_{unit} = 6$ 设计了一组共 6 条序列用于表面芯片杂交反应测试，如表 2。该组序列及其每条序列的互补序列都用于检测，以便确认非特异性的交叉配对和自身配对。以 a、c、e、g、i、k 为程序设计的序列的编号，b、d、f、h、j、l 为各自的互补序列的编号。

表 2　OligoPool 程序生成的一组寡核苷酸序列

编号	OligoPool 程序生成的寡核苷酸序列(5'-3')		左侧序列对应的互补寡核苷酸序列(5'-3')
a	TGAAGCGCGTTA	b	TAACGCGCTTCA
c	CAGACTAGCCTT	d	AAGGCTAGTCTG
e	AAGACGGGAAAC	f	GTTTCCCGTCTT
g	GCAGTATCCACA	h	TGTGGATACTGC
i	AGGAACTGAGCT	j	AGCTCAGTTCCT
k	CTCTGATTCGTC	l	GACGAATCAGAG

将以上 12 条序列在 5'端加 15bp 长的 polyA 尾及 NH_2 后固定在经异硫氰基修饰的玻片上，如下表 3 随机布点形成一个低密度的 24(16 寡核苷酸芯片阵列：

表 3　序列特异性杂交测试芯片点阵排布示意

aceaceaceaceaceaceaceace
gikgilgjkgjlhikhilhjkhjl
acfacfacfacfacfacfacf
gikgilgjkgjlhikhilhjkhjl
adeadeadeadeadeadeade
gikgilgjkgjlhikhilhjkhjl
adfadfadfadfadfadfadf
gikgilgjkgjlhikhilhjkhjl
bcebcebcebcebcebcebce
gikgilgjkgjlhikhilhjkhjl
bcfbcfbcfbcfbcfbcfbcf
gikgilgjkgjlhikhilhjkhjl
bdebdebdebdebdebdebde
gikgilgjkgjlhikhilhjkhjl
bdfbdfbdfbdfbdfbdfbdf
gikgilgjkgjlhikhilhjkhjl

对每一序列进行杂交检测。例如采用 a 序列，在其 5'端连接上荧光探针 FAM，对芯片进行杂交反应，图 5a。该图显示芯片上与 a 完全互补的序列 b 代表的点全部显示荧光，而其他点均未产生荧光，即 a 序列与其他序列未发生非特异性杂交反应。其他序列的检测也得到特异性杂交的结果(图像本文未列出)。

进一步地，对以上采用 OligoPool 程序设计的 6 条探针(a，c，e，g，i，k)同时进行检测，在其 5'端均连接上荧光探针 FAM，对芯片进行杂交反应，得到杂交后荧光图像图 5b。该图与芯片对应寻址，可以看出与 a，c，e，g，i，k 互补的 b，d，f，h，j 点全部显示荧光，表明发生了特异性杂交，其他点全部无荧光，表明没有发生自身非特异性杂交。

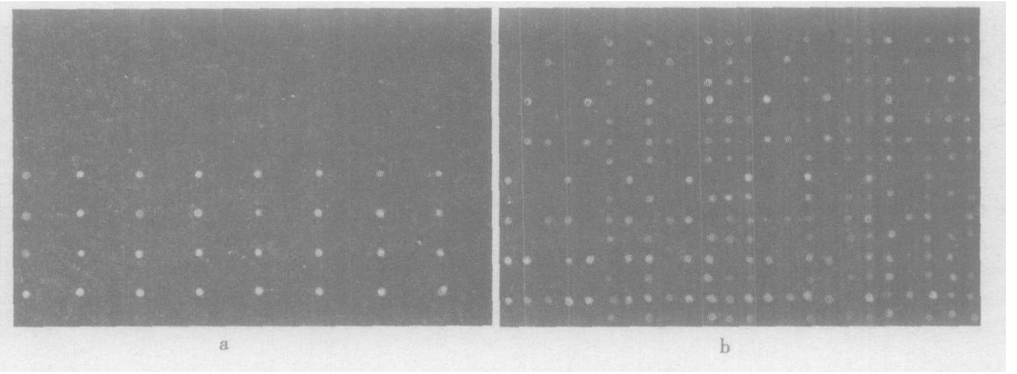

图 5　芯片杂交后荧光图像

Fig.5 The images of the fluorescence hybridization chips

5　结语

由于目前还没有寻找某一集合中最大相容子集的一种有效的算法，本文采用的是随机生成并测试的方法，这样就不能保证能得到相容的最大子集。程序中设定的 Number 参数只是期望得到的寡核苷

酸序列的个数,实际上并不一定能获得这么多的序列。调整最后两个参数可以解决不能获得足够多序列的问题。通常来说,特异性配对的碱基数为整个序列的长度,而非特异性配对只是其中部分碱基的配对,优化杂交反应的条件,则完全配对与部分配对杂交后显示的信号会相差很大,足以区分。因此,如果按某一参数不能获得足够多的序列,可以适当允许多一些非特异性的配对碱基数,提高相容得分的阈值。两条 GC 含量接近 50%的长度为 L 的随机序列,如果有一半碱基配对,它们之间的配对得分约为 $(3\times0.5+2\times0.5)\times L/2=1.25\times L$,相容得分阈值设定为此值以下就基本能满足区分假阳性要求。而对于配对单元最小得分参数,对于长序列,可以适当提高此值。

参考文献(References):

[1] Adleman L M. Molecular computation of solutions to combinatorial problems[J]. Science, 1994, 266(5187):1021 - 1024.

[2] 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.

[3] Gillmor S D, Rugheimer P P, Lagally M G. Computation with DNA on surfaces [J]. Surface Science, 2002, 500:699 - 721.

[4] Liu Q, Wang L, Frutos A G, et al. DNA computing on surfaces[J]. Nature, 2000, 403(6766):175 - 179.

[5] Wu H. An improved surface - based method for DNA computation[J]. Biosystems, 2001, 59:1 - 5.

[6] Frutos A G, Liu Q, Thiel A J, et al. Demonstration of a word design strategy for DNA computing on surfaces[J]. Nucl Acids Res, 1997, 25(23):4748 - 4757.

[7] Tanaka F, Kameda A, Yamamoto M, et al. Design of nucleic acid sequences for DNA computing based on a thermodynamic approach[J]. Nucl Acids Res, 2005, 33(3):903 - 911.

[8] Smith T F, Waterman M S. Identification of Common Molecular Subsequences[J]. J Mol Biol, 1981, 147:195 - 197.

[9] Altschul S F, Madden T L, Schaffer A A, et al. Gapped BLAST and PSI - BLAST: a new generation of protein database search programs [J]. Nucl Acids Res, 1997, 25(17):3389 - 3402.

[10] 肖位枢.图论及其算法[M].北京:航空工业出版社,1993.

(上接第 61 页)

可分成细小病毒亚科和浓核病毒亚科。其中浓核病毒亚科分为 3 个属:浓核病毒属、相同病毒属和简短病毒属。而 BmDNV - Ⅱ和河虾肝胰类似细小病毒作为浓核病毒亚科未分类的暂定种看待[11]。本文根据若干浓核病毒的结构蛋白的分子进化聚类分析表明可得到四大类浓核病毒结构蛋白,其中河虾肝胰类似细小病毒的结构蛋白可单独作为一类存在,BmDNV - Ⅰ与 BmDNV - Ⅱ的结构蛋白可以归为一类,具有同一进化来源,但已有一定的进化距离,说明这两种结构蛋白差异较大,需进一步实验验证。

参考文献(References):

[1] Seki H, Iwashita Y. Histopathological features and pathogenicity of densonucleosis virus (Yamashi isolate) in the silkworm, Bombyx mori, isolated from sericultural farms in Yamashi prefecture[J]. J Seric Sci Jpn, 1983, 52:400 - 405.

[2] Watanabe H, Kawase S, Shimizu T, et al. Difference in serological characteristics of densonucleosis viruses of in the silkworm, Bombyx mori[J]. J Seric Sci Jpn, 1986, 55:75 - 76.

[3] Li Y, Zadori Z, Bando H et al. Genome organization of the densovirus from Bombyx mori (BmDNV - 1) and enzyme activity of its capsid[J]. J GenVirol, 2001, 82: 2821 - 2825.

[4] Zadori Z, Szelei J, Lacoste M, et al. A viral phospholipase A2 is required for parvovirus infectivity[J]. Dev cell, 2000, 1:291 - 302.

[5] 许光治,郭锡杰,等.家蚕浓核病毒(镇江)株主要结构蛋白基因的克隆及表达[J].病毒学报,2004,20(3):279 - 282.

[6] Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M. R., Appel R. D., Bairoch A. *Protein Identification and Analysis Tools on the ExPASy Server* [A]; (In) John M. Walker (ed): The Proteomics Protocols Handbook: Humana Press, 2005, 571 - 607.

[7] Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences[J]. Proteins, 2000, 41(2):224 - 237.

[8] Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P. SMART 4.0: towards genomic data integration[J]. Nucleic Acids Res, 2004, 32(1): D142 - 4.

[9] Jaime Prilusky, Clifford E. Felder, et al FoldIndex(c): a simple tool to predict whether a given protein sequence is intrinsically unfolded [J]. Bioinformatics, 2005, 312 - 331.

[10] Ohno S. Repeats of base oligomers as the primordial coding sequences of the premeval earth and their vestiges in modern genes[J]. J Mol Evol, 1984, 20:313 - 321.

[11] Van Gegenmontel MHV, Fauquet C M, et al. Virus taxonomy classification and nomenclature of viruses (The seventh reports of the international committee on Taxonomy of viruses)[A]. San Diego: Academic Press, 2000, 1653 - 1724.