基因表达芯片探针选择的一种有效方法

倪青山, 王正志, 李冬冬

(国防科技大学 机电工程与自动化学院,中国湖南 长沙 410073)

摘 要:为了减少基因表达芯片探针选择的计算量,采用最大共有序列准则大大减少了需要计算的候选探针数量,并通过不断扩大进行计算的探针数目保证探针选择的效果. 结果表明该方法是十分有效的,探针选择的结果也是令人满意的.

关键词:基因芯片;探针选择;基因表达;最大共有序列

中图分类号: TP391.7 文献标识码:A 文章编号:1007-7847(2004)04-0371-06

An Effective Method of Selecting Probes for Gene Expression Arrays

NI Qing-shan, WANG Zheng-zhi, LI Dong-dong

(*College of Mechantronics Engineering and Automation, National University of Defense Technology, Changsha* 410073, *Hunan, China*)

Abstract: To reduce the calculation in the process of probe selection for gene expression arrays, the amount of candidate probes is reduced with the longest common sequence rule. In the same time, the effect of probe design is guaranteed by continuously increasing the probe number calculated. Applying to test data, it shows that the method is effective, and the result of probes selected is satisfied.

Key words: microarray; probe selection; gene expression; longest common sequence

(*Life Science Research*, 2004, 8(4): 371 ~ 376)

基因芯片的出现使人们能够同时对大量的基因序列进行分析,研究基因在不同的生物体、不同的组织以及生物体不同的发育阶段的表达差异,从而对基因的结构、功能进行深入地研究. 要对基因的表达水平进行检测,芯片上的探针必须能够在特定的杂交条件(主要指实验温度)下将目的基因和非目的基因区分开.

一般的,特异性探针与目的基因的杂交应该满足 Watson-Crick 准则,也就是说探针序列应该是目的基因某段序列的互补序列,但是由于探针长度较短,一般包含的碱基数目在 100 以内,而基因序列包含的碱基数目可达上万个甚至更多,这使得候选探针的数量相当的大,又由于基因序列之间具有一定的相似性,符合条件的探针数量相对较少,因此探针的选择是比较困难的. 探针如果选择得不恰当,就会和非目的基因杂交,产生错误的实验结果. 目前大多数探针设计的方法都是基于探针与基因的序列信息的,这些方法能够较快的找出探针集,但是仅仅利用序列之间的关系来描述探针的杂交行为是很不精确的,由于芯片

收稿日期:2004-06-14;修回日期:2004-09-03

基金项目:军队基础研究项目(JC-02-03-021)

作者简介:倪青山(1979-),男,黑龙江人,硕士研究生,从事生物信息学、生物芯片技术研究,Tel:+86-0731-4574991.

实验是在特定的温度下进行的，从而芯片上的探针应该具有一致的杂交温度，这使得基于序列信息的方法在实际实验中有时并不实用，因此 Lars Kaderali 等[1]利用热力学比对的方法，通过计算候选探针与非目的基因之间的最大杂交温度来确定特异性的探针集，但是由于该方法对所有符合条件的候选探针都进行了热力学比对计算，探针设计所需的时间较长 .

综合以上两种方法的优点，我们提出了一种有效的探针选择方法，首先利用最大共有序列准则估计候选探针的杂交特性，然后选择杂交特性好的探针作为新的候选探针，通过这个过程将进行比对的序列数量大幅度的减少，接着对新的探针集进行热力学比对，进而确定最终的探针集 . 测试结果表明最大共有序列准则能够很好地反映探针的杂交特性，用该方法选择探针的结果是令人满意的 .

1 问题描述

两个碱基序列的杂交行为可由它们的杂交解链温度表示，杂交解链温度越高，形成的双链结构越稳定 . 可以简单的认为当实验温度大于杂交解链温度时，这两个碱基序列不杂交，处于分离的状态；反之，这两个碱基序列杂交，形成稳定的双链结构. 两个碱基序列的最大杂交解链温度可以利用 NN 模型(Nearest Neighbor Model)，通过热力学比对[1]的方法计算得到，我们将在后面详细介绍.

定义 1：匹配温度　探针与目的基因的杂交解链温度，该温度就是探针与目的基因在满足 Watson-Crick 准则时的杂交解链温度 .

定义 2：误配温度　探针与所有非目的基因的最大杂交解链温度的最大值就是误配温度，也就是说在此温度下，探针与至少一种非目的基因进行了杂交 .

在芯片杂交完成后，要对芯片进行洗涤，除去未杂交的样品序列，为了保证目的基因能够依附在指定的探针上，芯片上所有探针的匹配温度与误配温度必须有一定的差异；由于芯片上的探针是在相同的条件下进行杂交的，因此还必须保证杂交行为的一致性 . 如图 1 所示，在杂交温度为 T 的情况下，芯片上的任意探针 P_i 与目标序列的 S_i 杂交温度均应落在 H 区，而与非目标序列的杂交温度应该落在 L 区，其中 ΔT 是设定的温度变化范围，用于补偿计算杂交温度 T 时产生的误差和实验的温度误差，只有这样才能在杂交温度为 T 时，保证探针与目标序列充分的杂交，而避免探针与非目标序列的误杂交 .

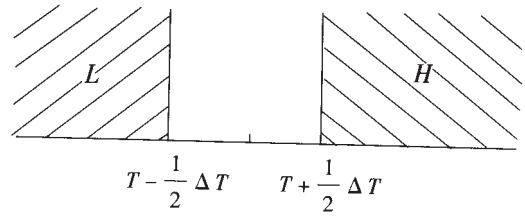

图 1　杂交温度

Fig. 1　Melting temperature

因此，探针的选择问题可以描述为：

给定 n 个基因序列 $S_1, S_2, S_3, \cdots S_n$ 和温度误差 ΔT，找到一个温度 T 和 n 个探针 $P_1, P_2, P_3, \cdots, P_n$，其中 P_i 的互补序列是 S_i 的子序列，要求()成立

$$\begin{cases} T_{\max}(P_i, S_i) > T+\frac{1}{2}\Delta T \\ T_{\max}(P_i, S_j) < T-\frac{1}{2}\Delta T \end{cases} \quad i=1,2,3,\cdots,n \quad j=1,2,3,\cdots,n \quad j\neq i \tag{1}$$

其中 $T_{\max}(a, b)$ 表示序列 a 与序列 b 的最大杂交温度 .

1.1 最大杂交温度的计算

当两个序列的杂交形式确定后，其杂交解链温度可由 NN 模型 (Nearest Neighbor Model) 计算得到，计算公式如下：

$$T_m = \frac{\Delta H}{\Delta S + R\ln(C_T/4)} - 273.15 \tag{2}$$

其中，ΔH、ΔS 分别是杂交反应中焓和熵的变化量，可由文献[2]得到，$R = 1.987$ cal/k·mol 为玻尔兹曼 (Boltzmann) 常数，C_T 为参加反应的序列的摩尔浓度，其值是未知的，Li 等[3]建议其取值为 1×10^{6} mol/L.

当探针序列与基因序列进行杂交时，由于探针序列较短而基因序列较长，探针就可以在基因序列的不同位置上与基因序列进行杂交，如图 2 所示，序列 ATAG 与序列 GCTGTCATCCA 的两种

杂交形式,(1)中引入了一个错配,而(2)中引入了一个空位.

图 2　不同的杂交形式

Fig. 2　Different types of hybridizing

杂交形式的不同其杂交解链温度一般来说也是不同的,如图2(1)中两序列的杂交温度为 – 118.05 ℃,图2(2)中两序列的杂交温度为 – 84.82 ℃. 如果将所有可能的杂交情况都进行计算,大约要计算 C_{mn}^{n} 种情况 (m 为基因序列的长度,n 为探针序列的长度),其计算量是相当大的,由于我们要计算的是两条序列的最大杂交温度,并不需要将所有的情况计算出来,因此利用 Lars Kaderali 等[1]的方法对探针和基因之间的最大温度进行计算. 该方法借鉴了动态规划的思想,虽然不能严格的保证算得的温度是最优解,但是 Lars Kaderali 等[1]证明了该方法的可行性.

1.2　最大共有序列

定义 3:最大共有序列　序列 a、b 的共有序列是指序列 a、b 的公共子序列,也就是说共有序列既是 a 的子序列又是 b 的子序列,最大共有序列是指序列 a、b 的所有共有序列中长度最大的序列.

可以看出两个序列的最大共有序列有时并不是唯一的,由于我们只关心最大共有序列的长度,而长度必定是唯一的,我们用 $Lcs(a, b)$ 表示序列 a、b 的最大共有序列的长度. 例如图 3(1) 中两个序列的最大共有序列长度 $Lcs(a, b) = 3$.

图 3　共有序列 (共有序列长度为 1 的未画出)

Fig. 3　Common sequence (common sequence with length 1 isn't shown)

如果序列 b 在 i 位置的一个序列片断为 a,那么在 b 中删除 a 后,会得到两个序列 $b_{0,\ i-1}$ 和 $b_{i+|a|,\ |b|+1}$ ($b_{i,\ j}$ 表示序列中从 i 到 j 的一段子序列,$|a|$, $|b|$ 分别表示序列 a、b 的长度),此时 a、b 的最大共有序列长度 $Lcs(a, b) = \max(Lcs(a, b_{0,\ i-1}), Lcs(a, b_{i+|a|,\ |b|-1}))$. 例如图3 (1)中的序列 a、b,其最大共有序列的长度 $Lcs(a, b) = \max(Lcs(a, b_{0,3}), Lcs(a, b_{8,11})) = 3$.

对 HIV-1 subtypes 的测试表明探针序列 a 与基因序列 b 的最大共有序列的长度 $Lcs(a, b)$ 较小时,探针序列的匹配温度与误配温度的差也相应的较低,这个差值越大,说明探针区分目的基因和非目的基因的能力越好,我们将依此作为对探针进行筛选的准则,即优先选择那些具有最短的最大共有序列的序列作为候选探针.

2　算法描述

基因表达芯片探针选择的流程如图 4 所示.

1)Suffix Array 数据结构[4]是一个序列的所有后缀子序列的有序排列,利用该数据结构很容易算出探针序列的最大共有序列长度.

2)候选探针的前处理

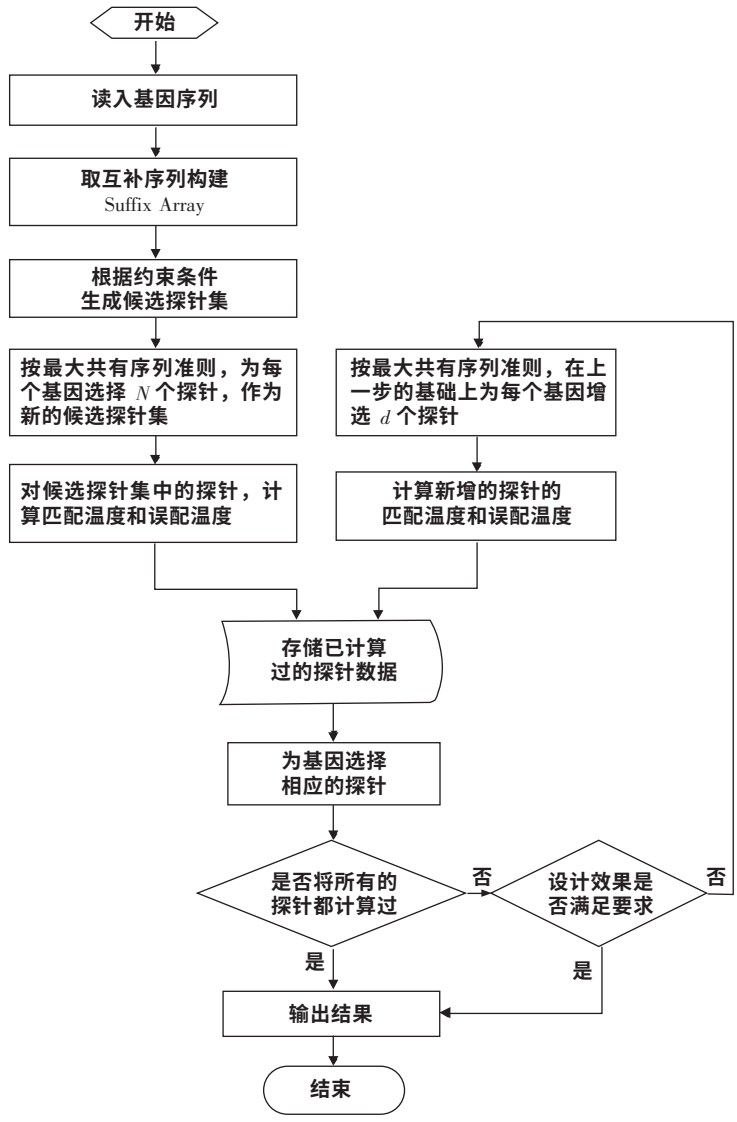

图 4 探针选择流程

Fig. 4 Flow chart for probe selected

探针首先要满足一些约束条件[5]，我们将不能满足以下条件的探针序列从探针集中删除掉.

① 探针序列在整个编码序列中必须是唯一的；

② 探针序列只能包含 A、C、G、T 4 种碱基；

③ 探针序列中同一种核苷酸（A、C、G、T）的数量不能超过序列总数的 50%；

④ 任何连续的 A、T 或 C、G 的数量不能超过数列总数的 25%；

⑤ G、C 总含量应该占序列总数的 40% ~ 60%，这个含量也可以根据基因组总的（$G + C$）的含量进行相应的调整；

⑥ 探针序列的任一长度为 15 的子序列在整个基因组中必须是唯一的；

⑦ 探针序列不能自杂交，即探针序列中互补片断的长度不能超过探针长度的 30%.

另外，根据实验的要求，探针的匹配杂交温度应大于给定的温度 Tc，上述规则也并不是绝对的，一般还要根据具体的实验情况和特定的生物样本进行相应的调整.

3）计算误配温度

探针与目的基因的杂交是满足 Watson-Crick 准则的，因而很容易利用 NN 模型计算出来. 但是探针与非目的基因的杂交由于并不满足 Wat-

son-Crick 准则，杂交位置和杂交的形式都是不确定的,需要计算杂交结构最稳定时的温度,即最大杂交温度. Lars Kaderali 等[1]利用启发式动态规划方法很好地解决了该问题，其主要思想借鉴了双序列比对的动态规划方法,首先构造比对表,在表的每个位置选择使杂交温度局部最优的杂交形式，整个表填完后便能得到最大的杂交温度. 从式 (2)可以看出杂交温度计算时,焓和熵的变化量分别处于除式的分子和分母的位置上，并不是它们的线性组合,因此,由局部最优推得的全局最优是近似的，也就是说，找到的并不一定是最优解，但是 Lars Kaderali 等[1]证明了该方法的有效性，况且我们在特异性探针的评价时还加入了温度误差 ΔT 见式 1).

在探针与非目的基因进行热力学比对计算的过程中，我们仿照 Lars Kaderali 等[1]的方法利用 Suffix Array 数据结构对计算量进行了削减. 基本思想是当两个探针序列有公共的前缀时，其中一个探针序列与非目的基因序列的比对表计算完成后,在计算另一个探针与该基因序列的比对表时,其公共的前缀序列就不必重新进行计算了，由于 Suffix Array 中存储着探针的有序排列，因此只需按此顺序对探针进行计算就能利用上述的思想来减少计算量.

4)通过计算探针与所有非目的基因的杂交温度,我们很容易得到探针的误配温度,利用 (2)式我们可以得到探针的匹配温度,接下来根据 (1)式的规则,将 T 值从探针的最高匹配温度开始逐渐降低,计算每个温度下探针的选择情况,找到最佳的一个作为实验温度,为每个基因选取合适的探针.

5)当每个基因上选择多少个候选探针时才能得到满意的设计效果呢？我们可以首先人为的估计一个值,如果在该值下已经得到了满意的效果，则探针设计完毕,若不能得到满意的效果,则按一定值扩大基因上候选探针的数目，由于前一次选择的探针一定在后一次选择的探针中，因此只计算新选入的探针即可,然后再计算实验温度,为基因选择合适的探针,若结果满意则结束,否则继续下去，直至得到满意的设计结果或全部的探针都已经计算过为止. 由于计算时间主要集中在探针的误配温度计算上,其它计算均可忽略不计,因此很容易看出我们的方法的计算量一定比 Lars Kaderali 等[1]的低,而探针的选择效果不会比 Lars Kaderali 等[1]的差.

3　结果分析

为了验证算法的有效性，我们对 HIV-1 subtype 进行了测试，数据来源于 Los Alamos 国家实验室[6]，该基因组中含有 58 个基因，每个基因约含碱基 9 000 个.

由于探针的匹配温度与误配温度的差值能够很好地反映探针的特异性，因此我们研究探针的最大共有序列长度与匹配温度和误配温度的差值之间的关系. 在探针长度为 20 时,按照最大共有序列的长度值任意选取了 12 组探针,每组 100 个探针，各组探针的最大共有序列的长度 L 分别为 8,9,10,…,19. 温度差值 dT 取平均数,探针的最大共有序列长度与温度的差值之间的关系如图 5 所示：

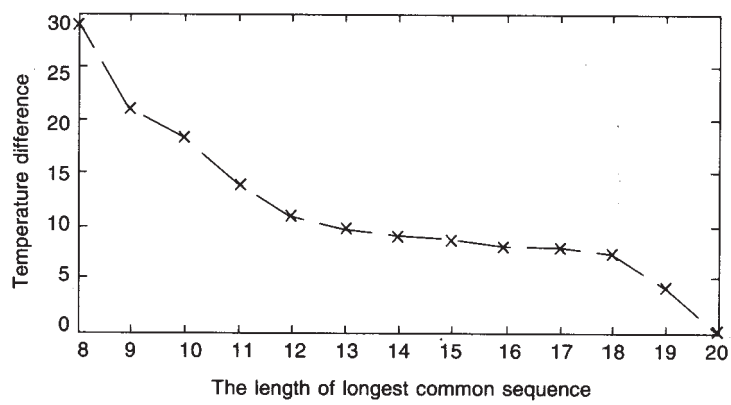

图 5　探针的最大共有序列长度与温度差值之间的关系

Fig. 5　The relation between the longest common sequence of probe and temperature difference

在 HIV-1 subtype 中序列的重复性很高，任意长度为 20 的候选探针序列的最大共有序列的长度值最小为 8，因此图 5 中横坐标是从 8 开始的，当最大共有序列的长度为 20 时，说明该序列在基因组中出现了重复，也就是该探针与非目的基因发生了匹配杂交，因此其温度差为零．从图中可以看出随着最大共有序列长度的增加，温度差的值逐渐减小，也就是说探针的区分目的基因与非目的基因的能力逐渐减弱，探针具有特异性的可能性逐渐减小．这说明了最大共有序列准则的有效性．

每个基因上选取的候选探针数目对最终的探针选择结果有很大的影响，即当每个基因上选取的候选探针数目较小时，有可能不能为所有的基因找到符合条件的探针．如果我们把探针设计效果表示为选出探针的基因的数目与基因总数的比值，记为：

$$\eta = \frac{\text{选出探针的基因数目}}{\text{基因总数目}}$$

对 HIV-1 subtype，我们在每个基因上选取的探针数目分别取为 5，10，15，…，100，探针的长度为 20，温度误差为 20 ℃，最低探针匹配温度为 70 ℃，进行实验．每个基因上选取的候选探针数目 N 与探针设计效果的关系如图 6.

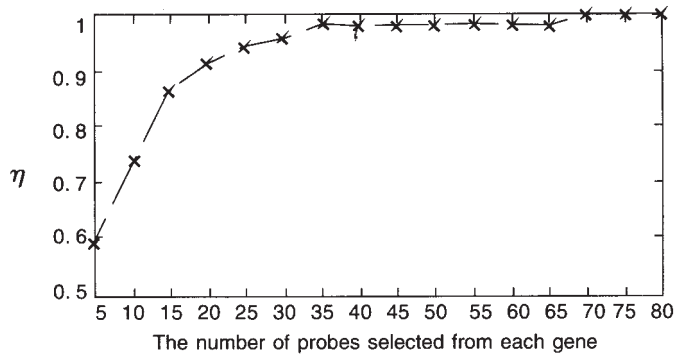

图 6 每个基因上探针选择的数目与探针设计效果的关系

Fig. 6 The relation between the number select from each gene and the effect of design

图 6 表明，随着每个基因上候选探针数目的增多，探针选择的结果越来越好．在 5 ~ 30 阶段，探针设计效果变化明显，而 30 ~ 80 变化缓慢，这也说明了我们用最大共有序列准则的有效性．从图中我们还可以看出，当每个基因上候选探针数目为 70 时，探针的设计效果已经为 1，即已经为所有的基因选择了合适的探针．

在用探针约束条件完成第一次筛选后，探针集中候选探针的数量大约为 80 000 个，而我们利用最大共有序列准则后，仅对 $70 \times 58 = 4\ 060$ 个探针进行温度计算，计算的探针数量减少了将近 20 倍，大大的减少了计算量，缩短了计算时间．

4 结论

由于基因表达片主要用于对特定基因的表达水平进行定性或定量的检测，其探针选择的优劣将直接影响实验的结果，选择的探针不但要满足特异性的要求，还要使芯片上探针的杂交行为具有一致性，其计算是相当复杂的，本文利用最大共有序列准则大大减少了进行计算的候选探针的数目，得到了满意的设计结果．

由于算法中计算量主要集中在误配温度的计算上，每个探针的计算都是可以独立进行的，因而，可以将程序改成并行计算的结构，在多台计算机上同时计算，进一步缩短计算时间．

参考文献（References）：

[1] KADERALI L, SCHLIEP A. Selecting signature oligonucleotides to identify organisms using DNA arrays[J]. Bioinformatics, 2002, 18: 1340-1349.

[2] KADERALI L. Selecting target specific probes for DNA arrays [D]. Germany: University of Cologne, 2001. 88-90.

[3] LI F, STORMO G. Selection of optimal DNA oligos for gene expression arrays[J]. Bioinformatics, 2001, 17: 1067-1076.

[4] MANBER U, MYERS E W. Suffix arrays: a new method for on-line string searches[J]. SIAM J Comput, 1993, 22: 935-948.

[5] LOCKHART D J, DONG H, BYRNE M C, *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays[J]. Nat Biotechnol, 1996, 14: 1675-1680.

[6] LANL, HIV sequence database - 1999 HIV-1 subtype reference alignments[Z]. Los Alamos National Laboratories, 1999.