浙江理工大学学报，第 29 卷，第 4 期，2012 年 7 月
Journal of Zhejiang Sci-Tech University
Vol. 29, No. 4, Jul. 2012

文章编号：1673-3851 (2012) 04-0599-05

影响 RNA 芯片探针杂交效率的因素分析

马琴琴，严国权，郭江峰，丁先锋
（浙江理工大学生命科学学院生物工程研究所，杭州 310018）

摘　要： RNA 芯片作为研究 ncRNA 的主要技术之一，已受到广泛关注。然而因 RNA 芯片中靶序列的单链特征，使得 RNA 芯片的探针设计比 DNA 复杂。本研究对加强型绿色荧光蛋白（EGFP）RNA 序列进行覆瓦式探针设计，制作了 RNA 原位合成芯片，研究探针/靶序列杂交双链 ΔG 和 T_m、靶序列二级结构以及探针的末端碱基对 RNA 芯片杂交信号强度的影响。结果表明，探针/靶序列杂交双链 ΔG 和 T_m 相同的探针间的杂交信号存在较大差异，探针/靶序列杂交双链 ΔG 和 T_m 与 RNA 芯片杂交信号之间未发现明显的规律性；靶序列二级结构的探针 PT-sc 参数与芯片杂交信号强度间呈较好的线性关系；探针 5'末端碱基组成对芯片杂交信号强度也有影响，5'末端碱基为 G/C 的探针杂交信号总体上高于其邻近的 5'末端碱基为 A/T 的探针的杂交信号，为 RNA 芯片的探针筛选提供一定的理论基础。

关键词： 覆瓦式探针；RNA 芯片；二级结构；探针/靶序列杂交双链

中图分类号： Q789　　**文献标识码：** A

0　引　言

在分子生物学研究中，生物芯片以其高通量的优点在研究基因的表达模式、绘制基因图谱、筛选特定靶基因和检测染色体拷贝数变化等研究领域有广泛的应用[1-2]。芯片数据的正确解读必须依赖于正确的设计方法，而高特异性和高灵敏度探针的选择，对于芯片样品的精确定量和芯片数据分析是至关重要的[3-5]，目前有很多程序可用于生物芯片探针的设计[6-7]。高质量的探针设计要求在复杂的体系里对芯片的探针杂交情况进行预测[8]；探针的特异性可以通过 BLAST 工具[9]来有效地检测；对低复杂度序列进行过滤，可以避免探针产生非特异性杂交信号；对探针的回文结构进行过滤[10]，可以提高探针和靶序列的杂交效率。Ratushna 等[11]证实靶序列形成的稳定双链二级结构会干扰探针与其相互作用。除探针自身因素外，研究人员还想出各种方法来优化探针的设计。在其他条件相同的前提下，选择 T_m（溶解温度）相近的探针组可以进一步加强芯片对靶序列浓度的区分能力[12]；Affymetrix 公司采用独特的 PM-MM 探针对设计以获得在复杂样品背景中检测的高灵敏度和特异性之间的优化平衡[13]。T_m 是影响核酸杂交的重要因素，探针 T_m 的微小改变可以导致芯片的信号值强度发生较大变化，T_m 较高的探针会产生相对较高的芯片信号值[5]；探针与靶序列形成的杂交双链能量（ΔG）也可以影响芯片信号值强度[14]。

目前关于芯片探针影响因素的研究主要是基于以 DNA 为靶序列的 DNA 芯片实验或者是基于小片段 RNA（如 miRNA）的 RNA 芯片实验，而 Non-coding RNA [15]的发现使得长片段 RNA 芯片[16]被人们广泛关注。由于长片段 RNA 芯片的靶序列呈单链状态，可以形成较稳定的二级结构[11]，会干扰探针与靶序列相互识别，因此 RNA 芯片的探针设计较 DNA 芯片更为复杂。本文通过加强型绿色荧光蛋白（EGFP）的 RNA 覆瓦式探针芯片实验数据

收稿日期：2011－11－22
基金项目：国家高技术研究发展计划项目（863 计划，2007AA02Z165）；广东省教育部产学研结合项目（2011B090400478）；浙江省自然科学基金项目（Y2100681）；杭州市科技发展计划项目（20110733Q04，20110733Q21）
作者简介：马琴琴（1986－），女，湖北荆州人，硕士研究生，从事生物化学与分子生物学研究。
通讯作者：丁先锋，E-mail：xfding@zstu.edu.cn

的分析，研究探针/靶序列杂交双链 ΔG 和 T_m、靶序列二级结构以及探针末端碱基对 RNA 芯片杂交信号强度的影响，为 RNA 芯片的探针筛选提供一定的研究基础。

1 材料与方法

1.1 材料

EGFP(加强型绿色荧光蛋白，gb|EU048697.1|，核苷酸序列 4448—5166)为本实验室保存样本。

1.2 方法

1.2.1 芯片探针设计及芯片合成

为明确靶序列二级结构对探针的影响，对 EGFP(719 bp)进行了覆瓦式探针设计。芯片上每个探针长度为 19 个碱基，且相对于其邻近的探针改变 1 个碱基。共含有 701 条探针，由 LC-Science 公司进行原位芯片合成。

1.2.2 靶序列制备及芯片杂交

以含有 EGFP 的质粒(载体为 pGEM-T Easy)为模板进行 PCR(引物序列为 T7-1F：TAATACGACTCACTATAGG；6*Bbs*I-2F：TGCAGTACGGAAGACCTCTCC)，扩增产物利用 AxyPrep™ DNA Gel Extrcation Kit (Axygen)进行目的片段回收。回收的目的片段 3'端含 *Bbs* Ⅰ限制性酶切位点的引物区序列，利用 *Bbs* Ⅰ(NEB)限制性酶切除该端引物区序列，酶切产物利用 AxyPrep™ Clean-up Kit (Axygen)进行快速 PCR 产物纯化回收，再将获得的 5'端含有 T7 启动子的 EGFP 序列利用体外转录试剂盒(Promega)进行体外转录，获得 EGFP 的 RNA 样本。最后将 EGFP 的 RNA 样本交由 LC-Science 公司进行标记及芯片杂交，LC-Science 公司返回的芯片数据用于实验数据分析。

1.2.3 靶序列(EGFP)二级结构预测及其探针参数计算

采用 Zuker 等编写的在线生物信息学软件“RNA mfold (version 3.2) web server”[17-18] (http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi)对 EGFP 序列进行 RNA 二级结构预测，使用缺省参数得到 34 个预测的 EGFP RNA 二级结构。由于 RNA 除了最低自由能结构外，还存在着大量的次优折叠结构[18-20]，而且 RNA 二级结构的预测也不是完全准确，所以我们构建并计算了每条探针的靶序列二级结构参数 PT-sc(Probe-Target-ss-count)，以分析 RNA 二级结构对探针信号强度的影响，其计算的方法为：

$$\text{探针 } PT\text{-}sc = \sum_{i=1}^{n} \frac{\text{探针互补靶序列碱基 } ss\text{-}count}{\text{靶序列预测的二级结构数} \times n} \times 100$$

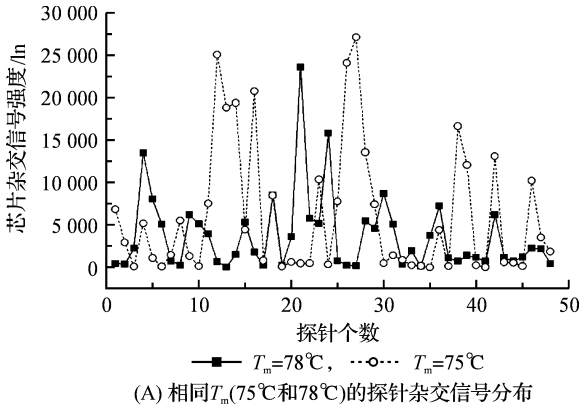
其中 n 为探针碱基长度(本研究中 $n=19$)，靶序列预测的二级结构数为在缺省条件下 mfold 程序给出的靶序列二级结构数(本研究中 EGFP 的二级结构数为 34)，探针互补靶序列碱基 ss-count 来自于 mfold 程序给出的 ss-count 文件，由于靶序列碱基 ss-count 代表此碱基在所有预测的二级结构中以单链状态存在的次数，所以靶序列探针的 PT-ss-count 参数可以综合 mfold 预测的所有靶序列二级结构信息，较为准确地反映与此探针互补的靶序列对应位置的二级结构状况。同时，使用在线的生物学分析工具“The DINAMelt Server：Prediction of Melting Profiles for Nucleic Acids”(http://dinamelt.bioinfo.rpi.edu/twostate.php)预测探针(DNA)/靶序列(EGFP RNA)杂交双链 ΔG 及其 T_m 值。

2 结果与分析

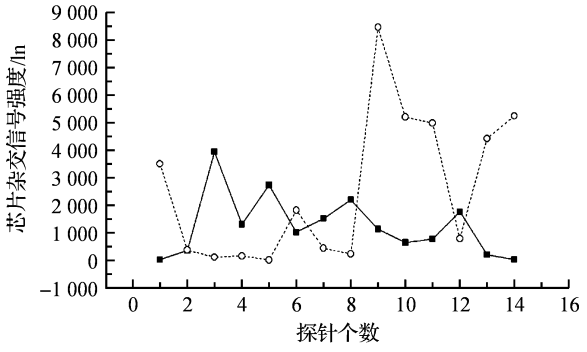
2.1 探针(DNA)/靶序列(EGFP RNA)杂交双链 ΔG 及其 T_m 值分析

传统的 DNA 探针设计方法通常要设定 T_m 值范围，以便使不同靶序列的探针具有相近的解链温度，来平衡不同靶序列的芯片数据。为研究该探针设计原则在全长 RNA 芯片实验中的作用，笔者预测了所有覆瓦式探针的 T_m，同时也预测了每条 DNA 探针与其 RNA 靶序列(EGFP)的杂交双链 ΔG，然后同每条探针的芯片实验信号进行比较。

图 1 A 给出了探针在相同 T_m 值(75℃和 78℃)下的杂交信号强度分布。从图中可以看出，在相同的探针长度(19 bp)、相同的靶序列浓度(1 ng/μL)和相同的杂交条件下(32℃)，同一 T_m 的探针信号强度变化很大，甚至可以达到 3 个数量级的差距。图 1B 显示出在相同的杂交双链 ΔG 条件下，探针的杂交信号强度变化也较明显。图 2 分别给出了 EGFP覆瓦式探针 T_m 值、杂交双链 ΔG 对杂交信号强度的散点图($r^2=0.048$ 和 $r^2=0.047$)，从图 2 中可以看出，在本实验条件下探针的 T_m 和其杂交双链的 ΔG 与杂交信号强度间不存在线性关系。杂交信号较高的探针，其 T_m 和杂交双链 ΔG 未发现明显的规律性，说明在全长 RNA 芯片的探针筛选过程中，限制探针的 T_m 或是杂交双链 ΔG 的范围可能会过滤掉一些高灵敏度的探针，同时也暗示了可能还存在其他影响探针杂交信号强度的因素。

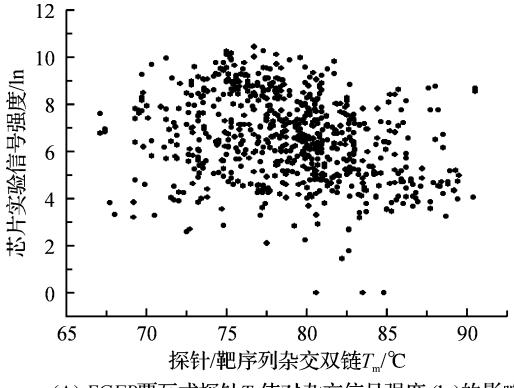


(A) 相同T_m(75℃和78℃)的探针杂交信号分布

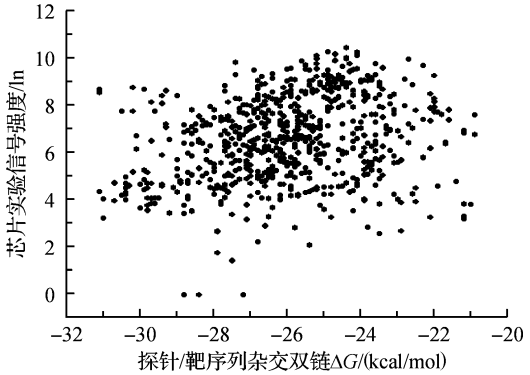


(B) 相同ΔG(−25.70 kcal/mol和−23.90 kcal/mol)的探针杂交信号分布

图 1　相同 T_m 值和杂交双链 ΔG 的 EGFP RNA 探针杂交信号分布



(A) *EGFP*覆瓦式探针T_m值对杂交信号强度(ln)的影响



(B) 杂交双链ΔG对杂交信号强度(ln)的影响

图 2　EGFP 覆瓦式探针 T_m 值、杂交双链 ΔG 对杂交信号强度(ln)的影响

2.2　探针 PT-sc 参数对杂交信号的影响

EGFP 的 RNA 覆瓦式探针的杂交信号强度(ln)与其对应的 PT-sc 关系如图 3 所示,从图 3 可以看出,701 条探针的杂交信号强度变化很大,且芯片杂交信号强度的变化趋势与探针的 PT-sc 的变化趋势相似,两条曲线变化几近平行,即 PT-sc 越高,芯片杂交信号强度越高,而 PT-sc 越低,芯片杂交信号强度越低。图 4 给出了探针 PT-sc 与杂交信号强度(ln)之间的散点图(r^2=0.349),图 4 显示探针的杂交信号强度随其 PT-sc 的升高而增大,当 PT-sc 大于 60 时,探针有稳定的较高杂交信号。

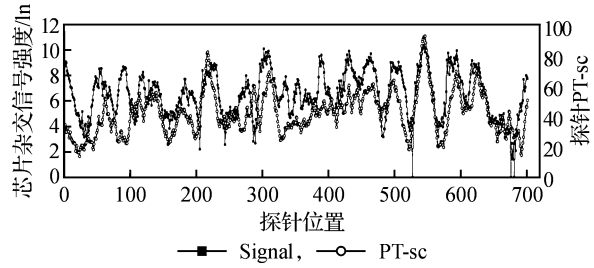


图 3　EGFP 覆瓦式探针的 PT-sc 与杂交信号强度的关系

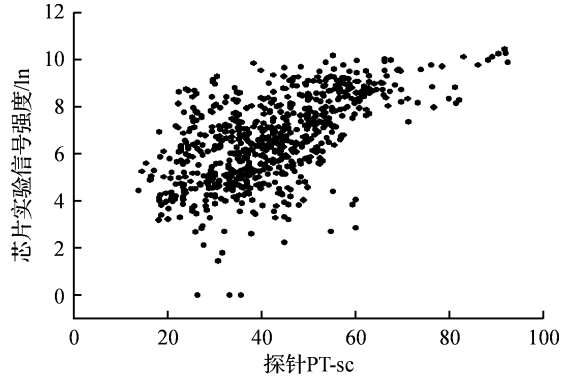


图 4　探针 PT-sc 对实验信号强度的影响

为进一步分析探针 PT-sc 对杂交信号的影响,将所有 701 条探针的信号强度按从低到高分为 7 组(除最后一组含有 101 条探针,每组均含有 100 条探针),然后分别取每组探针的平均杂交信号强度和平均 PT-sc 进行作图,结果如图 5 所示,从图 5 可以看出,探针的 PT-sc 与其杂交信号强度有很强的线性相关性(r^2=0.935)。芯片实验杂交信号强度产生的基础是探针和其靶序列形成异源双链结构,当单链的靶序列 RNA 自身折叠形成稳定的二级结构时,就会干扰探针和其靶序列分子间的相互作用。当探针的靶序列对应位置碱基已经和其自身的其他位置碱基形成双链结构时,探针就很难再与其相互杂交识别,产生较低的杂交信号。而当探针的靶序列对应位置碱基形成稳定的未配对的茎环结构时,其探针就很容易与其形成异源双链结构,从而产生

较强的杂交信号。探针的 PT-sc 值可以较准确地反映靶序列二级结构状况，能在一定程度上反映探针灵敏度的高低。

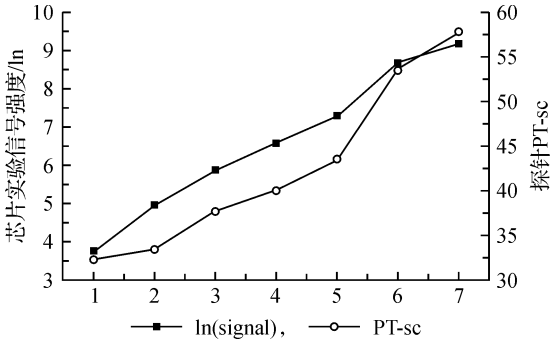


图 5 探针 PT-sc 对芯片平均杂交信号强度的影响

2.3 探针 5'末端碱基组成对芯片杂交信号的影响

图 6 是对 5'末端碱基为 G/C 和 A/T 的不同探针信号强度的统计分析，为比较探针末端碱基对杂交信号的影响，选择考察的探针对象为相邻的两个 5'末端碱基分别为 G/C 和 A/T 的探针，然后用 5'末端碱基为 G/C 的探针的杂交信号强度减去其相邻的 5'末端碱基为 A/T 的探针的杂交信号强度。从图 6 中可以看出，探针的 5'末端碱基的组成对探针杂交信号也有较大影响，5'末端碱基为 G/C 的探针的杂交信号强度总体上明显高于其邻近的 5'末端碱基为 A/T 的探针的杂交信号强度，这与 GC 配对比 AT 更加稳定相符合，从而使探针与靶序列形成的异源双链结构更加稳定。

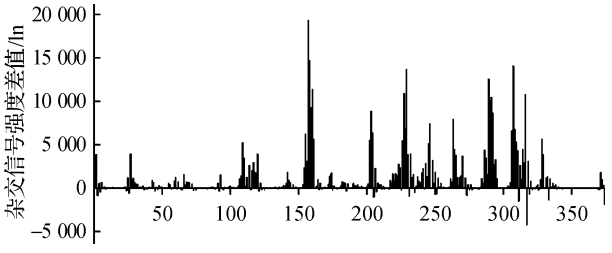


图 6 探针的 5'末端碱基对探针信号强度的影响

3 讨 论

在生物芯片技术体系中，关键步骤之一是为靶序列筛选高特异性、高灵敏度的探针。探针的特异性可以通过 Blast 程序很好地解决，但在特异性的基础上选择高灵敏度的探针却是芯片研究设计工作的一个难题。

本研究显示，相同探针/靶序列杂交双链 ΔG 和 T_m 探针间的杂交信号强度存在较大差异，探针/靶序列杂交双链 ΔG 和 T_m 与 RNA 芯片杂交信号之间未发现明显的规律性，因此在 RNA 芯片的探针设计过程中不宜将这个两个因素作为主要的筛选因素。探针的 PT-sc 和杂交信号强度之间存在一定的正相关性，PT-sc 大于 60 的探针有较高的杂交信号，说明靶序列的二级结构对 RNA 芯片杂交信号有较大影响，靶序列二级结构上的非双链结构区域更加容易完成同其探针的杂交识别过程，从而使其探针具有更高的杂交信号。探针 5'末端碱基对芯片杂交信号也有影响，5'末端碱基为 G/C 的探针杂交信号强度总体上高于其邻近的 5'末端碱基为 A/T 的探针的杂交信号强度。在 RNA 芯片探针的设计过程中，选择高 PT-sc 和 5'末端碱基为 G/C 的探针，可以在一定程度上提高探针的灵敏度。

虽然通过对 EGFP RNA 覆瓦式探针芯片的数据分析，得出了部分影响 RNA 芯片探针杂交信号强度的相关因素，但还不能较准确地预测探针的杂交信号变化。其原因可能是靶序列二级结构及其相关参数的预测与实际情况存在着一定的偏差，从而影响了对探针的杂交信号强度的分析。除以上提及的一些探针影响因素外，探针的长度、每条靶序列探针的个数以及 RNA 的三维结构的存在等也是芯片探针设计过程需要考虑的因素。因此在芯片探针设计过程中，必须依据各自实验的不同特性，综合考虑各种因素，才能设计出更合理有效的探针。

参考文献：

[1] He D, Zaitlen N, Pasaniuc B, et al. Genotyping common and rare variation using overlapping pool sequencing[J]. BMC Bioinformatics, 2011, 12(S6): S2.

[2] Kong W, Mou X, Hu X. Exploring matrix factorization techniques for significant genes identification of alzheimer's disease microarray gene expression data[J]. BMC Bioinformatics, 2011, 12(S5): S7.

[3] Lazaridis E N, Sinibaldi D, Bloom G, et al. A simple method to improve probe set estimates from oligonucleotide arrays[J]. Math Biosci, 2002, 176(1): 53-58.

[4] Saidi S A, Holland C M, Kreil D P, et al. Independent component analysis of microarray data in the study of endometrial cancer[J]. Oncogene, 2004, 23(39): 6677-6683.

[5] Wei H, Kuan P F, Tian S, et al. A study of the relationships between oligonucleotide properties and hybridization signal intensities from nimblegen microarray datasets[J]. Nucleic Acids Res, 2008, 36(9): 2926-2938.

[6] Chou H H, Hsia A P, Mooney D L, et al. Picky: oligo microarray design for large genomes[J]. Bioinformatics, 2004, 20(17): 2893-2902.

[7] Rouillard J M, Zuker M, Gulari E. Oligoarray 2.0: de-

sign of oligonucleotide probes for DNA microarrays using a thermodynamic approach[J]. Nucleic Acids Res, 2003, 31(12): 3057-3062.

[8] Luebke K J, Balog R P, Garner H R. Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts[J]. Nucleic Acids Res, 2003, 31(2): 750-758.

[9] Reymond N, Charles H, Duret L, et al. Roso: optimizing oligonucleotide probes for microarrays[J]. Bioinformatics, 2004, 20(2): 271-273.

[10] Wang X, Seed B. Selection of oligonucleotide probes for protein coding sequences[J]. Bioinformatics, 2003, 19(7): 796-802.

[11] Ratushna V G, Weller J W, Gibas C J. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design[J]. BMC Genomics, 2005, 6: 31.

[12] Nielsen H B, Wernersson R, Knudsen S. Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays[J]. Nucleic Acids Res, 2003, 31(13): 3491-3496.

[13] Dalma-Weiszhausz D D, Warrington J, Tanimoto E Y, et al. The affymetrix genechip platform: an overview [J]. Methods Enzymol, 2006, 410: 3-28.

[14] Held G A, Grinstein G, Tu Y. Modeling of DNA microarray data by using physical properties of hybridization[J]. Proc Natl Acad Sci USA, 2003, 100(13): 7575-7580.

[15] Matera A G, Terns R M, Terns M P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs [J]. Nat Rev Mol Cell Biol, 2007, 8(3): 209-220.

[16] Yin J Q, Zhao R C. Identifying expression of new small RNAs by microarrays[J]. Methods, 2007, 43 (2): 123-130.

[17] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction[J]. Nucleic Acids Res, 2003, 31(13): 3406-3415.

[18] Mathews D H, Sabina J, Zuker M, et al. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure[J]. J Mol Biol, 1999, 288(5): 911-940.

[19] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information[J]. Nucleic Acids Res, 1981, 9(1): 133-148.

[20] Mathews D H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization[J]. RNA, 2004, 10(8): 1178-1190.

Factors Analysis of Oligodeoxyribonucleotide Probes for RNA Microarray

MA Qin-qin, YAN Guo-quan, GUO Jiang-feng, DING Xian-feng

(Institute of Bioengineering, School of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: As one of main technologies on ncRNA research, the RNA microarray is receiving more attention, however, the probe design of RNA microarray is more complex than that of DNA microarray due to the single-chain characteristics of the target sequence. The RNA microfluidic picoarray for enhanced green fluorescent protein(EGFP) RNA transcript is designed using tiling probes in this study. The effects of free energy(ΔG) and melting temperature(T_m) of probe-target heteroduplex, secondary structure of target and base in the end of probe are evaluated, thus lay the certain basis of the screening of the probe of RNA microarray. The results show that the same free energy(ΔG) and melting temperature(T_m) of probe-target heteroduplex have different hybridization signal intensity, there is no significant regularity between free energy(ΔG) and melting temperature(T_m) probe-target heteroduplex and hybridization signal intensity is observed, there is a linear relationship between PT-sc of probe which reflect the secondary structure information of target and hybridization signal intensity. The base composition of the probe's 5'-end also affects the hybridization signal intensity. The stronger hybridization signal intensity are usually obtained when the 5'-end base of probes is G or C compared to the probes which 5'-end base of probes is A or T.

Key words: tiling probes; RNA microarray; secondary structure; probe-target heteroduplex

(责任编辑: 许惠儿)