科学通报 2008 年 第 53 卷 第 3 期: 271 ~ 280
评 述
《中国科学》杂志社
SCIENCE IN CHINA PRESS

Tiling Array 技术与应用研究进展

郎显宇[①②], 王俊[③], 迟学斌[①]

① 中国科学院计算机网络信息中心超级计算中心, 北京 100080;
② 中国科学院研究生院, 北京 100049;
③ 中国科学院北京基因组研究所, 北京 101300
E-mail: lxy@sccas.cn

2007-09-17 收稿, 2008-01-02 接受
国家高技术研究发展计划(批准号: 2006AA01A116)、国家自然科学基金(批准号: 60533020, 60673064)和科技部国家科技基础条件平台项目(批准号: 2005DKA64002)资助

摘要 Tiling Array 实验技术是从传统的微阵列(microarray)基因芯片技术发展而来, 在 Tiling Array 新技术产生发展的 5 年间, 它已经成为了全基因组生物信息挖掘的主要工具, 其高密度、高通量的特性使人们可以从全基因组水平考察生命过程和探索生命奥秘. 对 Tiling Array 技术和应用研究最新进展进行了较为详尽的描述, 其中包括 Tiling Array 技术概述、Tiling Array 应用研究、Tiling Array 重要的实验和发现以及对这些发现做出所有可能性的预测和解释. 除此之外, 对 Tiling Array 表达信号识别算法进行了简明的概述, 并对其中 3 种具有典型意义的算法给出了评价和比较.

关键词
Tiling Array
生物信息学
高通量
信号识别

随着越来越多物种基因组测序的完成, 下一步基因组研究的重点将是揭示物种基因组所隐藏的生物信息. 这一步的实现需要全面识别 DNA 所编码的基因、蛋白和其他功能元, 了解基因与蛋白的相互作用与调控机理, 以及如何配合而产生复杂的生物过程. 其中基础而重要的一步便是获得全基因组水平的表达产物, 进而开展功能元的研究.

对于人类基因组, 曾经估计编码蛋白质的基因数大约为 2×10^4~2.5×10^4 个, 除此之外, 还有许多不编码蛋白质的RNA基因, 如rRNA, tRNA, microRNA, snoRNA等. 全长cDNA测序(cDNA sequencing)方法识别了目前已知、高质量的编码蛋白基因, 但几乎所有类似的克隆方法都偏向于探测在生物组织中被充分表达的基因. 这类技术往往很难深入地探索基因组水平所有的表达信息, 也不适用于不同的组织和不同条件下转录信息的提取. 在刚刚过去的 5 年间, 从微阵列(microarray)基因芯片技术发展而来的新技术Tiling Array[1]使高通量、全基因组水平的表达探测在理论上得以开展; 随着芯片探针设计密度的不断增加, Tiling Array在高等真核生物全基因组水平的应用也得以逐渐实现.

1 Tiling Array 技术介绍

Tiling Array技术从微阵列发展而来, Tiling(或称tile path)指的是如瓦片一样覆盖基因组的探针序列. 如果说传统的微阵列技术是有偏向性的设计理念, Tiling Array可以说是无偏的芯片设计思路. 它把基因组染色体的双链序列无任何偏倚, 或按着一定的间隔规律、或者以序列交叠的方法、或者以序列头尾相接的方式制作成探针, 相邻探针中心位置之间的距离, 即定义为探针的步长(step或resolution). 如图1[1]是Tiling Array芯片设计工艺及其探针设计的几种方案.

一般来说, Tiling Array定义在基因组染色体水平, 它与传统微阵列的不同在于[2]: (1) 传统微阵列技术只检测染色体部分区域的生物特性, 如已知和预测基因的外显子部分; Tiling Array具有高密度、高通量的特点, 可以对全基因组水平的生物信息进行探测. (2) Tiling Array的探针筛选和芯片制备可以不依赖已有基因组注释信息; 传统微阵列通量较小、密度较

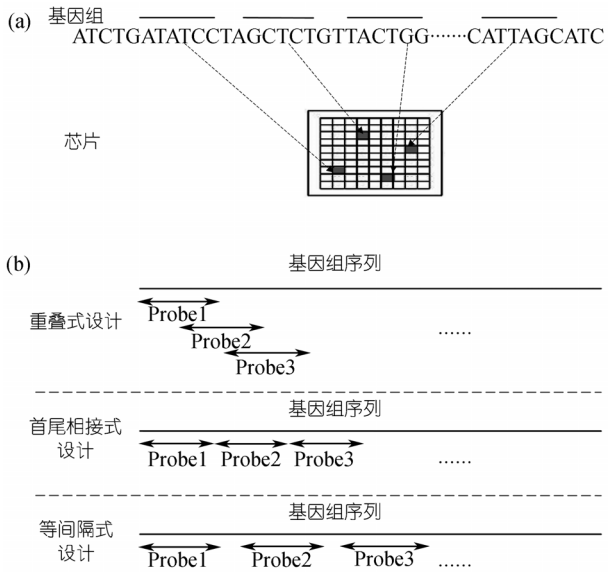

图 1 芯片与探针制备工艺

(a) Tiling Array 芯片设计工艺, 大部分芯片随机确定相邻序列在芯片上的位置以减少系统错误; (b) tile path 设计可以重叠, 首尾相接或者等距间隔

低, 测试需要更大的针对性, 其探针的选取需要基因组注释信息.

1.1 芯片与探针制备

芯片常用的制备方法主要有接触直接点样法[3]、照相平板术[4]、喷墨法等[5]. 接触直接点样法操作简单, 只要把事先准备好的探针点样到基板上, 探针取样通常依赖基因组注释. 点样的基板可以是玻璃、硅片、硝化纤维, 点样过程通过点样仪完成; 照相平板术多用于制作高密度芯片, 该技术不需要事先准备探针, 而是在基板上进行一次一个碱基的单链DNA原位合成; 喷墨法可用于原位合成, 也可以用作直接点样, 且都需要把用于合成的核苷酸或者合成好的探针喷射至基板预定位置, 其制作速度快、耗费低. 由于Tiling Array通量大、消耗高, 所以这类芯片多采用照相平板术和喷墨法制备.

除了上述芯片材质和制备方法比较关键之外, 加载在芯片上的探针则主导最终的芯片实验. 探针制备是个既老而又常新的研究方向, 这方面的研究之所以重要, 在于探针制备直接影响探针与靶标(target)杂交的结果, 探针制备的欠缺将导致杂交噪声的产生[6~10]. 随着对杂交现象的深入理解, 人们已了解假设与实验的差距, 探针制作需要从实验中提取可靠的资料[7,8], 而不是单纯停留于假设或想象. 目前的研究进展显示, Tiling Array探针制备一般需要两个步骤: (ⅰ) 探针设计, 即确定探针的长度与步长; (ⅱ) 探针筛选, 在Tiling Array探针设计后会得到大规模的探针集合, 需要最终筛选出不会产生杂交噪声的探针.

多数Tiling Array实验探针长度为 25~1000 nt, 普通的寡核苷酸(oligonucleotide, 简称oligo)芯片探针长度多为 25~70 nt, 也有更长的探针达到 100~150 nt. 某些探针设计方案提出, 寡核苷酸芯片采用较长的探针通常好于较短探针, 因为它能得到较好的杂交信号[6]. Tiling Array探针步长的大小则多取决于芯片的承载能力, 虽然探针步长越短其杂交结果越清晰, 但实验成本也随之升高.

杂交噪声就是指探针与非靶标序列发生交叉杂交, 从而导致信号出现噪声. 目前已知能够导致杂交噪声的原因有二: (ⅰ) 探针与非靶标序列相似; (ⅱ) 探针序列具有特性, 如某种碱基含量较高等. 为了排除这些可能引起杂交噪声的探针, 要对探针集合做大规模筛选, 这一步公认的考察因素[7~9]有: (ⅰ) 探针与非靶标相似性程度; (ⅱ) 探针与非靶标是否出现连续相同的序列片段; (ⅲ) 探针与非靶标杂交的自由能大小.

对于以上因素, 在实际筛选中, 需要根据不同长度探针的芯片实验来决定不同的参考阈值. 譬如Kane等人[10]对于探针长度为 50 nt的芯片实验测得, 当探针与非靶标序列全局相似性高于 75%, 或者探针与非靶标出现至少 15 bp连续相同的序列片段时, 探针会产生交叉杂交噪声; 而另一个相同长度探针的芯片测试表明[7], 相似性不高于 90%, 连续相同的序列片段不长于 20 bp, 杂交自由能高于−35 kcal/mol (1 cal=4.1868 J), 探针就不会产生显著的杂交噪声. 然而, 研究至今, 探针筛选的标准仍然悬而未决. 一方面, 相同长度的探针因为实验条件的不同而有不同的参考标准; 另一方面, 不同长度的探针需要相应的芯片实验, 测试各因素的参考阈值, 而当前研究较多的仅仅是探针长度为 50 和 70 nt两种情况. 所以, 各级别长度的探针筛选标准还有待进一步地完善和统一.

探针筛选标准决定了探针筛选算法, 而Tiling Array探针筛选所涉及的超大规模计算也不容忽视.由于芯片探针数量庞大, 通常以百万到千万计, 且每个待选探针都要与基因组所有表达的转录本做相似性比较, 不论利用启发式算法BLAST[11]还是Needleman-

Wunsch[12]动态规划的全局比对算法，这都是无比耗时的工作. 所以Tiling Array的探针制备必然需要高性能计算机以及计算方法的协助.

1.2 芯片类型

对应于微阵列芯片的不同类型, Tiling Array同样有两种类型的芯片[1]. 一种是高密度寡核苷酸Tiling Array芯片，这一类芯片探针的序列长度相对较短，多为 25~70 nt, 探针可利用照相平板术或者喷墨法直接合成在芯片上. 目前，此种生产工艺可以制造出高密度芯片，在小于 2 cm^2 的芯片上可设计超过 6600000 个特征区域，所说的特征区域就好比芯片阵列的一个个方格子，每个方格子都有百万计数相同的序列探针.

另一种是点样(spotted)平台，主要利用PCR产物或者细菌人造染色体(bacterial artificial chromosome)在玻璃片上制作探针[1]，PCR序列长度大约在 1 kb左右，这类芯片只能包含大约 1×10^4~4×10^4 个探针特征区域，其探针密度远小于oligo芯片. PCR芯片或BAC芯片的任何一处探针位置都包括目标序列以及目标序列的互补序列，这样还会导致测试中如果没有大量的额外实验，将很难确定表达序列是在双链的哪个链上. 例如, Rinn等人[13]所做的Tiling Array实验利用了PCR设计模式，其所需要的PCR产物要通过 2 万多次PCR反应获得，那么人类全基因组的PCR Tiling Array就需要将近 2 百万次的PCR反应.

由于第一类芯片的探针设计密度和通量更高，制作工艺简单，所以全基因组 Tiling Array 分析一般皆采用这一类 oligo 芯片.

1.3 芯片工作原理

在芯片与探针制作完毕后，接着便是RNA提取、杂交、扫描等工作流程，简单说明如下: (ⅰ) 表达产物提取: 基因组 RNA 表达的提取以及反转录 cDNA 的合成; (ⅱ) 杂交过程: 标记的 cDNA 与芯片杂交，形成“双螺旋”结构，其余的则被洗脱除去; (ⅲ) 扫描芯片: 此步骤将得到一张基因表达快照，芯片上的每个方格子会有不同的颜色或亮度显示; (ⅳ) 数据分析和处理: 对探针杂交信号，即亮度信息进行读取，并转化成数字信息. 对芯片探针后续的分析都是基于探针亮度分值的处理.

1.4 小结

综上所述，为了探索高等真核生物更长的全基因组(如人类基因组有 30 亿碱基)的生命信息，未来开发更高密度、低耗材成本的 Tiling 芯片是技术本身无法回避的问题; 同时，建立探针筛选的统一标准，以及研发高效率的探针制备软件是发展和提高 Tiling Array 技术的必经之路.

若从 Tiling Array 高通量的特点来看，它将是通往全局了解和把握基因组生物特性的高速公路，也正由于此其芯片造价十分昂贵. 若从 Tiling Array 技术和实验角度观察，芯片所产生的模拟信号总不如数字信号精准，其噪声的干扰也在所难免，所以经过 Tiling Array 所得到的结果需要经过另外的实验进行测准和验证. 不过没有任何技术是完美无缺的，Tiling Array 带给生命科学的震撼和期待仍然有目共睹.

2 Tiling Array 应用研究

目前的研究显示，这种高密度、高通量的Tiling Array实验技术可用来破译隐藏在基因组中的许多信息. 譬如，它可以检测基因组的转录特性、识别新编码或非编码基因、分析可变剪接、定位DNA模体(motif)、也可用于比较基因组研究，以及基因组重测序等[14].

对于转录层面的研究，Tiling Array 可用于新基因的发现、基因组的表达研究、可变剪接和 RBP (RNA-binding protein)目标识别. 对于基因组层面的分析, Tiling Array 可用于 ChIP-chip 研究、Methylome 分析、基因组重测序、基因组多态性和 CGH 研究等. 从目前涉及 Tiling Array 应用的各类研究看，发表于 *Science*, *Nature* 以及 *Genome Research* 等影响因子较高刊物上的 Tiling Array 应用主要集中在基因组的表达研究.

我们知道，仅仅拥有完整的基因组序列并不足以识别所有表达，因为一般所采用的编码基因计算预测方法充满了不确定性和各种错误，非编码RNA基因的计算预测目前尚无通用可行的方法. 虽然识别基因的分子生物学方法，如全长cDNA测序已经成功地识别了数以万计的基因，但此类方法由于无法甄别那些表达量低和在少数细胞类型或组织中表达的基因，所以 20 世纪后期出现的芯片技术成为了发现新基因的补充. 2001 年一个关于人类基因组的微阵列实验[14,15]利用喷墨(ink-jet)技术制作了oligo芯片，它仅仅是把 22 号染色体已知和预测基因的外显子(exons)制作成探针，检测结果显示，有大量的已知基

因和 57%左右的预测基因被表达, 但是这个微阵列的实验并没有考察基因组的无注释区域.

传统的关于基因的定义是基因组上那些编码蛋白质的区域, 这些区域的上游是调控序列, 而剩余的大部分区域被看作基因组的垃圾(junk). 直到不久以前, 利用无偏的全基因组Tiling Array实验探测整个基因组的表达, 使得这一传统观点受到了挑战. 这一类的全基因组芯片实验显示, 大量的表达在人类以及其他生物基因组中被发现, 原来通过EST, cDNA或基因预测得到基因组表达仅为其中的十分之一[14].尽管如此, 问题仍停留在探索这些发现是否是所谓的"转录噪声". 目前, 更多的研究正在继续确认这些新发现和试图揭示新发现表达之间的功能关联.

反义(antisense)表达也属于转录本的一种, 它们能同时控制基因组双链的基因表达. Kampa等人[16]在2002 年Affymetrix人类 21 和 22 号染色体研究基础上又精炼了实验和分析过程, 他们的结果显示, 大约 11%的表达是在已知外显子、mRNAs和ESTs的反义链上. 另一个点样平台实验显示有大约 50%内含子(introns)区域的表达也是反义表达[1]. 总体看来, 这一类研究表明, 人类 21 和 22 号染色体至少 20%的区域是反义表达[14], 其他真核生物的实验也同样证明了这一点[17], 但是反义表达的具体功能和它们在生物过程中扮演的角色还不为所知.

未来, Tiling Array技术应用于机体不同细胞类型、表型和组织的表达探测具有很广阔的前景[14], 这也是加速传统基因组注释的有力途径. 但同时这也对分析Tiling Array信号数据的算法与软件提出了更大的挑战, 即如何有效、准确地识别表达量很低或极低的转录本. 另外, 因为Tiling Array实验针对全基因组测试, 其实验发现的产物量较大, 所以发展更加快速精确的实验验证手段也非常关键.

3 Tiling Array 实验与发现

当前, 吸引众多目光的 Tiling Array 实验多集中于基因组的表达研究, 从 2002 年至今, 已经有几个物种、不同芯片设计模式的 Tiling Array 数据公布.

从芯片实验开展的次序来看, 传统的微阵列技术多用来检验一些已知和预测基因的外显子在某些细胞株或组织中是否被表达, 而后逐渐发展到用 Tiling Array 技术对物种全基因组水平编码和非编码基因进行挖掘和分析. 几年来, Tiling Array 实验的种种发现已让人类开始重新思考和审视物种基因组.

3.1 人类基因组 Tiling Array 实验和发现

2002 年 5 月发表于*Science*的一篇文章是关于人类基因组Tiling Array的实验[18], 它对人类 21 和 22 号染色体 11 个细胞株(cell lines)的表达进行了分析. 探针设计为染色体DNA非重复部分的序列(利用ReapeatMasker得到非重复基因组序列), 探针长度 25 nt, 每间隔 35 mer 设计一组探针, 即规格为 25nt-oligomers/35bp-step, 并采用Affymetrix芯片PM (perfect match)、MM (mis-match)设计模式, 其中MM探针与PM探针仅有中心的一个碱基不同且为互补, 利用细胞质poly(A+) RNA的cDNA序列与芯片杂交. 由于探针之间有间隔, 实验很难解释基因组的结构关系, 如外显子与内含子之间的分界, 而且由于实验样本无法区分正负链, 所以不能具体识别单链的表达和分析双链DNA转录的相互关系. 但从实验结果看来, 大约 90%表达产物在已知基因外显子以外.

2003 年Rinn等人[13,14]采用了点样平台, 利用PCR产物做成的探针覆盖了人类 22 号染色体几乎全部的无重复区域. 实验发现有至少注释基因 2 倍的转录本被表达. Rinn等人还把他们的实验结果同 2002 年Kapranov等人[18]所做的人类芯片实验结果做了对比, 并预测在这些实验中, 有至少 50%的表达在已知基因组注释之外. 他们通过比较人类与大鼠基因组, 发现相对保守的序列区域, 并对新发现的表达不纯粹是"转录噪声"而具有生物功能的假设提出了支持证据.

2004 年Merck实验室利用喷墨技术[19], 设计了探针规格为 60nt-oligomers/30bp-step的芯片, 并对人类 20 号染色体的 6 个组织和 22 号染色体的 8 个组织的表达情况进行了研究. 结果显示, 47%的阳性探针在目前注释的基因之外, 其中 22%的表达出现在内含子, 25%出现在基因间(intergene)部分.

同年, Bertone等人[20]设计了人类全基因组Tiling Array实验, 探针规格是 36nt-oligomers/46bp-step, 共制作了 134 张芯片, 仅对肝脏组织(liver)的表达做了研究, 分析出大约有 10595 个新表达为其他方法所没有发现, 大部分这样的表达落在基因间部分, 并一定程度远离原有的注释基因.

2005 年 5 月又一篇发表于*Science*的文章对于人类基因组 10 条染色体(6, 7, 13, 14, 19, 20, 21, 22, X和Y)8个细胞株的表达进行了探索[21]. 虽然芯片仍是

Affymetrix的设计方案，但其探针步长缩小为 5 bp，探针间出现了 20 bp的重叠. 这种设计密度所提供的基因组查询范围是 2002 年的 7 倍[18]. 信号识别结果显示，10 条染色体平均看来，有 31.8%的表达分布在基因间区域，42%~49%落在内含子部分；并且实验初步表明，转录系统是一个双链相互重叠关联的网络.

3.2 其他生物基因组 Tiling Array 实验和发现

2003 年发表于*Science*的拟南芥基因组Tiling Array实验[22]是第一个真核生物全部基因组的芯片实验，芯片的探针长度 25 nt，以首尾相接的方式覆盖了全基因组. 实验发现了很多新表达区域，其中包括2000个基因间位置. 另外，有大于30%的注释基因在反链位置发现表达，5817 个计算预测的基因被证实在某些生物过程中表达.

2004 年果蝇Tiling Array实验发现[23]，大约 41%的表达发生在内含子和基因间区域.

4 Tiling Array 表达预测

上述Tiling Array实验采用了不同的表达信号识别算法对表达样本进行采集，几乎每个实验所得到的大部分表达是在已知和预测外显子之外. 目前，对这些表达(简称dark matter)有很多可能性的预测和解释[2]，具体分析列举如下.

4.1 新的编码蛋白基因

全长cDNA测序技术对于新基因的发现提供了强有力的方法，一个全长cDNA测序的实验[24]发现了2000 个新cDNAs与Ensembl注释基因没有重叠，其ORF > 300 bp. LongSAGE实验[25]也发现了新的编码蛋白基因. 尽管cDNA测序等实验和预测方法表明仍有很多人类的编码蛋白基因等待发现，尤其是在某些组织或细胞株中表达量很低的转录本，但是编码基因发现的比率已经大为下降. 研究趋势表明，新编码蛋白基因不能解释大规模的dark matter[2].

表达的假基因(pseudogenes)也可能包括在dark matter中. 研究显示，人类基因组至少有 20000 个假基因，已知的一些假基因不但表达而且具有生物功能[26]. 由于这些假基因具有编码蛋白基因的序列特征，所以它们可直接被计算方法识别，且它们不会在Tiling Array实验发现的转录本中占很大比重[2].

4.2 新的非编码蛋白基因

很多功能已知的非编码蛋白基因(ncRNAs)，它们在高等生物组织中担任基因表达调控等重要职责. 目前，几乎所有关于Tiling Array的文章都认为ncRNAs最有可能解释部分被识别的dark matter[27].虽然已知的ncRNAs多在核子中，并且没有poly(A)，但有证据表明，发现的具有poly(A) 的ncRNAs在增加[27]. RT-PCR实验研究显示，大约 74%新发现的表达是ncRNAs[2]，虽然实验可以证明ncRNAs被表达，但是却不能肯定它们具有生物功能.

4.3 反义表达

一些dark matter在已知基因或者ESTs反链的相同位置表达. 在原核生物中，反义RNA在细菌中可能具有调控作用，*E. coli*大部分的基因都有反链表达，真核基因组也发现了反链表达[17]. 在用来确定dark matter属于哪个链表达的Tiling Array实验中，发现有大约 11%的表达，其反链的相同位置是外显子区域[2].

4.4 可变剪接新的转录本(isoforms)

研究显示，大约一半的人类 20, 21, 22 号染色体的Tiling Array表达在已知基因注释区域，这意味着这些dark matter有可能是已知基因的延伸或者是更复杂的表达[18,19]，可能是由于可变剪接而得到的新转录本. Kapranov等人[18]的人类染色体Tiling Array实验显示，仅有 5%发生在已知基因内含子的表达，与已知的可变剪接相对应. 这也暗示，人类基因组可能有很多可变剪接等待发现.

4.5 已知基因的延伸

因为cDNA克隆未必能获得全长的表达序列，基因发现软件也很难预测识别 UTRs (不翻译区)区域，所以很多已知的表达自然可能要长于目前的注释长度. 这可能可以解释一部分靠近注释基因并且在基因间的 dark matter，它们可能是 UTRs 区域，当然基因间部分的 dark matter 也可能是已知或未知基因的外显子或者内含子.

4.6 实验噪声

Tiling Array 实验无法免于噪声干扰，RT-PCR 和 Northern blot 对 Tiling Array 新发现表达验证的结果显示，部分新表达实际的表达水平很低，这也暗示是某些实验噪声导致了 dark matter 的出现.

影响之一是 RNAs 样本的污染. 因为 RNAs 样本至少包括一些残留的 DNA，这会导致一些 dark matter 的出现. 这种基因组水平上的污染取决于不同的

细胞组织和提纯手段.

影响之二是没有剪接的 mRNAs. 它们在 RNAs 提取过程中也被分离, 这样会导致某些已知基因内含子区域信号被判断为表达. 从细胞质 RNA 中提纯 mRNAs, 可以使这个问题的影响最小化.

影响之三是双链标记(double-stranded labeling). 由于杂交样本在标记过程中, 有可能出现假造的cDNA的互补链. 这种噪音主要表现为假的反链表达[28]. 但实验表明, 这个影响并不是dark matter的主要原因.

影响之四是交叉杂交(cross-hybridization)[29,30]. 前面芯片探针制备中已提到了杂交噪声, 此处所说交叉杂交与前述并无不同, 但由于探针筛选过程不完备, 会造成探针的交叉杂交噪声. Affymetrix公司为了控制和跟踪杂交噪声, 所以在Tiling Array设计中采用了PM和MM的设计模式, 其中MM探针即用来跟踪交叉杂交, 但这种设计是否有效仍说法不一[1,31]. 可见在芯片制备阶段, 尽可能排除可能产生噪声干扰的探针尤为重要, 如果筛选标准不完善, 在后续的表达信号识别或之后的数据分析过程中就不得不考虑交叉杂交的影响.

4.7 假阳性

所谓探针假阳性, 即此探针序列并非真实表达, 其主要是由实验噪声干扰和表达及信号识别算法的误判造成.

Tiling Array 技术缺少判断表达的真实参照, 也就是缺少基因组在哪些区域是不被表达的纪录. 这个局限性使得客观衡量信号识别算法的假阳性率具有较大困难. 尽管内含子和启动子有可能作为判断表达的依据, 但是因为基因组注释仍不确定, 它们也很难被准确定义.

5 Tiling Array 表达信号识别算法

在芯片制备、RNA 提取和杂交、芯片扫描一系列过程之后, 芯片的每个探针亮度都被计算机记录下来. 下一步的工作就是根据探针亮度来提取有效信息. 以 Tiling Array 应用于基因组表达挖掘而言, 就是通过探针的亮度信息判断探针序列是否表达.

在过去的几年中, 有一些用于Tiling Array表达信号识别的算法相继问世[32~34], 其中包括Kapranov等人[18]、Rinn等人[13]、Bertone等人[20]、Schadt等人[19]和Cheng等人[21]在其相应Tiling Array实验中使用的算法[32]. 可以说, 已知的信号识别算法多数面向各自的Tiling Array实验, 即算法与不同的芯片设计模式或具体实验相关. 譬如, Affymetrix芯片嵌入了细菌基因组的序列探针, 那么它相应的SW算法就可以借用细菌基因组提供的表达信息; Merck公司Schadt等人[35]所做的实验是多个组织的表达研究, 他们应用的信号识别算法也直接对多组织信号亮度同时分析, 不过这种方法不适合识别单个组织的表达. 另外, Bertone等人[20]采用的方法简单延续了微阵列技术的亮度分布来识别表达信号.

总体看来, 表达信号识别算法虽然不少, 但是具有普遍适用性的方法不多. 而且, 在目前探针筛选标准仍不清晰的情况下, 筛选出的探针仍具有制造杂交噪声的很大可能, 但已有的算法都不具备过滤杂交噪声和识别低表达量探针的能力, 这将造成识别假阳性率的升高, 也必然给信号识别之后的数据分析造成较大压力. 即使 Affymetrix 芯片设计考虑了除噪问题, 但效果并不确切.

下面对 3 个典型的 Tiling Array 表达信号识别算法进行介绍.

5.1 算法介绍

(ⅰ) 滑窗(sliding window, SW)算法[16]. SW算法是针对Affymetrix芯片的表达信号识别方法, 其优势在于利用查询探针和相邻探针的亮度值来确定这个查询探针的序列是否被表达(positive). SW算法可分为两个部分来描述.

第一部分, 探针的表达评估. 这一步主要通过一个探针和其相邻探针亮度确定此探针的表达分值. 首先, 设定一个固定的窗口长度, 其半径为BW, 则窗口长度为BW×2+1, 对于任意探针P_j, 设$Z_j = PM_j - MM_j$.对于某个观察探针P_i, 设其中心在染色体上的位置为PT_i, 则包含在窗口区间[PT_i–BW, PT_i+BW]中的所有探针两两之间做平均值计算$A_{m,n} = (Z_m+Z_n)/2$, m和n分别为窗口内包含的探针编号. 其次, 探针的表达评估值E_i = pseudo-median(P_i) = median ($A_{m,n}$: m<= n)为窗口内两两探针亮度平均值的中值.其中确定合适的滑窗长度很重要, 它有两个决定因素: (1) 探针间距离, 即探针序列中心间距离; (2) 探针所在染色体外显子的平均长度(一般约为 137 bp). 以 2002 年 Kapranov等人[18]的实验为例, 芯片探针长度为 25

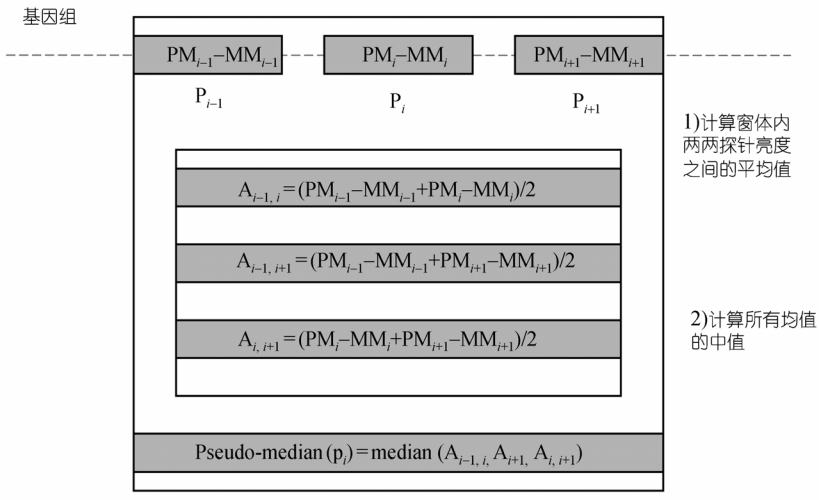

图 2　Kampa等人[18]的SW信号处理算法第一部分的计算过程

nt, 步长为 35 bp, 若设计窗口长度为 100 bp, 窗口内大约可包含 3 个连续探针. 其考察探针表达评估的计算过程如图 2 所示[1].

第二部分, 表达与非表达区域的划分. 通过第一部分的计算, 每个探针基于其亮度都得到了相应的表达评估分值, 我们需要一个表达评估的阈值来确定探针是否表达. Affymetrix芯片设计的另一个特色是在每张芯片上嵌入细菌基因组的序列探针, 这些探针与芯片上的人类基因组探针不具有相似性, 这样通过相同环境下提取的细菌RNA表达与芯片杂交, 并根据细菌的基因组注释可以计算出芯片的最高假阳性率, 从而给出探针表达的阈值. 如果探针的表达评估分值大于阈值, 则定义此探针是阳性的, 也就是表达的. 接下来, 根据表达的探针, 可以继续判断表达序列片段(transfrags). 这个连续表达的序列片段只有满足长度(minirun)至少 90 bp, 连接探针之间隔(maxgap)不大于 40 bp, 方可被称为transfrag[16].

(ⅱ) 亮度分布(signal distribution, SD)方法[20]. SD法是Bertone等人[20]用于识别人类基因组非注释区域表达序列片段TARs (transcriptional active regions)的方法. TARs和transfrags虽然称呼不同, 但是意义相同. 在Bertone等人的实验中, 探针的原始亮度达到整张芯片亮度的 90%以上才被认为是表达探针. 以当时Bertone等人的实验芯片为例, 探针长度为 36 nt, 步长 46 bp, 则要求至少 5 个连续表达的探针可被称为TARs. 任何不表达探针的出现都将中止TARs的延展. 在SD方法的实际应用中, 亮度百分数阈值可以根据实验定义.

(ⅲ) 基于HMM的信号识别算法. 基于隐马尔可夫模型(hidden Markov model), 针对Tiling Array信号识别问题, HMM[36]的观察序列可看作大规模的信号亮度值, 状态简单地说可分为两类: 一类是表达, 另一类是非表达. 那么, Tiling Array信号识别变成了对应每个探针的亮度值, 需要确定其相应的状态. 此问题可先从大规模的信号数据中训练和建立样本模型, 然后根据模型确定所有探针信号所对应的状态序列. 基于HMM的Tiling Array信号识别方法在文献中多有记载[37~40], 下面以Du等人[37]的HMM方法为例简单说明HMM如何应用于Tiling Array表达信号识别.

以图 3[37]标示的算法流程来看, 第一步当从大规模的探针亮度数据中, 根据一定的规则选取样本, 文中以最大熵法选取了从亮度显示上表达活跃和不活跃的探针作为训练样本; 第二步可以根据RefSeq或Ensembl基因注释, 为样本标注相应的状态, 表达或不表达; 第三步依据所选样本, 训练隐马尔可夫模型, 主要计算状态转移概率矩阵和给定状态下的观察值分布; 第四步利用Vitebi[41]经典算法, 根据HMM标记所有探针的状态.

基于 HMM 的信号识别算法是一个典型的机器学习(machine learning)方法, 它所涉及的问题一方面是学习样本的大小, 一方面是学习样本的纯度. 一般来说, 在没有样本噪声的情况下, 训练样本越大, 则

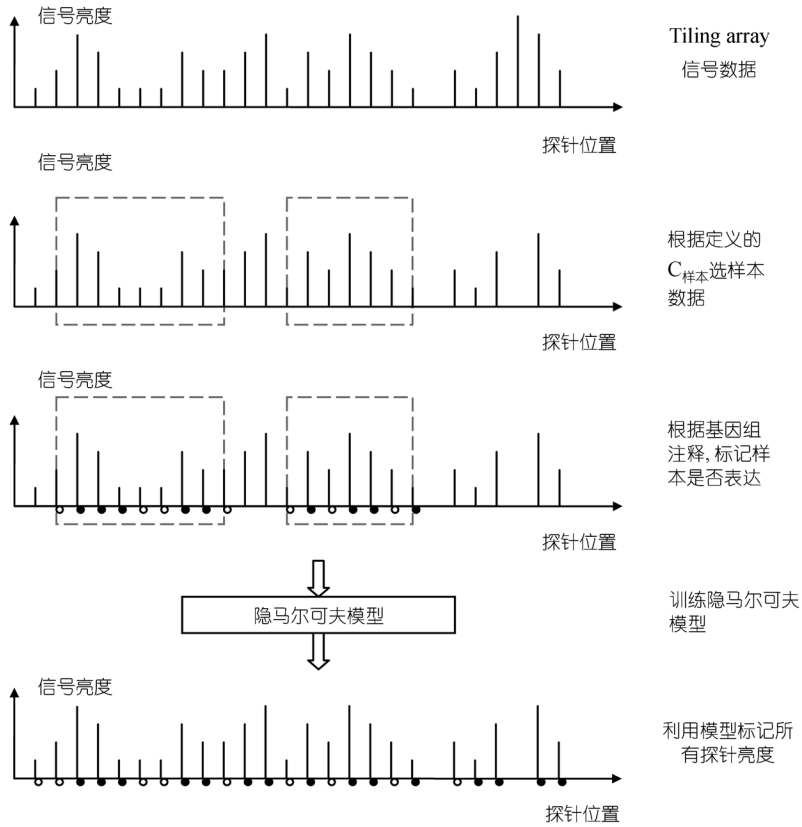

图 3 基于 HMM 的信号识别流程

结果越准确稳定.

5.2 算法比较

SW 算法适用于 Affymetrix 公司设计的芯片, 其芯片中嵌入了可作为表达参照的其他物种序列探针, 而 SD 方法并不拘于芯片设计. 另外, 二者在信号识别效果上也各有优劣. SD 方法在判断探针表达上有些粗糙, 容易受到交叉杂交的影响而增加识别假阳性率, 并且遗失表达量不高的探针. 而 SW 算法由于其相应的芯片探针 PM 和 MM 的设计模式, 本身对交叉杂交有一定制约作用. 此两种方法中, SW 的设计思路和精度更优.

以上两种算法都需要人为设定参数才能完成分析. SW 和 SD 方法必须通过基因组信息的帮助、有效的参数筛选, 来设定诸如窗口大小、表达阈值等一系列重要参数, 可以说参数的设定是两种方法的命脉所在, 它们直接决定着 Tiling Array 芯片表达识别的敏感度和假阳性率的高低. 而HMM机器学习方法完全不依赖人为设计参数来衡量探针是否表达, 它依靠数学建模直接分析和判断探针表达与否的状态. 几种方法比较而言, HMM算法的适用性更广, 准确性更高[32].

6 结论与展望

Tiling Array 技术产生发展的 5 年间, 已经成为了基因组生物信息挖掘的主要工具. Tiling Array 高密度、高通量的特性使人们可以从全基因组的水平考察生命过程和探索生命的奥秘. 也正是由于从整个基因组的角度出发, Tiling Array 仍然存在价格较贵、芯片技术需要不断发展的问题; 同时从微阵列基因芯片时代开始, 芯片技术就面临一个探针筛选和信号去噪的问题. 目前, 这类问题在 Tiling Array 研究中也并未得到很好的解决, 而且标准通用的 Tiling Array 信号识别方法也有待进一步的研究和发展.

致谢　感谢陆忠华老师和余晓哲同学在材料组织过程中提供的帮助.

参考文献

1 Royce T E, Rozowsky J S, Bertone P, et al. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet, 2004, 21: 466—475[DOI]

2 Johnson J M, Edwards S, Shoemaker D, et al. Dark matter in the genome:evidence of widespread transcription detected by microarray tiling experiments. Trends Genet, 2004, 21: 93—102[DOI]

3 Whitchurch A K. Gene expression microarray. IEEE Potentials, 2002, 21: 30—34[DOI]

4 Moore T R. Making chips to probe genes. IEEE Spectrum, 2001, 38: 54—60[DOI]

5 Hughes T R, Mao M, Jones A R, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol, 2001, 19: 342—377 [DOI]

6 Chou C C, Chen C H, Lee T T, et al. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Res, 2004, 32: e99[DOI]

7 Liebich J, Schadt C W, Chong S C, et al. Improvement of oligonucleotide probe design criteria for functional gene microarrays in enviromental applications. Appl Environ Microarray, 2006, 72: 1688—1691[DOI]

8 He Z L, Wu L Y, Li X Y, et al. Empirical establishment of oligonucleotide probe design criteria. Appl Environ Microarray, 2005, 71: 3753—3760[DOI]

9 Li X Y, He Z L, Zhou J Z. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res, 2005, 33: 6114—6123[DOI]

10 Kane M D, Jatkoe T A, Stumpf C R. et al. Assessment of sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res, 2000, 2: 4552—4557[DOI]

11 Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215: 403—410

12 Pearson W R, Lipman D. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA, 1988, 85: 2444—2448[DOI]

13 Rinn J L, Euskirchen G, Bertone P, et al. The transcriptional activity of human chromosome 22. Genes Dev, 2003, 17: 529—540[DOI]

14 Mockler T C, Chan S, Sundaresan A, et al. Applications of DNA tiling arrays for whole-genome analysis. Genomics, 2005, 85: 1—15[DOI]

15 Shoemaker D D, Schadt E E, Armour C D, et al. Experimental annotation of the human genome using microarray technology. Nature, 2001, 409: 922—927[DOI]

16 Kampa D, Cheng J, Kapranove P, et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosome 21 and 22. Genome Res, 2004, 12: 331—342[DOI]

17 Chen J J, Sun M, Kent W J, et al. Over 20% of human transcripts might form sense-antisense pairs. Necleic Acids Res, 2004, 32: 4812—4820[DOI]

18 Kapranov P, Cawley S E, Drenkow J, et al. Large_scale transcriptional activity in chromosomes 21 and 22. Science, 2002, 296: 916—919[DOI]

19 Schadt E E, Edwards S W, Debraj G, et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol, 2004, 5: R73[DOI]

20 Bertone P, Stolc V, Royce T E, et al. Identification of novel transcribed sequences in human using high-resolution genomic Tiling Arrays. Science, 2004, 306: 2242—2246[DOI]

21 Cheng J, Kaparanov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 2005, 308: 1149—1154[DOI]

22 Yamada K, Lim J, Dale J, et al. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. Science, 2003, 302: 842—846[DOI]

23 Stolc V, Gauhar Z, Mason C, et al. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science, 2004, 302: 655—660[DOI]

24 Ota T, Suzuki Y, Nishikawa T. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet, 2004, 36: 40—45[DOI]

25 Saha S, Sparks A B, Rago C, et al. Using the transcriptome to annotate the genome. Nat Biotechnol, 2002, 20: 508—512[DOI]

26 Hirotsune S, Yoshida N, Chen A, et al. An expressed pseudogene regulates the messenger RNA stability of its homologous coding genes. Nature, 2003, 423: 91—96[DOI]

27 Claverie J M. Fewer genes, more noncoding RNA. Science, 2005, 309: 1529—1530[DOI]

28 Castle J, Engele P G, Armour C D, et al. Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biol, 2003, 4: R66[DOI]

29 Flikka K, Yadetie F, Laegreid A, et al. XHM: A system for detection of potential cross hybridizations in DNA microarrays. BMC Bioinformatics, 2004, 5: 117—125[DOI]

30 Reilly C, Raghavan A, Bohjanen P. Global assessment of cross-hybridization for oligonucleotide arrays. J Biomol Tech, 2006, 17: 163—172

31 Wu C, Carta R, Zhang L, et al. Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Res, 2005, 33(9): e84[DOI]

32 Emanuelsson O, Nagalakshmi U, Zheng D Y, et al. Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. Genome Res, 2006, 17: 886—897[DOI]

33 Halasz G, van Batenburg M F, Perusse J, et al. Detecting transcriptional active regions using genomic tiling arrays. Genome Biol, 2006, 7: R59[DOI]

34 Huber W, Toedling J, Steinmetz L M. Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics, 22: 1963—1970

35 Ying L, Schadt E E, Svetnik V, et al. Identification of chromosomal regions containing transcribed sequences using microarrays and computational methods. In: 2003 Proceedings of the American Statistical Association Alexandria. VA: American Statistical Association, 2003: 4672—4677

36 Rabiner L. A tutorial on hidden Markov models and selected application in speech recognition. Proc IEEE, 1989, 77: 257—286[DOI]

37 Du J, Rozowsky J S, Korbel J O, et al. A supervised hidden markov model framework for efficiently segmenting Tiling Array data in transcriptional and chIP-chip experiments:systematically incorporating validated biological knowledge. Bioinformatics, 2006, 22: 3016—3024[DOI]

38 Munch K, Gardner P P, Arctander P, et al. A hidden Markov model approach for determing expression from genomic tiling microarrays. BMC Bioinformatics, 2006, 7: 1471—2105

39 Ji H K, Wong W H. TileMap: Create chromosomal map of tiling array hybridization. Bioinformatics, 2005, 21: 3629—3636[DOI]

40 Li W, Meyer C A, Liu X S. A hiddern Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics, 2005, 21: 1274—1282

41 Viterbi A J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inform Theory, 1967, 13: 260—267