基因芯片的数据处理与分析

张学工 凡时财 裴云飞
清华大学

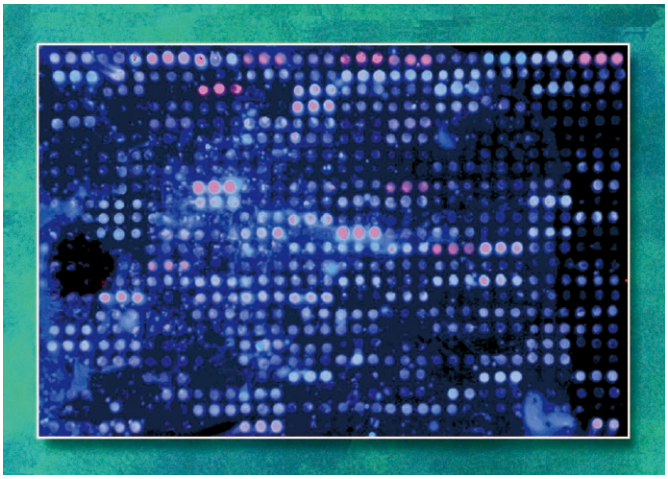
关键词： 基因芯片 数据处理

引言

本世纪初，人类基因组计划（Human Genome Project）的完成标志着生命科学的研究进入后基因组时代。如何理解这些海量的遗传信息成为后基因组时代的重要任务。分子生物学的中心法则告诉我们，细胞中的基因最主要是通过从DNA转录到RNA（mRNA）、再翻译成蛋白质来发挥作用的。根据目前的认识，人类基因组中编码蛋白质产物的基因的总数大约在20000～25000之间或者更多。这些基因在人体内不同组织的细胞中、在细胞不同的发育阶段有着不同的表达量，即所转录出的mRNA的丰度。而且基因的表达是受到调控的，众多基因在时间和空间上有规律地协调表达，是细胞和生物体正常生命活动的基础。

研究基因的表达无疑会对认识很多生命现象的规律具有重要意义。传统的用于研究基因表达的实验方法（如Northern-Blotting技术[1]），仅适用于单个或者较少几个基因。20世纪末，随着生物化学技术的发展，并受到计算机领域高密度芯片生产技术的影响，诞生了能够同时测量成千上万个基因的mRNA[2]表达量的新技术。通过这些技术，能够在一个几平方厘米的芯片上放置对应于成千上万个基因的DNA探针，从而同时测定这些基因在样品中的表达。这一类技术通常被称作DNA微阵列（microarray），中文更多地称为基因芯片，而英文的GeneChip由于已经被公司注册为专有的名称，因此只用来指特定的芯片类型。基因芯片是生物芯片[3]（BioChips）大家族中的一员，也是最重要的一员，其他类型的生物芯片还包括蛋白质芯片、组织芯片等等。与传统的分子生物学和生物化学实验只能一次得到很有限的数据相对照，这些能够同时获取大量生物分子数据的技术被统称为高通量（high-throughput）技术。

以基因芯片为代表的高通量分子生物学技术的产生，使得科学家获取实验数据的能力大



[1] 一种核酸印迹杂交技术

[2] 信使核糖核酸

[3] 由于常用玻片/硅片作为固相支持物，且在制备过程模拟计算机芯片的制备技术，所以称之为生物芯片技术。

大增强，在一张芯片上就能获得从上万个到上百万个探针的观测值，其中包含了数千个到数万个基因的表达值。这种海量的数据一方面为发现更多、更复杂的生物规律提供了可能，另一方面也对数据的处理和分析技术提出了前所未有的挑战。由于传统的数据分析方法力不从心，一系列新的适应高通量生物数据分析与处理的方法应运而生，其中，图像处理、模式识别与机器学习等现代信息处理技术发挥了重大的作用。

本文尝试对这一领域进行较全面的介绍，包括典型的基因芯片原理和实验流程，以及其中所主要涉及的图像处理、底层数据处理、高层数据分析与挖掘等方面的重要问题、代表性方法和应用等，对于一些近年来最新出现的新的基因芯片类型也给予扼要的介绍。

基因芯片的基本原理

基因芯片原理的基础是DNA的碱基配对原理，即在腺嘌呤（A）、胸腺嘧啶（T）、鸟嘌呤（G）和胞嘧啶（C）这四种核苷酸中，A和T、G和C分别能形成紧密的配对，这也是生物体内使得DNA能够复制和转录的基本机制。这种配对的形成过程称为杂交（hybridization）。利用这一原理，我们可以用一段特定的DNA序列作为探针（probe）来检测与之配对的DNA分子的存在及其丰度。基因芯片就是在一张面积很小的芯片上固定大量的DNA探针，将经过处理后的样品加入到芯片上，使样品中的核苷酸片断与相应的探针杂交，通过荧光成像获得每个探针上杂交的分子的浓度，再通过后期的处理即可获得相应的基因表达量。由于芯片的面积很小，能够保证各个探针与样品发生杂交反应的条件是一致的，而且保证可以用较少量的样品即可快速地获得基因表达量检测。

根据探针制备和固定技术的不同，基因芯片主要分为点样式cDNA芯片（printed cDNA microarray，简称cDNA芯片）和寡核苷酸芯片（oligonucleotide microarray）两大类。cDNA芯片是将实验室制备的探针通过点样固定在基片上，而寡核苷酸芯片则是采用原位合成技术在芯片上合成探针。下面简要介绍这两种芯片的原理及其实验流程。

cDNA芯片

cDNA是从mRNA通过反转录过程得到的DNA，cDNA芯片就是以这种反转录的cDNA片断作为探针的基因芯片。制作cDNA芯片时，首先需要构建cDNA文库（cDNA library），即从实验材料中提取将要研究的基因的mRNA，将它们反转录成cDNA，然后酶切成不同片段并克隆到载体里，形成所研究的基因对应的cDNA片断的一个分子库，即cDNA文库。从文库中选取特定的cDNA片断，利用PCR（Polymerase Chain Reaction，聚合酶链式反应）技术进行扩增和纯化，得到所需要的各个基因的探针，最后通过机械手将对应大量基因的探针以类似喷射打印的形式，精确地按照特定排列顺序滴加到经过衍生处理的基片上，从而完成芯片的制备。这一技术是由斯坦福大学的科学家发明的，其特点是成本相对低廉，而且芯片的探针可以根据生物学家的需要自行设计和提取，是一种可以实验室“自制”（home-made）的基因芯片。

采用cDNA芯片检测样品中的基因表达量时，除了待测试的样品外，还需要准备用于比较的对照样品，比如当待测试的样品是癌组织时，对照样品可以采用癌组织旁的正常组织样品，也可以把多个样品混在一起形成对照样品。从测试样品和对照样品中分别提取出mRNA，经过反转录得到cDNA，并进行荧光标记。一般对照样品和测试样品分别用绿色（Cy3）和红色（Cy5）进行标记，再等量混合后与cDNA芯片上的探针竞争杂交，杂交后的芯片经过洗涤，用激光共聚焦显微镜扫描。显微

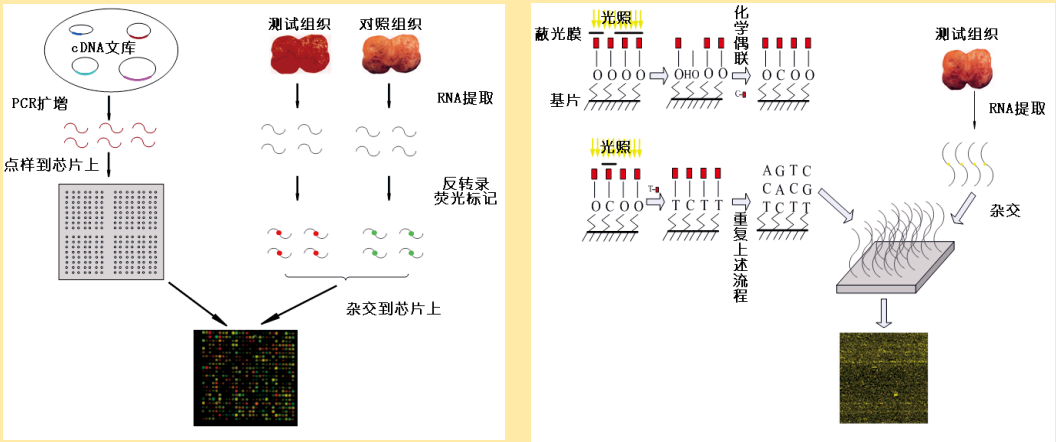


图1 基因芯片的基本原理示意图。（A）cDNA芯片；（B）寡核苷酸芯片。

镜通过发射两种不同波长的激光分别激发Cy3和Cy5，因此通过扫描可以得到对应荧光强度的图像。两图像合成以后，探针对应的基因如果在测试样本中相对高表达，则显示为红色；相对低表达则显示为绿色；表达相当则为黄色；均不表达则无色（黑色）。红绿颜色的相对强度则反映出了基因在两种样品中表达的数量之比。cDNA芯片的制作及实验流程见图1（A）。

寡核苷酸芯片

寡核苷酸芯片的探针是采用光引导聚合的原位合成技术合成在芯片上的，它的发明得益于照相平版印刷技术与DNA固相合成技术的结合。在特定蔽光膜的作用下，通过对可活化的基片进行选择性的光照脱保护，就能在芯片不同位点合成所需要的寡核苷酸序列。与cDNA芯片不同，寡核苷酸芯片是在公司里以工业化的形式生产出来的，因此可以进行质量控制，有利于不同实验室之间数据的比较。寡核苷酸芯片技术最早是由美国昂飞（Affymetrix）公司发明的，现在还有其他几家公司也提供类似的芯片，如安捷伦（Agilent）、GE医疗集团（GE Health Care）、NimbleGen等。以昂飞公司生产的芯片为例，其芯片上每个探针为一个包含有25个碱基的核苷酸片断，同时有一个只有在中间一位上与之不同的错配序列片断，构成PM/MM[4]探针对（probe-pair）。一个基因由多个探针对来代表（通常是11～20个），构成探针组（probe-set）。一个探针组中的探针是根据有关生物化学规律从所代表的基因的序列中抽取出来的，这个过程叫做探针的设计，不同的设计可能会影响杂交的效率和特异性。

用寡核苷酸芯片检测样品中基因表达量的方法与cDNA芯片类似，但是它不需要对照样品，而只对待测试的样品进行杂交和检测。在cDNA芯片中之所以要引入对照样品，是因为从cDNA文库中提取探针和点样到芯片上时，很难精确控制每一个探针在芯片上的浓度，因此需要经过对照样品来读取样品中的基因相对于对照样品的表达量；而寡核苷酸芯片的探针是在芯片上合成的，可以准确地控制其浓度，因此只需要测试样品即可检测出其中的基因表达量。具体过程是，从测试样品提取到mRNA经反转录标记后与芯片杂交，这里所用的标记通常是生物素标记。杂交后的芯片经洗涤后通过扫描得到图像，图像是单色的，每个探针在

[4] PM：perfect match完全匹配；MM：mismatch有一错误位点匹配

图像上的亮度反映了该探针检测到的mRNA的表达水平，通过对探针组中多个探针数据的综合即可得到相应基因的表达量。寡核苷酸芯片的制作及实验流程图如图1（B）所示。

由于cDNA芯片采用的是点样技术，各个实验室可以根据实际需要自行设计探针。但这一特点既带来灵活性，也意味着不同实验室、甚至不同操作员所得到的数据的异质性比较高，数据的质量主要依赖于实验室的技术水平。而寡核苷酸芯片都是直接从芯片公司购买的，因此标准比较统一，数据的可重复性好，当然，其成本也相对较高。另一方面，cDNA芯片由于在同一张芯片上对两种样品进行竞争杂交，用一张芯片就能得到对两种样品的比较；而对于寡核苷酸芯片，则必须用两张芯片才能完成一次比较。但是，由于cDNA芯片输出的只能是基因表达在两个样品间的比值，在对多个样品进行比较时，选择不同的比较样品会影响比较的结果。当然，由于两种芯片在原理和实验过程上的不同，因此在数据处理和分析上除了一些共性的问题外，也各自有很多专门的问题。

基因芯片的图像处理

上述两种基因芯片都需要通过扫描仪对杂交后的芯片进行扫描，根据图像上的亮度信息检测探针上信号的强度，因此，从扫描得到的图像中提取探针和基因的表达数据只是基因芯片数据处理的第一步。由于cDNA芯片并非一家生物技术公司的产品，因此存在多种与其相关的图像处理软件，既有基因芯片扫描仪所伴随的处理软件，也有一些专门的商业化软件、公开的免费软件和世界各地的研究者自己研发的软件。cDNA芯片的图像处理流程大致包括以下三个步骤：（1）标定每个探针所在的图像区域；（2）对探针所在的区域进行图像分割，得到前景（杂交信号）区和背景区的象素；（3）量化杂交信号与背景信号的强度值。现有的软件大多都能够输出各个基因的两个颜色通道的杂交信号强度与背景强度、数据可靠性估计等参数。一些学者仍在不断研究cDNA芯片的图像处理问题，有兴趣的读者可查阅相关文献[1]。

对于寡核苷酸芯片，由于昂飞公司已经将图像处理整合到芯片的实验设备中，所以用户通常可以直接获取图像处理后的探针数据，包括各探针对上的PM和MM信号强度及可靠性指标等。由于每个基因由多个探针组成，因此需要采用一些统计和模型估计方法才能得到基因的表达值。昂飞公司提供的基因芯片软件可以从探针数据计算出各个探针组（基因）的表达值，另一个常用的计算表达值的软件是dChip。考虑到探针间和芯片间可能的差异，从探针数据计算基因表达数据通常是与归一化一起进行的。其他类型的寡核苷酸芯片的图像处理情况与此类似。

芯片数据的低层处理

所谓低层处理（low-level processing）是指在得到图像处理数据之后如何更好地计算基因表达值的问题。由于实验过程中存在系统误差、实验误差等影响，因此必须对原始数据进行一系列低层处理，也就是数据的预处理，主要包括数据归一化、缺失值处理和野值剔除等。对于寡核苷酸芯片，从探针数据得到基因数据的过程也是一种低层数据处理。

数据归一化

在芯片实验中，由于存在染色效率差异以及实验方法固有的局限性等因素的影响，我们需要对芯片数据进行归一化。芯片数据归一化是对消除芯片系统误差、试验平台偏差等处理过程的统称，包括单个芯片内的归一化和多个芯片间的归一化。

cDNA芯片片内归一化的主要目的是减小由于两种颜色染色效率差异导致的系统误差。归一化采用的一个基本假设是：在测试样本与对照样本间大多数的基因是没有显著差异表达的，而在有差异表达的基因中，在测试样品中高表达的基因与低表达的基因在数量上也是大致相当的，因此芯片上所有基因的相对表达量应该是以0为中心的分布。这里的相对表达量就是两种颜色的表达量之比的对数。人们通常用两种颜色的对数比（相当于差）和对数积（相当于和）来考察每一个基因，画出如图2所示的M-A图。理想情况下，多数基因应该分布在图上水平的中心线附近，而在图中的例子里，由于染色效率的差异，该芯片数据的相对表达量分布有总体向上偏移的趋势，而归一化就是通过适当的运算，将M调整为以0为中心的分布。最简单的方法是将所有的芯片数据减去数据的均值或者中值[2]。但是，由于荧光染料的染色效率还受基因的实际表达量大小A的影响，因此有必要对不同A值的基因进行局部加权回归，这就是流行的Loess方法的出发点[3]，还可以以芯片上局部网格为单位进行基于信号强度的局部加权回归。在基因组上有一些基因是与最基本的细胞活动有关的，这些基因的表达在不同组织里和不同条件下通常变化不大或没有变化，被称作看家基因（Housekeeping genes），人们也研究了很多方法，以这些看家基因为基准来对两种颜色之间的差别进行归一化处理。

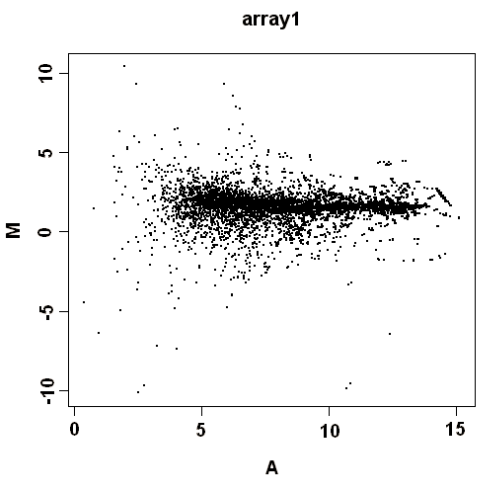


图2 cDNA芯片的M-A图举例

由于实验操作、样品准备等方面的差异，因此在不同芯片间有时会存在系统的差别，从而需要进行片间归一化。各种类型的基因芯片都存在这种归一化问题。片间归一化的核心思想就是确定一个基准芯片，以芯片之间的一些不变量为依据，对其他芯片的数据进行整体的拉伸或者压缩变换。采用不同的归一化依据就产生不同的方法：比如在最简单情况下，假设每一张芯片上基因表达的均值应该是相同的，这就是所谓的总体归一化方法；这种假设过于粗糙，后来便产生了基于看家基因等的归一化方法，即以这些特殊基因为基准对芯片间的表达值进行变换[4]；进一步的研究又发现，有时不易得到看家基因，而且有些看家基因的表达在各个芯片中也不是恒定不变的，因此又发展了以那些在多个芯片内排序比较固定的基因为基准进行归一化的方法[5]。

对于寡核苷酸芯片，李（C. Li）和王（W. H. Wong）提出了一种从探针对数据拟合基因表达值的模型（Li-Wong模型）[6]，并以某一用户选定的芯片为基准，通过从全部PM探针中选择排序不变子集来在芯片间进行探针的归一化，再利用归一化后的探针数据估计基因表达值。

缺失值处理和野值剔除

在芯片数据的获取过程中，由于实验、扫描或前期处理中的不完美因素，可能在某些探针上会得不到数据，或者得到一些比较奇异、与其他数据差异过大的值，前者称为数据的缺失，后者则称为数据中存在野值。

由于很多数据处理的方法不适用于有缺失值的情形，因此需要对缺失值进行合理有效的处理。现有的处理方法包括：直接去掉含有缺失值的基因，这种方法保证了余下数据的准

确性，但也可能因为一个数据的缺失而丢失一个基因的全部数据；另一种处理方法是将缺失值置零，这种方法虽然简单易行，但是把所有缺失的基因表达都算做0显然会带来很大的噪声；一些比较完善的方法采用最近邻和最小二乘的思想，用和缺失值所在基因表达模式相近的其他基因的数据，对缺失值进行插值估计。

野值的存在也会对数据的后期处理产生很大影响。现有的检测野值的方法不多，比较有代表性的是李和王提出的用寡核苷酸芯片探针表达模型检测探针、基因和芯片野值的方法[6]，能够有效地检测出多种类型的野值。在很多基因芯片数据分析中，人们会设置一定的基因表达值的上限和下限，高于上限和低于下限的数值都被认为是超出了仪器可信范围而被分别设置为上限值和下限值，这其实是最简单的检测和处理野值的方法。

芯片数据的高层分析

生物体能够发育成具有不同形态、不同功能的组织，源于基因在不同时间与空间上的特异性表达。如果有某些基因表达出现异常，则会影响机体的正常生理活动甚至引发病变。同样，如果机体出现某些病变，则通常也会在一些基因表达上有所反映。基因芯片在人类对象上一个最主要的应用是研究一些生理或病理状态下的基因表达变化。以癌症为例，典型的应用是通过基因芯片来比较癌变组织与正常组织的差异，或者癌症的不同亚型之间的差异，或者不同预后的病人之间的差异，等等。这种差异研究包括发现在所对照的两类（也可以是多类）中表达有明显差异的基因，发现与类别相关的基因表达模式，尝试用一些基因的表达模式来区分所研究的类。这些研究可以帮助人们理解疾病的机理、研究有效的疾病诊断和检测手段并且预测病人的预后，对于未来疾病的诊断和治疗有重要的意义。1999年，格鲁卜（T. Golub）等人在《科学》期刊上首次发表了用基因芯片研究癌症分类的成果，用寡核苷酸芯片得到的基因表达数据成功地将急性髓性白血病（AML）与急性淋巴细胞白血病（ALL）分开[7]，并列出了一些在两类中差异表达的基因，揭示出用基因芯片技术对癌症进行分类研究的可行性和巨大潜力。在此基础上，他们还通过对基因表达数据进行聚类分析，划分出了一个新的白血病亚型。目前，基因芯片已经被应用在几乎所有类型的癌症研究上，包括各种白血病及肺癌、肝癌、乳腺癌、前列腺癌、卵巢癌、淋巴瘤、恶性黑色素瘤、胃癌、直肠癌等等。

除了应用于人类疾病研究外，基因芯片还被广泛应用在对多种模式生物的研究中，其中有些研究与对癌症的研究思路相似，即通过基因芯片区分生物体的不同状态，并发现与这些状态相关的基因。另外一种典型应用是对基因功能和基因调控网络的研究。其中最有代表性的是对酵母细胞周期相关基因的研究，用基因芯片获得酵母全部基因在细胞周期的不同时期的表达谱，从这些表达谱中识别与细胞周期表达一致的基因，并进一步研究基因在细胞周期的不同相位的表达情况。还有一种典型应用是对多个基因进行逐一地敲除或扰动，用芯片观察其他所有基因在相应情况下的表达情况。利用这些基因表达的时间序列数据和/或基因扰动下得到的多种基因表达数据，人们开展了大量构建基因调控网络、基因关系网络的研究工作。

基因芯片杂交后扫描得到的是图像，经过图像处理后得到探针的杂交信号，再经过上面介绍的低层处理后得到基因的表达值（对于cDNA芯片来说，通常是测试样品与对照样品上基因表达的比值）。有时为了运算方便或者为了使数据分布更接近正态分布，人们可能对所有基因表达数据取对数。经过低层处理，一张基因芯片得到的数据成为一个向量，维数就是芯片上基因的个数。将一次实验的多张芯片数据放到一起就形成一个矩阵，通常其一列为一张芯片的数据，比如

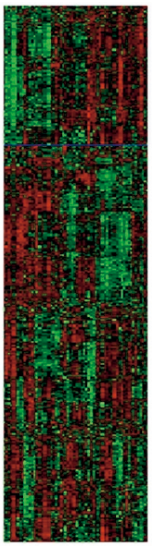


图3 基因芯片数据的热度图（heatmap）示例，图中小方块表示基因在样本中的表达值，每一行为一个基因，每一列为一个样本，红色表示表达值高，绿色表示表达值低，黑色表示表达值居中。

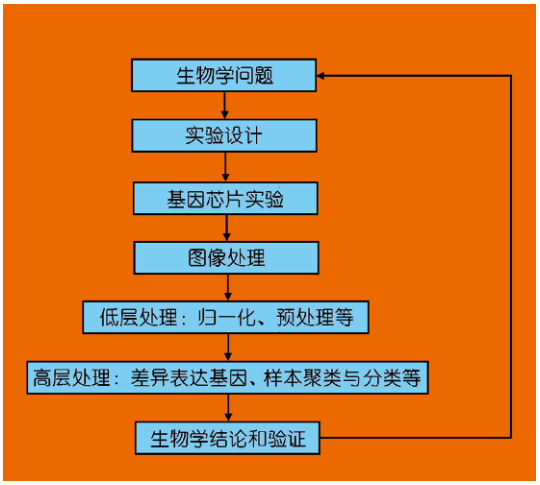


图4 典型的基因芯片研究流程[参考杜瓦德（S.Dudoit）和金特勒曼（R.Gentleman）的讲义]

一位病人或细胞周期的一个时间点上的基因表达值向量，而一行为一个基因。为了显示方便，有时人们用颜色代表基因表达值，把一个基因芯片数据集用如图3所示的彩色矩阵表示出来，有人将其称作热度图（heatmap）。所谓基因芯片数据的高层处理（high-level processing）就是指在得到基因表达值后对于疾病的分类、基因的差异表达、基因间的相关性、基因的表达模式和关系网络等方面的研究。所谓低层、高层只是习惯上的叫法，它们对于芯片数据的分析和应用同样重要。前者的主要目的是从芯片原始数据获得可靠的基因表达值，而后者的主要目的则是从基因表达数据分析生物现象并得出结论。图4给出了一个典型的基于基因芯片的研究的基本流程。

寻找差异表达基因

在两类样本之间寻找具有显著差异表达的基因是一种最常见的高层处理问题。最基本的方法是统计中的假设检验方法，即检验在两类中基因表达值的分布没有差异的空假设下，得到实际数据中所观察到的差异的概率（p值），比如t检验、秩和检验等。如果只检验一个基因，可以按照假阳性率5%来做判断。但是，由于基因芯片上有成千上万个候选基因，对这成千上万个基因做假设检验，即使假阳性率为5%，也可能会导致上百个虚假的发现，因此需要采用多检验的方法进行校正，比如以所有次检验中给出一个错误的概率作为显著性判断准则的邦弗朗尼（Bonferroni）校正[8]，这个概率称作家族错误率（family-wise error rate）。由于这种校正过于保守，人们又提出用所谓误发现率（False Discovery Rate，FDR）[9]来进行判断，即控制在所有所作出的判断中错误判断所占的比例。从基因芯片数据大规模涌现出来后，人们陆续提出了多种改进的检验差异表达基因的方法，比如萨姆（SAM）方法[10]等。

需要指出的是，在对cDNA芯片数据的处理中，从单张芯片上的红绿杂交信号即可算出两个样本上基因表达的差异。比如当红绿信号的比值在2以上或1/2以下时，判断该基因在两种杂交样品中的表达有差异，即所谓的二倍（2-fold）变化准则（当然也可以根据具体情况采用其他倍数作为准则）。由于这种差异表达是通过单次观测发现的，只能提供对差异表达的一个基本判断，无法估算置信度和显著性。cDNA芯片的数据也可以和寡核苷酸芯片一样，在两类样本间检测差异表达基因及其统

计显著性。

基因的过滤选择

由于基因芯片上的基因维数可能高达上千甚至上万维，给后续的很多处理带来困难，因此人们往往在后续处理之前首先对基因进行一定的筛选。利用上面所介绍的方法选择差异表达基因就是一种可以用来筛选基因的方法，比如在选择出的差异表达基因基础上对样本进行分类。需要注意的是，由于在选择差异表达基因时利用了样本分类的信息，在对这样构造的分类器进行性能评估时，如果采用交叉验证等策略，需要注意到前面的基因选择也需要被交叉验证，否则会因为信息的泄漏而导致评估结果偏于乐观，这种情况在样本数目较少时尤其严重。在部分已经发表的文献中也存在这种错误。

有一些不依赖于分类信息的基本的基因筛选方法，比如二倍变化准则，即考察基因在所有样本上的表达值的分布，如果一个基因的最大值比最小值的变化也不超过两倍，则可以认为该基因在所有样本上没有明显差别，因而可以被滤掉（根据具体情况也可以把倍数阈值调低或调高）；由于芯片和扫描的灵敏度限制，基因表达值低到一定程度可能就不再可靠，因而也可以把所有表达值低于某一阈值的基因认为是没有检测到的基因而删掉等等。不论后续的处理是样本的分类、聚类还是对基因调控网络的研究，适当的基因筛选通常是需要的，这一筛选往往能把维数从数万维降到几千维。

根据芯片数据对样本分类

基因芯片最多的应用就是对疾病样本的分类。在这类问题中，我们面对的是一些类别已知的训练样本，希望建立用基因表达数据来对疾病进行分类的分类器，是典型的监督模式识别问题。对于分类问题，所用的特征对分类器的效果都有着至关重要的影响。基因芯片包含成千上万个基因，而一般来说，由于各种因素的制约，样本量都相对较小，通常只有几百甚至几十个样本，这样，特征选择就变得尤其重要。

用于分类的特征选择可分为过滤（filtering）方法和包裹（wrapper）方法两大类。

上面介绍的基因筛选方法在用于后续分类时就是过滤方法，其特点是采用与后续的分类器算法无关的基因过滤准则。这种方法的优点是不针对专门的分类方法，因而可以适用于多种分类器；但是，这种单独的过滤准则有时与最终所要采用的分类准则不一定一致。作为过滤方法的特例，有人还采用主成分分析方法对数据进行变换，即选择最主要的主成分（由于它们是原有基因的组合，有人称之为“超基因”），然后在主成分空间里进行分类，或者根据在所选的主成分中原有基因各自的贡献大小选择对应的基因。

包裹方法是以分类器的性能作为特征选择的依据，通过迭代的方法选出一组对分类性能最好的特征。比如在用线性支持向量机（Support Vector Machine，SVM）作为分类器时，可以用训练后各维基因的权值大小，或者加权后各基因在两类间的分离程度作为依据对基因排序，选择一个基因子集进行下一步训练，然后再排序、再选择。代表性的方法有盖恩（I. Guyon）等的SVM-RFE[11][5]和张（X. Zhang）等的R-SVM[12][6]。对于非线性SVM或人工神经网络等其他分类器，也可以通过灵敏度分析得到分类结果对各维基因的依赖程度，依次作为选择基因的准则。

基因过滤方法所采用的准则通常是单基因的，无法考虑基因间的组合效应。包裹方法由于所采用的分类器考虑了基因间的线性或者非

[5] Recursive feature elimination based on support vector machine

[6] 递归支持向量机

线性组合关系，往往能达到更高的分类性能。但是考虑了组合关系后，基因选择的空间大大增大，使得假阳性选择的风险性也增加了，而且由于基因间组合模式的生物意义不如单个基因的差异表达更直接，因此也更难以进行实验验证。

基本上在模式识别领域中的各种常用分类方法都可以用在基因芯片数据的分类中。像费希尔（Fisher）线性判别、k近邻法[7]等通常与基因过滤方法结合使用，而支持向量机、人工神经网络等更复杂的机器学习方法则经常与包裹方法一起使用。当然，在一个实际应用中，可以多种方法混合使用，比如首先用一些一般性的准则进行基因的粗过滤，然后再用包裹方法进行精细的基因选择和分类。由于样本数少、维数高，很多传统方法容易出现过学习现象，需要特别注意。支持向量机方法因为在小样本下具有较好的推广性能而备受青睐，但与简单的差异表达基因分析相比仍然存在过学习的可能。

由于目前的基因芯片数据都是样本数很少，多数情况下没有独立的测试样本集，分类的性能一般采用交叉验证或者重采样的方法进行评估。需要注意的是，由于样本数极少，很少的样本变化也会影响基因的选择，所以在交叉验证时需要把基因选择步骤一起纳入验证范围内，防止信息泄漏而引起的过于乐观的估计。在采用重采样方法进行评估时，也可以采用一定的校正（如B.632+校正）来调整评估结果。

分类树在基因芯片数据的分类中也得到了比较多的应用，其特点是一边选择基因一边进行样本的部分划分。这种方法十分有利于结果的生物学解释，但“过学习”的风险较大。为此，布雷曼（L. Breiman）提出了所谓随机森林（Random Forest）方法[13]，通过样本重采样得到大量分类树，再将分类树融合，得到最后的分类器。

在用基因芯片数据对样本进行分类的研究中，一个突出的特点是基因选择的目的不仅仅是为了构造分类器，而且还为了发现疾病分类背后的基因表达规律。这一特点导致模式识别技术在这类问题中的应用与其他领域有所不同。目前情况下，用基因芯片进行疾病分类研究的主要目的还在于研究疾病的规律、大规模筛选可能的标志基因和为更进一步的实验研究提供指导，直接的临床应用研究尚不成熟。

聚类分析

非监督的聚类分析在基因芯片数据分析中有非常广泛的应用。有时会发现，同一类的癌症患者在基因表达模式上还可能存在着一些子类，发现这些子类对于更深入地理解这些疾病有非常重要的意义。聚类分析在挖掘这些信息中可以发挥重要的作用，在基因芯片数据上常用的聚类方法包括：分级聚类方法、K均值聚类方法和自组织映射（Self Organization Map，SOM）等。分级聚类算法的优点是非常直观，我们可以直接通过聚类树（dendrogram）观察到样本之间的相似程度，埃森（Eisen）等最早将分级聚类方法应用于基因芯片分析，并引出了将聚类数和颜色表示的芯片表达热度图共同显示的可视化方法[14]，如图5所示，成为基因芯片研究中最常用的图示方法。除了对样本的聚类，还可以分别对基因进行聚类，即所谓二维聚类，图5的例子就是对基因和样本分别举行聚类后的显示。如果同时对基因和样本进行聚类，就是所谓的双聚类（biclustering）问题。

与监督分类方法类似，应用聚类方法之前往往也需要首先进行基因的选择，降低基因维数，通常采用的选择准则是诸如二倍变化准则等非监督的准则。双聚类问题有些类似于监督分类时的包裹方法，即在聚类的同时选择对聚类最有利的基因组合。

[7] 最近邻法的扩展，其基本规则是，在所有N个样本中找到与测试样本的k个最近邻者。

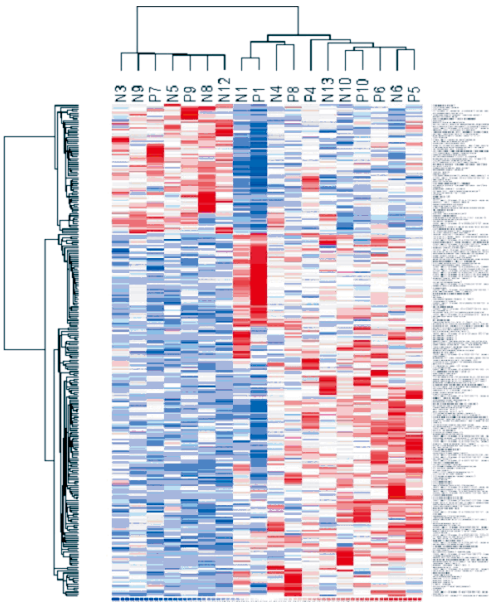


图5 对样本和基因进行分级聚类的例子，图中红色代表高表达，蓝色代表低表达。

基因芯片数据的可视化

由于基因芯片数据维数高，非常需要有效的可视化手段来展示数据的结构和样本之间的关系。前面的热度图、分级聚类数等都是常用的可视化手段。另一种常用的可视化手段是多维尺度变换（Multi-Dimensional Scaling，MDS），它把高维数据非线性映射到二维或三位空间来显示，使得在显示空间中样本之间的距离关系能够最佳地反映样本在原高维空间中的关系。图6是一个把样本的基因表达数据用MDS映射到三维空间显示的例子。

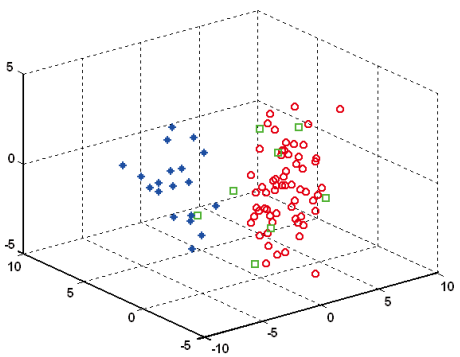


图6 用MDS展示样本在基因表达空间中的关系，图中红色圆圈、蓝色方点和绿色方框分别代表三种不同的癌症病例，它们之间的关系可以清楚地展现出来。

对基因间相互关系的研究

生命体是复杂的大系统，其基本的生命活动都不可能是单个基因作用的结果，理解基因之间复杂的相互作用，对于研究很多生命过程的规律都是至关重要的。基因芯片数据在这方面也发挥了重大的作用。对基因的聚类分析是研究基因间关系的最基本的手段，人们把具有相同或相似表达模式的基因称作共表达的基因，通过聚类分析把共表达基因聚类，根据共表达基因具有相似的功能，或者具有相同的调控元件等假设，从聚类的基因中寻找基因间的功能和调控关系，并且从共表达的基因上游调控区，寻找可能的共同的调控元件，进而构建调控网络。还有很多研究是根据基因在不同时间或者不同扰动因素下的表达谱，利用如贝叶斯网络等方法构建基因关系网络。

由于基因之间的调控关系错综复杂，基因芯片所能反映的信息有限，因此从基因芯片数据研究基因间关系通常还需要很多其他的先验知识，比如文献中对基因功能和调控关系的报道、GenBank数据库中的基因注释、GO数据库中的基因功能分类等等。由于文献和数据库中的基因注释信息非常复杂，研究用自然语言理解等智能化的方法自动挖掘生物文献和数据库中的知识，并且与基因芯片数据的分析相结合，是当前研究基因关系网络的一个重要方向。

芯片数据的数据库

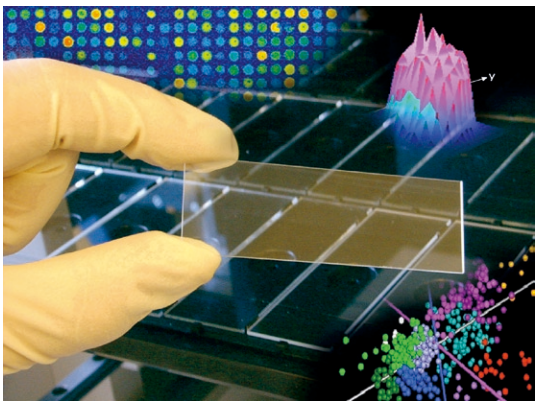
基因芯片的实验都是分布在世界各地的实验室分别进行的，很多实验数据在文章发表时就被公开了，也有很多数据库专门收集了各种芯片数据，这些公开的数据为人们的研究提供了很大的方便。这些数据库根据用途其基本

可以分为两类：一类是一些具有专门生物学意义的专用数据库，比如关于心血管疾病的专门数据库、关于乳腺癌的专门数据库等，这些数据库是人们在相应领域开展深入研究的重要资源；另一些数据库则提供用以测试各种分析算法的数据，并不专门针对特定的生物学问题，比如考普（L. M. Cope）等搜集了一些昂飞的芯片数据用来测试各种算法的性能（http://affycomp.biostat.jhsph.edu）。有些更大规模的数据库是根据芯片的类型来分类和收集的，比如斯坦福大学的SMD数据库（http://genome-www4.stanford.edu/MicroArray/SMD/）、麻省理工学院（MIT）的寡核苷酸芯片数据库（http://web.wi.mit.edu/young/chipdb）等，前者收集的都是cDNA芯片的数据，包括人、酵母、大肠杆菌、拟南芥等模式生物的约280组芯片实验的数据，大约包含20亿个探针；而MIT的芯片数据库收集的都是寡核苷酸芯片的数据，主要是一些涉及人类癌症，包括人的白血病、淋巴癌、肺癌、前列腺癌、上皮癌等近50组不同癌症实验的芯片数据。一些实验室还进一步开展了将多个数据集合进行联合研究的工作。

基因芯片的最新进展

一直以来，人们将注意力主要集中在DNA编码区的研究中，传统的基因芯片主要都是针对基因编码区域的研究来设计的。近年来，人们对基因组中的非编码区越来越重视，发现其中不但包含有基因的转录调控区域，还包含很多非编码的转录体，它们具有重要的调控功能；同时，选择性剪接的调控和功能一直得到很大的关注，一些其他的调控因素，如DNA甲基化等表观遗传学因素、染色体拷贝数异常等，也越来越显示出其重要性；而产生各种调控与表达差异和变化的原因是与人群基因型的多样性相关的。因此，最近两三年来国际上出现了多种新的基因芯片，包括覆瓦芯片（Tiling array）、外显子芯片（Exon array）、启动子芯片（Promoter array）和单核苷酸多态芯片（SNP array）等。

覆瓦芯片通过在基因组上“覆瓦式”高密度地设计DNA探针，能够有效地获得全基因组上的DNA或者RNA数据，可以用于研究编码基因、非编码区及编码区反链的转录活动，还可以与蛋白质免疫共沉降（Chromatin Immunoprecipitation，ChIP）技术联合检测全基因组上的转录因子结合位点、DNA甲基化等。目前的覆瓦芯片在全基因组或者整条染色体非重复区域每隔3个或5个碱基设计一个寡核苷酸探针，在检测全转录组的应用中，可以利用随机引物把所有RNA转录本在体外反转录成cDNA，标记后与芯片杂交，因此能够覆盖整个基因组或染色体上的所有转录区域，发现大量未知的转录体，是研究非编码RNA基因的最新重要技术平台。目前，已有报道利用覆瓦芯片技术研究人的10条染色体、水稻的全基因组、拟南芥和果蝇等物种部分基因的转录信息。另一种典型应用是所谓ChIP-on-chip应用，人们提出过采用蛋白质免疫共沉降技术获取全基因组上与所研究的转录因子相结合的DNA片断，然后通过覆瓦芯片检测这些DNA片断在基因组上的分布，从而得到这些转录因子所调控的全部基因。



在真核生物中，基因在转录后通常要经过剪接加工（splicing），剪掉成为内含子（intron）的部分，将保留的称为外显子（exon）的部分连接起来成为成熟的mRNA，然后才翻译蛋白质。一种重要的转录后调控机制是选择性剪接或可变剪接（alternative splicing），即同一个基因可能有不同的剪接方案，导致产生多种不同的蛋白质产物，这被认为是高等生物蛋白质复杂性的一个重要来源。在不同的组织或者不同的发育阶段，基因可能有不同的选择性剪接，即相同的基因在不同的组织或不同的发育阶段，由于可变剪接的作用而翻译出不同的蛋白质产物，从而发挥不同的作用。传统的分子生物学方法难以大规模研究这种可变剪接的机制，在cDNA或传统寡核苷酸芯片中也很难检测同一基因不同的剪接型。外显子芯片（exon array）在基因组上以外显子为单位设计探针，不但可以更精细地检测基因的表达，更可以帮助研究者用大规模实验手段来研究可变剪接事件。

启动子（promoter）是基因上游的一个重要调控区域，转录复合体结合到基因的启动子区后启动对基因的转录。最近，科学家用ChIP-on-chip技术专门设计的启动子芯片，成功地实现了全基因组上在人纤维原细胞中处于活动状态的启动子区域，为研究转录机制、染色体结构、发现新基因和新的转录体、研究基因表达间的功能关系等提供了有效的方法。

人类基因组上的单核苷酸多态性（Single Nucleotide Polymorphism，SNPs）是科学家研究的另一个热点，这是由于其中蕴含着揭示人群多样性规律、探索复杂遗传疾病机制和研究个体化医疗的重要信息。SNP芯片是高通量观测大量SNPs位点的新技术，目前已经能够在一张芯片上测50万个SNPs，为大规模人群基因组多样性研究提供了有力的工具。除了可以用来进行疾病的基因型遗传连锁分析，SNP芯片还可以用来高通量地检测染色体拷贝数异常，人们已经发现一些基因的扩增或缺失在癌症等疾病中扮演着重要角色。

除了这些较通用的芯片，还有很多各种类型的专用芯片，比如小分子RNA（microRNA）芯片、甲基化芯片、人类线粒体芯片等等。

小结

基因芯片技术可以称为上世纪末生物技术领域最重要的进展之一，也是人类基因组计划的直接成果之一，在新世纪开始的短短几年里又取得了突飞猛进的发展。基因芯片技术是推动生物学研究走进系统生物学时代的重要因素和技术基础。

基因芯片产生的海量数据大大超出了生物学家依靠传统研究手段所能分析的极限，把计算技术和信息技术推向了生物学研究的前台。功能基因组学、系统生物学和生物信息学都已经成为密不可分的概念。在广大信息科学家热切地希望在这一领域大显身手的同时，我们也必须意识到，我们试图用基因芯片以及其他技术，去研究包括人自身在内的生物对象，其复杂程度远远超过了以往我们所面对的任何技术研究对象，与生命的无穷奥秘相比，基因芯片等高通量实验技术所获得的海量数据仍然是非常有限的观测，这为信息科学家提出了前所未有的挑战。我们期待着信息科学家和生物科学家更紧密地结合在一起，共同迎接这些挑战。

张学工

博士，清华大学教授，清华信息科学与技术国家实验室（筹）生物信息学研究部主任，清华大学自动化系信息处理研究所所长，中国计算机学会高级会员。主要研究领域是模式识别和生物信息学。

凡时财

清华大学自动化系博士研究生。

裴云飞

清华大学自动化系博士研究生。

参考文献

[1] Yang, Y.H. et al, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 30(4):e15.1– e15.10, 2002

[2] Chen, Y., Ratio-based decisions and the quantitative analysis of cDNA microarray images, Journal of Biomedical Optics, 2: 364–374, 1997

[3] Ihaka, R. and Gentleman, R. R: a language for data analysis and graphics. J. Comput. Graph. Statist, 5:299–314, 1996

[4] Hill, A.A. et al, Genomic analysis of gene expression in C. elegans. Science, 290:809–812, 2000

[5] Li, C. & Wong, W.H.. DNA-chip analyzer (dChip). In: Parmigiani G., et al, editors. The Analysis of Gene Expression Data: Methods and Software. New York: Springer, 120–141, 2003

[6] Li, C. & Wong, W.H., Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc Natl Acad Sci USA, 98: 31–36, 2001

[7] T. R. Golub, T.R. et al, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286: 531–537, 1999

[8] Hochberg, Y., A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75: 800–803, 1988

[9] Benjamini, Y. & Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Statistical Soc. Ser. B – Methodological, 57:289–300, 1995

[10] Tusher, V.G., Tibshirani, R., Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci, 98(9):5116–21, 2001

[11] Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3):389–422, 2002

[12] Zhang, X. & Wong W. H., Recursive sample classification and gene selection based on SVM: method and software description, Technical Report, Department of Biostatistics, Harvard School of Public Health, 2001

[13] Breiman, L. Random forests. Machine Learning, 45(1):5–32 , 2001

[14] Eisen, B. et al. Cluster analysis and display of genome-wide expression patterns, Proc Natl Acad Sci USA, 1998, 95:14863

[15] Cheng J. et al, Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 308: 1149–1154, 2005

[16] Kim, T.H. et al, A high-resolution map of active promoters in the human genome. Nature 436: 876–880, 2005

[17] Li, L. et al, Genome-wide transcription analysis in rice using tiling microarrays. Nature Genetics, 38: 124–129, 2006

[18] Urban, A.E. et al, High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. PNAS, 103: 4534–4539, 2006

[19] Gershon, D., DNA microarrays: more than gene expression. Nature, 437: 1195–1198, 2006