第五讲　生物医学信息处理
——DNA 微阵列数据在医学中的应用*

马尽文[1,†]　　邓明华[1,2]

（1　北京大学数学科学学院　数学及其应用教育部重点实验室　北京　100871）

（2　北京大学理论生物中心　北京　100871）

摘　要　飞速发展的生物信息技术为现代医学提供了更为有效的工具. 特别是随着人类基因组计划的基本完成和逐步细化，人们已经试图从基因水平上来认识生命现象，特别是一些重要疾病的机理. 由于生物特性一般都涉及到多个基因的共同表达，这便出现了同时衡量成千上万个基因的表现水平的所谓 DNA 微阵列技术与数据. DNA 微阵列数据也被称为大规模基因表达谱. 根据这些微阵列数据，人们不仅能够对一些疾病进行分析，并且还能够发现一些新的生物特性与规律. 另外，利用微阵列数据能够选取出疾病的相关基因并进行疾病的分类与诊断. 这项研究无疑将推动医学的发展. 最近，人们还进一步通过基因表达水平值来发现基因之间的调控方式，这将为疾病病理的研究与治疗提供更科学的依据.

关键词　生物信息学，DNA 微阵列数据，相关基因，肿瘤诊断，基因调控

Application of DNA microarray data to medicine

MA Jin-Wen[1,†]　　DENG Ming-Hua[1,2]

（1　*School of Mathematical Sciences and LMAM, Peking University, Beijing*　100871, *China*）

（2　*Center for Theoretical Biology, Peking University, Beijing*　100871, *China*）

Abstract　The fast developing technology of bioinformatics has provided a new efficient tool for modern medicine. In particular, since the accomplishment of the human genome project we can now try to understand the phenomenon of life from the point of view of genes, especially with regard to the mechanism of major diseases. Since a biological feature generally involves many genes, DNA microarray chips have been developed to detect the expression levels of thousands of genes simultaneously, the so-called microarray data or large scale expression profile. With these data we can analyze the diseases, and even find new biological features or orderly patterns. Moreover, according to these microarray data we can select the informative genes of a certain disease and construct a classifier or diagonal system for it, which will certainly promote the development of medicine. Furthermore, microarray data have been applied to the analysis of regulation and control among genes, which will be significant for the pathology and treatment of diseases.

Key words　bioinformatics, DNA microarray data, informative genes, tumor diagnosis, regulation and control among genes

1　引言

随着科学技术的快速发展，人们对于自身的认识在不断提高. 然而，人类本身的诞生、成长、疾病等生命现象还存在着大量的谜团. 至今，一些重要的生命现象依然无法得到科学的解释. 特别是在医学方面，一些重要的疾病还很难进行诊断与治疗. 近年

* 国家自然科学基金（批准号：60471054，90208022，10271008）、国家高技术研究发展计划（批准号：2002AA234011）、国家重点基础研究发展计划（批准号：2003CB715903）资助项目

2004－10－09 收到初稿，2005－03－04 修回

† 通讯联系人. Email：jwma@math.pku.edu.cn

来,快速发展的信息技术与现代医学越来越紧密地结合在一起,并产生了生物医学信息处理技术. 目前,生物医学信息处理主要包括医学图像处理与分析、计算机辅助诊断与治疗系统、医学信号的检测与处理和基于基因技术的生物信息学. 而生物信息学是一门崭新的综合性学科,并在医学应用上有着广阔的空间和前景. 实际上,最近生物基因技术的快速发展,使得人们能够在分子水平上认识生命现象并在一些方面得到了突破. 由于人类或者生物所提供的基因数据是巨大的,并且按一定的方式进行着复杂的编码,破解这些生命的密码则需要数理科学和计算机科学的共同努力. 这样便产生了生物信息学和相应的生物信息技术[1]. 这门新的学科不仅为最终揭开生命之迷奠定了科学基础,并且为现代医学的发展提供了一些崭新的工具.

实际上,基因技术很早就被应用到医学诊断中. 一些遗传性疾病的根源大致可归结为一个或几个基因出了问题. 也就是说,遗传疾病是有一个或几个基因所控制的. 这些基因的缺失或没有得到充分的表达,则会导致了这种疾病的产生. 然而,这种分析方法很难推广到一般性疾病,如心脏病、癌症等. 实验表明,一般疾病不是由单个或几个基因的表现所控制,而是由许多基因的共同表现所确定的. 因此对于一般疾病的研究则需要观察众多基因的表达水平. 根据这种实际需要,DNA 微阵列芯片技术应运而生并得到快速发展. 这种生物芯片可以同时检测到几千甚至上万个基因的表达水平值. 这些生物数据不仅能够为生物特性、疾病的分析和发现提供依据,并且能够发现与疾病相关的基因,并应用于疾病的分类、诊断和治疗. 进一步,我们还可以通过这些数据分析发现基因之间的调控关系,为疾病的治疗提供依据.

本文将对于生物信息学中微阵列数据在医学中应用进行综述. 本文在第 2 节中,我们将介绍微阵列技术的发展和数据的采集. 第 3 节将介绍基于微阵列数据的疾病聚类分析和新规则的发现. 第 4 节将介绍如何利用微阵列数据选取疾病的相关基因和疾病的分类与诊断. 第 5 节将介绍如何利用微阵列数据来推断基因之间的调控方式. 最后,第 6 节给出了简要的总结和展望.

2 微阵列技术与基因表达谱数据

我们首先介绍一下基因的生物学定义和作用. 在我们的每个细胞中,都包含着完全相同的遗传物质,即 23 对染色体. 而每条染色体是由 4 种碱基通过双螺旋结构对偶连接而成的 DNA. DNA 是遗传的物质基础,它决定了蛋白质的合成. 基因被定义为能产生一个特定蛋白质的 DNA 序列片断. 它是人类遗传的基本单元. 科学家估计,人类共有 3—4 万条基因并且很稳定. 然而,这些基因在每个人中的表现不同,这就造了人类之间的千差万别. 这里包括了不同人之间的差别和每个人不同时期之间的差别. 为了检测基因的表达水平,人们于 20 世纪 90 年代开发了 DNA 微阵列基因芯片,即基因芯片. 它能够同时测量出成千上万个基因表达水平值. 这些数据被称为 DNA 微阵列数据或大规模基因表达谱.

我们知道蛋白质的合成是由 DNA 决定的,但是基因不能直接翻译成蛋白质,而是通过产生一个 mRNA 中间体来进行蛋白质的合成. 从基因到 mRNA 的过程称为转录(transcription),而从 mRNA 到蛋白质的过程称为翻译(translation). 在转录阶段,细胞核中 DNA 带有的遗传信息通过碱基配对原则转录到 mRNA(信使 RNA)上,这些 mRNA 再通过 tRNA 进行蛋白质合成. 微阵列实验就是将一些荧光标记的 mRNA(即 DNA 单链)通过配对杂交对应到微阵列芯片上的 DNA 探针上,从而测量细胞当中不同基因对应的 mRNA 丰度(redundancy),即基因的表达水平值(也称为基因表达谱数据). 根据这些数据,我们便可以分析不同的人或人在不同条件下的身体特征或健康的状况.

目前,微阵列芯片产品主要有两种:Oligonucleotide(寡核苷酸)阵列和 cDNA(互补 DNA)阵列[2]. 最早的 Oligonucleotide 阵列由 Fodor[3] 等所开发. 它是基于照相平板印刷技术(photolithography)的 DNA 芯片,后来由 Affimetrix 商业化. 该技术类似于电子芯片制作技术,将 Oligonucleotide 探针(通常的长度是 20 或 25bp)密集地排列在片基上,其密集度可以达到在一片 1cm 多见方的片基上排列几百万个寡聚核苷酸探针. 实际上,Oligonucleotide 阵列的成本目前还很高. 另一方面,cDNA 芯片由 Stanford 大学的 Brown P O 实验室所发明. 它是以全长 cDNA 作为探针(通常的长度为 100—5000bp),并将探针附着在固体表面(如尼龙薄膜)上制作而成. 这种方法大大降低了制作成本. 而且在 1996 年,Brown 在网上公布了制作 cDNA 芯片的详细步骤,使得该技术被广泛采用.

在 cDNA 阵列实验过程中,mRNA 先转录成为 cDNA,并加上红色 Cy5 荧光标记;另外有一个参考

样本的 mRNA 也转录为 cDNA，并以绿色 Cy5 荧光标记. 混合以后与 cDNA 阵列上的探针进行杂交，然后扫描检测两种荧光强度. 通常以 log（Cy5/Cy3）作为目标的表达水平值，如图 1 所示.

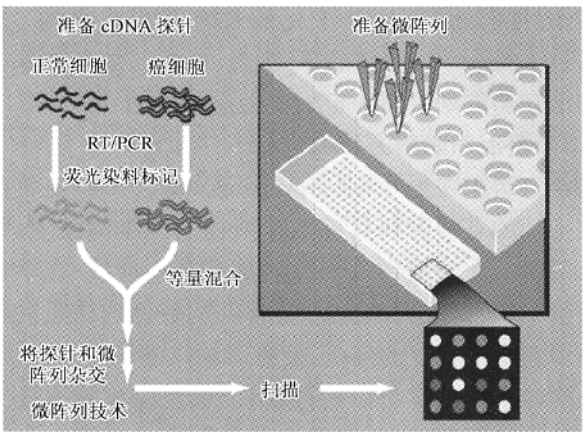

图 1　cDNA Microarray 技术示意图（来源于美国 NIH，National Human Genome Research Institute）

为了进行某项研究，我们常常需要不同的细胞样本进行 DNA 微阵列实验. 将这些实验的结果组合在一起，就得到一个大的数据矩阵. 我们通常以每行代表一个基因，每列代表一个样本（细胞组织），实验数据综合成一个矩阵 $A = (a_{ij})_{n\times m}$，其中 a_{ij} 表示第 i 个基因在第 j 个样本中的表达水平值. 这里所考察的基因个数为 n，而所具有的样本个数为 m. 微阵列数据有下述三个特点：首先，由于检测的仪器尺度的不同，我们一般需要对这些数据进行归一化处理. 目前已经有几种较成熟的方法，见文献[2]. 其次，由于实验条件不可能完全一样，数据中不可避免地存在着一定的噪声. 因此，在进行数据处理时，我们必须要考虑消除这些噪声. 第三，由于实验成本的原因，并且每个芯片只能使用一次，样本个数比较少，特别是相对于成千上万的基因个数. 在数据分析中，我们所面临的是一个非常典型的高维小样本问题. 这种情况无疑会给数据分析带来一定的困难.

3　数据聚类与知识发现

为了定量地分析某种疾病或生物特性，我们可以对所观察的研究对象做 DNA 微阵列实验，以获得一组微阵列数据或基因表达谱. 对于这些数据的分析主要是通过聚类分析方法进行的. 聚类分析又分为非监督的和有监督的. 非监督聚类分析方法是不需要微阵列数据之外的信息，直接对这些数据进行聚类，最终获得这些数据的分类规律. 实际中，这些所分出的类别会与一些生物特性相对应，为我们建立了它们的基因表达规律. 因此，我们将非监督聚类分类可看作一种生物知识发现或挖掘的技术. 另一方面，有监督聚类分析方法需要提供一定的指导信息，如是否有某种疾病等. 这样就可以将基因表达谱与指导信息之间的对应关系通过聚类分析来实现. 这将为疾病的诊断提供有效的模型和方法. 下面我们先介绍三种典型的非监督基因表达谱的聚类分析方法及其应用.

3.1　分层聚类（hierachical clustering）

分层聚类是应用最多的非监督基因表达谱聚类分析方法之一[4—8]. 根据分层聚类方法，我们将基因表达谱矩阵的每一列或者每一行看作一个向量（高维空间的一个点），根据这些向量之间的距离或者某种相关性度量进行聚类. 一开始，我们将每个向量看作一类，然后相近的两个向量逐步归并，直到将所有向量都归入一类结束. 分层聚类的结果是一个二叉树，其树叶对应于所有的向量. 这个树图表示了样本或基因之间的层次关系. 通过树图上的主要结果，我们可以发现样本或基因的分类规则. 上述聚类过程是从下向上进行的. 从算法上讲，我们当然可以采用从上到下的聚类过程，其结果是一样的.

分层聚类可分别对样本和基因进行处理，即将相似的基因聚类在一起或将相似的样本聚类在一起. 如果对每一节点上的两个类规定一个次序，便可以将所有点排列起来. 这样可将基因表达谱的行列重新排列. 若进一步将基因表达水平用不同的颜色表示，则重排后的表达谱将直观地显示出聚类的结果. 这种图示法更容易显示出基因和样本所具有的模式特征.

Eisen 等[4] 于 1998 年对芽殖酵母（saccharemyces cerevisiae）（大约 6200 个基因）和人体纤维原细胞（大约 8600 个基因）进行了自下向上的分层聚类. 他们引入了一种相似性度量，并选取不同时间点的细胞做 cDNA 实验，得到基因表达谱数据. 然后对行数据（与基因对应）进行分层聚类. 其结果排序采用简单方式，如平均表达水平. 他们发现了具有相似功能的基因确实聚集在一起了.

随后，Perou 等[5] 于 1999 年继续采用 Eisen 等的方法对乳腺皮细胞的基因表达谱进行了分层聚类分析. 他们所使用的微阵列中包含了大约 5000 个基因. 根据实验数据，他们对于基因进行分层聚类，并对某些重要基因的表达谱进行进一步的聚类，达到了对乳腺癌的合理分类. 另外，Alizadeh 等[6] 也采用了 Eisen

等的分层聚类方法,但所使用的距离度量是 Pearson 相关性. 他们研究了弥漫型大 B 细胞淋巴腺瘤[diffuse large B-cell lymphoma (DLBCL)],其中采用了正常和有病的样本共 96 个. 通过 cDNA 微阵列实验获得基因表达谱数据,其中基因个数为 4026. 分层聚类后发现正常和肿瘤的基因表达谱基本上是能够分开的. 进一步,他们使用了部分基因(即 germinal centre B-cell genes)的表达谱进行聚类,发现了 DLBCL 的两种类型,与临床观察是一致的.

Alon 等[7] 则采用了自顶向下的分层聚类方法. 他们使用 Affymetrix 微阵列进行实验,得到了 22 个正常样本和 40 个结肠癌样本. 基因个数达到 6500 多个. 在聚类时,他们挑选出了 2000 个表达最高的基因进行聚类. 在分层聚类过程中,他们还引入了确定性模拟退火机制. 在对不同样本的聚类中,正常和肿瘤的样本基本上被聚在不同的两个类中(其中有 8 个样本的聚类结果与实际不一致). 他们还考察了肿瘤样本与正常样本的差异是否取决于少部分基因的问题,结果是在只保留 500 个差别最显著(通过 t-检验)的基因的情况下,聚类仍然能够有效地将正常样本与肿瘤样本分开(有关具有差异的基因选取将在下节详细讨论).

Ding[8] 也采用了自顶向下的分层聚类方法. 但他将聚类过程转化为一个最优划分问题,并通过基于矩阵特征值的 Mcut 近似最优算法进行求解. 他对文献[9]中的白血病数据和文献[6]中的淋巴瘤数据分别进行了聚类分析. 对白血病数据集,他使用 t-检验挑选的 50 个差别最显著的基因的表达谱数据进行聚类. 结果将 38 个样本分为两类(一类含 11 个 AML,2 个 ALL;另一类是含 25 个 ALL). 对淋巴瘤数据集,96 个样本原来包含 9 个子类,去掉其中 3 个类(共 8 个样本),考虑剩下的 88 个样本. 他使用 F-检验选取了 200 个差别最显著基因,并根据这些基因的表达谱进行聚类分析. 结果这些样本被分为 6 类,只有 7 个样本出错.

3.2 *K*-means 聚类

K-means 聚类是一种传统的统计聚类方法. 在类别数 K 给定的情况下,该方法能够按欧氏距离将所有样本点自动地划分到 K 个类中. 该算法的基本思想是首先任意设定 K 个类中心的初始值,然后分别计算每个样本与各个类中心的欧氏距离,并将它归到距离最近的类中心代表的那一个类. 再计算每个类中样本点的平均点,并以此取代原来的类中心. 依次下去,直到类中心都不再变化,算法终止,并得到了分类结果.

实际上,Tavazoie 等[10] 便采用了 K-means 聚类方法对于酵母菌的基因功能进行了研究. 通过实验,他们获得了 15 个不同时间点上 6220 个基因的表达谱. 他们选取了其中变化最大的 3000 个基因,并通过 K-means 方法聚成 30 个类. 聚类的结果与基因的已知功能划分在一定程度上是一致的.

3.3 自组织映射方法

自组织映射(self-organizing map,SOM)是将高维空间的点映射到具有拓扑结构的二维网格上. 输入的每一个分量到每个结点都具有连接权. 在学习过程中,获得最大响应的结点以及附近的结点的权值得到不同程度的修正,使得对输入的响应增加. 对不同样本反复学习以后,网格能够反映出输入模式的结构特征,并自动完成样本或分类.

Golub 等[9] 则使用 2 个结点的 SOM 对 38 个白血病样本进行学习,得到两个类分别是 24 ALL + 1AML,10AML + 3ALL. 当他们再用 4 个结点的 SOM 学习同样数据时,发现将 ALL 的两种类型:B 细胞 ALL 和 T 细胞 ALL 也分开了. 另外,Tamayo 等[11] 用一个 6×5 阵列的 SOM 对酵母数据进行学习,得到了一些具有生物学意义的结果.

上述三种聚类方法是对基因表达谱数据进行非监督聚类方法的典型代表. 实际上,许多传统的非监督聚类方法都被应用到基因表达谱数据的分析上. 尽管在对疾病或生物特性方面已经取得了许多有意义的结果,但传统的非监督聚类方法在基因表达谱分析中却存在着下述三点不足:(1)当对不同样本进行实验获得基因表达谱时,无疑存在着噪声的干扰. 但现在对于噪声还没有很好的处理方法,仅能做的就是对每个样本的基因表达谱进行归一化处理. (2)在对基因表达谱数据进行聚类时,不管对基因还是对样本,所考虑向量的维数都相当高,而样本个数却相对较少. 对于这种情况,很多方法是无法使用的,而且即使能够直接使用,其效果也很不稳定,并且分类的性能也很难评价. (3)传统的非监督聚类都需要给定数据中的类别个数,否则聚类是无意义的. 而实际中会出现数据中的类别数是隐含的,但我们很难明确知道这一信息. 这种情况下的聚类就变得相当困难. 尽管已经有一些算法可以解决这一问题,但对于如此高维的数据依然很难执行. 这三点是目前非监督聚类方法无法或难于克服的问题. 因此,基因表达谱的分析迫切要求我们建立新的更有效的非监督聚类方法.

4 相关基因选取与疾病诊断

在进行疾病研究中,我们通常积累了一定数量的样本,并且知道它们的诊断结果(有病或无病,或属于哪一类型的疾病). 根据这些样本并通过 DNA 微阵列实验,我们可以得到它们的基因表达谱. 在这种情况下,我们自然希望通过对这些数据及其诊断结果的学习建立一个合理的疾病诊断(即分类模型),能够对新的病例进行诊断和分析. 这就是基因表达谱的有监督聚类或学习问题. 实际上,人们在这方面已经做了许多努力,一些典型的有监督的学习方法都被用到了. 这方面的研究主要是解决两个问题. 首先,如何在大批的基因中选取出与所研究的疾病相关的基因. 相关基因也被称为疾病的信息基因或具有在疾病与正常样本上表现出显著差异的基因(简称具有差异的基因). 但实际中,为了应用的广泛性,DNA 微阵列一般都设计了几千甚至上万的基因点. 对于一种疾病来说,与其相关的基因个数一般远低于阵列中的基因个数. 也就是说,许多基因的表达水平值对于该疾病的诊断不仅没有用,而且变成为一种噪声,干扰着诊断的结果. 因此,为了疾病诊断模型的合理建立,我们必须预先将相关基因选取出来,并将无关的基因及其表达谱数据去掉. 另外,相关基因的选取也是疾病病因研究的基础,具有重要的生物学意义. 其次,如何选取合理的非监督聚类方法,建立诊断模型. 目前,人们已经发现了一些较好的聚类方法可用于诊断模型的建立,并且获得了很高的正确率. 本节将从这两方面介绍一些典型的方法和应用结果.

4.1 相关基因选取

相关基因选取最常用的方法就是单基因打分法. 按照一定的度量,对每个基因与所考虑疾病的相关性打分,并按分数排列,最后选取出分数最高的一组基因作为相关基因组. 实际中,相关性的度量,即打分的方法,是衡量每个基因的表达水平值在两类样本上分布的差异性. 这显然是合理的,因为当一个基因与疾病相关性越强时,这两个分布的差异就越大. 相反,若这两个分布是一致的,这个基因的表现对于类别(或者说疾病)是不起作用的. 我们一般选用能够反映这两个分布差别的统计量进行比较或检验.

Golub 等[9] 提出的打分方法是衡量每个基因的表达水平值在两类分布的中心值(均值)的差,并进行适当的归一化. 也就是说,这两个分布的中心之间的差别越大,该基因对于疾病越重要. 他们从分数的正负两端选择相关基因. 随后,Furey 等[11] 采用了两类分布的中心值(均值)差的绝对值进行打分,选择最高分的基因为相关基因.

Dudoit 等[13] 提出的基因打分法建立于类别之间的方差与类别本身的方差之比,同样也反映了两个分布之间的差异,具有合理性. Bendor 等[14] 则根据每个基因表达水平值在两类样本上线性可分的程度来打分,也反映了两个分布的差异. 进一步,Bendor 等[15] 还提出了按两个分布的互信息量打分的方法. 这是因为当两个分布差异大时,呈现出相互独立的特性,而互信息量趋于零. 相反,当它们的差别很小时,则共性增大,则互信息量也变大. 最近,一种基于两个分布的 Kullback-Leiber 判别信息量的基因打分方法也被提了出来[16,17]. 由于 Kullback-Leiber 判别信息量能很好地反映了两个分布之间的差别,这种方法选取相关基因的效果更为显著.

基因打分的方法的缺点是很难判断选择多少个相关基因是合理的. 这往往需要许多实验进行判断. 但如果按某种置信水平,对于两个分布是否一致做出统计假设检验. 若是一致的,则认为这个基因是不相关的,否则是相关的. 这样则可以直接得到相关基因的个数. 显然,采用 t – 检验就可以达到相关基因选取. 但 t – 检验是需要正态假设的. 这一假设在一般情况下并不成立[17,18]. 为了克服这一缺点,邓林等[18] 提出了秩和相关基因选取方法. 实验表明,这种方法在一定的置信水平下是合理的和有效的.

从组合数学上讲,我们需要找到一个合理的基因集合,在分类器上取得最好的分类效果. 根据这种理解,人们就提出了基于某种分类器的搜索算法[19,20]. 搜索算法是利用某种分类器,尝试不同的基因集合,挑选接近最佳的基因集合. 因此,它的计算量很大. 最常用的是根据遗传算法进行搜索. 遗传算法是模仿生物进化的方式,将特征组合进行二进制编码,对这种编码进行交叉、组合等变换,试图寻找最优的特征组合的方法. 对每一个组合,需要用分类器进行评价. Liu 等[19] 应用遗传算法来进行基因选择. 他们对每一种组合定义一个“ 适应度 ”并依此进行搜索. 他们采用并行计算,并将 SOM 作为分类器. 实验结果表明,对于文献[9] 中的白血病数据集选出了 29 个基因,交叉检验正确率为 95%,测试集上正确率为 88%. 对于文献[7] 中的结肠癌数据集,则选出了 30 个基因,交叉检验正确率达 92%.

Xiong 等[20] 采用了逐步选择法和 Monte Carlo 方法进行搜索,并应用 Fisher 线性判别法. 逐步选择法是首先找出分类效果最好的两个基因的组合,然

后从剩下的基因中每次选择一个加入选择的基因集合,使得分类正确率达到最高. Monte Carlo 方法是随机模拟的方法,每次反复 200200 次选择出 k 个基因,测试出分类效果,记录分类效果最好的基因集合. 然后逐步增加 k 选择更好的基因集,直到增加 k 不能得到更好的结果为止.

实际中,人们还从其他角度得到了一些相关基因选取方法,如文献[21—27]所给出的方法. 它们也都通过一些数据验证了其合理性. 但它们没有直接使用基因表达谱在两类样本上的差异,因此是一些间接性的方法. 这些方法具有一定的局限性,还有待于进一步的研究和发展.

4.2 用于肿瘤分类和诊断的模型

在病理分析和肿瘤诊断中,我们希望能够根据一组已知的基因表达谱数据建立分类或诊断(二元分类模型)模型,用于对未知样本进行分类或诊断. 这些已知的样本一般包括某种疾病的和正常的(或其他病种)确诊信息. 根据这些数据,我们首先将与该疾病(或分类)相关的基因选择出来. 然后,利用一种有监督的分类器根据相关的基因表达谱和分类结果进行训练,最终建立分类模型. 当然,在一些早期的研究工作中,人们还没有考虑提取相关的基因,而是直接进行训练和分类. 这种直接建立分类模型的方法现在越来越被实验证明是不够精确的. 近年来,人们在这方面做了很多探索. 下面我们介绍三种典型的有监督的分类模型和学习方法.

4.2.1 加权基因投票法

加权基因投票法是一种很直接的肿瘤分类方法[9,28]. 对于每一个基因 g, 设第 1 类和第 2 类样本的表达水平值的均值分别为 $\mu_1(g)$,$\mu_2(g)$,标准差分别为 $\sigma_1(g)$,$\sigma_2(g)$,则定义两类之间的相关性为 $P(g,c)=[\mu_1(g)-\mu_2(g)]/[\sigma_1(g)-\sigma_2(g)]$. 对于一个新样本,根据基因 g 的表达水平值 χ_g ,可以得到此基因的投票值:

$$P(g,c)\left(\chi_g-\frac{\mu_1(g)-\mu_2(g)}{2}\right).$$

将所有相关基因的投票结果中的正值累加表示对第 1 类的投票结果,而所有负值累加则表示对第 2 类的投票结果. 这两个投票结果的绝对值的较大值所对应的类别就是新样本的分类结果. 同时,还需要定义一个预测强度. 当预测强度低于一定程度时,可以不做分类,待进一步分析研究. 这样便可减少错误分类的机会.

Golub 等[9]在实验中使用 Affymetrix 微阵列,得到了 38 个白血病样本(27 个 ALL,11 个 AML)的 6817 个基因表达谱,然后根据加权基因投票法建立分类模型,并对另外的 34 个测试样本进行分类. 投票结果作出了 29 个分类,全部都正确. Slonim 等[27]对于同样的数据集进行了研究,在 38 个样本的训练集上作了交叉检验,作出了 36 个分类,全部正确. 另外,他们还对 ALL 进行了分类(T 细胞,B 细胞)实验,在 33 个 ALL 样本上做交叉检验,作出了 32 个分类,全部正确.

4.2.2 Fisher 线性判别法

Fisher 线性判别法是一种重要的分类方法. 其基本思想是寻找一个方向,使高维空间的样本点投影在这个方向上,能最容易地分开. 我们首先定义各类样本的均值为 $m_i=\frac{1}{N}\sum_{x_j\in C_i}x_j$,类间离散度矩阵 $B=(m_1-m_2)(m_1-m_2)^T$, 类内离散度矩阵 $W=\sum_{x_j\in C_1}(x_j-m_1)(x_j-m_1)^T+\sum_{x_j\in C_2}(x_j-m_1)(x_j-m_2)^T$. 然后,寻找一个最佳方向 ν, 使样本点投影到方向 ν 上以后,两类样本的类间离散度尽量大,而类内离散度尽量小. Fisher 线性判别法就是优化目标函数 $J=\nu^TB\nu/\nu^TW\nu$. 该问题可转化为求 $W^{-1}B$ 的特征值问题. 实际上,该矩阵的最大特征值对应的特征方向就是最佳投影方向. 这样,高维空间的分类问题转化到一维上,只需要确定一个点即可分开两类样本. 由于基因表达谱是一个典型的高维数据,Fisher 线性判别法也常被用来解决基于基因表达谱的肿瘤分类问题[13,20]. 实验结果表明,Fisher 线性判别法能够得到比较满意的分类结果.

4.2.3 支持向量机

支持向量机是根据统计学习理论建立的一种有监督的二元分类学习机器[29]. 其目标是通过学习寻找到划分两类样本的最优分类线(或面),达到最小风险或最大的推广能力. 在操作中,我们可以通过优化分类线在两类样本点上的间隔来实现. 实际上,支持向量机已经有许多成熟的软件可供我们使用,并且可以通过参数和核函数的选择增大其灵活性和适应行性. 当然,随着软件技术的提高,支持向量机的学习能力会越来越强.

Brown 等[30]用 SVM 对酵母基因表达谱数据[4]中的基因进行分类,学习样本包含了 2467 个基因,待测的有 3754 个基因. 他们对比了其他分类方法,发现 SVM 的分类结果优于其他的方法. Mukherjee 等[31]采用线性 SVM 在白血病数据[9]上进行了肿瘤

分类. 在训练集上的交叉检验结果完全正确. 在 34 个样本的测试集上,则有 0—2 个分类错误(使用不同数量的基因).

Furey 等[12] 也使用 SVM 对 Ovarians 数据集(97082 DNA clones, 31 个样本)进行分类,交叉检验的最好结果是在选取了 50 个基因的时候出现的(有 5 个分类错误). 另外,他们试验了白血病数据[9],38 个样本的训练集上交叉检验全部正确,测试集中的 34 个样本测试结果有 30 到 32 个正确. 另外,在结肠癌数据集[7] 上,62 个样本作交叉检验的结果有 6 个分类错误. 他们对比了感知机形式的分类器,发现 SVM 的分类结果更好些.

5 基因调控关系分析

在上述研究中,我们未考虑基因之间的相互作用. 实际上,基因之间存在着一定的调控关系. 这些关系对于生物机理的分析和理解尤为重要. 许多学者在这方面做了研究,提出了一些从基因表达谱数据中发现基因之间的调控关系的方法. 比较典型的方法有布尔网络(Boolean network)、贝叶斯网络(Bayesian network)和决策树(decision tree). 下面我们逐一介绍这三种方法的基本模型和算法思路.

5.1 布尔网络

布尔网络是一种单元间的逻辑分析网络. 其中单元状态是二元变量,而单元之间则通过一定的逻辑关系相互联系. 应用布尔网络的关键是根据给定样本的二元表格来建立单元或变量之间的逻辑关系. Liang 等[32,33] 首先将布尔网络的概念引入到基因表达的时间序列谱数据分析之中,以基因作为布尔网络的单元,单元之间的连接则表示基因之间的相互作用关系. 他们期望从表达数据中发掘出基因之间的调控关系以及相应的调控模式. 首先,将每个基因的表达水平值与平均表达水平值对比,进行二值离散化,即得到高表达(对应于 1)和低表达(对应于 0);其次,列出所考虑基因在不同时刻的表达值(状态转移表),即一个输入/输出表格;最后,根据状态转移表格推断出变量之间的逻辑关系. 他们使用了一种基于互信息的推断方法. 下面我们简要介绍其基本思想.

给定逻辑变量 X 与 Y 的一组数据. 根据其状态出现的频率,可以定义信息熵以及互信息:

$$H(X) = -\sum_{x=0,1} p(x)\log p(x),$$

$$M(X,Y) = H(X) + H(Y) - H(X,Y),$$

其中 $p(x)$ 表示状态 x 出现的频率. 如果 $M(X,Y) = H(X)$,则变量 Y 完全被变量 X 所决定. 这一简单事实是他们进行逻辑推断的依据. 也就是去搜索满足 $M(X,Y) = H(X)$ 的模式,从而找出 X 决定 Y 这样的逻辑规则. 当然,这里 X 可以不限于单个基因. 它可以是两个基因的联合,也可以是三个基因的联合甚至更多基因的联合. 相应地可推导出多个基因决定某个或多个基因的表达值的逻辑规则. 不过,相关个数越高,计算量就越大,并且计算量是相关基因个数的超指数函数. 因此,在实际计算中,相关基因个数不可能太大. 所幸的是,生物学知识告诉我们,绝大多数基因只受到一小部分基因的调控,如果我们将讨论范围限制在一定规模的相互作用之上,所需要的计算量还是可以容忍的.

5.2 贝叶斯网络

贝叶斯网络是目前在生物信息学中广泛应用的一种概率图模型[36,37]. 它把要分析的对象间的相互关系表示为一个有向无环图. 在基因调控网络分析中,图的节点是基因或者某个影响基因表达的条件,而有向边则表示两个节点之间的关系. $A\rightarrow B$ 表示节点 A 对节点 B 有直接的影响,通常称 A 为 B 的父节点,即 $A = \mathrm{parents}(B)$. 我们可以用概率语言来描述这个图,即把节点看成随机变量,节点之间的相互关系通过条件概率来表达,并且假定在父节点状态给定的条件下,每个节点与其非子节点之间相互独立. 如果用 $X = \{X_1, X_2, \cdots, X_n\}$ 表示这些随机变量,则这些随机变量的联合概率 $Pr(X)$ 可以分解为

$$Pr(X) = \prod_{i=1}^{n} Pr[X_i \mid \mathrm{parents}(X_i)].$$

这种因子分解特性为概率计算以及参数估计带来了极大的便利. 不妨假设 X_i 只可能取 0,1 两个值. 首先,如果每个节点 X_i 最多有 k 个父节点,那么最多有 2^k 个参数来描述局部的相互作用关系. 根据以上公式,最多有 $n2^k$ 个参数就可以描述全部随机变量的联合分布,即参数的个数是变量个数的线性函数. 其次,由于联合概率可因子分解,当给定完全学习样本时(即每个节点对应的随机变量的状态已知),似然函数同样可以因子分解,变为形如 $\theta^{N_T}(1-\theta)^{N_F}$ 的表达式的乘积,其中 N_T 和 N_F 是在与该参数对应的条件下,随机变量 X_i 处于状态 1 与状态 0 的次数. 从而得到该参数的估计量为 $N_T/(N_T + N_F)$,就像在掷硬币的实验中,通过出现正面的频率来估计硬币出现正面的概率一样简单.

贝叶斯网络更为精彩的一面是其结构学习功

能. 也就是说,在给定学习样本的情况下,怎样的基因间相互关系才能够最好地匹配目前的观测数据?当然,一个首要的问题是如何衡量结构与数据的匹配程度. 一种量度就是 Heckerman[36] 所定义的分数 BDe,即

$$\mathrm{BDe}(G:D) = \log\int Pr(D\mid G,\Theta)\,Pr(\Theta\mid G)\,d\Theta + \log Pr(G)$$

其中 D 代表数据,G 代表网络结构,Θ 代表参数. 然而,由于结构空间维数是变量个数的超指数函数,要搜索一个分数最大的网络结构是一个非常困难的问题. 即使限制每个节点最多有 k 个父节点,而当 k 大于 1 时,该问题仍然是一个 NP 完全问题. 如何在实际过程中寻找‘足够好’的结构,人们给出了一些方法. 一种非常简单的方法就是只搜索树形结构,利用图论中的算法可以有效地得到结果. 但这样得到的结构可能与真实结构相差甚远. 另一种方法是可采用一些启发式的搜索方法,从一个简单的网络或根据已有的生物知识所确定的某个初始网络出发,对其进行一些微小的变动,计算其对贝叶斯分数的改变量,根据改变量来确定变化后的网络结构的取舍. 如此迭代直到某个适当的终止条件得到满足. 最终得到的网络结构就是学习的结果. 还有一种选择是只看每个特定的子结构(比如从某个节点到另外一个节点之间的关系或者某个特定的通道)是否存在,如果用一个函数 $F(G)$ 表示这个特定的子结构在网络结构 G 中存在与否,那么就是要计算如下的后验概率:

$$Pr(F\mid G) = \sum_{G} F(G)\,Pr(G\mid D).$$

然而用这个公式很难直接计算. 一个比较简单的近似方法就是选择一些(比如 1000 个)后验概率比较高的结构 G,根据它们的平均效果来估计子结构的存在. 在具体的实现中,可以从某个已知的结构出发,对其进行微小的扰动,根据各自的得分保留一定(比如 1000 个)的得分高的结构 逐次扰动,逐次更新高分结构集合. 直到满足一定的终止条件结束迭代过程. 最后根据所保留下来的高分结构集合对子结构给出一个分数. 当所有的子结构都确定下来时,那么最好的结构也就确定下来了.

Friedman 等[34] 利用贝叶斯网络分析了酵母细胞周期基因表达谱数据[37]. 这个数据包含了 76 个表达谱数据,对应于 6 个实验条件下的时间序列(6177 个基因). Spellman 等[37] 检测到 800 个基因的表达量在整个细胞周期中有明显变化. Friedman 等[34] 对这 800 个基因的表达谱数据应用贝叶斯网络学习算法,得到了一个庞大的网络. 很多具有生物意义的基因间调控关系都能够被该算法恢复出来. Hartemink 等[35] 用贝叶斯网络分析了另一组酵母基因表达数据. 这组数据包含了 320 个样本在不同的生长环境下的表达谱(6135 个基因). 他们挑选了 32 个参与酵母 Pheromone 或 Mating response 基因,应用贝叶斯网络学习算法,得到了许多与生物学知识相符合的基因间调控关系.

5.3 决策树

决策树是信息科学中常常采用的一种机器学习方法. 该方法从学习样本出发,逐次选择样本特征,使得该特征能够最好地划分当前样本集合(用信息熵的原则来评价样本划分的好坏),将当前样本划分成两部分. 当然,这种分割过程并不能无限地进行下去,通常需要采用所谓的最小描述长度(MDL)原则来决定以上迭代过程是否中止. 这些原则都被实现到决策树算法 C4.5 中.

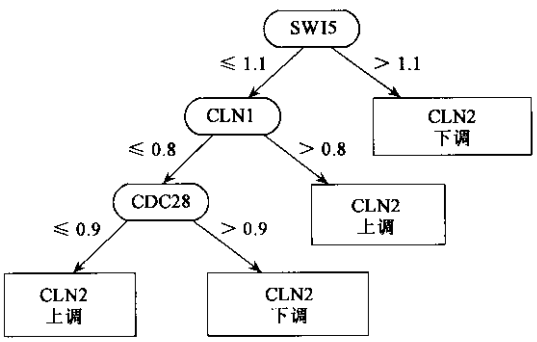

图 2 决策树示意图(原图见文献[38])

Soinov 等[38] 将决策树应用于基因表达数据分析之中. 基于 Spellman 酵母细胞周期基因表达实验数据[35],他们利用决策树生成算法 C4.5 对参与细胞周期的基因建立了决策树. 图 2 是一个简单的决策树例子. 从图 2 中我们可以清楚地看出基因间的相互影响关系,而且还能看出这种关系是如何具体表现的. 例如,上游基因的表达在一个什么样的范围时会引起下游基因的表达水平的变化,是上调了其表达量还是下调了其表达量等等. 与前两种方法相比,该方法不要求事先对每个基因的表达水平进行离散化,而是通过数据本身来决定离散化水平的. 另外,根据学习建立起的决策树,我们则可以推导出一些规则. 比如,基因 A 和基因 B 的表达水平值同时上调时引起基因 C 的表达水平值的下调. 这样的规则正是生物学家所需要的. 当然,决策树模型本身具

有一定的局限性,特别是它对数据比较敏感. 而基因表达数据往往包含了大量噪声. 因此,要想得到稳健的相互作用关系,需要有多个数据的分析加以证实.

6 总结与展望

本文对于 DNA 微阵列技术和数据的处理及其在肿瘤分类与诊断方面的应用进行了综述和分析. 目前,DNA 微阵列芯片已经越来越可靠,精度也越来越高. 价格也变得越来越便宜. 数据的获得也相当容易. 对于 DNA 微阵列数据的研究已经持续了近 7 年的时间,在生物特性的发现、相关基因的提取、肿瘤的分类和诊断等方面取得了重要的研究成果. 这些结果不仅受到生物学家的重视,也越来越被其他领域的科学家和医学界的重视,因为它提供了一种分析基因功能和认识生命过程的新方法. 我们相信这项技术必将给生物学和医学带来一场新的革命. 然而,微阵列数据的分析和处理又是一项艰巨的任务,因为过去人们很少遇见和处理如此高维的数据. 即使采用最好的降维方法,如何控制差错依然是一个大问题. 因此,这方面的研究才刚刚起步,还需要对这些数据做进一步的分析和研究,寻找它们的规律和数学模型,才有可能广泛地应用到医学中去.

另一方面,人们也开始研究蛋白质的表现水平,因为这对于生命过程的分析和理解更为重要. 实际上,测量蛋白质表达值的生物技术已经出现,即蛋白质芯片技术(protein array, protein chip)[39]. 但由于该技术成本太高,目前还无法满足大规模测量的要求. 2004 年初,UCSF 的一个研究小组给出了一种实验方法可以测量约 80% 的酵母蛋白质表达水平值[40]. 我们相信,随着生物技术的飞速发展,在不久的将来将会有更多的蛋白质表达数据公布. 随着 mRNA 表达数据和蛋白质表达数据的不断积累,生物信息学必将发挥越来越重要的作用.

致谢 感谢葛菲同学为本文查阅了大量的资料.

参 考 文 献

[1] Attwood T K, Parry-Smith D J. Introduction to Bioinformatics. Essex, England: Longman, 1999

[2] Speed T. Statistical Analysis of Gene Expression Microarray Data. Boca Raton, FL: Chapman & Hall/CRC, 2003

[3] Fodor S P, Rava R P, Huang X C *et al*. Nature, 1993, 364: 555

[4] Eisen M, Spellman P, Brown P *et al*. Proc. Natl. Acad. Sci. USA, 1998, 95: 14863

[5] Perou C M, Jeffrey S S, Rijn M V D *et al*. Proc. Natl. Acad. Sci. USA, 1999, 96: 9212

[6] Alizadeh A A, Eisen M B, Davis R E *et al*. Nature, 2000, 4051: 503

[7] Alon U, Barkai N, Notterman D A *et al*. Proc. Natl. Acad. Sci. USA, 1999, 96: 6745

[8] Ding H Q. Proc. Fifth Int. Conf. on Computational Molecular Biology (RECOMB2002). ACM Press, 2002. 127—136

[9] Golub T R, Slonim D K, Tamayo P *et al*. Science, 1999, 286 531

[10] Tavazoie S, Hughes J D, Campbell M J *et al*. Nature Genetics, 1999, 22: 281

[11] Tamayo P, Slonim D, Mesirov J *et al*. Proc Natl. Acad. Sci. USA, 1999, 96: 2907

[12] Furey T, Cristianini N, Duffy N *et al*. Bioinformatics, 2000, 16(10): 909

[13] Dudoit S, Fridyand J, Speed T P. Journal of American Statistical Association, 2002, 97(457): 77

[14] Bendor A, Bruhn L, Friedman N *et al*. J. Computational Biology, 2000, 7: 559

[15] Bendor A, Friedman N, Yakhini Z. Agilent Technical Report AGL-2000-13, 2000

[16] 葛菲,马尽文. 信号处理,2005,21(4) 待发表[Ge F, Ma J W. Signal Processing, 2005, 21(4), in Press(in Chinese)]

[17] Yan X, Deng M, Fung W *et al*. Journal of Theoretical Biology, 2005, 234 395

[18] 邓林,马尽文,裴键. 科学通报,2004, 49(13): 1311 [Deng L, M J W, Pei J. Chinese Science Bulletin, 2004, 49(13): 1311(in Chinese)]

[19] Liu J, Iba H, Ishizuka M. Genome Informatics, 2001, 12: 14

[20] Xiong M, Li W, Zhao J *et al*. Mol Genet Metab, 2001, 73(3): 239

[21] Park P J, Pagano M, Bonetti M. Pacific Symposium on Biocomputing, 2001, 6: 52

[22] Li W, Grosse I. Proceedings of the Seventh International Conference on Research in Computational Molecular Biology. 2003, 217—223

[23] Hellem B T, Jonassen I. Genome Biology, 2002, 3(4): Research 0017.1

[24] Hastie T, Tibshirani R, Eisen M *et al*. Genome Biology, 2000, 1(2): Research 0003.1

[25] Jaeger J, Sengupta R, Ruzzo W L. Pacific Symposium on Biocomputing, 2003, 53—64

[26] Guyon I, Weston J, Barnhill S *et al*. Machine Learning, 2002, 46(1—3): 389

[27] Feng J, Shi J, Shi Q. High dimensional feature selection for discriminant microarray data analysis. In: Advances in Data Mining and Medoling. World Scientific, 2003

[28] Slonim D K, Tamayo P, Mesirov J P *et al*. Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB'00). 2000, 263—272

[29] Vapnik V. Statistical Learning Theory. New York: Wiley, 1998

[30] Brown M P S, Grundy W N, Lin D *et al*. Proc. Natl. Acad. Sci., 2000, 97(1): 262

[31] Mukherjee S, Tamayo P, Slonim D *et al*. AI Memo 1677, Massachusetts Institute of Technology, 1999

[32] Liang S, Fuhman S, Somogyi R. Pacific Symposium on Biocomputing, 1998, 3: 18—29

[33] D'haeseleer P, Liang S, Somogyi R. Bioinformatics, 2000, 16: 707

[34] Friedman N, Linial M, Nachman I *et al*. J. Comput. Biol., 2000, 7: 601

[35] Hartemink A J, Gifford D K, Jaakkola T S *et al*. Pacific Symposium on Biocomputing. 2001, 422—433
[36] Heckerman D. A tutorial on learning with Bayesian networks. In: Jordan M I. ed. Learning in Graphical Models. Kluwer, Dordrecht, Netherlands, 1998
[37] Spellman P, Sherlock G, Zhang M *et al*. Molecular Biology of the Cell, 1998, 9: 3273
[38] Soinov L A, Krestyaninova M A, Brazma A. Genome Biology, 2003, 4(1): Research 6
[39] Zhu H, Bilgin M, Bangham R *et al*. Science, 2001, 293: 2101
[40] Ghaemmagham S, Huh W K, Bower K *et al*. Nature, 2004, 425: 737

·物理新闻和动态·

$SrTiO_3$ 在微电子器件中取代硅材料的前景

据预测,到 2007 年,信息产业的发展将遭遇基本物理限制的阻滞. 例如,基于 Si 和 SiO_2 的晶体管,其单位面积的集成数将不再能以 Gordon Moore 定律的速度(每两年翻一番)增长. 鉴于氧化物材料 $SrTiO_3$ 有可能以纳米精度控制其电子特性,人们看好这类材料在微电子领域的应用前景.

满足理想化学计算比的 $SrTiO_3$ 是绝缘晶体. 在 20 世纪 60 年代,$SrTiO_3$ 晶体因其具有宝石一样光泽而被用作金刚石装饰品的替代物. 如果在 $SrTiO_3$ 中移去一些氧原子,它将变成暗蓝色,并从绝缘体转变为导体. 这是因为,每当出现 1 个氧原子空位,就相当于大晶体中添加了 2 个电子. 或者说,氧空位的作用等效于电子的施主掺杂. 通常,1%—4% 的氧空位掺杂,便可将 $SrTiO_3$ 的低温载流子迁移率提高到 $10^4 cm^2 V^{-1} s^{-1}$. 制备 $SrTiO_3 - SrTiO_{3-\delta}$ 超晶格将有助于增长电荷载流子的寿命. 问题是:(1) $SrTiO_{3-\delta}$ 组分的尺寸可以做到多小,并且保持组分不扩散. (2) 在氧空位浓度很低的条件下,如何从成像角度表征材料. 最近,美国贝尔实验室的 Muller D A 等,以精巧的实验回答了这两个问题.

研究者使用脉冲激光淀积法,在 $SrTiO_3$ 衬底的 TiO_2(001)端面上外延生长出 $SrTiO_{3-\delta}$ 均匀薄膜,并制备出 $SrTiO_3 - SrTiO_{3-\delta}$ 超晶格结构;使用扫描透射电镜的‘电子能量损失谱’和‘圆环(annular)-暗场像’等手段,表征样品断面,进而以前所未有的精度定位了氧空位. 他们的结果证明:$SrTiO_3$ 区和 $SrTiO_{3-\delta}$ 区两者之间,可以是突变方式的渡越,转变区小至 1 个单胞层,并且没有因扩散而导致边界轮廓模糊.

(戴闻　编译自 Nature, 2004, 430: 620 和 657)