三、大规模研究癌症的方法

几十年来，传统方法对癌症的研究是首先选择一些癌症基因、基因组区域或蛋白质，然后将它们与健康组织或健康细胞进行比较。然而，随着大规模数据的产生和分析技术的来临，研究癌症的模式正在发生变化。基因组学、转录组学、蛋白质组学和生物信息学的应用，让人们得以测试大量的新假说，从而促进了癌症研究的发展。例如，这些大规模技术的应用，扩大了与特定类型癌症发展相关的遗传变异的可检测数量，并能够整合分子特点从而预测癌症和治疗反应。图2显示了生物信息学、基因组学、转录组学和蛋白质组学结合在一起研究癌症、预断病情的模型。

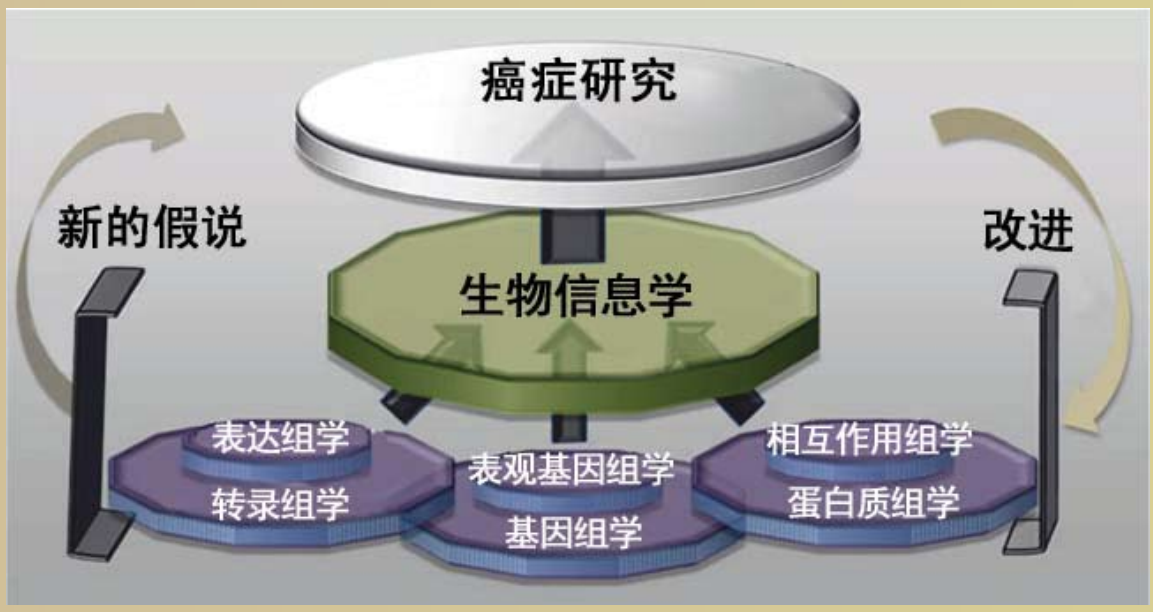
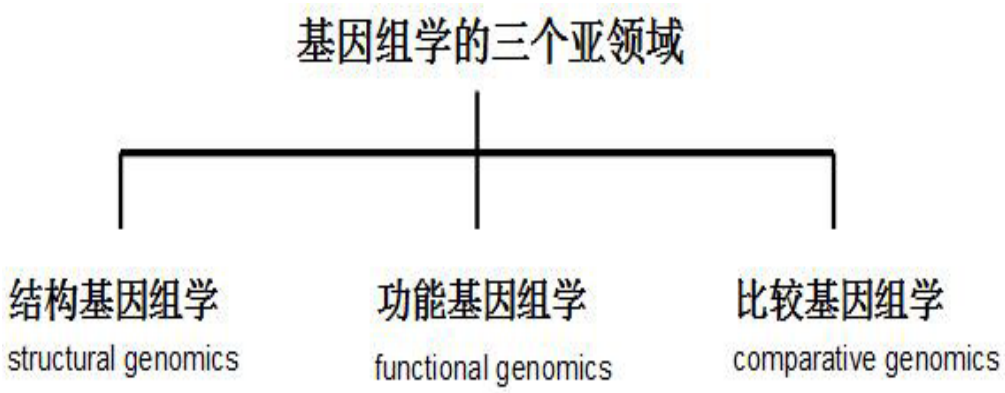


图2 生物信息学、基因组学、转录组学和蛋白质组学结合在一起研究癌症、预断病情的模型。

1．基因组学

所谓基因组学（Genomics）是指发展和应用DNA图谱、测序新技术以及计算机程序分析生命体（包括人类）全部基因组结构及功能的学科。



从历史上看，在Frederick Sanger开发出测序技术之际，就为基因组学奠定了坚实的基础。他们测定和储存了病毒X174的DNA序列。而基因组学的诞生则与破解流感嗜血杆菌（Haemophilus influenza）基因组序列的关系更为密切。

现在，基因组学已成为生物科学的前沿学科，它的发展主要依靠以下两个因素：

(1) 上世纪80年代发展起来的大规模测序技术；
(2) 分析、储存和整合这些资料的计算机方法的发展。

从广泛的基因组意义来讲，这些资料来自加州大学圣克鲁斯分校（UCSC）（http://genome.ucsc.edu）、美国国立生物技术信息中心（NCBI ）（http://www.ncbi.nih.gov）和欧洲生物信息学中心（EBI）（http://www.ebi.ac.uk）。

基因组学技术一个重要的新内容就是纳入了生物医学研究，且其中最相关的是“癌症基因组学”。它整合了大量的数据和计算机资源，有助于研究癌症细胞或癌症组织的基因组结构变化。

下面就从几方面举例说明基因组学方法如何革命性地改变了癌症的研究。

在最近的研究中，Campbell等人在碱基对水平确定和描绘了DNA删除、串联重复、反向重复、倒位和染色体重排等改变在两个肺癌细胞株基因组中的情况。他们基本上利用了生物信息学方法来确定癌症基因组的变异，如以基因短序列为基础和以人类公布的DNA序列作为参考。Sjoblom和Wood等人则搜索了乳腺癌和结肠肿瘤样本的突变。他们首先利用生物信息学工具，在可以利用的参考序列数据库（http://www.ncbi.nlm.nih.gov/RefSeq）中选取了18191个人类基因。接下来，他们测定了所有肿瘤样本中这些编码基因外显子的序列，结果找到了大约80万个潜在突变。随后，他们又使用其它生物信息学方法消除人为假象和正常变异，并进行同义替换，旨在在这些肿瘤中找出可靠的体细胞突变。

最后，他们平均在每个乳腺癌或结肠肿瘤样品中发现了80个突变基因。领导小组还发表了一份类似的分析报告，分析了胶质母细胞瘤和胰腺肿瘤。表2描述了在癌症研究中生物信息学和基因组的应用。

如在乳腺癌研究中，高通量基因组技术的潜力已经迅速应用于该领域。这些方法之所以吸引人，是因为它们可以同时测量数以千计的DNA序列、mRNA转录物、多肽和代谢物（图3），从而使我们对细胞过程机制有一个整体了解。

乳腺癌是一个极其复杂的疾病，它有许多危险因素，从遗传上的易感性到生活方式的选择，例如饮食和运动都可促使其发病。此外，乳房组织研究起来非常困难。这是由于它是由多种组织组成的，并且在月经周期、妊娠期和衰老期结构都会发生变化。从常见的临床特点，如肿瘤大小、淋巴结转移、亚型和雌激素受体（ER）的表达来看，乳腺癌是一种高度异质性的疾病。分子图谱已经确认了这一点，并强调其复杂性，因此不同的肿瘤对应着不同的治疗方法。

当考虑高通量研究的价值时，要把多种可能的易混淆因素和局限性考虑进去，这主要集中体现在基因表达谱，当中所产生的许多问题都适用于其它“组学”技术，例如CGH阵列、miRNA和蛋白质组学为基础的方法。

表2基因组学、转录组学和蛋白组学生物信息学技术在癌症研究中的应用

方法	可以检测的内容	应用的工具
基因组生物信息学	突变、多态性、甲基化的变化、染色体扩增、删除和重排	DNA测序、下一代测序技术、单核苷酸多态性阵列、比较基因组杂交、基本局部比对搜索工具（BLAST）、FASTA、其它特定的软件和软件包、网络工具和公共数据库
转录组生物信息学	新外显子和内含子区域的定义、基因表达、转录后修饰、疾病状态目标（基因）的优先表达	微阵列为基础的技术、基因序列表达分析技术（SAGE）、大规模平行测序技术（MPSS）、下一代测序技术、PCR技术为基础的基因表达、BLAST、序列比对程序Blat、Sim4工具、其它特定的软件或软件包、网络工具和公共数据库
蛋白质组生物信息学	蛋白质鉴定、蛋白质水平的测定、翻译后修饰、蛋白质之间的相互作用以及酶活性	免疫组织化学、萤光免疫检验法、质量光谱、二维凝胶电泳（2D gel）、蛋白基因芯片、具体的软件、网络工具和公共数据库

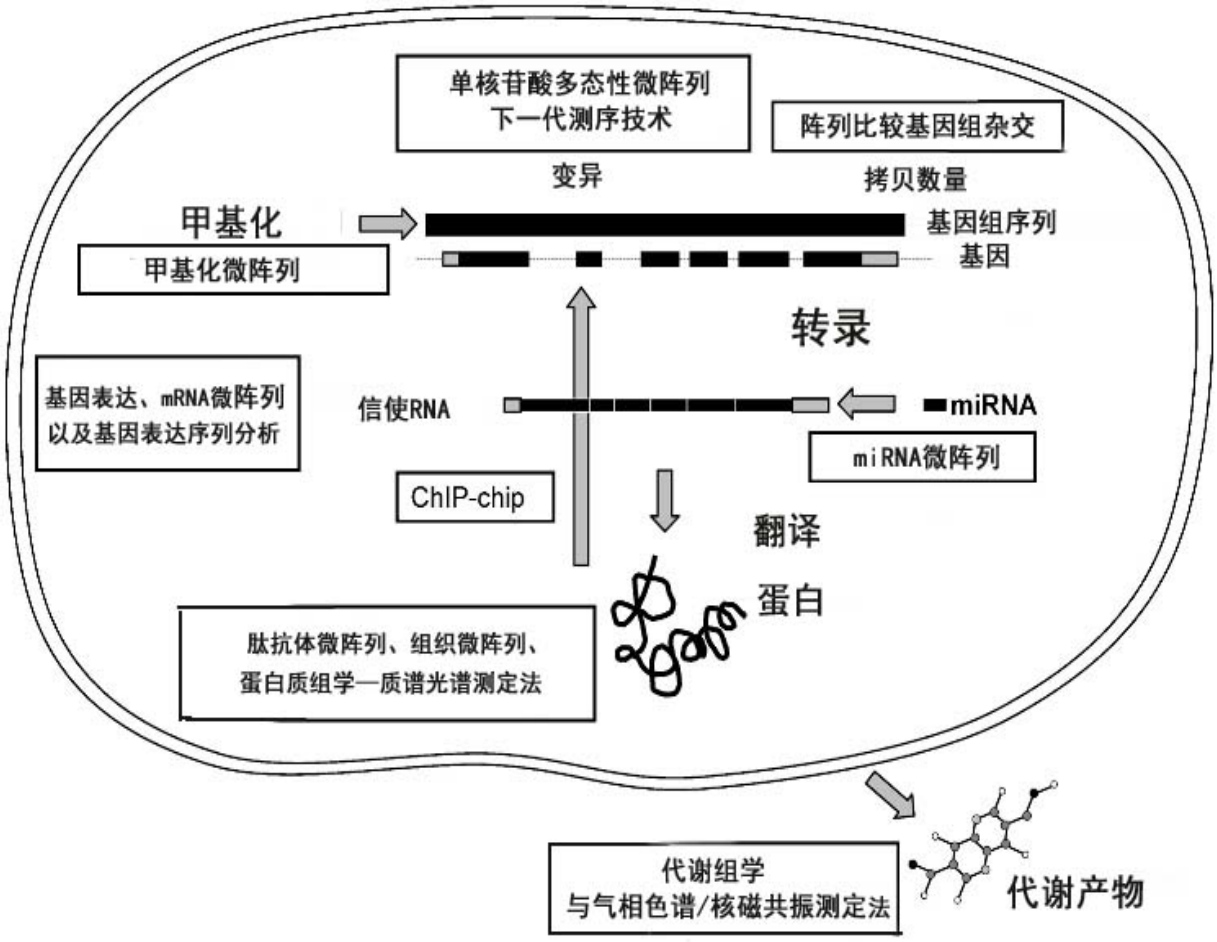


图3 不同类型的高通量基因芯片技术汇总以 及它们的测量内容。现在很多基因芯片能以整体和系统的方式测量许多分子的变异。

1.1实验设计

从测量大量病人的单个转录产物或蛋白质水平，到同时测量数以千计的基因或蛋白的技术转变，需要更好地理解多种测试的概念和错误发生的几率。高通量方法一个吸引人的地方是，它们靠数据驱动而不是靠假说，因此不会被以往的知识所限制。虽然这些假说可以有效地证明或否定一个给定的假说，但它的真正价值是产生一个新的假说。结果拥有的特征数量，例如基因、转录产物、单核苷酸多态性和多肽等比样本的数量还要多，很多明显的差异表达特征可能是偶然的，而不是真正的生物学差异。考虑到样品的异质性和多变性，在低维视野中的数据不能分成明确界定的类群（图4）。为了得到更详细的数据集以及多种测试问题的知识，读者可以参阅Clarke 等人的文章。最大的基因特征差异来自于群体选择（生物差异）和实验偏差（技术差异）造成的差异[1]。

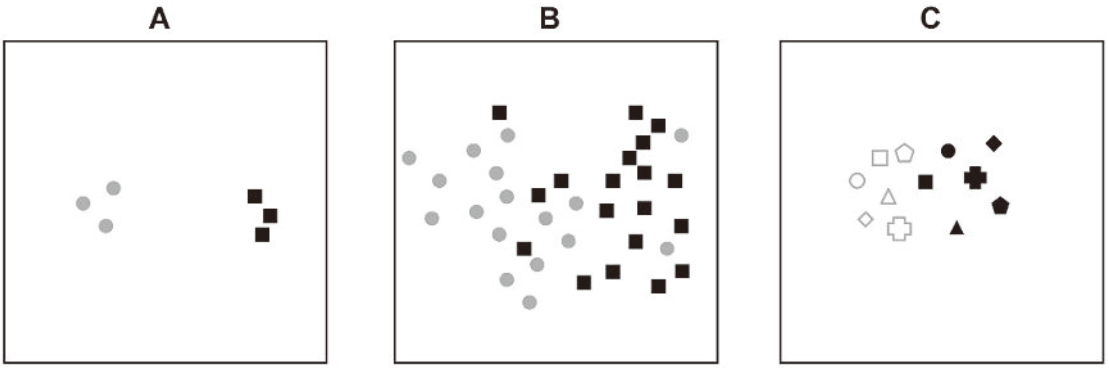


图4 重复或样品之间的差异决定了表达差异的基因数量，这些基因可以分亚组。A：灰色的圆圈重复和黑色的方框重复内部各自紧密成簇，但它们之间却严格分开。B：灰色的圆圈重复并没有和黑色的方框重复严格分开，所以与（A）相比，需要大量的个体以鉴定出它们之间的差异。C：在测量前（实体图标）后（空心图标），与特定疗法或步骤相关的常见表达差异在匹配的样品中很容易被鉴定出来。

1. 2 生物差异

当基因芯片试验的目的是鉴定实验组和对照组基因表达差异的时候，被调查的表型，包括疗法、高表达基因、肿瘤或病人特征必须代表变异的最大来源。如果不是这样，由于混杂因素（无论已知的还是未知的）的存在，其结果必定会受到影响。减少这种潜在问题的一个方法就是确保两个样品组除了被测表型外，在其它各个表型中都尽可能相似。另一个可选择的方法是利用大量的样品增加试验差异和两个小组之间的差异[1]。

用细胞系做实验、通过相同的方法准备样品、遵守严格的协议以保证重复高度相似和减少差异的来源，这些都是相对容易做到的。在这种情况下，为了区分实验组和对照组，相对少的重复实验是需要的（图4A）。但对应用初级患者的样品，分子异质性是一个更大的问题。基因表达受这么多因素的影响，为了区别实验组和对照组，需要高度区分选择的标准和大量的种群（图4B）。另一种适合于观察不同个体变化的方法是观察处理前后同一个体的正常组织或肿瘤的匹配样品（图4C）。这些配对的研究增加了统计的力量并可以预测哪个种群会对处理产生潜在的反应。另一个重要因素是肿瘤组织的组成，这些肿瘤组织是提取DNA和RNA等许多研究的起点，因为很多分子的分析都可以采用激光捕获显微切割技术。Perou’s和Sorlie’s研究小组提到，对类似正常的亚型的再分析需要这样的精度[1]。

1. 3技术差异

虽然互补序列退火和杂交的基本原则是相同的，但在基因芯片的设计和生产上存在根本的不同。早期的基因芯片是在单个实验室由PCR产物、克隆的cDNA或合成的寡聚核苷酸印到玻璃板上而形成的。随后，该技术被迅速发展和扩大，并用来评测许多其它的变异（图3）。一些企业已将商业性的基因芯片设计得简单易用，这些企业包括Affymetrix公司（美国加州圣塔克拉拉县）、安捷伦科技有限公司（Agilent Technologies，美国加州圣塔克拉拉县）和 Illumina公司（美国加州圣地亚哥）[1]。

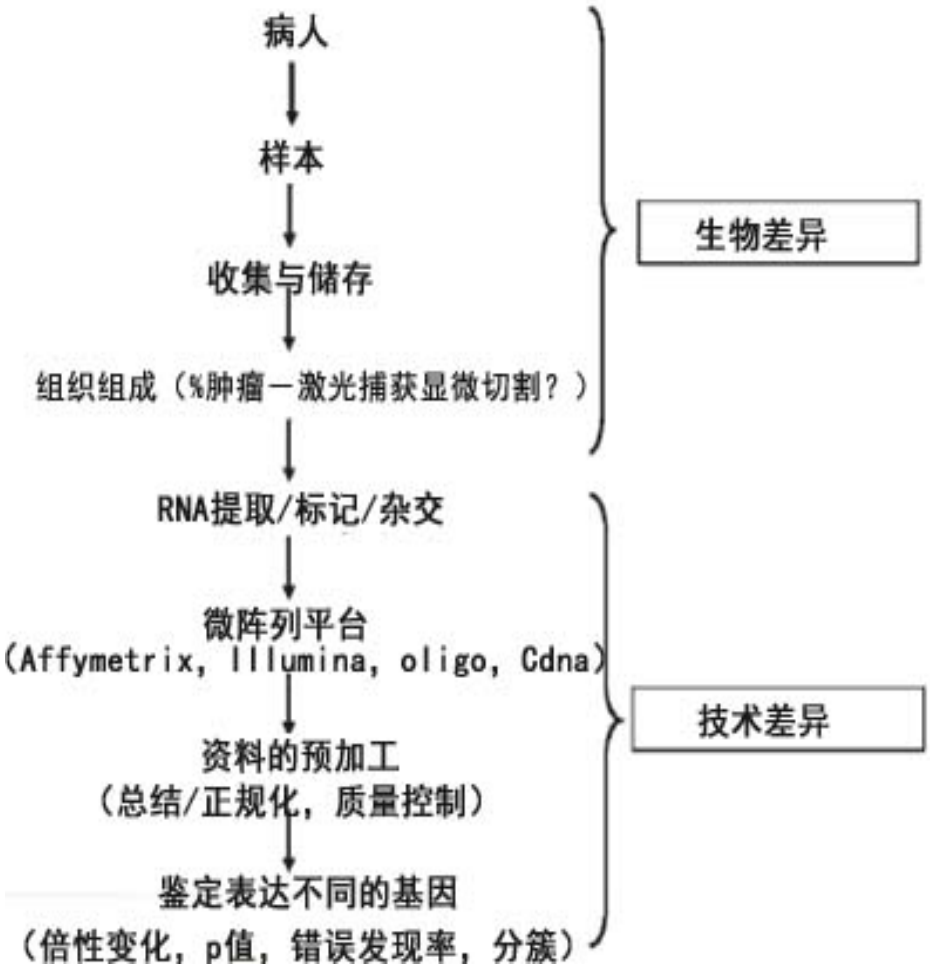


图5 微阵列实验的关键步骤。多个阶段引入了生物差异和技术差异，这些差异都将对最后的结果产生影响。重要的是，所有步骤都已被明确描述。

大多数同行评审期刊的出版前提是基因表达数据可以被广泛利用，并且数据库（如ArrayExpress数据库和NCBI Gene Expression Omnibus数据库）可以令研究人员更为方便地获得这方面的信息。除了原始数据，作者还必须根据MIAME（文后小词典1）的指导原则提供使用样品、平台和方案的细节。这项规定提高了基因芯片研究的透明度，并可以进行荟萃分析（meta-analysis）。但是，典型芯片技术流程的许多步骤在研究中经常变化，这可能会引起偏差。一系列重大基因的相似研究缺少重叠，这些差异通常归因于基础技术方面的不同，例如探针序列的设计或实验步骤。尽管如此，Sorlie等人证明，在三个不同的芯片平台，不用检查就可以区分乳腺癌的亚型（图5）。每一种芯片平台鉴定的差异最明显的表达基因，通常都会有一个非常显著的路径重叠。重要的是，所有芯片显示的结果都是高度依赖它们最初被设计时的信息的。借助最新的基因注解图谱可以发现，先前无论用哪种方法鉴定的差异表达的基因都存在30~50%的差异[1]。

1.4 亚型、分类和预后

乳腺肿瘤可以通过多种组织病理学和分子病理学方法分离，以预后或预测它对各种治疗方法的反应。在乳腺癌领域已经有三种分析基因表达芯片的途径（图6）[1]。

第一种方法是无需检查的分析方法。人们通过一组固有的基因将肿瘤分成亚组。这组固有基因反映了肿瘤间，而不是肿瘤内的基因表达差异，无需使用选择标准。Luminal乳腺癌和basal-like乳腺癌亚型之间最显著的分子差异，已经被不同的技术及平台多次鉴定和验证。“molecular apocrine”肿瘤的鉴定，和ER阴性肿瘤被分为五种亚型的研究也已经完成。molecular apocrine亚型的鉴定与显著不同的临床结果相联系，这些临床结果可能是对不同治疗方法的反应[1]。

另外两种方法使用监督途径。这两种方法的基础是个体临床随访的资料或肿瘤生物学的特征，例如雌激素受体状态、等级或扩散等情况（图6）。在Amsterdam组（cDNA阵列）的70个基因特征和Rotterdam组（Affymetrix寡聚核苷酸阵列）的76个基因特征之间缺少重叠（3个基因），这可能说明了依据随访数据分析的基因组方法是不可靠的。但是从逻辑上讲，无监督方法证明的异质性，会排除样品不同群体适度研究的重复结果。当检查入选标准，如年龄、淋巴结状态、肿瘤直径和辅助治疗等平台，如基因和寡聚核苷酸阵列或定量逆转录PCR以及不同数据的分析方法时，特征之间的差异得以解释。尽管方法明显不同，而且对基因特征缺乏共识，但上面阐述的三种途径（如图6）都能准确预后。利用一个单独数据集测试几个特征，证明对单个病人的预测结果表现出高度一致性[1]。

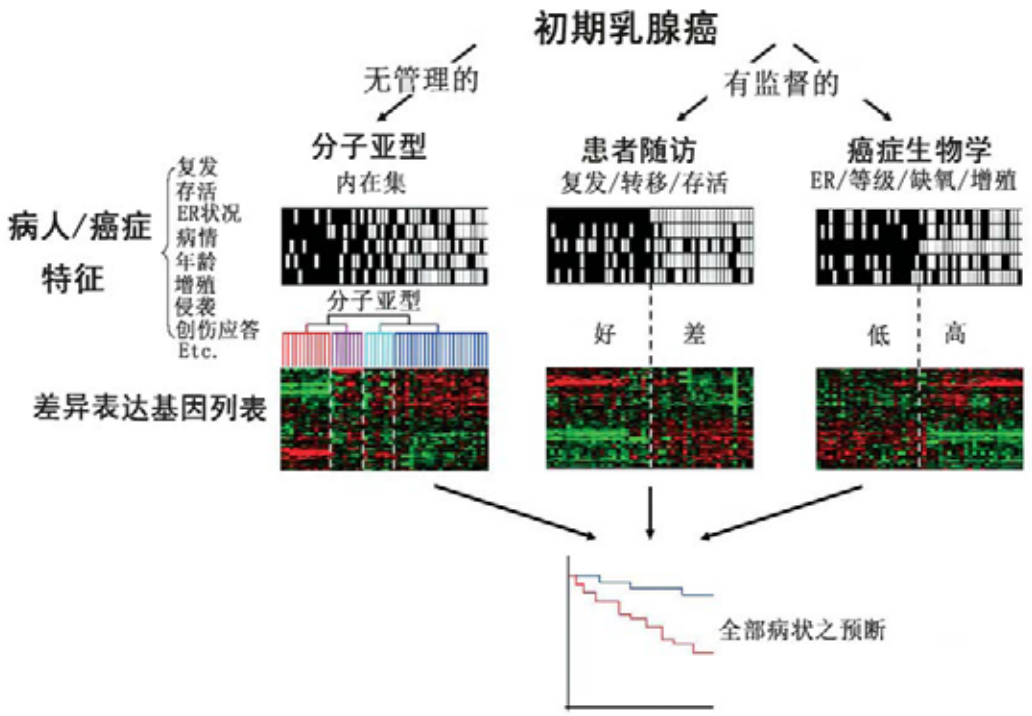


图6 不同方法预后的一致性。尽管用来鉴定表达特征的策略不同，每种方法都能预后。这些形态是互相关联的，可能比现有的单个标记更精确。ER：雌激素受体。

1.5 验证数据集和特定的数据偏差

构成基因表达特征的基因本质上取决于病人和肿瘤的特点、微阵列平台、标准化的方法和基因选择的统计阈值或使用的分类算法等（图5）。使用一个特定的数据集所形成的预测，可能会存在数据库内在的偏差。Ein-Dor等人证明许多不同但可预测的70个基因可以简单地通过改变“训练组”和“测试组”的成员来产生。验证数据集没有用来产生基因特征的数据集完美，这似乎是不可避免的。当验证的平均随访时间（14年）比原先的（8年）长时，研究人员发现76个基因的Rotterdam症状具有较强的时间依赖性。众所周知，许多病人的自身情况会影响其基因表达（和其它肿瘤特征），如年龄和种族。Anders等人证明乳腺癌在年轻女性中发病率升高，这种情况可能与PI3K、Myc和β-酪蛋白相关，而老年妇女肿瘤的发生却与Src的激活和E2F的失活有更多的联系。只有在特定的病人小组中机制明显不同的基因才会在相似的群体中再现[1]。

最近在使用重复的验证数据集比较了乳腺癌细胞和正常乳腺细胞系（MCF7和MCF10A）之后，研究人员检测了数据集之间的差异。这些数据集由不同数量起始RNA、不同的方案、不同代的Affymetrix基因芯片或不同代的扫描硬件产生。他们发现系统的以及增加的偏差，早在基因芯片的RNA、杂交和影像捕获等阶段时就产生了[1]。

1.6 数据整合和META分析

利用独立数据验证新结果，对确定新发现的真实性是必需的。例如，不同平台多个实验的meta分析产生了新预测特征，这与该平台的具体特征同样好，甚至更好。这些方法可删除单个芯片平台的固有偏差，并能够集中在表达差异的基因。但是，跨平台的meta分析可能会受共同基因的数量限制。为了比较不同基因芯片的数据，人们对跨平台的规范化和可测定距离的方法提出了要求[1]。

克服上述异质性的一种方法是通过结合数据集来扩大研究规模，但是这会使数据分析的问题更加令人望而生畏。现在公共领域中，乳腺癌基因表达的数据集代表着meta分析的宝贵资源。但是，特定的数据集偏差在原始数据水平排除了已公布研究的整合，但目前没有合适的解决办法（图7）[1]。

研究中，简单的数据均值化（mean-centring）预处理足以使验证的细胞系与公布的乳腺肿瘤数据集相吻合，表现优于距离加权（distance-weighted）偏差，并且产生了与ComBat类似的结果[1]。

ComBat是一种经验贝叶斯方法（empirical Bayes method），可以用来调整批处理（batch）的影响。一些meta分析研究已经出版，通常验证以前的结果，强调“真实”结果性、一致性和提高清晰度[1]。

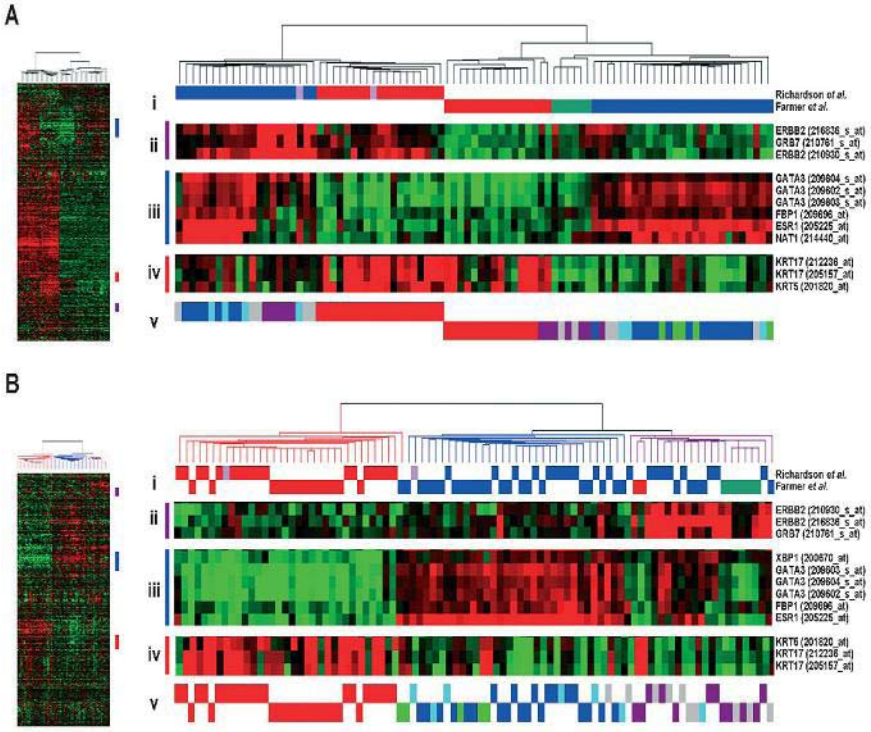


图7 整合基因表达数据时，数据集的特定偏差必须删除。结合两个已公布的乳腺肿瘤基因表达谱。A：在batch-centering之前；B：在batch-centering之后。基于Sorlie等“固有”基因的640探针集，进行分层次肿瘤聚类。缩略图显示了全部的640探针集，（i）Richardson等人的肿瘤分类——红色：basal-like乳腺癌亚型；蓝色：非basal-like乳腺癌亚型；粉红色：BRCA1乳腺癌亚型。Farmer等人的肿瘤分类——红色：basal-like乳腺癌亚型；蓝色：luminal乳腺癌亚型；绿色：顶泌（乳腺癌家族新成员）。与“Sorlie亚型”联系的基因簇分类将如下所示突显出来；（ii）ERBB2基因簇；（iii）luminal A基因簇；（iv）basal基因簇；（v）用质心（Centroid）预测将肿瘤分成五个Norway/Stanford 亚型——红色：basal亚型；深蓝色：luminalA亚型；浅蓝色：luminalB亚型；紫色：ERBB2亚型；绿色：normal-like亚型；灰色：未分类。

最近研究证明，整合高达六个乳腺癌的Affymetrix基因芯片，不仅可以增加预后的精确度，而且还能通过删除系统的、多重的偏差得到进一步的提高。当测试组和训练组中病人和肿瘤有相同特征时，可以产生最精确的预后推断。构建一个更大的、能够代表所有乳腺癌病人的数据集的途径，将是根据常用的临床病理参数产生单个病人小组的基因表达分类。严格的进入条件将严格限制适用的病人或肿瘤的数量，这种方法无需考虑未知的复杂因素。临床实践中，单个样品的预测程序应该能够适用于大量的患者。研究表明，这些程序最好来自于最大可能的群体（或整合的数据集）。研究者需要知道数据库组成不同在多大程度上影响了meta分析，如果数据库容易受到进入标准和处理方法的影响，那么把不同的数据库结合起来就不一定总是恰当的[1]。

临床医师必须选择最适合病人的治疗方案，但是现在许多疾病参数可以定性而不能定量分析。预后模型，如瑞士St.Gallen早期乳癌辅助治疗会议和诺丁汉预后指数（Nottingham Prognostic Index, NPI）可以用来指导治疗决定。虽然这些模型能够精确预测人类预后的结果，但不能鉴定发生在哪个妇女身上。这不可避免会造成过度治疗，或者治疗不足[1]。

根据National Comprehesive Cancer Network（一个由美国最好的17间癌症治疗和研究机构所组成的组织）的准则，一些好的预后小组，也会导致80%的非必需化疗[1]。

为了使分子特征在治疗方案选择中更有价值，他们必须优于或增加现存的临床指导方案的价值。肿瘤的传统分类可以在高风险和低风险情况下选择治疗方案，但是，肿瘤的状况常常刚好处于高风险和低风险的中间状态，所以治疗的方案急需改善。在这些情况下，相对安全的选择是过量治疗。病例中的一小部分可能受益，但对其余的可能有一定的副作用。相反，一个保守的方法可以避免不必要的治疗，而且减少成本，但是，一些能够受益的妇女可能无法得到治疗。研究基因表达和获知预后因素之间的联系，可能有利于中间组[1]。

目前，基于基因表达谱研究的两个临床测验已经商业化，而且正在多中心、多国家的试验中进行评估。乳腺癌患者已越来越依赖另一种基因检查方法——OncotypeDx。它可确定乳腺癌复发的风险，并可通过化疗药剂最为有效地杀死肿瘤细胞。OncotypeDx着眼于21种基因，以此判断化疗是否会对早期乳腺癌病人有帮助。OncotypeDx从石蜡包埋的切片中提取RNA，然后用实时定量反转录聚合酶链式反应检测21个基因的表达。该项测定方法适用于评价激素受体阳性、淋巴结阴性的Ⅰ/Ⅱ期乳腺癌患者的复发风险。这项研究将登记多达10,000激素阳性（ER阳性和/或PR阳性）、ERBB2阴性和淋巴结阴性乳腺癌的女性，以决定哪些女性应接受激素辅助化疗[1]。

对4964个没有进行过化疗的淋巴-淋巴结阴性乳腺癌档案材料的研究发现，ER阳性、他莫昔芬（tamoxifen）治疗和未治疗的病人的复发率与乳腺癌死亡有紧密的联系。在B-20研究中，复发率不仅可以量化淋巴结阴性、ER阳性乳腺癌女子的复发率，而且可以预测化疗的益处。但是，在对149个未受到辅助疗法治疗的病人的研究中发现，21基因依赖的复发率不能预测差异大的疾病的复发率，说明群体选择是非常重要的。OncotypeDX已加入美国临床肿瘤学会的批准列表中，2008年已经完成了60万个OncotypeDX测试。尽管这将减少不必要的治疗成本，但是必需的成千上万美元的测试成本可能会对卫生服务提供商造成一定影响[1]。

美国食品和药物管理局已批准了Mammaprint临床测验，这项试验是由Agendia公司（美国加州亨廷顿湾）利用70个肿瘤基因（70-基因）特征发展来的，且对用于设计和分析原始的70-基因特征的关注越来越高。这些问题很大一部分已纳入6000个病人的MINDACT临床研究实验。乳腺癌国际协作组（BIG）的转化性研究系统（TRANSBIG）旨在通过将分子生物学方面的研究与早期乳腺癌治疗处理相结合，从而改良个体化治疗。TRANSBIG的一个设计严谨的前瞻性研究正在对70-基因表达进行检验。此项研究将独立确定70-基因的预后价值是否可以在多中心、更广泛的总体中得到重现，以及是否能够像先前描述的那样在不久的将来替代传统的临床病理学指标。近来，人们利用一个不同的微点阵平台对76-基因表达进行回顾性分析发现，它可被用来预测所有年龄组淋巴结阴性乳腺癌患者的预后[1]。

“个体化治疗”的一个特征是它难以确定经过同样方法治疗的、具有类似特征的患者的适合数量，以有效地统计这一项研究。虽然高通量表达谱分析方法有很大的潜力，但在它们成为标准预后工具之前，还需要仔细加以验证。在此期间，它们产生的大量宝贵数据将改善我们对乳腺癌发展、进程和治疗相关分子变异的了解[1]。

2. 转录组学

一个基因组在任何时候的全部的转录物被称为转录组。与基因组学概念相仿，转录组学被定义为全面研究转录组的科学。

与基因组不同的是，转录组是非常动态性的，不仅同一个生物的不同组织之间不一样，并且病态，如癌变组织与健康态的组织之间也不一样。基于这种结果，很多研究者研究了大量基因的表达谱，并试图鉴定癌症组织中基因的表达情况。下面描述了两个转录组学被用来研究癌症中基因表达情况的事例。

Rhodes等人开发了一种计算协议，旨在几种类型的癌症中，发现基因表达的一个meta标记（meta-signature）。他们分析了40个公布了的癌症组织微阵列数据集，这些数据集涉及了超过3700个癌症样本的大约3800万个基因表达谱。结果，他们发现多于10个的癌症患者中有60个基因的表达水平超过了正常组织的相同基因的表达水平。很多其它的实验也使用几乎同样的策略以试图鉴定出肿瘤中基因表达的标记。使用序列为基础的方法，Sugarbaker等人整合了计算机方法和下一代测序技术，并研究了6个患者的间皮肿瘤。结果，他们发现了间皮瘤的一些转录组学的特征，从中鉴定15个非同义突变。

美国克莱格凡特研究院（J. Craig Venter Institute）、Lugwig 癌症研究所的三个分所以及纽约斯隆-凯特琳癌症纪念研究中心（Memorial Sloan-Kettering Cancer Center）的研究人员利用Roche 454转录组测序仪来寻找高度重排的乳腺癌细胞系中的基因组易位。转录组是指细胞中转录本的集合。转录组反映了活跃的基因组，它包含了从不同的基因剪接形式到基因表达的一切信息。但是Strausberg等人还想从转录组中了解其它信息——有关基因组易位事件的线索。为了检测基因重排的活跃基因产物，他们利用了Roche 454-FLX焦磷酸测序仪来评估高度重排的乳腺癌细胞系HCC1954的转录组。研究小组从该细胞系中获得了510,703个cDNA序列读数。在这些读数中，超过384,900 个读数与9,221 个RefSeq 基因mRNA 配对。剔除掉这些与RefSeq配对的序列，研究人员尝试将剩余的序列与人类参考基因组比对，剩下47,370 个序列与两者都不能配对。

研究人员随后将剩余的序列放入计算机分析流程，提取出496 个含有至少两个不同的基因组位点信息的潜在的嵌合体转录本。大体上一半的转录本代表了同一个染色体上的重排，而另一半则包含了不同染色体间的重排。研究小组从中挑选了33个假定的嵌合体用于后续研究，并于实验中验证了13个嵌合体cDNA。小组检测到的大部分变异也同样存在于同一个体的血细胞对照系中。这促使研究人员利用长片段PCR（Long Range PCR）、Sanger末端测序和荧光原位杂交来区分反式剪接事件和真正的基因组重排事件。

尽管他们发现的一些重排与细胞系的已知变化有重叠，但其它都是新的。比如，他们在最初的实验中发现了4个染色体间转位和1个染色体内重排。接下来，研究小组再回去寻找那些在基因组上可定位于不止一个定位点的嵌合体转录本。在此过程中，他们又发现并证实了两个染色体间重排。

研究中总共发现了7个重排，据统计至少影响了9个不同的基因。其中5个重排产生的转录本编码了截短蛋白，包括具有致癌功能的蛋白。例如，研究人员检测到MRE11A和NSD1的截短，MRE11A基因编码的蛋白参与了几种乳腺癌中检测到的双链断裂修复，而NSD1基因则编码了可能的转录调控因子，有时存在于急性髓细胞白血病中。

正如上述基因组学指出的那样，这些报告都以例子说明了生物信息学和转录组学可以被整合，并可以相互协作从而在与癌症的斗争中产生有用的成果（表2）。

3. 蛋白质组学

蛋白质是生物体的重要组成部分，参与几乎所有生理和细胞代谢过程。此外，与基因组学和转录组学比较，对一个细胞或组织中表达的所有蛋白质，及其修饰和相互作用的大规模研究称为蛋白质组学。

蛋白质组学通常被认为是在基因组学和转录组学之后，生物系统研究的下一步。然而，蛋白质组的研究远比基因组学复杂，这是由于蛋白质内在的复杂特点，如蛋白质各种各样的翻译后修饰所决定的。并且，研究基因组学的技术要比研究蛋白质组学的技术强得多，虽然在蛋白质组学研究中，质谱技术的研究已取得了一些进展。

尽管存在方法上的挑战，蛋白质组学正在迅速发展，并且对癌症的临床诊断和疾病治疗做出了重要贡献。几项研究鉴定出了一些蛋白质在乳腺癌、卵巢癌、前列腺癌和食道癌中表达变化。例如，通过蛋白质组学技术，人们可以在患者血液中明确鉴定出肿瘤标志物。表1列出了更多的蛋白质组学技术用于研究癌症的例子。

另外，高尔基体功能复杂。最新研究表明，它除了参与蛋白加工外，还能参与细胞分化及细胞间信号传导的过程，并在凋亡中扮演重要角色，其功能障碍也许和肿瘤的发生、发展有某种联系。根据人类基因组研究，约1000多种人类高尔基体蛋白质中仅有500~600种得到了鉴定，建立一条关于高尔基体蛋白质组成的技术路线将有助于其功能的深入研究。

蛋白质组学是一种有效的研究方法，特别是随着亚细胞器蛋白质组学技术的迅猛发展，使高尔基体的全面研究变为可能。因此研究人员希望能以胃癌细胞中的高尔基体为研究对象，通过亚细胞器蛋白质组学方法，建立胃癌细胞中高尔基体的蛋白质组方法学。

研究人员采用蔗糖密度梯度的超速离心方法分离纯化高尔基体，双向凝胶电泳（2-DE）分离高尔基体蛋白质，用ImageMaster 2D软件分析所得图谱，基质辅助激光解吸离子化飞行时间质谱（MALDI-TOF MS）鉴定蛋白质点等一系列亚细胞器蛋白质组学方法建立了胃癌细胞内高尔基体的蛋白图谱。

最后，人们根据分离出的纯度较高的高尔基体建立了分辨率和重复性均较好的双向电泳图谱，运用质谱技术鉴定出12个蛋白质，包括蛋白合成相关蛋白、膜融合蛋白、调节蛋白、凋亡相关蛋白、运输蛋白和细胞增殖分化相关蛋白。通过亚细胞器分离纯化、双向电泳的蛋白分离及MALDI-TOF MS蛋白鉴定分析，研究人员首次成功建立了胃癌细胞SGC7901中高尔基体的蛋白质组学技术路线。

3.1 蛋白质功能预测工具

也许生物信息学方法在癌症研究中最常用的就是基因功能预测方法，但是这些数据库只存储了基因组的大约一半基因的功能。为了在微阵列资料基础上完成功能性的富集分析，基因簇的功能注解是非常重要的。近几年生物学家研发了一些基因功能预测的方法，这些方法旨在超越传统的BLAST搜索来预测基因的功能。基因功能预测可以以氨基酸序列、三级结构、与之相互作用的配体、相互作用过程或基因的表达方式为基础。其中最重要的是基于氨基酸序列的分析，因为这种方法适合于微阵列分析的全部基因[2]。

在表3中，前三项列举了三种同源搜索方法。FASTA方法虽然应用还不太广泛，但它要优于BLAST，或者至少相当。FASTA程序是第一个使用的数据库相似性搜索程序。为了达到较高的敏感程度，程序引用取代矩阵实行局部比对以获得最佳搜索。美国弗吉尼亚大学可以提供这项程序的地方版本，当然数据库搜索结果依赖于要搜索的数据库序列。如果最近的序列数据库版本在弗吉尼亚大学不能获得，那么就最好试一下京都大学（Kyoto University）的KEGG站点。PSI-BLAST（位点特异性反复BLAST）是BLAST的转化版本，PSI-BLAST的特色是每次用profile搜索数据库后再利用搜索的结果重新构建profile，然后用新的profile再次搜索数据库，如此反复直至没有新的结果产生为止。PSI-BLAST先用带空位的BLAST搜索数据库，将获得的序列通过多序列比对来构建第一个profile。PSI-BLAST自然地拓展了BLAST方法，能寻找蛋白质序列中的隐含模式，有研究表明这种方法可以有效地找到很多序列差异较大而结构功能相似的相关蛋白，所以它比BLAST和FASTA有更好的敏感性。PSI-BLAST服务可以在NCBI的BLAST主页上找到，还可以从NCBI的FTP服务器上下载PSI-BLAST的独立程序。在检查PSI-BLAST的搜索输出时，也有一些注意事项，因为假的匹配记录很容易污染分析结果[2]。

表3 蛋白质功能预测工具[2]

预测工具	类型	所在地	网站
BLAST	同源搜索	NCBI：美国国立生物技术信息中心； NIH：美国国家医学研究院	http://www.ncbi.nlm.nih.gov/BLASTselect protein-protein BLAST
FASTA	同源搜索	美国弗吉尼亚大学、 日本京都大学	http://fasta.bioch.virginia.edu http://fasta.genome.jp/
PSI-BLAST	同源搜索	NCBI：美国国立生物技术信息中心； NIH：美国国家医学研究院	http://www.ncbi.nlm.nih.gov/BLASTselect “PSI- and PHI-BLAST”
Pfam	蛋白质家族鉴定	华盛顿大学	http://pfam.wustl.edu

续上表

预测工具	类型	所在地	网站
SMART	保守结构域搜索	EMBL：欧洲分子生物学实验室	http://smart.embl-heidelberg.de
PROSITE	功能模体搜索	瑞士生物信息研究所	http://us.expasy.org/prosite http://motif.genome.ad.jp
ELM	真核生物功能结构域搜索	ELM 联合体	http://elm.eu.org
STRING	通过比较基因组学进行功能预测	EMBL（欧洲分子生物学实验室）	http://string.embl.de
PSORT	亚细胞定位预测	人类基因组中心东京大学	http://www.psort.org
PFP	通过发掘PSI-BLAST结果进行功能预测	美国普渡大学	http://dragon.bio.purdue.edu/pfp

Pfam数据库（Protein families database of alignments and HMM, 蛋白质家族比对和HMM数据库）是基于HMM模型（隐马尔可夫模型）构建并拓展起来的。它实际上是一个涵盖了生物蛋白质序列中常见结构域的序列及其相对应的隐马尔科夫模型的数据库，由英国的Sanger Institute维护。Hmmpfam的工作原理简单来说，就是将用户所提交的查询序列在Pfam库中做比对计算，然后预测出查询序列中所隐含的结构域信息[2]。

表4中描述的三个数据库资源——简单模块构架搜索工具（simple modular architecture research tool, SMART）、Motif数据库（PROSITE）以及 ELM是具有不同特点的数据模体数据库。SMART储存有蛋白质家族的保守区域，可以作为每一个基因家族的特征标记。SMART可以说是蛋白结构预测和功能分析的工具集合。简单点说，SMART就是集合了一些工具，可以预测蛋白的一些二级结构，如跨膜区（Transmembrane segment）、复合螺旋区（coiled coil region）、信号肽（Signal peptide）和蛋白结构域（PFAM domain）等。另一方面，PROSITE中的序列模体是一些重要的生物学位点，包括功能位点和容易被修饰的位点。ELM是真核生物功能位点数据库[2]。

PROSITE数据库是多序列比较而得到的单一保守序列片段，或称序列模体。PROSITE数据库是基于对蛋白质家族中同源序列多重序列比对得到的保守性区域，这些区域通常与生物学功能有关，例如酶的活性位点、配体或金属结合位点等。因此，PROSITE数据库实际上是蛋白质序列功能位点数据库。通过对PROSITE数据库的搜索，可判断该序列包含什么样的功能位点，从而推测其可能属于哪一个蛋白质家族。Prosite数据库实际上包括两个数据库文件：一个为数据文件，即Prosite，该文件给出了能进行匹配的序列及序列的详细信息；另一个为说明文件，即PrositeDoc。PrositeDoc说明文件中给出该序列模式的生物学功能及其文献资料来源。PROSITE数据库使用正则表达式来表示序列模式[2]。

STRING是一个已知和预测基因间功能联系的数据库。STRING一个有趣的特点是，一个查询序列的功能是利用比较基因组学方法预测的。例如，假设一个要查询的基因是几个基因组中功能已知的基因，这几个基因组进化上相关，那么预示着要查询的基因与相邻基因可能涉及相同的途径或功能[2]。

具有相同的系统发生的那些基因，或同时存在和同时消失的那些基因也预示着他们的功能是相互联系的。SMART也利用微阵列中的共表达来分析，用户可以利用SMART站点进行功能预测，基因功能之间的联系资料也可以免费获得[2]。

PSORT工具可以预测基因的亚细胞定位。从根本上说，PSORT工具基于其氨基酸序列预测蛋白质亚细胞定位。它利用机器将要查询蛋白质的特殊序列（如信号肽序列）检测和分类并定位到已知位置。PSORT II是广泛使用的蛋白质亚细胞定位分析软件，通过输入的氨基酸序列，能够预测出其在亚细胞结构中可能位置[2]。

PFP（蛋白质功能预测）服务器是最近研发的。不同于传统的PSI-BLAST，PFP利用序列采样数可以发掘更多的功能信息[2]。

在列出的蛋白质功能预测工具中，BLAST、FASTA和Pfam最可靠，但它们无法提供关于已经储存在公共数据库中的已注解基因的更多的信息。其它方法都优于上述三种方法，且有更广的覆盖率，但是使用时要小心，因为有相对较高的假采样。为了避免这种情况发生，应该多采样几种方法，检查获得结果的一致性[2]。

表4 蛋白质结构预测工具[2]

预测工具	类型	所在地	网址
PSIPRED	二级结构	伦敦大学	http://bioinf.cs.ucl.ac.uk/psipred/
PORTER	二级结构	都柏林学院	http://distill.ucd.ie/porter/
SAM-T02	二级结构	加州大学圣塔克鲁兹分校	http://www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html
SABLE	二级结构和溶剂可接近性	美国辛辛那提儿童研究基金会儿童医院医疗中心	http://sable.cchmc.org/
PredictProtein	二级结构和其它	美国哥伦比亚大学	http://cubic.bioc.columbia.edu/predictprotein/
COILS	两个或以上的α螺旋组成的超螺旋结构区域（卷曲螺旋区域）	瑞士，EMBnet	http://www.ch.embnet.org/software/COILSform.html
GlobPlot	无规则区域	欧洲分子生物学实验室	http://globplot.embl.de/
PONDR	无规则区域	印地安纳大学	http://www.pondr.com/
TMHMM	跨膜结构域	丹麦科技大学	http://www.cbs.dtu.dk/services/TMHMM-2.0/
HMMTOP	跨膜结构域	匈牙利科学院	http://www.enzim.hu/hmmtop/

续上表

预测工具	类型	所在地	网址
SWISS-MODEL	3D结构 同源建模法	瑞士生物信息研究所	http://swissmodel.expasy.org
HHPred	3D结构 同源建模法	马克斯•普朗克科学促进协会	http://protevo.eb.tuebingenmpg.de/toolkit/index.php?view=hhpred
MODELLER	3D结构 同源建模法	加州大学旧金山分校	http://salilab.org/modeller/
FUGUE	3D结构, 指认方法（线引法或穿线法）	剑桥大学	http://www-cryst.bioc.camac.uk/~fugue/
Phyre	3D结构, 指认方法（线引法或穿线法）	帝国理工学院（伦敦大学（University of London）的独立学院）	http://www.sbg.bio.ic.ac.uk/~phyre/
SPARKS	3D结构, 指认方法（线引法或穿线法）	纽约州立大学水牛城分校	http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sparks2.html
Robetta	3D结构 从头计算法（ab initio）	华盛顿大学	http://robetta.bakerlab.org/

3.2 蛋白质结构预测工具

当候选基因被选择出来通过微阵列进行实验检测时，生物信息学蛋白质预测工具就对设计生物化学实验非常重要。例如，一个基因预测的二级结构就对预测结构域很有益处，因为这对设计有限的蛋白质水解以鉴定基因的功能区域非常重要。当今二级结构预测算法的准确率大约75%，足以达到平常预测的要求。表4中列举了5种二级结构预测工具，它们都利用机器学习的技术来识别大家熟悉的二级结构，如α-螺旋和β-折叠。机器学习就是要使计算机能模拟人的学习行为自动通过学习获取知识和技能，不断改善性能，实现自我完善。机器学习研究的就是如何通过识别和利用现有知识来获取新知识和新技能。 PSI-PRED、PORTER、SABLE和PredictProtein seartificial利用人工神经网络（Artificial Neural Networks，简写为ANNs），而SAM-T02则利用HMM。 SABLE和 PORTER 被认为是这个领域中最准确的预测工具，它们的准确率分别可以达到78.4%和79%。尽管列表中的工具在蛋白质预测方面准确率还相对较低，但这些工具不仅能够预测二级结构，也能够预测其它结构信息，包括混乱区域、两个或两个以上的α螺旋组成的超螺旋结构区域、每个残基的可溶解程度以及要搜索序列的模体结构，所以他们可以用来“一步完成”蛋白质序列分析。COILS通过序列中周期出现的疏水残基来预测蛋白质的卷曲螺旋。卷曲螺旋是存在于多种天然蛋白质中的结构模式。近年来，人们通过对天然蛋白质中的卷曲螺旋结构以及根据已有知识设计合成的卷曲螺旋结构的研究，已基本掌握了这类结构模式的特点，并将特异的卷曲螺旋结构应用于生化分析、工业和医药卫生等领域[3]。

GlobPlot和PONDR是预测蛋白质固有无规则区域的工具，这些区域的天然构象中，没有稳定的蛋白质二级结构。这些无规则区域的重要性最近才刚刚被人们发现，因为它们是很多重要的功能性位点，例如其它蛋白质和配体的结合区域位于球状蛋白质的结构域的外部，所以本质上是没有规则的[3]。

HMMTOP是预测蛋白序列的跨膜螺旋与拓扑结构工具，TMHMM是预测蛋白的跨膜螺旋工具。它们都利用了HMM。跨膜结构域预测是生物信息学在蛋白质结构预测中应用得最成功，HMMTOP预测的跨膜蛋白质98%的结构域和85%的拓扑结构是正确的。且上述两种工具是基于网络的，HMMTOP也提供本地拷贝[3]。

表5的后面列举了5种预测三级结构的工具。预测蛋白质三级结构的方法在最近几十年中有了较大的改进，并且有些方法的准确率已经足以应用于实践[3]。

结构预测方法大致分为三类

(1) 同源性建模（homology modeling）方法：这类方法的理论依据是，如果两个蛋白质的序列比较相似，则其结构也有很大可能比较相似。有实验表明，如果序列相似性高于75%，则可以使用这种方法进行粗略的预测。这类方法的优点是准确度高，缺点是只能处理和模板库中蛋白质序列相似性较高的情况；

(2) 从头计算（Ab initio或*de novo*） 方法：这类方法的依据是热力学理论，即求蛋白质能量最小的状态。生物学家和物理学家等认为从原理上讲这是影响蛋白质结构的本质因素。然而由于巨大的计算量，这种方法并不实用，目前只能计算几个氨基酸形成的结构。IBM 开发的 Blue Gene 超级计算机就是要解决这个问题；

(3) 穿线法（Threading或fold recognition）：由于 Ab Initio 方法目前只有理论上的意义，Homology方法受限于待求蛋白质必需和已知模板库中某个蛋白质有较高的序列相似性，因此对于其它大部分蛋白质来说，有必要寻求新的方法，于是Threading应运而生。

以上三种方法中，Ab Initio方法不依赖于已知结构，其余两种则需要已知结构的协助。通常将蛋白质序列和其真实三级结构组织成模板库，待预测三级结构的蛋白质序列，则称之为查询序列（query sequence）。

SWISS-MODEL和HHPred 是基于网络的同源建模工具，HHPred软件也可以下载。MODELLER是这一类型软件中应用最早和最广泛的软件。MODELLER和SWISS-MODEL有同源建模数据库。接下来三个工具FUGUE、Phyre和SPARKS属于穿线法。穿线法可在数据库中搜寻和待测蛋白结构非常匹配的模板蛋白质。与同源建模法不同，穿线法中模板蛋白质和待测蛋白质明显的序列相似性并不是必要条件。穿线法可以检测一个数据库中亲缘关系很远的蛋白质，Z-score作为一个统计学值，可以显示模板蛋白质和待测蛋白之间的匹配程度，当Z-score较低时，就意味着没有匹配搜索的结构[3]。

最后，Robetta工具属于从头计算法。它利用从数据库收集的序列片段来组装模型，是一个自动化的蛋白质结构预测服务工具。它由贝克实验室提供，用于非商业性质的从头计算和比较建模[3]。

3.3 蛋白质-蛋白质相互作用数据库

表5列出了蛋白质之间相互作用（protein-protein interactions, PPI）的数据库。在过去的几年中，有大规模实验开始研究蛋白质之间的相互作用，并且很多相关资源可以在互联网上得到。了解一个基因编码蛋白质与其它蛋白质之间的关系，对于推测这个基因发挥功能所需的背景关系具有重要意义。BIND（biomolecular interaction network database）数据库是BOND（biomolecular object network databank）数据库的一个子数据库，它是现在最大的PPI数据库。BIND数据库收录了1500种生物分子之间的200,000种相互作用的数据。这种相互作用不仅包括蛋白质之间的相互作用，还包括蛋白质与DNA、RNA、小分子、脂质以及糖类物质之间的相互作用。BIND数据库每日更新、覆盖面广，包含人、果蝇、酵母、线虫等物种的PPI[3]。

表5 蛋白质-蛋白质相互作用数据库和数据库工具

工具	类型	所在地	网址
BIND	蛋白质-蛋白质相互作用途径	加拿大多伦多西乃山医院	http://bind.ca/
DIP	蛋白质-蛋白质相互作用	加州大学洛杉矶分校	http://dip.doe-mbi.ucla.edu/
MIPS	哺乳动物的蛋白质-蛋白质相互作用	慕尼黑蛋白质序列信息中心	http://mips.gsf.de/proj/ppi/
HPRD	人类蛋白质参考资源	美国约翰霍普金斯大学	http://www.hprd.org/
GRID	酵母、果蝇和线虫的遗传和生理作用	加拿大多伦多西乃山医院	http://biodata.mshri.on.ca/grid/
IntAct	蛋白质相互作用数据库的db系统和工具的开发资源	欧洲生物信息学中心	http://www.ebi.ac.uk/intact/
Ospray	蛋白质相互作用的可视化工具	加拿大多伦多西乃山医院	http://biodata.mshri.on.ca/osprey/

在 BIND 数据库中，PPI 被分成三大类：二元分子相互作用（binary interaction）、分子复合物（molecularcomplexe）以及生物途径（biological pathway），它们分别从不同层面呈现了分子间的相互作用关系[3]。

DIP（database of interacting protein）数据库专门存储经实验证实的来自文献报道的二元PPI，以及来自 PDB（protein data bank）数据库的蛋白质复合物。目前DIP收录了18,000种相互作用的数据，DIP的目的在于建立一个简单、易用、高度可信的PPI公共数据库[3]。

MIPS（mammalian protein-protein interactiondatabase）数据库同样利用文献挖掘技术，专门存储哺乳动物的PPI，主要包括人、大鼠和小鼠等物种。该数据库详细记录了蛋白质相互作用的类型、实验证据及其结合位点。同时，它还提供蛋白质名称、实验方法和物种等多种查询方式[3]。

HPRD（human protein reference database）数据库是包含蛋白质注释、PPI、转录后修饰和亚细胞定位等多种信息的综合数据库[3]。

IntAct也是一个存储和分析生物分子间相互作用的公共数据库。它主要记录二元相互作用及其实验方法、实验条件和相互作用结构域，包括人、酵母、果蝇和大肠杆菌等物种。 IntAct 数据库分为基本查询和高级查询：基本查询可以根据蛋白质名称、PubMedID等进行简单搜索；高级查询根据实验方法和IntAct自定义的控制词汇进行查询。GRID存储了酵母、果蝇和线虫的遗传和生理作用。Osprey蛋白质相互作用网络可视化系统是加拿大多伦多大学一个生物信息学研究组开发的，其目的在于更好地研究蛋白质相互作用网络和蛋白质复合物[3]。

四、大规模研究癌症的技术及其应用

如前所述，大规模研究对现代生物学研究的发展至关重要。然而，这些方法，如基因组学、转录组学和蛋白质组学的存在，是因为实验技术产生和开发了大量的资料，如DNA测序技术、微阵列技术、基因表达序列分析技术（SAGE）和质谱技术等。这些技术也促进了生物信息学的发展。值得注意的是，所有的这些技术原来几乎都主要用于对癌症的研究。下面将描述这些技术及其是如何用于生物研究的。

1. 表达序列标签技术

序列表达标签（expressed sequence tag, EST）技术是快速高效认识生物体基因和基因组的研究手段之一。此技术以大规模cDNA测序为基础，即提取生物体的mRNA，进一步获得cDNA文库，挑选其中300~500bp的部分（即EST序列）进行序列测定，然后通过生物信息学手段得到全长的基因序列。通常EST序列位于一段cDNA的3’ 端非翻译区，也可以在该cDNA中随机选取。利用EST技术克隆基因及进行功能分析，使克隆和定位新基因的策略发生了巨大变革。EST的产生流程是从特定状态的组织或细胞中分离出mRNA，将mRNA逆转录成cDNA并亚克隆到载体中，然后再利用载体上的引物对插入片段测序，测序出来的片段结果即称为EST。这项技术最早是作为一种替代的测定表达基因的方法，主要是代替缓慢、小规模和费时费力的Northern杂交。1991年，Adam等人在609个序列中得到了第一套EST，并用来研究中枢神经系统基因表达。在此之后，很多研究都利用EST来测定不同组织和器官的基因表达情况，EST的量也迅速增加。截至目前，在NCBI为基础的EST公共数据库和dbEST公共数据库中，载有来自1587个生物体的超过54万个序列（dbEST发表072508）。

从本质上讲，EST分析有着明确的生物信息学路径

第一步，序列修剪：低质量的碱基或受污染序列都将被删除；

第二步，根据相似性对EST进行分类；

最后一步，在序列水平，基于它们的重叠注解EST，带有已知序列的基因或基因组区域。在这些初步处理之后，产生的结果可以用于基因发现、基因表达分析或转录后的变异分析，如选择性剪接或替代多聚腺苷酸化研究等。