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。

EST与癌症之间关系的研究早就证明是富有成果的。首先，许多工作都采用EST技术研究癌症；其次，两个不同的癌症计划突出了EST对转录组学研究的重要性。癌症基因组剖析计划（CGAP）从超过10个癌症类型中得到了大约150万个EST。人类癌症基因组计划（HCGP）利用一种称为开放阅读框EST的新技术，从多于10个癌症类型中得到了超过100万个流行肿瘤的EST。多年来，癌症的EST数量主要存储在表达序列标签数据库（database EST, dbEST）。

2．基因表达序列分析技术和大规模平行测序技术

基因表达序列分析（SAGE）和大规模平行测序技术（MPSS）是两个基于短标签测序的方法。它们可以通过短的序列标签（SAGE短的序列标签为10或17bp，MPSS大规模平行测序技术短的序列标签为13或20 bp），能够量化基因的表达，这些标签在多聚腺苷酸转录物内，毗邻NlaIII（SAGE）或DpnII（MPSS）的3’端。SAGE和MPSS实验的结果是短序列集，某一特定标签的频率代表对应转录物的丰度。

虽然SAGE和MPSS可以产生类似的输出结果，即一系列的标签，但它们的实验方案是完全不同的。例如，SAGE使用传统的克隆及DNA测序方法，MPSS则基于酶消化和杂交，利用新的克隆和专属的平行测序方案。因此，SAGE平均输出包含10万个标签，而MPSS能够输出包含超过1,000,000个标签。

一个在SAGE和MPSS数据分析中的重要步骤就是正确指定/定位标签到基因。基本上有三种策略来完成这个过程：

(1) 基于数据库的注解，数据库通过内部电脑方案构建；

(2) 注解，标签对标签，通过网络站点完成，如SAGE Genie数据库和SAGEmap数据库。SAGEmap是美国国家癌症研究所（NCI）提供的一个在线SAGE分析工具，其数据库包含多种正常组织、癌前病变组织和肿瘤组织的SAGE文库。SAGE Genie数据库在SAGEmap的基础上提供了更为友好的界面，通过查询SAGE Genie数据库，可以获知基因在正常组织和癌变组织中的相对表达量；

(3) 利用参考数据库大规模注解。

SAGE分析的另一个重要步骤是识别每个样本中的不同标签。为此，有几种方法可以采取，与Baggerly等人、Thygesen和Zwinderman以及Zuyderduyn所描述的一样。

SAGE的显著特点是快速高效地、接近完整地获得基因组的表达信息，它可以定量分析已知基因及未知基因表达情况。在疾病组织、癌细胞等差异表达谱的研究中，SAGE可以帮助获得完整转录组学图谱，发现新的基因及其功能、作用机制和通路等信息。MPSS是对SAGE的改进，它能在短时间内检测细胞或组织内全部基因的表达情况，是功能基因组研究的有效工具。但它需要的配套软硬件较为昂贵，目前国内外相关的应用报道不多。MPSS技术对于致病基因的识别、揭示基因在疾病中的作用、分析药物的药效等都非常有价值，该技术的发展将在基因组功能及其相关领域研究中发挥巨大的作用。

SAGE和MPSS技术自出现以来，就被人们广泛应用（特别是SAGE技术）。许多公布的实验，尤其是癌症研究实验都已经开始用这些方法来分析基因的整体表达。最早对人类癌基因组进行全面基因分析的技术是SAGE。

有三个特点令SAGE和MPSS成为癌症中基因表达分析的有效方法

(1) SAGE与MPSS不用事先选择基因进行研究就可以提供mRNA群体的特点，这样可以发现新基因，例如致癌基因；

(2) 在一个实验中获得的数据可以直接比较任何其它实验室的数据，或比较现有公共数据库的数据，允许一个大规模的基因表达比较；

(3) 生成的数据，标签的频率是以数字格式展现，能够稳健地进行基因表达统计分析。

3．微阵列技术

微阵列自上世纪90年代中期出现之后，就迅速在研究领域得到了广泛的应用。近年来，基因芯片实验已成为很多生物领域研究性文章中“必有”的名词。

虽然有很多方案和系统类型适用于微阵列，但其基本技术涉及4个主要步骤
(1) 从生物样品中提取RNA；
(2) 复制RNA成为 cDNA，其中包括插入荧光核苷酸或标记，目的是以后可以荧光染色；
(3) 杂交的标记RNA（cDNA）到微阵列芯片；
(4) 在激光灯下扫描微阵列芯片并测量基因表达。

现在，最常见的微阵列平台是“cDNA微阵列”和“寡核苷酸微阵列”。寡核苷酸基因微阵列的优势在于所有探针设计有类似的杂交温度和亲和力。尽管存在差异，对于每个样品，两个平台都能够测量超过10k基因的表达。

微阵列实验中的一个关键步骤是处理和分析结果。即使有许多软件包可以利用，但仍然很难找到一个单一的软件来处理原始数据（如背景校正和规范化）和进行分析（如识别和呈现的基因表达差异）。可能最好的办法是发展生物信息学协议（途径），整合软件包等。

除了在基础研究中的使用，微阵列技术也已广泛用于确定疾病相关基因的研究中。例如，微阵列技术已被用于确定：(1) 肿瘤与健康的组织中基因表达的差异；(2) 与肿瘤的进展相关的基因；(3) 以及能够准确从正常状态区分癌症的基因，甚至是肿瘤的多个亚型。下面列举了集中分析功能簇的微阵列软件（表6）、利用统计学方法的微阵列分析软件（表7）以及微阵列数据资料库(表8）。

表6 集中分析功能簇的微阵列软件[2]

软件	所在地	网站
GoMiner	NCI（美国国家癌症研究所）与 NIH（美国国家医学研究院）	http://discover.nci.nih.gov/gominer/
GoSurfer	美国哈佛大学	http://www.biostat.harvard.edu/complab/gosurfer/
GenMAPP	加州大学旧金山分校	http://www.genmapp.org
ArrayTrack	美国食品药品监督管理局（FDA）	http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm

GoMiner是一个利用Ontology来解释基因组和蛋白数据的软件包，它可以很好地利用GO的层次结构自动地对任意长度的基因列表进行功能的富集。它的开发主要是为了对芯片数据进行生物学解释。GoMiner可以通过输入来源于表达谱的一个差异表达（包括上调和下调）的基因列表以及一个基因全集列表，从而计算出发生差异表达的基因集在GO中的富集情况，因此它可以应用于对组学数据进行快速有效的分析和组织。GoSurfer在分析基因集（来自于基因组范围的计算、芯片分析或其它相应的高端方法）时，使用GO信息对调查微阵列分析或染色体广泛计算的结果非常有用。

为了帮助生物学家更好地理解基因组学和蛋白组学中的路径问题，格莱斯顿心血管疾病研究所研发了GenMAPP。GenMAPP又名Gene Microarray Pathway Profile，是一个microarray表达式数据形象化工具。GenMAPP允许查看表示基因分组和生物路的映射数据，还可以对大量的基因表达数据进行可视化处理以及生物途径分析。数据库（生物途径图）来源于KEGG数据库。

通过Genmapp软件分析可对任何感兴趣的生物过程进行基因差异表达分析。GenMAPP被整合到了最流行的生物信息学工具——由生物信息组织开发的Reactome和Cytoscape之中。Reactome和Cytoscape软件是系统生物学研究所、加利福尼亚大学圣地亚哥分校、美国斯隆凯特琳癌症中心以及巴斯德研究所共同合作研发的，这两个工具能够帮助研究人员对基因组规模的数据进行观察和分析。这些观察和分析并不是孤立的，而是与生物学路径问题等内容联系的。ArrayTrack软件既可以分析差异表达的生物通路，也可以用于毒理分析以及药理分析的生物芯片分析系统。ArrayTrack软件是生物芯片应用领域最重要的工具之一。

表7 利用统计学方法的微阵列分析软件[2]

软件/连接	所在地	网站
caGEDA	美国匹兹堡大学	http://bioinformatics.upmc.edu/GE2/GEDA.html
SAM	美国斯坦福大学	http://www-stat.stanford.edu/~tibs/SAM/
NUDGE	美国华盛顿大学	http://www.bioconductor.org/
Microarray Software Comparison	香港中文大学	http://ihome.cuhk.edu.hk/~b400559/arraysoftrpackages.html

表8：微阵列数据资料库[2]

数据资料库	所在地	网站
GEO	美国国立生物技术信息中心与美国国家医学研究院	http://www.ncbi.nlm.nih.gov/projects/geo/
ArrayExpress	欧洲生物信息学研究所	http://www.ebi.ac.uk/arrayexpress/
CIBEX	日本Nat. Inst. Genetics	http://cibex.nig.ac.jp
Standard Microarray Database (SMD)	美国斯坦福大学	http://smd.stanford.edu/
基因表达数据库达（GXD）	The Jackson Lab	http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml
Oncomine	美国密歇根大学	http://www.oncomine.org

caGEDA 可以为微阵列分析的每一步提供很多多用途的统计学工具，如预处理、特征分析和患者预测模型开发。在它们的数据集中，用户可以很便利地进行各种比较分析。基因芯片显著性分析（SAM）方法是微矩阵显著性分析软件和EXCEL软件的插件，由Stanford大学编制而成。NUDGE和SAM都用到了R语言。Microarray Software Comparison站点包含了很多基因表达统计分析的网站链接。

GEO是由NCBI在 2000 年研发的一个基因表达和杂交微阵列数据库。它可同时作为获取来自不同生物体的基因表达数据的在线资源。截至2004 年 3 月，数据库中包含的内容有605个Platform、14391个Sample以及816个Serial。

Platform 是关于物理反应物的信息，Sample是关于待检测的样本信息和使用单个Platform产生的数据，而Series则是关于样本集的信息，它反映了样本间的相关性和组织性。ArrayExpress基于基因表达数据的微阵列公共知识库，目的是存储被注释的数据。目前它包含多个基因表达数据集和与实验相关的原始图像集。ArrayExpress数据库接受MAGE-ML格式的数据递交，或者通过MIAMExpress的基于Web界面注释和递交的数据。 ArrayExpress提供一个简单的基于Web的数据查询界面，并直接与Expession Profiler数据分析工具相连，它可以进行表达数据聚类和其它类型的Web数据挖掘，并将进一步研发多个实验和数据库间的交叉查询。

ArrayExpress数据库中的数据将与所有由 EBI维护的或在线的数据库相联接。SMD是一个使用Oracle作为数据库管理软件的关系数据库。 SMD存储微阵列实验的原始数据、归一化数据和对应的图像文件。另外，SMD还提供数据获取、分析和可视化的界面。目前，SMD包括层次聚类和自组织映射等方法，还将加入k-平均聚类、单值分解和丢失值归纳等方法。CIBEX是日本建立的DNA数据库。它作为公共数据库，存放了高通量的基因表达实验数据，包括以微阵列为基础的实验测量基因、基因表达系列分析（表达序列标签）以及质谱蛋白质组数据。Oncomine数据库的目的是收集、标准化、分析和发布已经公开发表的癌症基因表达数据。该数据库目前包含16,656张芯片和49种癌症类型。GXD基因表达数据库还存放了小鼠的基因表达和遗传资源。

4．下一代基因组测序技术

DNA（和cDNA）测序经历了一个巨大的变化，在此过程中，DNA测序反应的成本不断下降，被测序的核苷酸的数量则不断增加。现在，我们正处于下一代（或第二代）测序技术时期，并且第三代（下下一代）即将到来。然而，是什么能够使下一代测序方法更便宜、速度比传统的测序方法更快的呢？相对于“Sanger”测序（传统方法）可以在同一时间内阅读并处理96个序列，新一代测序流程能够平行处理数百万序列，并在几个小时内产生大量的序列。

第三代测序技术是基于纳米孔（nanopore）的单分子读取技术。它有着更快的数据读取速度，应用潜能也势必超越测序技术。例如，Roche-454测序、Illumina-Solexa高通量测序以及ABI公司推出的自主研发的SOLiD 测序仪（ABI SOLiD）平均每次运行分别可以产生120万个读取数据（平均阅读长度400 bp）、40~50个万读取数据（平均读取长度36 bp）以及大约1.00亿个读取数据（平均读取长度25~35 bp）。

新一代测序很快就应用到了和医学相关的领域，如癌症基因组学。这些技术也许有利于分析突变、改进基因表达的量化研究和发现/研究调节RNA分子（也就是非编码RNA），其许多功能都有助于了解癌症的特点。然而，尽管有这些巨大的潜力，下一代测序技术的应用现在仍然被一些因素阻碍，如每次运行上涨的成本（至少8000美元）和每次实验都会产生巨大的数据，而生物信息学处理这些数据仍面临着一定的难度。

5. 质谱技术

质谱分析技术是一个强大的分析技术。它可以用来识别未知化合物和量化已知化合物，并阐明一个分子的化学性质。质谱（MS）是几乎所有蛋白质组学实验的核心。基本上，占主导地位的MS工作流程开始于一个特定位点的酶消化，使蛋白质变成肽。接下来，肽被处理成挥发性物质，质谱仪产生每个样品的光谱图。最后，把光谱图与数据库里的肽序列进行比较，推断蛋白质序列。

MS实验的第一步，酶消化是由1个蛋白酶（如胰蛋白酶）完成的。为了使肽变成挥发性物质，人们通常采用两种方法——基质辅助激光解吸/电离（MALDI）和电喷雾（ESI）。MALDI用于挥发含有少量肽的混合物；电喷雾用于挥发含有大量的肽的肽混合物。最后是生物信息学方案，即光谱（观察到的峰）的处理和与可以消化的蛋白质序列数据库的肽序列比较。由于蛋白质鉴定依赖于与序列数据库匹配，目前蛋白质组学主要限制那些综合序列数据库可用的物种。

MS可直接应用的是检测蛋白质或肽的峰。它们的质量和所带电荷不同，从而可以将癌症患者的样品与正常个体的样品进行比较。例如，采用这一方法，Nakagawa等人确定了两个乳腺癌相关的多肽；Hao等人确定了一些有关胃癌的多肽；而Sun等人则确定了116个可以用来区分肝癌和正常肝细胞的蛋白质。

即使MS的应用对癌症研究有巨大的潜力，但由于个体间的基因变异和血浆蛋白质组动态变化是多种因素（如性别、年龄和健康状况）综合的结果，所以这种类型的分析仍然存在一些大的障碍。

五、各种“组学”信息的整合

正如以前指出的那样，在基因组学、转录组学和蛋白质组学方面的进展为生物学家提供了大量要处理的数据。然而，这些数据大多来自不同的平台或存储在不同的数据库中，数据整合是非常不容易的，甚至是不可行的。鉴于此，人们希望形成一个生物信息学的特定领域，以便解决这些很重要的问题，这就是“整合基因组学”。