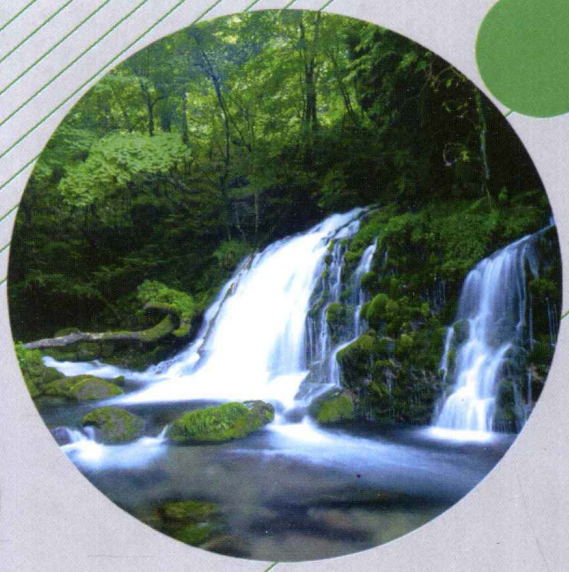


生物信息学中的数据挖掘方法及应用

梁艳春　张　琛　杜　伟　吴春国　曹忠波　著

科学出版社

生物信息学中的数据挖掘方法及应用

中国科学院科学出版基金资助出版

生物信息学中的数据挖掘方法及应用

梁艳春　张　琛　杜　伟　吴春国　曹忠波 著

科学出版社

北京

内 容 简 介

本书针对生物信息学领域的一些前沿课题，以数据挖掘算法为中心，系统地介绍了机器学习、统计学习及多种智能算法在生物信息学相关领域的应用，为生物信息学方向的初学者提供了入门知识，也为相关研究人员在特定方向深入研究提供了参考信息。主要内容包括操纵子预测、原核生物系统发生树的构建、基于数据扰动的误标记样本检测、差异表达基因识别以及基因表达数据的特征选择等。

本书可以作为高年级本科生或研究生的生物信息学课程教材，也可供相关研究领域生命科学工作者和计算机应用人员参考。

图书在版编目(CIP)数据

生物信息学中的数据挖掘方法及应用 / 梁艳春等著. —北京：科学出版社，2011

ISBN 978-7-03-032658-4

I. ①生… II. ①梁… III. ①生物信息论-数据采集-研究
IV. ①Q811.4

中国版本图书馆 CIP 数据核字(2011)第 222710 号

责任编辑：杨 锐 胡 凯 / 责任校对：刘亚琦
责任印制：赵 博 / 封面设计：王 浩

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com
源海印刷有限责任公司印刷
科学出版社编务公司排版制作
科学出版社发行 各地新华书店经销
*
2011 年 11 月第 一 版 开本：B5 (720 × 1000)
2011 年 11 月第一次印刷 印张：13 1/2
印数：1—2 500 字数：260 000
定价：48.00 元
(如有印装质量问题，我社负责调换)

前　言

生物技术的飞速发展使人类逐步加深了对自身遗传信息的了解，尤其是随着人类基因组计划(human genome project, HGP)的实施，人们逐步掌握了破解遗传密码的基本元素。现代的高通量测序技术，使得人们可以非常容易地获得现存物种的全基因组核酸序列或蛋白质组的氨基酸序列。然而，正如我们拿到了一份密文，却无法破译其中蕴含的信息一样，尽管我们可以拿到基因组的核酸序列，但是距离真正破解基因组遗传密码还有相当长的路要走。这是因为基因组本身是庞大的，而其中蕴含的编码机制、转录机制和调控机制等又是极为复杂的。每时每刻，来自不同物种的基因组编码信息都会从世界各地研究机构的测序仪中源源不断地喷薄而出，摆在科研人员面前的重要任务就是解读出海量组学数据中埋藏的关于生命的基本规律。正是在这种背景下，生物学与计算机科学共同孕育了生物信息学这门前沿交叉学科。

生物信息学的研究领域十分广泛，从序列比对到基因发现与功能研究，从基因表达分析到蛋白质结构与功能预测，乃至于更为复杂的调控网络、代谢网络及蛋白质相互作用网络等，都有生物信息技术大显身手的舞台。本书将致力于使用生物信息学手段来进行基因组相关的数据分析，我们所关注的重点是基因表达信息。检测和分析基因的表达情况是揭示基因组调控机制的重要前提，而许多基因功能的分析工作也依赖于基因表达分析。随着科学研究的进步，传统的着眼于单个基因或单个蛋白质的观点已经过时，取而代之的是整体性地研究基因组中的所有基因、mRNA、蛋白质及各种代谢产物，这种整体性的思想正是系统生物学(systems biology)的起源。

本书将围绕生物信息学，对基因组和基因表达分析中的几个典型问题进行深入而具体的阐述，并介绍如何利用相关的数据挖掘方法来解决这些问题。本书所阐述的内容主要集中在基因组分析和基因表达数据分析两个方面。对于基因组分析方法，本书主要围绕原核生物的基因组，重点阐述操纵子预测方法和原核生物系统发生树构建方法。对于基因表达数据，本书将按照基因表达数据分析的流程，对误标记样本检测算法、差异表达基因的识别方法，以及基因芯片特征选取方法进行详细的介绍。在本书所介绍的生物信息学数据挖掘方法中，既有传统的经典算法，又有最近提出的新颖算法，同时给出了这些方法的模拟实验，通过比较研究展现了各种方法的特点。

本书的主要内容分为 7 章。

第 1 章阐述了生物信息学相关领域的发展现状，对生物信息学数据挖掘分析所研究的对象和领域进行了介绍。

第 2 章对操纵子预测的研究背景及现状进行了介绍，并分析了操纵子预测中使用的相关数据和模型构建基础，同时重点阐述了两种基于计算智能方法的操纵子预测方法。

第 3 章介绍了原核生物系统发生树构建的研究背景及现状，并论述了构建原核生物系统发生树的相关数据和方法基础，最后重点介绍了基于连续直系同源基因的系统发生树构建方法及基于全基因组序列和注释信息的系统发生树构建方法。

第 4 章主要阐述基于数据扰动的基因芯片误标记样本检测算法，对基于 LOOPC 矩阵的 CL-Stability 算法和 LOOE-Sensitivity 算法，以及基于数据扰动的基因芯片误标记样本检测算法做了全面的描述，并利用数值实验对这些算法进行了比较。

第 5 章主要阐述基因表达数据中的差异表达基因识别，重点论述了基于统计学显著性分析的 T-test 方法、SAM 方法、RankProd 方法，以及基于总体数据集统计评估的检测等一系列方法。

第 6 章介绍了基于微阵列数据的特征选择方法的相关基础，重点阐述了多阶段特征选择算法、双向局部化特征选择算法和基于改进遗传算法的特征选择算法，并给出了比较全面的数值实验比较分析结果。

第 7 章介绍了双聚类算法的相关基础，主要针对癌症基因芯片数据的相关分析这一热点研究问题，应用改进的双聚类算法对其进行分析和讨论，验证了改进双聚类算法的有效性。

本书获得中国科学院科学出版基金、国家高技术研究发展计划(863)项目(2009AA02Z307)和国家自然科学基金项目(60803052、61073075、10872077)的资助。由于时间仓促和作者水平所限，书中不当之处在所难免，请读者批评指正。

作　者

2011 年 9 月

目　录

第1章 绪 论

1.1 什么是生物信息学

生物信息学(bioinformatics)是一门新兴的前沿交叉学科，它的研究焦点主要集中于使用统计学和计算机科学工具，分析和解释海量分子生物学数据信息。生物信息学作为一门专门的学科，发轫于20世纪80年代。“生物信息学”这一名称首先是由Paulien Hogeweg和Ben Hesper于1978年提出的[1]，而直到20世纪80年代末才被广泛使用。在这一时期，由于测序技术的飞速发展，尤其是随着人类基因组计划(human genome project, HGP)的实施，海量的DNA序列信息从世界各地研究机构的测序仪中源源不断地喷薄而出。基因组信息就像是一本密码书，尽管这本书中只有A、T、C、G四个字母，然而其容量却是惊人的。例如，人类基因组序列一共约含有3 310 004 815个碱基对，而正是以这些信息为蓝本，造就了自然界生命和物种的多样性。那么，如何解读这样海量的信息，就成为摆在生命科学研究工作者面前的一道难题。幸好，计算机信息技术已经有了长足的发展，而处理海量数据的挖掘工作正是计算机科学的专长。因此，大量的计算机科学家和统计学家进入了这一领域，帮助生物学家来完成大量生物信息数据的处理工作，从而产生了生物信息学这一学科。

到了21世纪，人类步入了信息化时代，分子生物学也有了更进一步的发展，包括基因组学、蛋白质组学、转录组学等学科都在分子生物学领域扮演了重要的角色；相对而言，各种组学数据所呈现的多维度、多粒度、海量庞杂等特点也让生物信息学作为一种分析和研究的手段有了不可替代的用武之地。有人做了一个有趣而形象的类比：计算机科学如今对于分子生物学的作用，就如同17~18世纪数学对于物理学的作用。当时，由于物理学的发展，大量的计算工作需要完成，由此，诸如微积分等一系列新的数学方面的研究工作随之完成，可以说数学作为工具帮助物理学的发展，反过来物理学也作为驱动力促进数学的进步。同样地，如今分子生物学的海量数据分析需要计算机科学以及统计科学等学科的帮助，相应地，大量新型的计算机算法理论也随之不断涌现。因此，我们有理由相信，在21世纪中期，随着人类对于生命本身的不断了解，生物信息学一定会扮演更为重要的角色，一方面它将为人类逐渐解开生命的密码大显身手，另一方面也将为人类的信息技术书写新的篇章。

1.2 生物信息学的研究对象

生物信息学致力于对各种分子生物学数据的收集、分析和处理，其研究的主要对象包括基因组数据、蛋白质组数据和基因表达数据等。

1.2.1 基因组数据

对于大部分生物，DNA 组成了细胞中的基因组，而噬菌体和病毒的基因组则有可能由 DNA 及 RNA 组成。一般来讲，双链 DNA 基因组的信息都是正向读取的，即从分子的 5′端至 3′端。基因组的信息对于 DNA 而言主要包含在其 4 种碱基中：鸟嘌呤(简记为 G)、腺嘌呤(简记为 A)、胞嘧啶(简记为 C)和胸腺嘧啶(简记为 T)。对于基因组而言，除了 DNA，RNA 也是具有遗传信息的物质，其中包括含有蛋白质信息的 mRNA，以及具有调控作用的 miRNA 和 siRNA 等非编码 RNA。

基因组数据的累积速度是惊人的，在目前最权威的基因组数据库 GenBank[2] 中，1982 年时包含 606 个序列、680 338 个碱基对；截至 2011 年 1 月，这一数字已攀升到 135 440 924 个序列、126 551 501 141 个碱基对。其中，代表性物种的数据量见表 1.1。

表 1.1 GenBank 数据库中现存代表物种数据量[2]

代表物种	包含数据量/亿 bp
人 (*Homo sapiens*)	147.9
小鼠 (*Mus musculus*)	88.6
大鼠 (*Rattus norvegicus*)	64.4
牛 (*Bos taurus*)	53.6
玉米 (*Zea mays*)	50.4
猪 (*Sus scrofa*)	47.8
斑马鱼 (*Danio rerio*)	31.4
紫色球海胆 (*Strongylocentrotus purpuratus*)	13.5
水稻 (*Oryza sativa*)	12.0
烟草 (*Nicotania tabacum*)	11.9
热带爪蟾 (*Xenopus tropicalis*)	11.5

1.2.2 蛋白质组数据

蛋白质组是与基因组对应的数据。蛋白质是构成生命的重要物质，由 DNA 经过转录、翻译和表达而产生。蛋白质是由氨基酸组成的，构成生命的氨基酸一共有 20 种：非极性、疏水性氨基酸有甘氨酸、丙氨酸、缬氨酸、亮氨酸、异亮氨

酸、苯丙氨酸和脯氨酸；极性、中性氨基酸有色氨酸、丝氨酸、酪氨酸、半胱氨酸、甲硫氨酸、天冬酰胺、谷氨酰胺和苏氨酸；酸性氨基酸有天冬氨酸和谷氨酸；碱性氨基酸有赖氨酸、精氨酸和组氨酸。这 20 种氨基酸是构成蛋白质一级结构序列的基本单元，具体信息见表 1.2。

表 1.2 20 种氨基酸基本信息表

中文名	英文名	缩写	相对分子质量	侧链	类型
丙氨酸	alanine	A 或 Ala	89.079	CH_3—	脂肪族类
精氨酸	arginine	R 或 Arg	174.188	HN═C(NH_2)—NH—$(CH_2)_3$—	碱性氨基酸类
天冬酰胺	asparagine	N 或 Asn	132.104	H_2N—CO—CH_2—	酰胺类
天冬氨酸	aspartic acid	D 或 Asp	133.089	HOOC—CH_2—	酸性氨基酸类
半胱氨酸	cysteine	C 或 Cys	121.145	HS—CH_2—	含硫类
谷氨酰胺	glutamine	Q 或 Gln	146.131	H_2N—CO—$(CH_2)_2$—	酰胺类
谷氨酸	glutamic acid	E 或 Glu	147.116	HOOC—$(CH_2)_2$—	酸性氨基酸类
甘氨酸	glycine	G 或 Gly	75.052	H—	脂肪族类
组氨酸	histidine	H 或 His	155.141	N═CH—NH—CH═C—CH_2—	碱性氨基酸类
异亮氨酸	isoleucine	I 或 Ile	131.160	CH_3—CH_2—CH(CH_3)—	脂肪族类
亮氨酸	leucine	L 或 Leu	131.160	$(CH_3)_2$—CH—CH_2—	脂肪族类
赖氨酸	lysine	K 或 Lys	146.170	H_2N—$(CH_2)_4$—	碱性氨基酸类
甲硫氨酸	methionine	M 或 Met	149.199	CH_3—S—$(CH_2)_2$—	含硫类
苯丙氨酸	phenylalanine	F 或 Phe	165.177	Phenyl—CH_2—	芳香族类
脯氨酸	proline	P 或 Pro	115.117	—N—$(CH_2)_3$—CH—	亚氨基酸
丝氨酸	serine	S 或 Ser	105.078	HO—CH_2—	羟基类
苏氨酸	threonine	T 或 Thr	119.105	CH_3—CH(OH)—	羟基类
色氨酸	tryptophan	W 或 Trp	204.213	Phenyl—NH—CH═C—CH_2—	芳香族类
酪氨酸	tyrosine	Y 或 Tyr	181.176	HO—Phenyl—CH_2—	芳香族类
缬氨酸	valine	V 或 Val	117.133	CH_3—CH(CH_3)—	脂肪族类

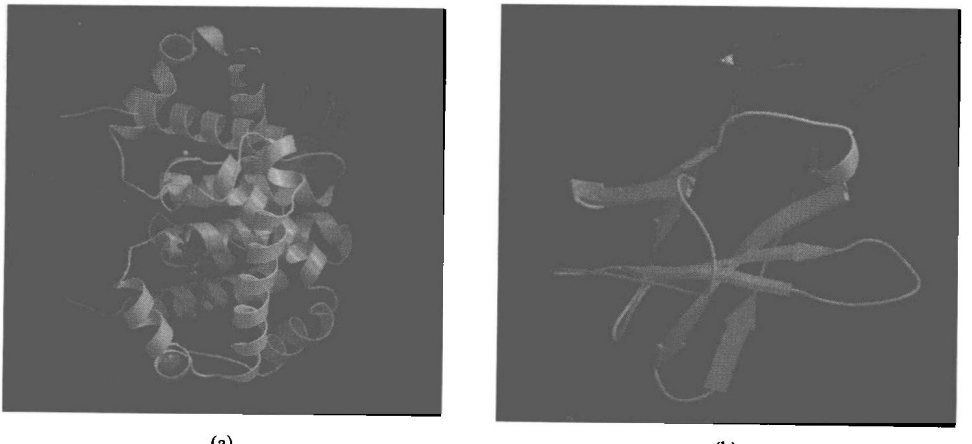
与基因组相应的，公共数据库中的蛋白质一级结构数据的增长也是十分迅猛的。根据权威蛋白质序列数据库 SWISS-PROT[3]统计，截至 2011 年 9 月，SWISS-PROT 数据库中包含 532 146 条序列，共 188 719 038 个氨基酸，其中代表

性物种数据量见表 1.3。

表 1.3 SWISS-PROT 数据库中现存代表物种数据量

代表物种	序列数
人 (*Homo sapiens*)	20 248
小鼠 (*Mus musculus*)	16 388
拟南芥 (*Arabidopsis thaliana*)	10 708
大鼠 (*Rattus norvegicus*)	7 641
酿酒酵母 (*Saccharomyces cerevisiae*)	6 620
牛 (*Bos taurus*)	5 860
殖裂酵母 (*Schizosaccharomyces pombe*)	4 976
大肠杆菌 (*Escherichia coli*)	4 430
枯草杆菌 (*Bacillus subtilis*)	4 244
黏菌 (*Dictyostelium discoideum*)	4 118
线虫 (*Caenorhabditis elegans*)	3 335
热带爪蟾 (*Xenopus tropicalis*)	3 321
果蝇 (*Drosophila melanogaster*)	3 125
水稻 (*Oryza sativa*)	2 797
斑马鱼 (*Danio rerio*)	2 755

可见，蛋白质氨基酸残基序列的多样性是相当丰富的。然而，蛋白质的多样性并不仅仅表现在它的一级结构上，蛋白质还可以通过不同的螺旋和折叠方式构成二级结构。图 1.1 所示为蛋白质常见的二级结构。



(a) (b)

图 1.1 蛋白质的两种基本二级结构

(a)蛋白质的 α 螺旋(α-helix)；(b)蛋白质的 β 折叠(β-sheet)

蛋白质的二级结构也是生物信息学研究的主要对象。包含蛋白质结构的数据库最为权威的是 PDB(Protein Data Bank)，截至 2011 年 9 月，PDB 数据库中包含

蛋白质结构 73 065 个。显而易见，蛋白质结构数据要远远少于蛋白质序列的数据，这是由于测定蛋白质三维结构的 X 射线结晶方法或是 NMR 技术从实验工作量上都是十分困难而艰巨的。尽管蛋白质二级结构预测技术已相对成熟，但是目前利用计算的方法预测蛋白质三维结构，仍然是生物信息学领域的重要挑战之一。

1.2.3 基因表达数据

基因表达数据是生物信息学研究的重要内容之一。通常情况下，基因表达数据是通过基因芯片(microarray)高通量地测定出来的海量基因的定量表达水平数据。由于基因表达数据的分析是本书的重点内容，我们将详细地对基因表达数据和基因芯片加以介绍。

1. 基因表达

既然基因芯片是检测基因表达水平的工具，那么首要的问题是要搞清楚什么是基因表达。在 1.2.1 节和 1.2.2 节两节中提到，基因是遗传物质中最主要的成分，是生物体形成的蓝本，而蛋白质则是组成生命的主要物质，也就是说，如果将基因组比喻成烹制晚宴的菜谱，那么蛋白质就是晚宴所用的最主要的食材。摆在人们面前的问题是，饭菜是如何通过菜谱烹饪出来的？Francis Crick 于 1958 年提出的中心法则从框架上初步回答了这一问题(图 1.2)。

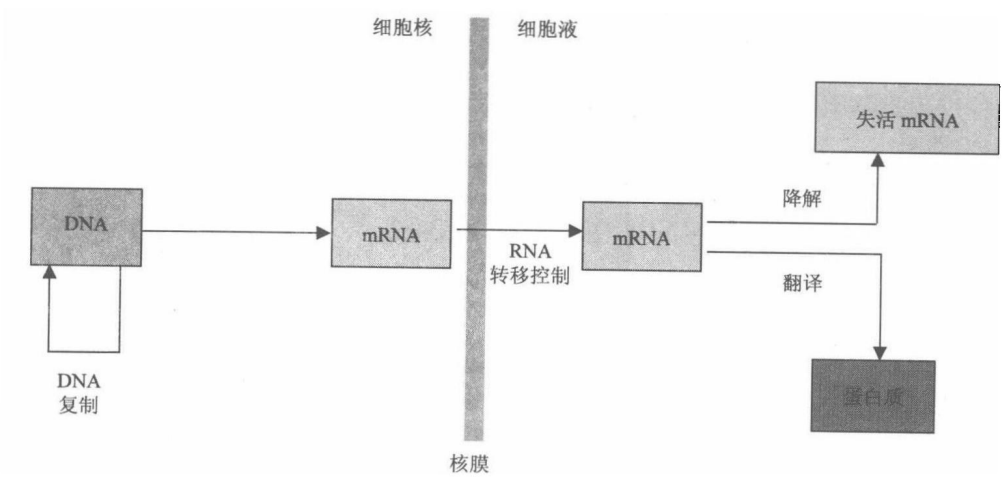


图 1.2 中心法则示意图

中心法则是分子生物学领域的一个最基本的法则，它描述了遗传信息是怎样流动的。DNA 通过转录过程把 DNA 片段的信息转载到信使 RNA(mRNA)上，这个过程由 RNA 聚合酶和转录因子共同作用完成。RNA 经过了编辑和剪切，并从细胞核内被转移到细胞核外。最后，成熟的 mRNA 到达核糖体，在这里 mRNA

被翻译成蛋白质。当然，以上过程是针对于真核细胞而言的，原核细胞由于没有细胞核，它的转录和翻译是同时进行的。这个从 DNA 转化为蛋白质的过程就被称为基因的表达。DNA 是可以复制的，通过 DNA 的复制，遗传物质可以将遗传信息从亲代转移到子代。

在中心法则中，能够通过转录翻译成蛋白质的 DNA 片段被称作基因，这部分 DNA 片段只占基因组总数的不足 2%，而其余大量的基因间区域占据整个基因组的绝大部分。20 世纪 90 年代初期，人们对这些庞大的基因间的 DNA 片段并没有认识，认为它们并没有任何作用，是“垃圾区域”。而随着研究的日益深入，研究人员发现，恰恰相反，这些基因间区域不但不是垃圾区域，反而具有非常重要的功能，其重要性完全不亚于基因。

再回到前面的比喻，如果将基因和蛋白质比喻成菜谱和食材，那么细胞就是一个主厨，负责创造出丰富的菜肴。但是与现实不同的是，对于同一个人而言，他的所有细胞大厨都只有一份相同的菜谱，然而它们做出来的菜肴却是十分多样的。下面的一些例子就足以说明这一点：① 不同种类的细胞之间是有差异的，如脑细胞和肝细胞；② 不同年龄的人的细胞是有差异的，如婴儿和成人；③ 不同健康状况的细胞是有差异的，如癌细胞和正常细胞。到底是什么促成了这些差异的存在呢？事实上，正是无限庞杂的基因间区域负责控制基因表达的变化。基因就像菜谱中的名词，每个基因对应固定的蛋白质(在选择性剪切的调节下可能会对应多个蛋白质)；基因间区域就像是菜谱中的动词和介词，负责告诉细胞在什么条件下表达哪些基因，表达量是多少，这一过程被称为调控(regulation)，而调控过程是基因组参与的最复杂的过程之一，对于调控的认识是人类破解遗传密码的关键。

另外，搞清楚不同状态下的基因表达水平，是了解基因调控的重要途径，因此发展出检测基因表达水平的技术就是顺理成章的。

2. 基因表达水平检测技术

基因表达检测最直接的方法是检测蛋白质的丰度，但是，这种看似直接有效的方法却存在一些问题：

(1) 有的蛋白质的活性非常高，很有可能一经产生就马上被使用，因此测定结果很难反映真实的表达水平；

(2) 最常见的测定蛋白质丰度的技术是质谱技术，但是质谱技术难以高通量地同时测定大量蛋白质的表达水平，这使其在表达检测方面受到了限制；

(3) 目前已出现蛋白质芯片技术，可以高通量地测定基因的表达水平，然而由于存在蛋白质相互作用的问题，如何使芯片上的蛋白质不相互干扰还是一项需要突破的技术性挑战。

由于直接检测蛋白质丰度的方法存在着以上这些局限，因此，目前主流的基因表达水平检测方法并不是直接检测蛋白质，而是退而求其次，通过检测 DNA 所转录成的 mRNA 的丰度来间接测定基因的表达水平。由于 mRNA 相对稳定，容易测量，并且根据碱基互补配对原则，mRNA 的探针可以有效地控制非特异性结合。这种技术的主要原理是利用 4 种核苷酸之间两两配对互补的特性，使两条在序列上互补的单核苷酸链形成双链，这一过程被称为杂交。将其中一条链作为探针(probe)，就可以特异性地结合对应的 mRNA，从而通过测定结合 mRNA 的数量来得到其对应基因的表达水平。

根据探针的不同，基因芯片可以分为两种：一种探针为 mRNA 逆转录得到的互补 DNA(cDNA)，这种芯片称为 cDNA 芯片；另一种探针由寡核苷酸(oligonucleotide)构成，因此被称为寡核苷酸芯片。无论哪种基因芯片，其制备过程都是相似的：

(1) 首先在特定条件下培养细胞，收集和提取细胞中的 mRNA 样本，并为其做标记，一般情况下使用荧光标记；

(2) 准备基因芯片，芯片上会有高密度的序列特异性探针；

(3) 样本与芯片探针杂交；

(4) 洗去非特异性 mRNA 的结合；

(5) 根据标记信号测量 mRNA 丰度。

3. cDNA 芯片

cDNA 芯片在 1995 年诞生于斯坦福大学 Pat Brown 实验室(图 1.3)。

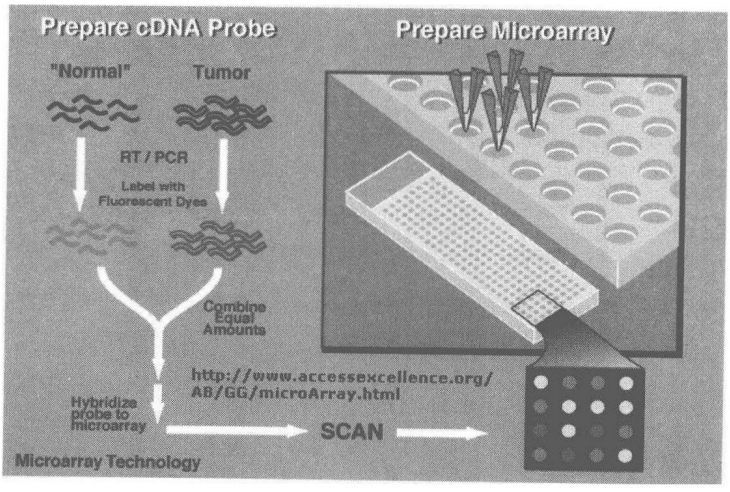


图 1.3 cDNA 芯片流程模式图

cDNA 的主要特点如下：