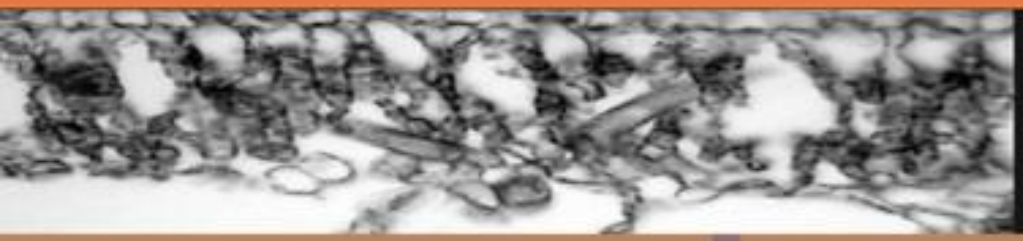
国家科学技术学术著作出版基金资助出版

系统生物学



林标扬　编著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

第1章 复杂系统的研究方法

对一个复杂系统的研究，我们经常会问这样的问题，这个复杂系统是由什么构成的？怎么会是这样一个结构？为什么会是这样？要解答这些问题，我们必须有一个很清晰的思路，包括了解以下几个方面：①系统的逻辑性——这个系统应该遵循一个逻辑规律；②系统的因果关系——这个系统有因果关系的规律，即一个物体、一个过程、一个现象在逻辑上造成这种现象的原因；③有机结构的原理——即一个系统有一个有机的组织形式，所以每个系统的各个组成部分的过程都可以形成一个有机的结构，即形成一个有机的系统；④系统的可分性——在逻辑上一个系统可以分为亚系统和组成部分，这些亚系统的组成部分是可以研究的，它们的特征是可以被定义的，但是为了防止无限制地将系统进行分解，一般会假定系统有一个最小的组成部分——不可分的组成部分，如在化学上把原子作为最小的组成部分。

在历史上，研究复杂系统的方法有三种，包括还原论（Reductionism）、神圣论（holism）和一般系统理论（General systems theory）。

1.还原论

在过去的两三个世纪中，还原论（Reductionism）尤其运用于物理学和化学，并取得了极大的成功。还原论可以分为几种不同的形式。

第一种还原论形式是根本性的还原论（Ontological reductionism）。根据这一理论，任何一个复杂系统都可以用其组成部分的特征来描述，所以严格意义上说仅仅是其组成部分的总和。持这种理论的人认为，所有系统的组成部分都按照自然界的法则组合在一起，假如有新的特征而不能用它的组成部分来解释，那是因为对其组成部分的了解还不彻底，或者有新的组成部分还未发现，因此必须继续寻找它的组成部分。这种根本性的还原论，在科学研究方面已经很少有人相信，但它在过去的几个世纪中，对机械理论甚至包括生物学上的解剖学、形态学、细胞学等的发展起了很重要的作用。

第二种还原论形式是在方法学上的还原论（Methodological reductionism），这是一种研究的手段，利用一个替代系统来研究所要

研究的系统，即用一个模型来研究和描述复杂系统。持这种还原论观点的人，也接受系统的新的涌现性。该观点的一个典型的例子就是hertz和Rosen所提出的hertz□Rosen模拟关系（hertz□Rosen modeling relation）（hertz，1994；Rosen，2000），他们指出一个复杂系统常常不能用一个单一的和最大的模型来进行足够的描述和模拟，要对一个系统进行满意的模拟，常需要多个互补的模型对该系统进行模拟和解释。

2.神圣论

神圣论（holism）的观点认为，把一个系统的组成部分放在一起并不能产生它的整体（如有机生命体）或解释它所有的行为和特征。极端神圣论认为，一些超自然的创造力决定一个复杂系统的整体性和特征。然而，在科学界很少有人持极端神圣论。目前，比较常见的神圣观认为，一个复杂系统的各个组成部分之间是有一系列联系的，如果把这些组成部分单独地分开，它们的联系就不存在了。

目前，很多生物学家采用还原论的方法来研究复杂系统，但是当他们分析和解释各自数据的时候，又常抱着一定的神圣观。有人认为，在生物学中把生命的各个组成部分分开进行研究时，破坏了生命整体的一些最重要特征，使整体特征研究变得非常困难。

3.一般系统理论

一般系统理论（General systems theory）包括它的变体（如控制论）、系统研究和系统思维。它是研究系统功能和组织的一门学科。一般系统理论通常有如下假定：

①作为一个整体来说，系统的行为和功能是有一定规律的；

②系统行为和功能的规律不能由组成部分的特征来显示，所以各组成部分的规律并不能体现整体系统的行为和规律。

系统理论在于克服还原论和还原论方法的一些缺陷。Ludwig von Bertalanffy最早建立了一般系统理论的学说。他在《一般系统论》一书中指出，系统的普遍性质有系统整体性、关联性、动态性、有序性和预决性。

从20世纪50年代起，系统论在实践中得到了很大的发展。目前，一般系统理论包括以下四种具体的理论：

◆混沌理论（Chaos theory）；

◆元胞自动机、细胞自动机（Cellular automata）；

◆突变论或灾变论（Catastrophe）；

◆等级层次理论（hierachical system）。

其中，前三种理论（混沌理论、元胞自动机和突变论）是动态系统的一些表达形式，可以用一系列微分方程来体现。等级层次理论是指系统有有机层次结构，即系统有不同的层次，但在某一层次上的功能并不能由它下一层次的功能来体现。

4.混沌理论

混沌理论（Chaos theory）是由美国气象学家E。N。洛伦茨（Lorenz）在20世纪60年代初研究天气预报中大气流动问题时首先发现的。他在计算机上模拟地球大气的研究中发现，只要计算机模拟出发点的初始值有一个很微小的差异（小数点后第3位数），模拟的结果就截然不同。由于在技术上不可能无限精度地测量初始值，因此我们不可能预言任何混沌系统（在这里指长期天气预报）的最后结果。但是，洛伦茨还发现，混沌系统尽管看起来杂乱无章，但其实具有某种规律（patterns）。对混沌系统的模拟，计算机可输出几千个可能的预测，这些预测在某种状态范围内是随机分布的，但也有一定的模式。正如每日的天气可以变化多端，不可对它进行长期的预测，但逐年的气候还是保持某种稳定性的。

1972年，洛伦茨做题为“predictability：Does the Flap of a Butterfly’s Wings in Brazil set off a Tornado in Texas？”（预测性：是否巴西蝴蝶的一个偶然的扇动将会在德克萨斯州制造一次龙卷风？）的会议报告，也说明气候变化的复杂系统对起始的条件是非常敏感的。

混沌理论是研究描述一个系统随着时间的变化而变化（即动态系统行为）的一种理论（Ruelle，1991）。混沌理论认为混沌中无序现象的

两个基本特点是不可预言性和对于初始值的极端敏感性，这种敏感效应通常称为蝴蝶效应。

混沌理论充分认识到复杂系统是一个动态的、非线性的、不平衡的系统，系统未来的状态不能根据过去的或者现在的状态来预测。在一个混沌的状态中，系统的行为也是不可预测的，同时又具有一定的规律。混沌理论在各个学科应用非常广泛，如气候的预测、股票市场的预测以及医学上癫风的发作等。

在系统生物学中，混沌理论也有一定的作用，比如，如何从高通量数据中找出规律，区分随机的数据和混沌的数据（Chaotic data）。在对细胞进行模拟的时候，一些小分子的随机变化可能会对细胞的命运产生重大的影响，所以随机的模拟（Stochastic modeling）在系统生物学中也是很重要的一种模拟方式。

5.元胞自动机，细胞自动机

自动机是一种数学的演算方法。其原始概念是指，在一个网格结构上，将每一个网格定为一个细胞，然后每一个细胞可以根据其邻近细胞或空间的特性进行自动演化。该种细胞自动机包括两个部分：各个细胞或格子的初始状态，即整个自动机的初始状态，以及从旧细胞产生新细胞的规律（Walfram，1984）。元胞自动机是由计算机之父纽曼（John von Neumann）在20世纪40年代初提出的，最初目的是发展具有自我复制能力的计算机，以发展自动生成、自动复制的自动机。1970年，John Conway把该概念进一步发展成电脑上的生命游戏（Game of life）（Sigmund，1995）。

细胞自动机（Cellular automata）的运作包括以下几个主要组成部分。①网格：即细胞，细胞自动机是由一组网格形成的。②网格的状态：每个网格具有一定的状态（States），这种状态可以是二元制的（如有和无，活的和死的），也可以是多种状态的。在任何一个时间，每个网格将呈现这组状态中的一种状态。③邻近网格或者邻居：细胞自动机中每个网格的状态是根据其邻近网格的状态进行变化的。④设定它的演化规则：每个网格在一定时间中的演化，即在下一个时间点的状态，是由当前的状态以及邻近网格对该网格的影响的总和所决定的。因此，在一定时间和空间的影响下，细胞自动机将形成一系列离散的状态（Discrete states），即所有网格或细胞自动机将根据这个演化规则进行同步更新。

规则30（Rule 30）的细胞自动机在一维图上的演变，规则30的细胞自动机在二维图上的进化演变。一个最简单的细胞自动机，即一维的细胞自动机，其中每个细胞只有两种状态，每个细胞的邻居定为它左右两边的邻居，因此一个细胞与其两个邻居就形成一个三个细胞的邻居，总共有8种可能（2的三次方）。三个细胞组成一个演绎规律，所以这种细胞自动机的演绎规律称为规则30，它是在1984年由Stephen Wolfram提出的。

规则30的细胞自动机（Stephen Wolfram等，1984）

在系统生物学中，细胞自动机的研究方法也有一定的应用。比如，在蛋白□蛋白相互作用网络上，一种蛋白的功能会受其邻近蛋白的影响，这些蛋白所形成的网络结构是由网络中蛋白□蛋白相互作用的规律来决定的。

6.灾变论

灾变论（Catastrophe theory），或称突变论，是指在非线性系统中，某些参数的微小变化就可使整个系统失去平衡，使系统发生重大的、突然的变化。

在20世纪60年代末，灾变论是由法国数学家R。托姆（RenéThom）为解释胚胎的成胚过程而提出来的（Thom，1972）。70年代以后，E。C。塞曼（Christopher Zeeman）等人进一步发展了灾变论，并把它应用到生物学、生态学、医学、经济学等领域。灾变论研究跳跃式转变、不连续过程和突发的质变。灾变论建立在结构稳定性的基础上。结构稳定性反映同一物种在千差万别形态中的相似性。稳定结构的丧失，就是突变的开始。灾变论是研究不连续现象的一个新兴数学分支，也是一般形态学的一种理论，能为自然界中形态的发生和演化提供数学模型。

第2章 等级层次理论

等级和层次普遍存在于我们的社会、生物系统以及生物分类等。等级层次理论（hierarchy theory）就是从数学角度把一个系统分成有等级、有层次的不同部分（pattee，1973）。在不同等级间，有一定的非对称关系（Asymmetric relationships），这种非对称关系是指上一层的等级高于下一层的等级，并且每一等级与上面层次的关系和与下面层次的关系是不对称的；从生物学角度来说，也就是更高一层次的功能并不能在另外一个层次上被还原。根据等级层次理论，一个系统的复杂性（Complexity）与复合性（Complicatedness）是不同的：若一个等级系统由许多低水平的层次所构成，并且有相当简单的组织结构，这种层次不丰富的等级结构不属于复杂（Complex）系统，而被认为是复合（Complicated）系统。即假如一个很大系统的组织结构非常简单，则综合在一起的行为还是比较简单的。反之，假如一个复合系统的结构比较复杂，则其行为也会比较复杂。

1.生物系统的特性

系统生物学的基本概念和原理：生物系统是一个复杂的系统，包括不同的组分以及它们之间的相互作用，其可以用网络系统来模拟。生物系统的主要特性有如下几个方面。①系统涌现性（Emergence）：即系统作为一个整体，可以产生各个组成部分所没有的新功能，即整体大于各个组成部分的简单加和。②稳健性（Robustness）：生物网络具有负反馈和多通路生物途径等，因此系统具有一定的稳健性。③无尺度性（Scale□free）：生物系统的网络结构还具有无尺度网络结构的特性，即少数大的枢纽（hub）和多数小的链接。

2.系统涌现性

系统涌现性（Emergence）是指一个系统自动形成一些新的系统特性，这些特性不能从其组成部分的特性中预测出来。因此，系统涌现性有如下三个重要的特征：①原来各组成部分并不存在的特征；②新的、可定性的，新涌现的特性具有质的突变；③不能从其组成部分的特性中预测，所以系统涌现性有别于系统的预测性。系统的预测性（Anticipation）是指系统可被预测的一些特性。如某些系统的组成部分、特征以及环境的相互作用有一定的规律性，给出一定的参数后，

即可预测系统的特性。此时，即使产生新的系统特征，也是可被预测的，有别于系统涌现性所产生的特征。

3.稳健性

生物系统都是动态的系统。动态系统理论中，一个很重要的概念就是系统状态（System state）。系统状态是指用某一时点的足够多的信息来预测未来系统行为的系统描述，常用一组变量来表示。如在代谢物网络的微分方程模型中，系统状态就是每一种化学物质浓度的集合；在随机模型中，系统状态是一个概率分布或者每种生物分子数的集合。一个系统的稳定态（Steady state），或称稳定状态（Stationary state）或不动点（Fixed point），指的是在时间上所有系统变量的值都保持相对不变的状态。

生物系统的稳健性是指生物系统能够抵抗内部和外部干扰，并维持其功能的一种特性（Kitano，2004；Kitano，2007）。理解生物系统的稳健性是深刻理解生命现象的一个基础。生物系统的稳健性基本可以体现在以下三个方面。①适应性（Adaptation）：即生物体对环境条件变化的适应。②参数不敏感性（parameter insensitivity）：即系统对某些动态参数是相对不敏感的。③逐渐降解性（Graceful degradation）：指在一般的条件下，单个系统的功能受到损害后，整个系统表现为慢慢破坏和降解，而不是灾难性的破坏。

在工程系统中，可以通过如下几个方式达到系统稳健性。①通过系统的正反馈和负反馈作用，达到对一个系统稳健性的控制。如在基因调控方面有正反馈和负反馈现象。②累赘性或者重复性，即把执行同样功能的多个组成成分引入系统中作为系统的备份。如在基因水平上，很多基因常常具有相同的功能（累赘性），这也是造成基因敲除失败的常见原因，即敲除一个基因后，还有备份基因。③结构的稳定性，即把稳定的结构构建到系统中，以达到系统的稳定性。如生命系统中，蛋白□蛋白相互作用的网络系统具有很稳定的结构。④模块性，即把系统分成不同的亚系统，这些亚系统在物理位置和功能上是分开的，是相互绝缘的，所以一个模块功能的失调并不会很快导致其他模块的失调，也不会导致整个大系统的灾难性的功能失调。如生命系统也具有模块性，包括细胞的细胞器以及每个信号传导通路都可分别形成不同的模块。

要指出的是，稳健性（Robustness）、稳定性（Stability）或者是体内恒定理论（homeostasis），概念相近，但又有所不同。稳健性是一个更广泛的概念，它主要是指维持系统功能的稳定性；而稳定性或者体内恒定规律是指维持系统状态的稳定性（即稳定态）。一个稳健的系统可以有几个不同的稳定态，只要在不同的稳定态下，该系统都能维持它的功能，就称为系统的稳健性；一个系统可以在不同稳定态之间变化，但仍维持了系统的功能，这也称为系统的稳健性。比如一个细胞在极端的环境，如热休克的状态下，经常会产生其他蛋白（如热休克蛋白）来维持细胞的活性，使细胞进入另一个新的稳定状态，也称为细胞的稳健性。又如细菌在抗生素作用下会产生抗药性，所以细菌就由不抗药状态变成抗药状态，即细菌有系统的稳健性，可以在抗生素条件下生存。再如艾滋病病毒能够以很高的突变率来应付机体的免疫系统以及综合疗法，即艾滋病病毒可以根据DNA的突变产生无穷多的稳定状态来维持其生命和致病性。

稳健性（Robustness）、稳定性（Stability）或者体内恒定（homeostasis）。假定系统的起始状态在稳定态1的中心，一个系统扰动可以把系统推到稳定态1的边缘，但系统仍可回到稳定态1，这就是系统的稳定性和体内恒定。如在扰动后，系统转折到稳定态2，系统即丧失稳定态1的稳定性，并在稳定态2状态下达到新的稳定性。如果系统在稳定态2的功能与稳定态1相比是不变的，则可以说系统具有稳健性。在极端的情况下，系统可以在多种不同的稳定态中转变而保持其稳健性。

必须指出在系统稳健性和脆弱性（Fragility）之间是存在一定折扣（Trade□offs）的。例如要维持系统生物学的稳健性，常需利用很多资源，如：用两个基因执行同样的功能，比用一个基因执行具有更多、更强的稳健性，但前者也需要使用更多的资源。当用更多的资源来维持系统的稳健性时，也引入了系统的脆弱性，如：一个系统需要很多资源才能维持其功能和稳健性，假如没有资源或资源缺乏，系统就非常脆弱；反之，一个简单系统不需要很多资源，在没有资源或资源缺乏的条件下，简单系统反而具有更强的竞争力。

稳健性的数学方程式：

稳健性可以定义为系统在外部和内部扰动下维持一个或多个功能函数的特性。因此，一个系统（S）在扰动条件（p）下维持其功能（a）的稳健性（R）可以用数学公式描述：

RSa，p=∫pψ（p）DSa（p）dp（1.1）

函数ψ（p）是扰动p发生的可能性，当所有的扰动同等概率发生时，该值为1.D（p）是对扰动（p）的评价函数，p是整个扰动的可能空间。当系统（S1）在一系列扰动因素（Y）下，对于维持某特定功能（a）比另外一个系统（S2）更具稳健性时，可以用公式描述：

RS1a，Y>RS2a，Y

系统S1和S2间稳健性的差别（ΔR）可以用公式（1.3）描述：

ΔRS1，S2a，p=∫pψ（p）（DS1a（p）-DS2a（p））dp=RS1a，p-RS2a，p=0

公式是波德积分公式的衍生，由此也可推断稳健性与脆弱性之间的平衡和折扣关系。如果S1和S2稳健性相同，则方程式应该为零。但是，假设S1和S2是对不同扰动参数的最优化稳定系统，上述方程也可说明增加系统对一些扰动参数的稳健性，会导致系统对另外一些扰动参数的不稳定性（即脆弱性）。

第3章 系统生物学是一门整合不同数据的交叉学科

1.系统生物学是结合自上而下和自下而上研究方法的学科

从系统角度来研究生物学，也可根据研究出发点分为两种研究方式，即自上而下（Top down）和自下而上（Bottom up）的研究方法（Bruggeman和Westerhoff，2007）。自上而下的系统生物学研究方法基于高通量组学技术的发展，目前已可对DNA、RNA、蛋白等进行比较全面的高通量组学技术分析。它实际上是先用组学的办法来收集实验数据，然后对实验数据进行分析、整合，研究不同分子之间的关系，最后形成一个假说，以说明分子之间的相互关系。这些假说可以预测一些新的相互关系，而后者可以用新一轮的试验来验证，然后进行分析，因此这也是一个重复迭代的过程。由此可见，自上而下的系统生物学方法是从大量的数据开始的，尽量进行没有偏见的分析，用大规模的数据来分析分子之间的相互关系，而不依赖或应用以往的知识或人为的判断来进行数据分析。

但是，目前要完全依靠自上而下的系统生物学研究方法还具有很大的局限性。自上而下的系统生物学研究方法有点类似于工程的反向工程。生物系统很复杂，若要完全依靠自上而下的方法来研究系统的复杂性，则挑战性很大。如要研究人的3万～4万个基因，对每个基因进行扰动（如敲除试验或过度表达），然后再把敲除或者过度表达的基因放在不同的环境条件下（如不同的营养、激素或药物等），进行多个时间点的观察，数据量将非常大。另外，要整合生物系统所有分子（包括DNA、RNA、蛋白质和代谢物）在不同环境条件下的表达数据，目前挑战性也很大。因此还要用到自下而上的系统生物学研究方法。

自下而上的系统生物学研究方法以系统功能或机制为出发点。它与自上而下的系统生物学研究方法的区别是：“自下而上”是根据机制进行研究，而“自上而下”是根据数据、现象、表象提出模型。自下而上常常对一个系统的亚系统进行详细分析，来构建模型，再与不同的亚系统结合在一起，最后产生一个对整体的（如一个细胞）模型。如用一棵树来做比拟，自上而下的研究就像是从一棵成年树的树叶和树枝开

始来推断这棵树的成长过程；而自下而上的研究就相当于从一颗种子开始，来推断这颗种子可以长成什么样的树，及在不同土壤、气候条件下的成长状况等。因此，目前系统生物学需要结合自上而下和自下而上两种研究方法进行研究。

2.系统生物学是综合发现性科学和假说性科学研究手段的学科生物系统是极其复杂的，它的复杂性可以体现在以下四个层次。第一层次：表现在基因、RNA、蛋白以及代谢物水平的变化。目前，高通量的分析技术，如高通量DNA测序技术、基因表达分析技术、高通量蛋白质组学质谱分析技术以及代谢组学技术，使得我们对第一层次的分析能力大大提高，但是如何把基因组数据、转录组学数据、蛋白质组学数据和代谢组学数据整合在一起分析，还具有很大的挑战性。第二层次：基因、蛋白、代谢物等可以形成一些功能性连接，如基因调控网络、代谢网络和信号传导途径等。第三层次：各种网络和途径又可以进一步组合成功能性模块，来完成细胞内一定的功能。第四个层次：不同的功能模块可以形成一个复杂的系统，包括细胞内不同细胞器之间的相互联系、细胞与细胞之间的通讯、细胞与间质细胞的联系，最后形成组织、器官和个体。金字塔的底部显示细胞的基本功能组，如基因组、转录组、蛋白质组、代谢组（层次一）。在此基础上，不同层次在结构和调控功能上都有不同的整合，如调控网络、代谢途径（层次二）、功能模块（层次三）和阶层性网络（层次四）。不同有机体的构成不同，但形成的网络系统有着共同的特征，有些特性也存在于其他大型网络结构（如社会网、计算机网）中。

多层次之间的联系也是系统生物学研究的内容之一，对生物系统这个复杂系统的研究，需要对数据进行整合、模拟，它就需要各个学科的参与，如生物学、化学、物理学、数学、计算科学、工程学等。不同学科人员之间的沟通，是系统生物学成功的关键之一，所以我们必须培养交叉学科的人才。如使计算机学科的人也能够理解基本的生物学概念；使生物学、化学方面的人员也能够听懂计算机学科的语言，以及解释数据、理解预测模型等。所以，系统生物学是用交叉学科的研究手段来研究复杂的生命系统的学科。

生物系统的复杂性体现的四个层次。

近年来，高通量技术的发展，特别是基因测序技术、现代化蛋白质组学技术、DNA芯片技术等的迅猛发展，以及大规模数据分析模拟、计算机科学计算的发展（包括控制论、非线性动态理论的发展），使系

统生物学实现了真正的发展。但系统生物学是一门刚刚起步的学科，还面临巨大的挑战。现代系统生物学有以下六个特征。

（1）对所有的生物分子进行定量分析和测量，即对所有的DNA、RNA、蛋白质和代谢物进行精确的测量分析，而不仅仅对一部分DNA、RNA、蛋白质进行分析。这是对系统结构的研究，找出系统的所有组成部分，以及这些组成部分的相互关系、网络结构。

（2）精确测量所有DNA、RNA、蛋白质代谢物等在不同的状态和变化条件下的动态变化。研究网络系统或者复杂系统的动态性，对系统动态的理解是系统生物学的关键内容。

（3）用计算机和数学的方法来整合不同的数据，包括DNA、RNA、蛋白质及其相互作用的数据，并把这些数据与它们在不同的环境因素（如细胞所处的微环境等）、疾病的表现型等结合在一起进行整合分析。

（4）对现代生物系统进行动态分析，并把动态分析提高到有机体水平，包括在不同的发育阶段、不同的生理条件、不同的健康与疾病状况下以及在不同的环境条件下对不同大分子（DNA、RNA、蛋白质等）的动态测量过程。系统动态包括时间上的动态，如发育过程或者疾病发生、发展过程中的动态，也包括空间上的动态，如细胞与细胞相互作用、肿瘤细胞的转移等。

（5）建立一个研究系统控制机制的模型，以正确理解系统、精确建模和进行模拟，然后提出对系统进行扰动的假说。

（6）发现性科学（Discovery science）和假说性科学（hypothesis□driven science）结合的手段。如对一个系统建模，提出假说，进行扰动试验，进一步分析测量，提出进一步的模型，再提出一种新的假说，进行新的扰动试验，形成一个循环迭代的过程，最后系统模型完全能够预测扰动试验的结果，即测量结果和模型所预测的扰动结果是一致的。

最后，在对系统完全理解之后，可以对系统进行重新设计，包括修改原来的系统或构建全新的系统，以达到我们想要的生物特征。如可以改良细菌和酵母菌来生产我们所需要的药物和工业产品，也可以修改

疾病扰动的网络系统，使系统回归到原来的状态或达到与原来同样功能的状态，从而恢复或保持健康状态。

第4章 网络和网络系统生物学

生物系统是指由各种分子（蛋白、基因、小分子等）组成的复杂网络系统。传统的生物系统研究通常是指对单个分子或不同分子组成的生物通径的研究。但在生物系统中，不同分子和不同通径是存在相互作用的，即构成网络。因此，可以说，决定生物系统特性和结果的是网络，而不是单个分子或生物通径。单纯用还原论法来研究生物学问题，找出各个基因以及每个基因单独的相互关系是不可能理解生命现象的。要真正理解生命现象，就必须了解各个不同组成部分的相互关系以及它们的动态关系，也就是它们之间所形成的网络结构，即用网络系统学的方法来研究生命现象。

本章从网络的角度来研究复杂的生物系统。我们先介绍几个用来描述网络拓扑结构的基本概念。网络的拓扑结构描述网络的结构，包括节点及链路。节点（Nodes；网节，顶点）就是网络单元，而链路（Links，Edges；链接，网边）是两个节点间的连线。链路分“物理链路”和“逻辑链路”两种，前者是指实际存在的物理连线，后者是指在逻辑上起作用的网络通路。链路容量是指每个链路在单位时间内可接纳的最大信息量。通路（pathways）是指从起始节点到信息接收节点之间的一串节点和链路，即指一系列穿越网络而建立的节点到节点的链路。

近几年，网络结构在复杂系统研究中被广泛应用，如利用网络系统来研究全球的计算机网络（万维网）。把网络服务器或网站作为网节，则它们之间的相互连线、电缆或光纤之间的联系就称为链路。网络系统还可用来研究社会学，如研究人与人之间的联系，把人作为节点，把人与人的联系作为节点与节点之间的联系，这种联系包括友谊、合作、家庭关系或者在科学上共同发表文章等。

这里我们首先讨论一下网络系统的一些特征，特别是网络系统的拓扑学特征。

1.无标度网络

网络包括随机网络（Random network）和无标度网络（Scale□free network）（Barabasi和Bonabeau，2003）。对一个网络最简单的定量

尺度是节点度分布（Degree distribution）。节点的节点度（Degree）是指与某一节点直接连接的邻居节点的数量。平均节点度是指一个节点直接连接的邻居节点的平均数。对随机网络来说，它的平均节点度遵循一个简单的参数。如一个简单的渔网，每个节点具有同样的平均节点度，即与其他节点的链路数目相同。相反，无标度网络是指在网络中某些节点与其他节点有很多相连的链路，但是大多数节点与其他节点的链路很少。有很多链路的节点被称为集散节点，其常有几十至几百个与其他节点连接的链路。无标度网络的例子有因特网、电力网、运输系统、人类社会和人体细胞代谢网络等。近期的研究表明，复杂生物网络，如酵母、线虫、果蝇的蛋白相互作用网络也是无标度网络。

无标度网络的一个特征可用数学方程描述如下，假如这个网络有k个链路，平均节点度的分布就遵循一个幂次定律（power law），公式如下：

p（k）～k-n

公式表明，一个网络的节点度和每个节点与其直接连接的节点的平均数之间存在幂次方的反比关系，即一个节点与k个其他节点相连的概率与1/kn成比例。k是节点度的变量。n是一个与标度无关的常数，即与网络大小无关。因此将具有这种特性的网络称为无标度网络。从直观上看：大多数节点只有少量的连接，而极少数节点却有大量的连接。无标度网络系统的一个典型例子是一个国家的航空网系统，其有几个大的空港，大的空港与很多中小城市间有许多航线，而中小城市之间直接的航线较少。无标度网络系统的一个重要特征是，一些连线的失败并不影响整个网络系统，比如某家航空公司的某个航班取消并不影响整个运输系统。但是它对有组织的攻击的抵抗力很弱，如在战争时期，假如对所有的大空港（即重要的集散节点）进行攻击，就会使整个运输系统处于瘫痪状态。这就与随机网络不一样，在随机网络中，大多数节点的连接点只有很少的几个，大多数节点有相同数目的连接点，它没有高度连接的集散节点。典型的例子如一个国家的公路运输系统，大多数城市有为数不多的高速公路和区间公路相连。

无标度网络节点连线的分布，与随机网络节点连线的分布大相径庭。随机网络节点连线的分布为泊松分布（poisson distribution）（呈钟形），钟形曲线在中间有一个峰值，无标度网络节点连线的分布遵循幂次定律，幂次定律的曲线是用连续递减的函数来描述的。

计算机模拟实验证明，无标度网络比随机网络具有更强的稳健性。如在一个万维网网关的模拟实验中，即使把80%的网关去掉，剩下的网关还能够形成网络系统，在任意两个网结之间可以找到一个通径。而在随机网络中，只要一部分网结被切断，整个网络系统就被分为许多个很小的互相不能通信的小区。然而，无标度网络对集散节点的被攻击是非常敏感的，只要把几个主要的集散网结去掉，整个网络就会处于瘫痪状态。因此，对于无标度网络系统来说，一个很重要的问题是到底有多少个集散节点。模拟实验表明，5%～15%的集散网结被去除之后，整个无标度网络系统就会瘫痪。

在生物系统研究和药物开发中，无标度网络系统具有很重要的潜在意义。如要找到最有效的药物，一般要找到药物可作用的集散节点，但是这些集散节点往往会产生很大的副作用。前列腺肿瘤细胞中雄性激素受体的基因调控网络就具有无标度网络的结构特征。该网络系统的重要集散网结包括雄性激素受体和pIK3R1这两个基因，而其他很多基因的连接点就相对较少，所以雄性激素受体和pIK3R1是该网络两个比较重要的药物靶点。

优先连结：无标度网络系统形成的过程中有一种优先连结（preferential attachment）机制，即新链路将在已有许多链路的节点与新节点之间产生，实际上也就是一个“富者更富”的过程，最终产生无标度网络的集散节点。

无标度网络系统形成过程中的优先连结（preferential attachment）机制，（a）中有12个节点形成了一个网络结构，其中节点A和B具有较多的连接点。因此，当一个新网节点X加入到原来网络中时，它会优先与A和B形成连接，使A和B的连接点越来越多，最终形成集散节点

第5章 阶层网络

复杂的生物网络不是随机网络，其常形成一定的层次和结构。以下主要介绍用于描述网络层次和结构的数学参数和模型。

1.聚合系数

生物网络常形成网络簇（Network cluster）。网络簇的稠密连通分支具有簇内连接稠密、簇间连接稀疏的特点。网络簇的连接可用聚合系数来描述。

聚合系数（Clustering coefficient）：一个节点的聚合系数可描述这个节点与其邻近节点聚合在一起的程度（Watts和Strogatz，1998）。一个节点i的聚合系数Ci是节点i与其邻近节点的实际连接数（ni）和所有可能连接数的比值。数学公式为

Ci=2ni/［ki（ki-1）］

ki为节点i的节点度。

一个有n个节点的网络的平均聚合系数为

=1n∑ni=1Ci

聚合系数的算法。灰色的节点i有三个白色的节点邻居，粗实线表示节点i的邻居之间的实际连接，而虚线表示节点之间没有连接。

合系数的算法。

研究一个含n个节点的网络的平均聚合系数，可以了解该网络可能的模块性。

2.小世界网络

如果一个网络的平均聚合系数远远高于有同样节点的随机网络的平均聚合系数，同时网络的平均最短路径（Shortest path length）都很短，则称该网络为小世界网络（Small world network）（Watts和Strogatz，

1998）。生物系统的很多网络属于小世界网络。小世界网络是介于规则网络和完全随机网络之间的一类模型。

3.阶层性的网络

要理解网络的阶层性，可以先看看如何构建一个阶层网络（hierarchical networks）。由小网络的重复来构建阶层网络的过程。

小网络的重复形成了阶层网络。从（a）的完全连接的五个节点形成的聚合结构开始。（b）是在（a）基础上加了四个完全相同的（a），把各自边界上的四个节点与原来聚合结构（a）的中心节点相连，形成了一个有25个节点的网络。（c）则利用同样的原理，把聚合结构（b）重复四次，同样把边界上的节点与最原始的聚合结构（a）的中心节点相连，于是构成一个具有125个节点的网络。如此，该过程可以一直进行下去。

无标度网络和阶层网络的网络平均聚合系数与网络链接数k的关系。阶层网络（用黑点表示）的C（k）约等于k-1，而无标度网络的平均聚合系数为一特定的常数，不依赖于网络的链接数k。需要指出的是，在生物网络（如代谢网络、蛋白□蛋白相互作用网络）中，无标度（Scale free）和网络的聚合（Clustering）这两个特征并不是相互排斥的，而是可以共存的。两者共存则形成阶层网络。无标度网络和阶层网络的区别在于阶层网络平均聚合系数依赖于网络的链接数k（约等于k-1）；而无标度网络与随机网络的平均聚合系数一样，均不依赖于网络的链接数k。

三种复杂网络系统的结构。（a）描述一个无标度网络，而（b）和（c）则描述加上网络的聚合而形成的模块结构，最终形成了阶层网络。

三种复杂网络系统。（a）中，描述了一个无标度网络，它的平均节点度分布遵循一个幂次定律（power law）。在该类网络中，有几个高度连接的网节，即集散节点（蓝色的圈）。为将整个网络组织在一起，这些集散节点发挥了很重要的作用。一个具有256个网节的无标度网络系统的典型结构。这个网络是根据优先连结（preferential attachment）原理，用pajek大规模网络系统分析软件，根据没有模块结构的（群聚）聚类算法来构建的。一个具有四个模块的网络系统，四个模块间的连接点比较少；由256个网节点所形成的具有四个模块的网络系统。

三种复杂网络系统。图（c）所示为一个有阶层性的网络系统，它具有无标度网络的结构，并含有模块性，可见阶层网络是由模块性和无标度网络结合在一起形成的，其阶层性由蓝色到红色四种不同颜色来显示。

衡量复杂网络节点的重要性通常需要用不同的中心度指标。典型的中心度指标包括点度中心度（Degree centrality）、中间中心度（Betweenness centrality）、接近中心度（Closeness centrality）、特征向量中心度（Eigenvector centrality）和子图中心度（Subgraph centrality）等。

在图论和网络分析中，有许多关于点在中心度的测量，它体现了一个点在的相对重要性（例如，个人在社会网络中的地位，或者空间句法理论中一个房间在整个房屋中的重要性，又或一条道路在城市道路网中的利用度）。

在网络分析中，常用的中心度度量有4种，即点度中心度、中间中心度、接近中心度及特征向量中心度。关于中心度的综述可以参看Opsahl等的文献（Opsahl T等，2010）。

4.点度中心度

中心度指标中，最简单的是点度中心度。点度中心度（Degree centrality）是指一个节点关联的事件数（即一个节点所含的边数）。度通常解释为节点捕捉的任何通过网络的信息的能力。如果网络是有向的（即链路是有向的），那么度可分为入度和出度。入度是指向节点的链路数，出度则是指从节点导出的链路数。对于社交关系，我们通常用入度来表示知名度，用出度表示合群性。

对于一个包含n个节点的G=（V，E），即有V 个节点数，E 个链路数。节点v的点度中心度CD（v）定义如公式：CD（v）=deg（v）n-1用的稠密邻接矩阵来表示计算所有点V在点度中心度需要Θ（V2）的时间，而对于所有链路E 用稀疏矩阵表示来计算则需要Θ（E）的时间。中心度的定义可以推广到多张上。令v*表示G中点度中心度最大的节点，令G′=（V′，E′）为n点并使以下量达到最大：h=∑|v′|j=1CD（v′*）-CD（v′j）G的点度中心度可定义为：CD（G）=∑|V|i=1［CD（v*）-CD（vi）］h如果G仅包含一个节点，它连接其他所有节点，同时其他点仅和这个中心点相连，那么h将达到最大，在

这种情况下：h=（n-1）1-1n-1=（n-2）所以G的点度中心度可以退化成：CD（G）=∑|V|i=1［CD（V*）-CD（vi）］n-2一般来说，点度中心度较高的节点是网络中活跃的节点，常是一个共同连接点或一个集散节点。节点A 具有最大的点度中心度，因为与它连接的节点最多。但是，它并不见得是网络中最重要的节点，因为与它连接的其他节点集中在网络的一个局部。

中间中心度（Betweenness centrality，BC）是一个节点的中心度的测量（边的中间性这里不予讨论），衡量一个节点是否是一个网络的中心。中间中心度表示所有的网络连接中通过某个节点的最短路径条数。中间中心度很好地描述了网络中一个节点可能需要承载的流量。一个节点的中间中心度值越大，流经它的数据分组越多，成为连接网络中不同部分的桥梁，也意味着它越容易拥塞，成为网络的瓶颈。

位于许多其他点的最短路径中的点，比不在最短路径中的点有更高的中间中心度。

对于一个包含n个节点的G=（V，E），节点v的中间中心度定义为CB（v）=∑s≠v≠t∈Vσst（v）σst这里，s和t指任一对节点，σst是从s到t的最短路径数，σst（v）是从s到t的所有最短路径中通过节点v的路径数。中间中心度与节点对的数量相关，它可以通过除以不包括节点v的节点对的数量进行规范化：对有向图，除数是（n-1）（n-2）；而对于无向图，除数则是（n-1）（n-2）/2.

计算所有节点的中间中心度和接近中心度涉及所有节点对的最短路径的计算。在计算所有点的中间中心度和接近中心度时，总假设是无向的，并允许与环和多重边相连，为保持简单的关系，通常假设图是不含环和多重边的（这里“边”代表连接两个节点的链路），利用Brandes的算法需要将最终值除2，因为每个最短路径都计算了两次（Brandes，2001）。

中间中心度较高的节点常在网络中居重要的位置，如把它去除，会导致网络的分散，使不同的局部网络之间脱离连接。节点B具有最大的中间中心度，因为它连接了由节点A和节点B形成的小的局部网络。

5.接近中心度

接近性是对节点的一种中心度的测量。“浅”于其他点的节点（即与其他点趋向于有较短的几何距离的点）有更高的接近性。网络分析中更倾向于用接近性来表示最短路径距离，因为它给更中心的点赋予了更高的值，所以通常与网络的其他度量指标（如度）有正相关。

在网络论中，接近性（Closeness）是一个复杂的中心度的测量。它是由节点v及与其相连的其他节点的平均几何距离（即最短路径）来定义的：∑t∈V＼vdG（v，t）n-1这里n≥2.接近性可视为信息在网络中从一个定点传播至其他相连点的一种时间度量。有时接近性也被定义为这个量的倒数，即计算信息从一个定点传播至网络中其他相连点的速度（而非时间跨度）。节点v的接近性CC（v）是其到具有V个节点的图中所有其他点的距离和的倒数：CC（v）=1∑t∈V＼vdG（v，t）通常地说，一个接近中心度（Closeness centrality）较大的节点可以很快到达网络中的其他节点，即它与其他节点的距离较短。节点B具有最大的接近中心度，因为它可在最短的距离内到达最多的其他节点。

节点表示蛋白，黑色的双向链接代表蛋白□蛋白相互作用（ppI），灰色有方向的链接表示转录调控作用（TRI）。来自Yeger□Lotem 等的图一A，版权许可复制Yeger□Lotem 等研究了蛋白□蛋白相互作用和基因转录网络中的网络模体，并提出了两个新概念。①节点的延伸度（Extended degree）：是指进入和离开一个节点的每一种链接的数量。如节点a有一个蛋白□蛋白相互作用（ppI）、一个入边（进入节点的）转录调控作用（TRI）、一个离边（离开节点的）转录调控作用（TRI），依此类推。②链接的分布（Edge profiles）：是指连接两个节点的一套链接（包括其类型和方向）。如节点a和b之间有一个蛋白□蛋白相互作用和一个进入节点a的转录调控作用。

6.特征向量中心度

特征向量中心度（Eigenvector centrality）是对网络中节点重要性的度量。如Google公司对万维网网页的排名就采用了特征向量中心度的概念。它是指如果某节点与网络中重要的节点相连接，则该节点的重要性就比它与网络中不太重要的节点的连接要大，得分也较多。

如使用邻接矩阵来寻找特征向量中心度，令xi代表第i个节点的得分，令Ai，j表示网络的邻接矩阵，即如果第i个节点与第j个节点连接，则Ai，j=1，否则Ai，j=0.

对于第i个节点，令特征向量中心度的得分与所有连接节点i的节点得分和成比例。所以xi=1λ∑j∈M（i）xj=1λ∑Nj=1Ai，jxj这里M（i）表示与节点i连接的节点集，N表示节点总数，λ是常数。用向量符号可以写成x=1λAx，或用特征向量方程表示Ax=λx，对节点B来说，节点A与B相连接，则该连接对B的重要程度要大于节点A1与B的连接。

第6章 生物网络模体

生物网络不是随机产生的，而是有一定结构特征的。生物网络系统具有模块性（Modularity），这种模块性是由进化或者细胞、组织、器官在结构上的分区所造成的，在其他人工构建系统中也很常见，比如大规模集成电路或软件系统中也经常存在这种模块性。随机网络与实际网络的不同。

（a）为一个实际的生物网络，（b）为随机网络。在随机网络中，任一节点进出的网边（链路）数量都是相同的。而在实际生物网络中，则存在一个前馈（Feed□forward）网络模体（由虚线显示），在（a）中出现五次。

网络模体是指：相比于随机网络，复杂网络中网结之间显著增多的连接的图形和规则。最简单的网结X和Y连接的生物学例子。网结三种可能的连接方式和规则。

（a）为网结X和Y连接的生物学例子，如：基因转录网络中，转录因子X结合在基因Y的启动子上，调控基因Y的表达；细胞网络中，神经细胞X和Y通过神经突触相连；生态食物链中，大鱼吃小鱼。图（b）为三个网结构成的13种可能的子图。子图的可能性组合随节点数目的增加而增多，如四个节点则有199种可能的子图。

Alon U实验室提出网络模体（Network motifs）的概念（Milo等，2002；Shen□Orr等，2002）。Alon研究组在生物网络和人造技术网络中找到了许多网络模体，包括前馈（Feed□forward）、双双（Bi□fan，即两个起始节点交叉调控两个靶向节点）、双平行（Bi□parallel）、三链式（Three chains）等网络模体。

Yeger□Lotem等研究了蛋白□蛋白相互作用和基因转录网络中的网络模体（Yeger□Lotem等，2004）。两个、三个、四个基因或蛋白的可能的相互作用。他们首先用严格的参数（对蛋白□蛋白相互作用，数据至少需要来自两个实验，对转录调控作用，他们必须不仅仅是用来自全基因组水平的研究的数据，如染色质沉淀结合芯片或测序的数据），找出较为可信网络，称为严谨网络（Stringent network）的酵母

菌1385个蛋白的1832个相互作用网络及128个转录因子和591个靶基因的1351个相互作用。

两个蛋白或基因之间可能的相互作用。点表示基因及其蛋白产物，有方向的灰色链接表示基因转录调控的相互作用，黑色的双向链接则表示蛋白□蛋白相互作用。

三个蛋白、基因间可能形成的网络模体。（a）蛋白小集团（protein clique）；（b）相互作用的转录因子共同调控第三基因；（c）前馈连环（Feed□forward loop）；（d）被同一基因调控的相互作用的蛋白；（e）两个相互作用的转录因子共同调控一个基因的混合前馈连环。点表示基因及其蛋白产物，有方向的灰色链接表示基因转录调控的相互作用，黑色的双向链接则表示蛋白□蛋白相互作用。

在四个蛋白的多种可能的相互作用图谱中，找到了63种网络模体，其中都包含一个或一个以上的由三个蛋白构成的网络模体。因此，小的网络模体是网络的构建单元。

在酵母菌严谨网络（Stringent network）中找到的四个蛋白相互作用的网络模体。（a）三个蛋白相互作用的网络模体排列组合后可能形成的网络模体。（b）不能三个蛋白相互作用的网络模体排列组合形成的网络模体。Ⅰ：双双模体（Bi□fan motif，即两个起始节点交叉调控两个靶向节点）；Ⅱ：包含前馈连环的模体；Ⅲ~Ⅵ：右边模体是左边模体的延伸，例如左边的蛋白相互作用网络中由两个蛋白之间的相互作用扩展到右边的三个蛋白之间的相互作用。点表示基因及其蛋白产物，有方向的灰色链接表示基因转录调控的相互作用，黑色的双向链接则表示蛋白□蛋白相互作用。

第7章 生成随机网络的算法

检测网络模体的最简单方法是与含有相同数量的节点和链接的随机网络进行比较。Milo等总结了随机网络生成的三种方法。

1.交换算法

交换算法（Switching algorithm）对一个网络执行一系列的蒙特卡罗交换。例如选择一对链接（X1→X2；Y1→Y2）进行交换，形成（X1→Y2；Y1→X2）。若该交换导致了多重链接或自回链接，则取消该交换。

交换算法（Switching algorithm）。通过链接交换产生随机网络，交换不改变节点度。

2.匹配算法

匹配算法（Matching algorithm）把网络的每个节点的入边（链接）和出边（链接）想象成中间断开的短棒，即入边短棒和出边短棒；然后随机地选择一个节点的入边短棒和另一个节点的出边短棒，进行连接。若该组合导致了多重链接或自回链接，则取消该组合，上述过程则重新开始。

匹配算法（Matching algorithm）

3.赢家算法

赢家算法（“Go with the winners”algorithm）对每一个网络的操作过程与匹配算法相似。为了弥补在匹配算法过程中多重链接或自回链接而造成的网络数不断减少的情况，赢家算法过一段时间就将剩余的网络复制，重复匹配算法的过程，直至所有节点的入边短棒和另一个节点的出边短棒都形成连接。然后选择一个随机网络作为输出。

子图搜索法

生物网络中，网络模体的最简单的发现算法首先是在实际网络中挖掘子图（Subgraph）。子图搜索法包括穷尽递归搜索算法（Milo等，

2002）、采样算法如网络链接取样法（Edge sampling，ESA）子图枚举法（Enumerating subgraph，ESU）。

对找到的子图或模体的统计意义可以用Z评分（Z score）来评价。比较真实网络和随机网络（≥1000）子图的发生频率，如Nreal代表真实网络子图的发生频率，Nrand代表随机网络子图的发生频率，那么Z score=（Nreal -Nrand）/SD。SD为随机网络子图的发生频率的标准差。

这些方法已被广泛应用于大肠菌、酵母菌的基因转录调控（Transcription□regulation interaction，TRI）和蛋白□蛋白相互作用网络（protein□protein interaction）。

第8章 常用的网络分析软件

□Osprey：一个相互作用网络的可视化系统。

Osprey 的工作效果图

□MAVisto：一个探索网络模体的工具。

□FANMOD：一个快速检测网络模体的工具。

□NeMo（Network Module identification in Cytoscape）：一个在Cytoscape 软件中检测网络模体的工具。

□hCCA（heuristic Cluster Chiseling Algorithm，启发式聚簇砍凿法）：网络模体发现软件。

□MCODE：网络模体发现软件，是一个Cytoscape插件。

□Mfinder：网络模体发现软件，mDraw为网络的可视化软件。

□pajek：大型网络分析软件。

□Kavosh：网络模体发现软件，节约计算中央处理器和随机内存。

□用强力的图形分析（power graph analysis）挖掘蛋白网络：用图形分析发现蛋白网络的软件。

□VANTED（利用生物网络关联性的高级的数据展示和分析系统）：网络模体发现和可视化软件。

□CFinder：蛋白小集团（Cliques）和生物网络中重叠网络模体发现软件。

第9章 新兴的高通量测序技术

DNA（或RNA）编码生命体的遗传信息。DNA（或RNA）测序是指通过各种技术方法确定特定DNA（或RNA）片段中核酸碱基的排列顺序。DNA双螺旋结构发现后不久，就有了核酸测序方法的报道（Whitfeld，1954）。但是，真正意义上的核酸测序始于20世纪70年代。通过RNA测序，Fiers和他的同事们发表了第一个完整基因的序列（Min Jou等，1972）和首个噬菌体的基因组（Fiers等，1976）。同一时期，出现了几种不同的DNA测序方法（Gilbert和Maxam，1973；Sanger和Coulson，1975；Maxam和Gilbert，1977）。1977年，Sanger发明了具有里程碑意义的末端终止法（Sanger等，1977）。末端终止法，又称为Sanger测序法，比同时期的测序方法效率更高，毒性更小，因此很快被广泛应用。

末端终止法测序以DNA单链为模板，用DNA聚合酶从测序引物延伸互补链。延伸反应体系的脱氧核苷三磷酸（Deoxynucleotide triphosphate，dNTp）混合物中混入1%的2′，3′□双脱氧核苷三磷酸（Dideoxynucleotide triphosphate，ddNTp）。由于后者不能与下一个脱氧核苷三磷酸形成磷酸二酯键，从而延伸过程可随机地终止，最终形成长短不一的DNA片段（单链）。测序过程由四个延伸反应组成，每个反应混入四种ddNTp（ddATp，ddGTp，ddCTp和ddTTp）中的各一种。反应后，利用电泳分离技术，根据DNA随机终止形成的片段，从小到大地读出对应延伸终止末端的碱基。

末端终止法经过不断的优化和改良，在自动化（Automatic）、并行化（parallel）方面获得了重要的提升。以人类基因组计划为例，通过对Sanger方法的不断优化，测序成本已经从之前的平均10美元完成1个碱基，降到了平均1美元完成10个碱基。近30年来，末端终止法一直是应用最为广泛的测序技术，被誉为DNA测序的金标准。期间还有一些其他的测序方法，如连接测序法（Sequencing by ligation，SBL）、焦磷酸测序法（pyrosequencing）（hyman，1988）、杂交测序法（Sequencing by hybridization，SBh）（Drmanac等，1989）等，但局限于当时的技术条件，并没有获得大规模的应用。

美国应用生物系统公司（Applied Biosystems，AB）将荧光标记的ddNTp用于Sanger测序的pCR反应，并配合光学检测技术和自动化设

计，提高了测序通量并降低了测序成本。利用自动化的Sanger测序法，人们已经完成了多个物种的全基因组测序，为现代生物医学开启了基因组（Genomics）时代的大门。然而，人们也清楚地看到，每一个物种全基因组的测定，所耗费的人力、物力以及时间都相当巨大。1990年启动的国际人类基因组计划，花费了约3亿美元的巨额投入和多国科学家十几年的共同协作，才完成了这一举世瞩目的生命科学“登月计划”。

人类基因组序列图谱的完成，引起了学术界和商界精英们对高通量（high□throughput）测序技术的兴趣。以花费1000美元测个人全基因组序列为目标（Service，2006），人们开始致力于开发成本更低、速度更快的测序技术。同时，随着人们对功能基因组学研究的不断深入，研究热点也从宏观共性的研究逐步深入到对微观个体差异的研究，个体化医疗（personal medicine）的理念开始盛行。

正是基于上述需求，在多个科研机构和商业公司的共同努力下，一批高效率、高通量的测序仪或测序原型应运而生。新兴的高通量测序技术发展迅猛，在技术原理与开发潜能上渐渐地细分为第二代测序技术和第三代测序技术。鉴于测序技术的日新月异和技术细节的错综复杂，很难对这些新的技术进行完整的划分。为了便于理清思路，本文定义第二代测序技术为运用冲洗与扫描技术（Wash□and□scan techniques）确定DNA单分子克隆共识序列（Consensus）的高通量测序方法。所谓冲洗与扫描，就是测序时每合成一个碱基或新增一个检测信号后，将DNA聚合酶和其他试剂冲去后进行扫描。第三代测序技术是指连续碱基读取的直接检测由单个分子测序信号的单分子测序（Single□molecule sequencing，SMS）方法。

1.第二代测序技术

第二代测序技术与Sanger测序法相比，大大提高了测序通量，一次能并行对几十万甚至几百万个DNA分子克隆进行测序。从技术细节上，第二代测序技术都包含模板预备、测序、成像和数据分析四个环节。每个测序平台在个别或一些环节的技术差异形成了各有特色的测序方法。为了实现测序的高通量，第二代测序方法通过模板预备过程将待测序的DNA模板附着或固定在固体表面或支持物上。分散的待测模板占据各自非常狭小的位点，保证成百上千个位点能够同时进行测序反应而互不干扰。

模板预备不同于样品准备，即测序文库准备。模板预备是指将待测DNA克隆的单链固定在测序反应支持物上，并与测序引物杂交以备测序反应的开始。因为第二代测序平台的成像系统并不能灵敏地检测单个DNA模板所发出的荧光信号，所以在测序前需要对每个待测序模板进行扩增。同时为了使每个单独的测序反应互不干扰，一般采用乳液pCR（Emulsion pCR，empCR）（Dressman等，2003）或固相扩增（Fedurco等，2006）的方法将模板的克隆限制在各自的物理空间内。

第二代测序技术的模板预备原理。

第二代测序技术经过六年多的商业化推广，已经有了很好的科研基础，应用的领域也相当广泛。目前，第二代测序技术主要有三个商业化的平台代表：Roche公司的454焦磷酸测序，Illumina公司的Solexa测序和Life Technologies公司的SOLiD测序。

2.1454测序技术

2005年，454Life Sciences公司（2007年被Roche正式收购）推出了454Genome Sequencer 20（GS20）焦磷酸测序仪。454测序技术得以广泛地应用于各个科研领域，使得这一测序平台在新一代测序技术领域中占据举足轻重的地位。

焦磷酸测序法早在1985年就有报道（Nyren和Lundin，1985），后来经过不断改进（hyman，1988；Ronaghi等，1996；Ronaghi等，1998；Ronaghi，2000；Ronaghi，2001），最终成就了首个实现广泛应用的第二代测序平台。焦磷酸测序的基本原理是：由DNA聚合酶（Klenow大片段）、ATp硫化酶、荧光素酶、腺苷三磷酸双磷酸酶等四种酶及其底物5′□磷酰硫酸和荧光素组成的反应体系共孵育，在四种酶共同催化下的酶级联化学发光反应。在每一轮测序反应中，只在反应体系中加入一种dNTp，若该dNTp与待测模板配对，DNA聚合酶就可以将其掺入到合成链中并释放出等摩尔数的焦磷酸基团（ppi）。释放出的焦磷酸基团经硫酸化酶催化形成等摩尔数的ATp，生成的ATp和荧光素酶共同催化荧光素转化为氧化荧光素。氧化荧光素发出的可见光强度与生成的ATp量成正比，光信号经CCD摄像机捕获，并通过计算机软件转化为一个峰值。因此，峰值的高度与反应中掺入的核苷酸数目成正比。光信号被淬灭，测序反应室内再加入腺苷三磷酸双磷酸酶将ATp和剩余的dNTp降解，从而再生反应体系。然后进入下一个循环，加入下一种dNTp，继续DNA互补链的合成，产生相应的光信号

（Ronaghi，2001）。需要指出的是，焦磷酸测序在读取连续的单碱基重复时存在缺陷，不能准确地测出长度。

焦磷酸测序的基本原理。

454测序仪正是基于焦磷酸测序法而开发出来的高通量测序平台。454测序步骤如下。①测序文库的构建。将待测样品的DNA用物理方法打碎成400~800bp的片段，经末端修复，并在两端分别加上特异性的锚定接头。②乳滴pCR。将带有接头的DNA片段分别结合到特殊设计的DNA捕获磁珠上，使大部分的磁珠携带一个DNA片段。磁珠被单个油水混合小滴包被后，在这个小滴内进行独立扩增。乳滴pCR可以避免pCR产物间的交叉和减小扩增偏好的影响，确保每个磁珠上的DNA模板为单个克隆。由于这些磁珠的表面布满了上百万条与文库构建时加入的接头互补的寡核苷酸序列，因此，可以保证每一个待测序片段与磁珠表面的寡聚核苷酸杂交，孵育的油水混合小滴中含有pCR反应体系。待测样品在磁珠上经过孵育pCR，每一个磁珠表面将获得超过一百万倍的原始DNA片段的拷贝，以达到测序反应所需的检测信号值。③pTp载样。将经过孵育的磁珠加至454公司发明的454pTp板（454picoTiterplate，pTp）表面，pTp板上的每一个孔只能容纳一个经过孵育的磁珠。④测序。将包含有焦磷酸测序反应激活液的微磁珠加入pTp孔内，通过检测到的光信号确定待测DNA的序列。

目前，Roche/454的测序平台已经更新到GS FLX+和GS Junior两大系统。前者的最大读长可达1000bp（主要为700bp），为二代测序平台中读长最长的，与Sanger测序的读长（800～1000bp）相当。GS FLX+根据读长要求的不同，可运行10～23个小时完成测序反应。后者是面向中小型实验室推出的新的普及型测序平台。GS Junior系统仪器大小与普通的激光打印机类似，每次运行能产生10万个读数，平均读长约400bp，准确率能达到99%。在1～2天的时间就能完成全部的测序流程，测序反应仅需10个小时。

3.2Solexa测序技术

2006年，当时正迅速成长的基因芯片产商Illumina公司收购了正在研发第二代测序仪的Solexa公司。收购完成后不久，Illumina/Solexa便推出了以Genome Analyzer命名的高通量测序仪。由于该测序技术的核心在Solexa公司时期便已完成，因此人们习惯于把这种测序方法称为Solexa测序技术。Solexa测序以边合成边测序（Sequencing by

synthesis，SBS）自称，主要是区别于Sanger测序，后者需合成好一系列大小的DNA后方才读取每个位置的碱基。更确切的定义可以把Solexa测序称为循环可逆终止测序（Cycling reversible terminator sequencing），即测序反应中dNTp的特殊修饰阻断下一个碱基的延伸，并在成像完成后消除阻断基因进入下一个测序循环。Illumina/Solexa将四种dNTp标记上不同的荧光，利用DNA聚合酶合成互补链时，每添加一种dNTp就释放出不同的荧光，根据捕获荧光信号获得待测片段的序列信息。

Illumina/Solexa测序反应与碱基读取的原理。

Solexa测序实验流程如下。①测序文库的构建。利用物理方法将待测样品DNA打碎，在DNA碎片两端加上接头。②锚定桥接。Solexa测序时利用微注射系统将已经加过接头的待测片段添加到玻璃流动栅元（Flowcell）内，每一个流动栅元又被分为8条通道（Lane），每条通道的内表面上能以共价键的形式随机固定单链接头序列和带接头的单链待测DNA片段。③桥型扩增循环获得多拷贝待测DNA片段。在流动栅元内加入未被标记的dNTp和酶，起始固相桥型扩增。所有的单链桥型待测片段被扩增成为双链桥片段，通过变性，释放出互补的单链，锚定到附近的固相表面。通过不断循环，将会在流动栅元的固相表面上获得上百万条成簇分布的单链待测片段。④测序。加入带有不同荧光标记的dNTp，并且dNTp是经过抑制基团修饰的，功能类似于ddNTp，可以使得DNA合成终止，不同点在于该抑制基团可消除，所以称为可逆终端循环技术。在抑制基团的作用下，DNA聚合酶每次只能装配一个带荧光标记的碱基到DNA模板上，然后洗脱未装配的dNTp，接着荧光成像来检测装配上去的碱基信息，再去除碱基上的抑制基团和荧光标记，洗脱后可以进行第二个碱基的装配。通过这样的循环反复，完成测序工作。

目前，Illumina/Solexa推出了四种通量不同的测序平台，以满足不同的测序需求和硬件条件。通量从小到大分别为Miseq（1Gb），Genome Analyzer IIX（95Gb），hiScanSQ（150G），hiSeq Systems（300~600G）。Illumina计划于2012年下半年推出读长更长、速度更快的测序解决方案：Miseq的读长将实现250bp，而hiSeq有望实现在30小时内完成2×100bp的测序，并最长可测150bp。

4.SOLiD测序技术

2006年，Applied Biosystems（AB）公司（现已与试剂巨头Invitrogen合并为Life Technologies）收购了正在开发另一种第二代测序平台的Agencourt personal Genomics（ApG）公司，并在2007年正式推出SOLiD（Sequencing by Oligonucleotide Ligation and Detection）测序仪。与454和Solexa不同的是，它不是利用DNA聚合酶在合成互补链的过程中产生测序信号并读取序列信息，而是利用DNA连接酶通过连接反应来确认与待测模板互补的碱基。

SOLiD “双碱基”连接测序原理。SOLiD的关键技术是通过连接来获得“双碱基”的荧光数据，并通过多次连接以及连接引物的多次移位，达到完整测序与校验的目的。SOLiD 连接反应的底物是8碱基单链荧光探针混合物。在连接反应中，这些探针按照碱基互补规则与单链DNA模板链配对。探针5′末端分别标记4种颜色的荧光染料，并且这四种颜色与探针3′端第1、2位构成的碱基对有对应关系；探针3′端3～5位的“n”表示随机碱基；6～8位的“z”指的是可以与任何碱基配对的特殊碱基。第一轮连接测序：利用DNA连接酶将与待测片段第1位、第2位碱基匹配的8碱基单链荧光探针连接上去；激发荧光后拍照成像；然后通过化学方法将第6位到第8位的碱基切除。洗涤后，加入探针和酶启动下一个连接反应。在经过了5～7个循环的连接测序以后，最终延伸的产物再从模板上面解链移除，至此第一轮的连接测序完成。由于每次都是得到8碱基单链荧光探针上第1位、第2位碱基的序列，在一连串的连接测序以后，得到的测序结果为模板DNA的第1、2、6、7、11、12等位置的碱基信息。第二轮的测序过程，测序引物的3′端比第一轮的引物缩短了一个基因，从n-1的位置开始启动新一轮的连接测序。SOLiD运行过程中，一共需要启动五轮这样的连接测序反应。总之，这五轮测序的连接引物相同，但在p1引物区域的位置依次相差一个碱基。

SOLiD测序的基本流程如下。①测序文库准备。利用物理方法将待测样品DNA打碎，在DNA碎片两端加上一对接头p1、p2.②乳液pCR。将已加好接头的待测DNA片段，接头引物p1、p2（其中p2引物的量远远大于p1引物的量），DNA聚合酶以及直径为1μm的表面附满与p1接头互补序列的小磁珠，通过乳滴pCR，使待测片段被锚定到磁珠表面，然后扩增获得大量待测DNA片段的拷贝。③边连接边测序。

从其测序原理可知，SOLiD测序的读长一般是5的倍数，目前可以达到最大读长75bp。SOLiD测序采用双碱基编码法进行测序，测序过程

中，对每个碱基分析判读两遍，能够在序列测定过程中过滤原始数据错误，提供内部错误校正。因此，SOLiD可以用于低覆盖率的突变分析。

5.第二代测序技术的缺陷

第二代测序技术的特点决定了其自身的缺陷（Schadt等，2010）：第一，冲洗与扫描过程的反复使得测序耗时过长；第二，冲洗与扫描过程的反复耗费大量的反应试剂，不利于节约成本；第三，测序依赖于DNA分子的pCR克隆而非单分子，使测序前测序文库的准备工作变得繁琐；第四，pCR扩增的错误和偏好使测序结果不能如实地反应原始DNA的序列信息和丰度对比；第五，pCR克隆内的分子在测序反应过程中同步性逐步降低，限制了测序的读长。第二代测序技术检测DNA克隆整体发出的测序信号，这些测序信号组成的序列信息就是共识序列。在冲洗与扫描的循环中，克隆内部的各个分子间逐渐失去同步性。另外，由于模板预备过程中pCR扩增的错误导致克隆内的非一致性。

第10章 第二、三代测序技术间的过渡

如前所述，学界对测序技术的划分尚未清晰。依照本文的划分标准，尚有一些测序方法介于第二代与第三代之间，如helicos Biosciences公司开发的heliScope Sequencer和Life Technologies旗下的Ion Torrent测序平台。从单分子测序的角度讲，两者都实现了对单个DNA分子序列的直接检测。但是，在相邻两个信号的读取之间，碱基信号的释放都被目的性地中断以清除当前循环对下一个信号的影响，即没有脱离前述的冲洗（与扫描）过程。

1.离子半导体测序技术

半导体测序技术（Semiconductor sequencing）也是一种边合成边测序技术，即在合成模板DNA互补链时检测相应完成聚合的碱基。与第二代测序技术利用合成（或连接）时的发光反应不同，半导体测序技术通过半导体材料感应DNA聚合时释放的特定信号。

2010年，Ion Torrent公司（同年被Life Technologies公司收购）发布了第一台半导体测序仪。Ion Torrent测序平台的核心是一块新型的半导体芯片，其中包含了数百万个孔和相应的底部感应器。测序反应时，该芯片所使用的电子检测系统简化了测序过程，并大大降低了测序仪的费用。仪器中没有光学组件，而主要由电子读取器、微处理器和流体系统组成的。与第二代测序技术相同，Ion Torrent测序反应检测的是最小单元为DNA模板的pCR克隆，而非单个分子。事实上，除利用半导体检测碱基的合成外，Ion Torrent测序平台的设计与Roche/454非常相似。其模板准备过程通过乳液pCR扩增DNA模板，确保每个磁球上有待测DNA模板的单分子克隆。测序时，每个合成反应只加入一种未标记的普通的dNTp，根据是否检测到小孔内ph值的变化来判断是否在延伸链上加入了相应的核苷酸。Ion Torrent测序的实验操作流程与Roche/454的焦磷酸测序平台相似。

Ion Torrent半导体测序基本原理。

离子半导体测序的优势在于测序的合成反应中只需使用普通的非标记的dNTp，并且节约了除聚合酶外各种生物酶的消耗，故试剂成本相当低廉。同时，感应器对信号的实时检测，有利于提高测序的速度。遗

憾的是，由于测序模板DNA克隆同步性随测序循环下降，因此离子半导体测序的读长依旧不能突破第二代测序的极限。截至2012年3月，Ion Torrent测序平台读长仅有400bp。

此外，Ion Torrent平台的专利来自DNA Electronics公司。据报道，Roche正在与后者合作开发基于半导体测序技术的大读长、高通量、低成本的测序平台。

2.首个单分子测序平台

helicos Biosciences公司开发的heliScope Sequencer可谓是单分子测序的先锋。在开发前期，由于heliScope的测序读长一直处于很低的水平，严重阻碍了它的发展。直至2006年，经过众多科学家的不懈努力，终于在读长上取得了突破性进展，经过对其体系的进一步优化，推出了目前的heliScope单分子测序方法。2008年4月，Science杂志以Reports的形式介绍了这一方法。heliScope测序仪主要利用DNA聚合酶装配双链DNA的原理，每次加入DNA聚合酶和一种荧光标记的碱基，通过检测荧光来监测各处模板的合成情况，从而达到测序的目的。

实事上，heliScope测序的原理与Solexa方法非常相近。两者的主要区别是heliScope为单分子测序，模板预备并不通过pCR形成单分子克隆，而是直接将单分子的DNA模板（单链）固定在支持物表面。因此，heliscope的模板预备分为引物固定型、模板固定型两种。①引物固定型：事先将引物探针以共价键连接到固体支持物上，再把随机片段化后连有接头的模板与引物杂交。DNA杂交后达到固定待测模板的目的，固定的引物亦是后续测序反应的测序引物。②模板固定型：在上述引物固定的测序反应完成后用变性的方法洗去原先的待测模板，留下固定在支持物上的由引物探针延伸而来的互补模板用于第二轮测序。helicos BioSciences公司2010年上市的测序仪同时采用这两种模板预备方法，分别对应于它的第一轮和第二轮测序过程。heliScope测序反应时每次只加入一种荧光标记的碱基，通过检测荧光来监测各处模板的合成情况，从而达到测序的目的。

helicos Biosciences模板固定方法与测序反应原理。

heliScope测序的基本流程为：①利用物理方法将待测样品打断，在待测DNA样品的3′末端加上poly（A），利用末端转移酶在接头末端加上Cy3荧光标记。②将已加polyA的待测DNA样品加到一平板表面上，平

板表面固定有50个T的寡聚核苷酸链，通过退火，待测DNA样品可以被锚定在表面。③单分子测序。

目前，heliScope测序法的平均读长在25bp左右，最长可达到30bp。经过简单的改造，在第一轮测序结束后，通过加热解链，可以对同一DNA分子进行二轮测序（Two□pass），其错误率可以控制在0.2%～1%，每小时可获得大约25～90Mb的数据（harris等，2008）。与第二代测序仪相比，heliScope测序最大的优势在于它不需要经过pCR进行模板的扩增。它的荧光捕获系统检测灵敏度非常高，能检测一个单分子合成所释放的荧光信号。由于不需经过预扩增，它可以避免很多因为pCR扩增而引入的不确定因素，而且可以通过两轮测序来提高测序的准确性。同时heliScope测序所消耗的试剂量也会大大降低，因此它的测序成本相较于第二代测序平台要低很多。另外，heliScope也能够用于RNA测序，只需以逆转录酶替代DNA合成酶，这是对第一、二代测序的重大突破。

3.第三代测序技术

在本章关于第三代测序技术的定义之下，可将第三代测序技术进一步分成三类。①边合成边测序技术：以单个单链分子在DNA聚合酶作用下合成互补链的反应过程为检测基础，如pacific Biosciences公司的实时单分子测序技术；②纳米孔测序技术：以单个碱基通过纳米孔时释放的信号为检测对象，如Oxford Nanopore公司在2012年2月展示的最新研发进展；③使用先进的显微技术对DNA分子直接成像并读取碱基序列，包括透射电镜和扫描隧道显微镜。这些技术正处在不同的研发和应用阶段，并且在不同的应用中各有优点和缺点。由于显微测序技术的发展尚处于萌芽阶段，尚未有公司公开演示真正的测序实例，因此本章着重介绍前两类的测序方法。

4.实时单分子测序技术

第三代测序技术中，以合成为基础的单分子测序是指通过固定合成酶，限定DNA的合成反应过程，并实时地记录下每个延伸碱基的信号。为了实现测序的实时性，即DNA合成时核苷酸的掺入过程不间歇，荧光标记核苷酸采用了不同的策略。与Illumina/Solexa和helicos BioSciences不同，第三代测序技术中的荧光基团与核苷酸的磷酸键相连，而非与碱基相连。其原理是，合成酶在掺入碱基时切断磷酸键而释放出荧光基因，留下未修饰的DNA片段可继续延伸下一个碱基。因

此，第三代的合成测序技术摆脱了繁琐的冲洗与扫描的过程，加快了测序反应。

实时单分子测序的模板固定和核苷酸修饰。

2011年，pacific Biosciences公司正式推出了其开发的第三代DNA测序仪pacBio RS。pacific Biosciences将自己发明的测序技术命名为单分子实时测序（Single□molecule real□time sequencing，SMRT），这是首个实现直接观测单个聚合酶合成DNA互补链为测序基因的第三代方法。如上所述，通过磷酸键修饰荧光核苷酸已经摆脱了冲洗与扫描的枷锁，打下了实时性的基因。

实时观察DNA聚合酶的另一个挑战是如何能在DNA合成期间检测到单个核苷酸的掺入。因为在显微成像实时记录DNA链上荧光的时候，反应体系内荧光标记的核苷酸形成了非常强大的荧光背景。这种荧光背景使单分子的荧光探测成为不可能。pacific Biosciences测序方法的创始人从微波炉上得到了灵感。微波炉门上的金属筛布满了小洞，它们比微波的波长要小得多，因此这些洞能阻止微波通过并穿透玻璃。但是，波长更小的可见光就能够通过，让我们能够看到正在加热的食物。SMRT技术应用了相同的原理，不过规格就缩小至纳米级。

pacific Biosciences的“金属筛”称为ZMW（Zero□mode waveguide，零模式波导）。ZMW是一个直径为几十纳米的小孔，它阻止可见的激光（波长大约为600nm）完全透过ZMW。从底部射入的激光在进入ZMW后迅速衰减，因此只有下面的30nm空间被照亮成为检测敏感区。在每个ZMW中，单个DNA聚合酶分子利用专利技术锚定在底部玻璃的表面。随后核苷酸涌入ZMW中，并在阵列表面扩散。当聚合酶检测到正确的核苷酸时，便将其掺入新生链中，这个过程需要几毫秒，而单纯的扩散只需要几微秒。这种时间差使掺入的核苷酸产生了很高的信号强度，类似于脉冲信号。因此，ZMW有能力在荧光标记核苷酸的背景下检测单个掺入事件。

pacific Biosciences实时单分子测序原理。

由于pacBio RS实现了DNA聚合过程的实时检测，所以研究人员能在短短的几分钟内对长片段DNA进行测序。从样本制备到测序，所需的时间还不到一天。典型的测序运行时间低至30min。这在之前是无法想象的。此外，pacBio RS目前的平均读长可达3kb，最长可达10kb，比

第二代测序及Sanger测序都要长得多。试剂消耗和样本制备也极少，不需要常规的pCR扩增，失误也大大减少。由于其单分子测序的实时性，pacBio RS也能直接用于分析DNA模板的甲基化。

5.纳米孔测序技术

Oxford Nanopore Technologies公司正在研究的纳米孔单分子技术是一种基于电信号的测序技术。其优势在于它不需要对DNA进行标记，也就省去了昂贵的荧光试剂和CCD照相机。纳米孔测序预计能满足大部分测序用户的需求，有99.8%的准确率，且错误很容易通过计算来纠正。产物的延伸也没有问题，因为纳米孔记录每一个碱基，而不管其前后的碱基，读长也会很长。Bayley研究组认为纳米孔技术可以读取数千个碱基，序列质量也不会下降。即使中途有一些小差错，它也可以重新开始。

纳米孔测序技术从1995开始发展（Kasianowicz等，1996），意在确定DNA链中的核苷酸顺序。纳米孔是一个极其微小的洞，直径为 1 nm，采用某些细胞跨膜蛋白，如Oxford Nanopore Technologies公司以α□溶血素（Stoddart等，2010）来设计纳米孔，并将环式糊精共价结合在孔的内侧。当核酸外切酶消化单链DNA后，单个碱基落入孔中，瞬间与环式糊精相互作用，并阻碍电流穿过小孔。纳米孔测序原理就在于纳米孔导电性的变化，即由离子通过纳米孔的导电作用可以观察到有轻微的电流产生，而电流量与纳米孔的大小及形状有关。单一核苷酸或DNA链穿过纳米孔时，会对纳米孔产生特有的变化。A、T、G和C每个碱基以及甲基胞嘧啶都有自己特有的电流振幅，因此很容易转化成DNA序列。每个碱基也有特有的平均停留时间，它的解离速率常数是电压依赖的，+180mV的电位能确保碱基从孔的另一侧离开。

纳米孔单分子技术的另一大特点是能够直接读取甲基化的胞嘧啶，而不像传统方法那样必须要用重亚硫酸盐（bisulfite）处理，纳米孔单分子技术的准确率能达到99.8%，而且一旦发现替换错误也能较容易地更改，因为4种碱基的电信号差异很明显，因此只需在与检测到的信号相符的2种碱基中做出判断，就可修正错误。

利用电泳原理及纳米孔上的酶催化反应可以使DNA的核苷酸逐个地穿过纳米孔。这是Oxford Nanopore测序仪仍面临两个重要的技术问题之一，也就是：如何将核酸外切酶更好地附着在孔上，让它每次只掉入一个碱基，这是一个巨大的挑战。

纳米孔技术提供了一个无标记的、导电的单分子DNA 测序技术平台，无需pCR扩增或化学标记的步骤，因此无需识别化学信号。目前，正在开发的纳米孔技术可以简单归纳为以下几种。

（1）核酸外切酶技术。目前，Oxford Nanopore研究用核酸外切酶技术来进行测序。此技术用核酸外切酶从DNA链上切下单个碱基，并使后者通过蛋白纳米孔。

但核酸外切酶技术的困难在于其需要使酶按正确的顺序切下每个碱基，并要使碱基进入纳米孔，整合了核酸外切酶和纳米孔检测系统；尤其当核酸外切酶水解DNA核苷酸间的磷酸二酯键时，能否保证随后释放的碱基直接进入附近的α□溶血素纳米孔。因此，这种方法是将核酸外切酶附着于纳米孔上，可以通过生物素连接到β、γ□溶血素。可以让蛋白的中心核布满带电的物质，如此正负电荷就出现在孔的相反两面。但其不能形成一个将核苷酸导向特定轨迹的机制。

（2）DNA链直接测序技术：这是Oxford Nanopore目前研究的另一种技术，其使DNA链经过蛋白纳米孔，读取该DNA链上的单个核苷酸序列。

（3）固态测序技术：采用合成材料而不是蛋白做成纳米孔。

Roche和IBM也在合作开发以纳米孔为基础的测序技术平台。该平台将大大降低测序成本，人类基因组测序的费用降到每基因组100～1000美元。该平台将以IBM的DNA晶体管为技术，能在DNA链通过硅片的纳米孔时读取DNA链上的单个分子。模拟图表示带有电极的单链DNA移动穿过硅片中间，由IBM Blue Gene supercomputer发明。褐色与黄色带分别表示为金属盒绝缘层。

第11章 总结

从Sanger测序技术诞生至今，测序技术不断推陈出新，形成了现在三代同堂的鼎盛局面。新兴的测序技术一直在追求更高的通量、更长的读长、更高的准备性、更快的分析以及更低廉的价格。虽然第三代测序技术已初露峥嵘，但是第二代测序技术在一定时期内的高通量、自动化方面的技术优势，使测序通量大大提高，从而降低了单个碱基的平均成本，并由此占据了应用的市场。第二代测序技术平台已应用在众多领域，并积累了丰厚的经验，如全基因组的全新（De novo）测序或重测序（Reseqeuencing）、染色质免疫共沉淀测序（Chip□Seq）、RNA全测序（RNA□Seq）、基因组甲基化测序（methyl□Seq）等。本书将在下一章中详细介绍第二代测序在各个领域的应用。虽然第三代测序技术和第二、三代测序间的过渡平台，展现了单分子测序、大读长、短时间等不同优势，但由于其发展时间短，故本书暂不系统地介绍这些技术平台的实际应用。

读累了记得休息一会哦~

公众号：古德猫宁李

- 电子书搜索下载
- 书单分享
- 书友学习交流

网站：沉金书屋 https://www.chenjin5.com

- 电子书搜索下载
- 电子书打包资源分享
- 学习资源分享

第12章 新兴高通量测序技术

在系统生物学中的应用从早期Frederick Sanger 的手工测序，以及基于Sanger 法开发的第一代自动化测序仪，到目前的新兴高通量测序平台，这一领域发生了巨大的变化，测序费用也大幅度降低（Mardis等，2008；Shendure等，2008；Ansorge 等，2009），使人类向1000美元测定一个人类基因组的目标迈出了一大步。近几年市面上出现了很多新兴高通量测序仪产品，例如美国Roche Applied Science公司的454基因组测序仪、美国Illumina公司和英国Solexa technology公司合作开发的Illumina测序仪、美国Applied Biosystems公司的SOLiD测序仪、Dover/harvard公司的polonator测序仪以及美国helicos公司的heliScope单分子测序仪。

新兴高通量测序技术发展之迅猛、速度之快令人瞩目，无疑为系统生物学的研究提供了新的手段、平台和应用。与其他学科技术相比，新兴高通量测序技术是第一个真正系统地、全面地研究一个系统的方法，已成功应用于众多领域并得到广泛认可。本章将着重介绍新兴高通量测序技术在系统生物学中的应用，尤其在癌症研究领域的应用，并详细论述了最新的文献内容与相关信息，读者如需进一步了解和掌握新兴高通量测序技术的新方法、新应用与新研究，可同时登陆相应公司的网站和有关新兴高通量测序技术的网站和博客。

1.新兴高通量测序技术在单核苷酸多态性鉴定中的应用单核苷酸多态性（Single nucleotide polymorphisms，SNp），即指基因组DNA中某一特定核苷酸位置上发生转换、颠换、插入或缺失等变化，而导致的核酸序列的多态性。依据排列组合原理，SNp一共可以有6种替换情况，即A/G、A/T、A/C、C/G、C/T 和G/T，SNp在CG序列上出现最为频繁，而且多是C→T，因为CG中C（即胞嘧啶）常为甲基化的，自发脱氨后即变为胸腺嘧啶。SNp具有数量多、分布广和遗传稳定等特点。SNp分析对于群体遗传学、疾病相关基因的研究、新药研究、临床检验和分子诊断等领域有着重要作用。

SNp的鉴定一般通过完整基因组重测序或特定目标区域的测序来完成。随着读序长度的增长、双端测序技术的发展、测序密度的提高以及多样本带标签混合测序技术的成熟，完整人类基因组重测序（包括癌症基因组重测序）的费用已不是很高。pleasance等（2010）利用

Illumina GAII基因组分析测序仪对来自同一个体的恶性黑素瘤和淋巴细胞系进行基因组测序，得到了40倍以上深度的测序数据，获得了第一份个体癌症组织突变概况。他们总共鉴定出33345个位点的碱基替换，其中32325个位点是单碱基替换，510个位点为双碱基替换（即连续两个碱基同时突变）。

另外，还可以有针对性地对癌症基因组中特定区域（如外显子或感兴趣的基因区域）进行测序，该方法可以大大提高目标区域的测序深度。Yeager等（2008）利用Roche公司454测序仪在欧洲人的39例前列腺癌和40例对照组织中对一个136kb的基因区域（8号染色体：128473000-128609802）进行重测序，该区域为胸腺癌、大肠癌和前列腺癌的易感区域。他们鉴定了这一区域中的常规SNp［主要等位点频率，MAF（major allele frequency），>1%］，其中442个为全新SNp，这些SNp的发现可以帮助精确诊断前列腺癌，并且分析突变引起的功能影响。Sugarbaker等（2008）则应用454测序平台对4例恶性肋膜间皮瘤组织、1例肺腺瘤组织和1例正常肺组织的cDNA进行测序，每例组织平均得到266Mb的数据，并对恶性肋膜间皮瘤特有突变及RNA表达差异进行分析，得出的结果也验证了已报道的表达差异RNA。另外他们还对点突变位点进行了筛选，设定的筛选标准为：该位点必须至少被4条读序覆盖并来源于正负双链，同时没有出现在已有数据库、对照组织或正常组织中。基于此标准，他们在4例肿瘤组织中鉴定出15个全新非同义突变，其中7个为点突变，3个为点缺失，4个为特有表观抑制突变，还有1个突变属于理论上推测的RNA编辑（RNA editing）现象。其7个点突变中有3个位点在其他49例恶性肋膜间皮瘤病人中，至少有1例得到了验证。同时，这些点突变所在的基因与癌症也是相关的，包括XRCC6、pDZK1Ip1、ACTR1A和AVEN。

最近，我们报道了通过长片断pCR扩增和NimbleGen序列捕获对一组关键基因的完整区域进行富集的技术，这些区域包括启动子、外显子、内含子和下游区域，并利用新兴高通量测序仪鉴定SNp位点（Cheng等，2010）。以123kb的ApC基因组区域为例，我们在27对直肠癌和癌旁组织中进行捕获富集测序，利用针对新兴高通量测序数据的分析软件MAQ和SAMtools，鉴定出210个高可信度SNp，其中69个是全新SNp。其中在外显子区域的SNp有11个，1个是全新SNp，另10个SNp虽是同义突变，但可能影响外显子增强因子或外显子抑制因子结合位点，并进一步影响转录剪切形式。ApC基因是进行测序研究最多的基因之一，但应用我们的上述方法仍可鉴定出很多全新SNp，这有力说

明了新兴高通量测序仪的优越性。通过等位基因特异pCR，我们验证了8个随机挑选的全新SNp，进一步验证了这些结果的可靠性（Cheng等，2010）。

相比于外显子测序，大片段目标基因组序列捕获技术可以用来鉴定影响剪切形式的内含子区域SNp，以及启动子和增强子区域SNp。人类基因往往被多种顺式作用元件所调控，从而在正确时间位置转录出正确剪切形式的mRNA。这些作用元件包括内含子增强子（ISE）、外显子增强子（ESE）（Cartegni等，2003）、内含抑制子（ISS）和外显子抑制子（ESS）（Wang等，2004）。外显子增强子（ESE）是一种能指引和促进RNA前体进行正常剪切的顺式作用元件，能结合SR蛋白和固定剪切所需的蛋白；而外显子抑制子（ESS）则通过结合hnRNp蛋白组合物来抑制邻近剪切位点及外显子的使用。SNp能够影响这些作用元件的结合位点，进而影响基因转录剪切，而基因转录剪切形式的变异是导致许多疾病的原因之一（Cartegni等，2002；Wang，Cooper，2007）。我们还在ApC基因上游区域鉴定出7个SNp，其中3个为癌症组织特有，6个经预测能影响转录因子结合位点（Cheng等，2010）。例如，位于5号染色体112070456nt的SNp，由G变为A（参考序列为人类基因组hG19），从而丢失了c□Ets□1［T00112］和R2［T00712］两个转录因子结合位点；位于5号染色体112064475nt的SNp，由A变为T，丢失了pRB［T00696］和pRA［T01661］两个转录因子结合位点，但获得了c□Myb［T00137］的结合位点；位于5号染色体112063970nt的SNp，由A变为T，获得了WT1I□KTS［T00900］、WT1□KTS［T01839］和ETF［T00270］三个转录因子结合位点。Bond等（2004）研究也发现一个在人类MDM2基因启动子区域的SNp，能削弱p53肿瘤抑制通路从而加速肿瘤形成。当然，我们新发现的ApC启动子区域SNp是否在大肠癌的发生和生长过程中起作用，还需要进一步实验研究。

由此可见，相比于全基因组测序，目标序列捕获技术结合新兴高通量测序仪的实验方法可以大大增加测序深度，而且可以用来鉴定罕见的等位基因。然而用商业化技术，诸如Nimblegen和Agilent的芯片捕获技术来获取目标序列也可能会丢失部分序列。我们测试使用Nimblegen芯片捕获30个基因的完整基因组区域并测序，得到覆盖率大约为60%～95%。同时，还以ApC基因为例，比较了长片断pCR扩增技术和序列捕获技术的差异。当只取唯一匹配的测序读数来统计ApC基因区域的测序深度时，发现两种技术所得到的平均测序深度几乎一样。但在高

GC含量区域中，基于pCR富集技术的测序效果比基于Nimblegen捕获技术的测序效果更好。基于pCR富集技术的测序深度大大高于基于Nimblegen捕获技术的测序深度。同时，低测序深度的区域往往是重复区域或低复杂度序列。

ApC基因区域的测序深度覆盖统计。红线表示相应位置的测序深度，绿线表示GC含量，以100bp为窗口。蓝色方块为低复杂度区域，由Cross match软件定义。pCR_C：癌症组织的长片断pCR扩增；pCR_N：癌旁组织的长片断pCR扩增；NG_C：癌症组织的NimbleGen序列捕获；pCR_N：癌旁组织的NimbleGen序列捕获

在新一代测序技术的支持下，科学家们也开始致力于区分驱动突变（Driver mutation）和随从突变（passenger mutation）（Ley等，2008；Mardis等，2009）。但这不是一项简单的研究，目前关于驱动突变和随从突变的界定尚不清晰。重复出现的突变往往被认为是可能的驱动突变。Shah等（2009）报道了对粒层细胞肿瘤（GCTs）RNA序列多性样分析的研究，粒层细胞肿瘤是最常见的一种恶性卵巢性索间质肿瘤（SCST）。对最初3例患者肿瘤组织RNA测序发现，在FOXL2基因存在一处错义点突变，402位置C变成G（氨基酸改变，C134W），并且在89例成人粒层细胞肿瘤中验证到86例（97%），在14例泡膜细胞瘤中验证到3例（21%），在10例青少年粒层细胞肿瘤中验证到1例，但同时在49例其他种类恶性卵巢性索间质肿瘤和329例不相关卵巢肿瘤和乳腺肿瘤中却没有得到验证。由此，有力说明这例突变是粒层细胞肿瘤的驱动突变。

4.2新兴高通量测序技术在检测基因组结构改变中的应用肿瘤基因组常经历体细胞染色体改变：一类为DNA拷贝数变化（即通过获得或者失去DNA的方式改变基因的剂量），包括插入和缺失，这就是通常所指的拷贝数变化（Copy number variations，CNV）；另一类则不影响DNA的拷贝数，包括倒位、基因融合（即染色体易位）。与基因组内其他的变异形式（如SNp点变异）不同，结构变异难以通过二代测序得到的某一小片段DNA序列读码来找到，要通过基因组DNA序列对比分析来完成，因此结构变异比SNp更难检测。对它们的研究是随着2001年人类基因组测序的完成、基因芯片技术的应用和新一代DNA测序技术的出现而兴起的。这些技术能够认证用常规细胞遗传学技术难以分析的亚显微结构变异。

检测单拷贝数获取和缺失的理论覆盖率。（a）图示利用序列检测拷贝数变化的全过程。（b）和（c）检测单个拷贝数获取和缺失的能力计算。我们将窗口中L的长度数值固定为10~100kb。线条表示基于正常分布的随机变化的比例的分布所产生的近似能力。例如L=30kb时，我们将泊松分布的随机改变（蓝色点所示）同能力的百分比进行作图，最后得到一条曲线。曲线的近似值精确到10%以内（见由蓝点组成的曲线），经Nature publishing Group许可复制

随着生物信息学的兴起，新兴高通量测序技术极大促进了基因组结构变异的研究，研究者们通过计算机对不同来源基因组DNA序列的比较分析，发现大量的结构变异。Chiang等（Campbell 等，2008；Chiang等，2009）对采用新兴高通量测序技术来鉴定癌症拷贝数变化的研究进行了详细分析，假设人的基因组长度为A=2.2×109，总的可比对上的序列为1000万条，读长为36个碱基，那么在一段50kb的区域中能比上的序列数为：50000×107/A=～230.此结果表示在上述条件下，如果在该区域中约有230条序列读数，那么就表示该区域有2个拷贝（因为人是2倍体）。依此类推，约115个读数代表一个拷贝；约345个读数代表3个拷贝（Chiang等，2009）。他们随后计算了检测单个拷贝变化所需要的理论覆盖率，结果表明：为了检测一个50kb区域内所需要的单个拷贝数，需要1500万条比对上的序列；如果要分析单个拷贝数的缺失，则需要约600万条比对上的序列。

他们利用新兴高通量测序和昂飞的SNp芯片分析技术（Affymetrix SNp Array 6.0），分别检测了三组配对的正常和肿瘤细胞系的CNV情况，并比对了两种技术的差异。结果表明，两种方法在证实已经存在的拷贝数变化时展示出了高度的一致性，但新兴高通量测序技术在定位结构的断裂点方面要优于SNp芯片分析技术。同时，他们的结果表明，在估计拷贝数变化方面，新兴高通量测序技术取得了更高的动态分布范围，而芯片分析技术由于杂交探针饱和等原因而次于新兴高通量测序技术（Chiang等，2009）。

利用新兴高通量测序技术，Wood和Schweiger等分别提出了从福尔马林固定的石蜡包埋样本中提取纳克级的DNA量进行高分辨率多样品分析拷贝数变化的方法（Schweiger等，2009；Wood等，2010）。新兴高通量测序得到的高通量数据与基因组杂交芯片（Array□based comparative genomic hybridization，aCGh）的结果进行比较，得到了

一个很高的皮尔逊相关系数0.9362277（Schweiger等，2009），该数值表明新兴高通量测序技术能替代aCGh进行CNV的研究。

新兴高通量测序技术与aCGh方法检测拷贝数分析的实例比较。（a）利用新兴高通量测序分析3号染色体的结果；（b）利用新兴高通量测序分析5号染色体的结果；（c）利用aCGh分析3号染色体的结果；（d）利用aCGh分析5号染色体的结果。

随着末端配对测序等技术的发展，新兴高通量测序技术可进一步应用于肿瘤基因组中进行更精细的结构改变分析。在末端配对测序中，产生两个配对的序列并且比对到参考基因组上，如果实际配对间的距离（即所代表的长度）与预期的长度有很大的差异，或者方向反常，则暗示着结构的改变。这种末端配对比对（paired□end mapping）技术已被成功用来揭示结构改变，并得到了优于芯片技术的高分辨率结果（Medvedev等，2009）。

pEM技术识别标志说明。相互配对的序列从供体样品中得到，在供体样品中配对序列相向排序（蓝色的接着橙色的并配对），而且配对序列都比对到参考序列上。基本的识别标志包括（a）插入（Insertion）和（b）缺失（Deletion），供体配对序列与参考基因组（Ref）中的配对序列的距离不同。（c）倒置（Inversion）：即两个配对的序列顺序与参考基因组（Ref）中的序列相同，但其中一个改变了方向。（d）连接（Linking）：几个不协调的末端配对，它们有相似的距离，但在供体序列上他们之间的距离（橙色虚线箭头所示）比在参考序列中小。配对末端序列的方向和顺序依赖于参考基因组上的两个片段的方向和顺序；它们都没有改变。（e）连接插入（Linked insertion）：由两个连接的标志构成，当被插入的序列从基因组另一个位置拷贝而来时，连接插入识别标志得以形成。（f）倒转复制（Everted duplication）：一个串联顺序的复制将产生一个倒转外翻的复制连接识别标志，其中配对序列的顺序颠倒了但仍保持合适的方向。匹配的末端序列将复制区域的末端串联到它的开始处。（g）锚定解切配对（缺失）［Anchored split mapping （deletion）］和（h）锚定解切配对（插入）［Anchored split mapping （insertion）］，在锚定解切配对中，配对序列中的一条完好地匹配上，但剩下的一条是断裂匹配。比如（g）锚定解切配对（缺失），断裂匹配的那条序列的前、后部分都围绕在缺失的地方；但对（h）锚定解切配对（插入）而言，断裂匹配的那条序列的前、后部分都比对到邻近位置，但其中间部分比对不

上。（i）悬挂插入（hanging insertion）：当一个全新的基因组片段插入后，一个悬挂插入识别标志可以被创造出来，在其中仅有一条读取序列完全配对。来自Medvedev 等的图一（Medvedev 等，2009），经Nature publishing Group许可复制Clark等使用U87MG细胞系（一种脑瘤细胞），采用双端测序法得到了平均插入长度为1.4kb、读取序列长度为50个碱基的匹配末端，超过30倍的基因组覆盖率的数据。他们生成了总数为1014984286的配对末端读数，证实了35处染色质间的易位事件，1315个结构改变（长度>100bp），191743个小的（21bp）插入和缺失（Indel），以及2384470个单核苷酸改变（SNV）。而且，他们的结果表明512个基因是纯合子突变，其中包括154个单核苷酸改变，178个小的缺失，145个大的微缺失，以及35个染色质间的易。Campbell等利用大规模平行测序，将两个肺癌患者的基因组打断成短DNA片段，产生了匹配末端的读取序列（Campbell等，2008）。他们研究了那些没有相互正确地匹配到对应的人类参考基因组的配对序列，分析了单个碱基配对的分辨水平，从而发现306个生殖细胞结构突变以及103个体细胞重组（Campbell等，2008）。

这些研究说明末端配对测序在检测结构改变包括小的缺失、微缺失以及染色体易位等方面的进一步应用。

2.新兴高通量测序技术在识别基因融合中的应用

在各新兴高通量测序平台中，由于454测序平台的读长较长，使得它能很好地应用于识别基因融合（Gene fusion）。Zhao等利用454测序法，对乳腺癌细胞系hCC1954的cDNA进行测序，获得了120Mb的数据，包含平均长度245bp的510703条cDNA序列，发现了MRE11与NSD1的融合基因，以及其他7个新的基因融合（Zhao等，2009）。Maher等利用双端转录组测序（paired□end transcriptome sequencing），发现了一些高频率融合的基因（如BCR□ABL1，TMpRSS2□ERG）和一些低频率的基因融合。他们在4个常用细胞系中发现了12个原先没发现的基因融合，以及前列腺癌中新的ETS基因融合（Maher等，2009）。Berger等利用双端测序和高分辨率的基因拷贝数分析方法，发现了黑色素瘤11个由于基因重组产生的基因融合和12个全新的相邻基因融合（Readthrough transcript）。相邻基因CDK2和RAB5B的融合（在10例的黑色素瘤样本中有4例）。CDK2是一个蛋白激酶，在细胞周期中的G1/S期转变过程中起着重要的调控作用。新编码的融合基因中包含一个缺少22个氨基酸的CDK2蛋白（Berger等，2010）。

新的相邻基因融合。（a）CDK2□RAB5B融合；（b）CDK2在10例黑色素瘤中的表达水平。RpKM：每kb外显子在检测到的百万条序列中的拷贝数。

为了研究一些在功能上重要的基因融合，Wang等开发了一个算法，利用高通量测序数据，并整合分子相互作用、代谢途径、功能注释等信息，来评估与癌症基因分子机制相关的重要基因融合现象。通过分析肺癌的转录组测序数据和基因组数据信息，他们在h1792细胞系中发现了新R3hDM2□NFE2基因融合现象（Wang 等，2009）。此外他们还结合了芯片上的数据，提出癌症中基因融合现象的系统分析方法，并指出在寻找新基因融合时的关键因素。

除了在整个转录组的测序数据中能研究基因融合外，对特定序列的测序数据中也能发现基因融合现象。Levin等先通过基因芯片捕获了467个肿瘤相关基因的序列，然后进行高通量测序，结果显示该方法具有更高的灵敏度，更容易发现新的基因融合现象，同时还能提供相应基因的表达水平的数据（Levin等，2009）。

另一种是Chmielecki等发明的基于基因组DNA上目标序列的测序方法，系统地分析了人类90多个酪氨酸激酶的融合情况。他们先通过探针捕获GXGXXG所在的外显子，及其上游的两个外显子和三个内含子的序列，然后通过454测序。通过生物信息学分析，发现所有的酪氨酸激酶的融合都与一个保守的GXGXXG片段有关，而且80%的酪氨酸激酶的基因融合的位点在编码GXGXXG的外显子上游的3个内含子之内，这些发现使得他们能系统地分析酪氨酸激酶的重排。通过该方法他们第一次发现TpC□1细胞系中CCDC6□RET融合，以及KG□1细胞系中FGFR1Op2□FGFR1融合（Chmielecki等，2010）。

3.新兴高通量测序技术在肿瘤基因表达谱研究中的应用RNA末端标签测序与RNA测序已被广泛应用于癌症转录组表达分析中。RNA末端标签测序（End□tag sequencing，end□taq□Seq）技术可用于检测RNAs的3′端序列，该片段通常由限制性酶（例如NlaⅢ）切割靠近3′polyA端的序列得到，再用MmeI，切取靠近3′端的17bp的片段用于测序。

RNA末端标签测序技术类似于SAGE或者LongSAGE，只是后两种技术在标签形成后没有加接头这一步骤，而RNA末端标签测序技术是在片段两端加上接头，然后利用高通量测序技术来完成。我们为首批将RNA末端标签高通量测序技术应用于癌症转录组分析的研究组之一

（Lin等，2005；Grigoriadis等，2006；Cheng等，2010）。随着新兴高通量测序技术的发展，Morrissy等也采用RNA末端测序技术分析正常人类组织的转录组表型，他们指出end□Tag测序技术由于测序深度比LongSAGE更好，能检测到低拷贝的转录本。例如，他们利用两种方法分别检测相同的RNA，结果显示LongSAGE检测到7055个基因，而RNA末端测序检测到11165个基因，其中包括93.5%LongSAGE 检测到的基因。另一方面表现在检测转录因子（TFs）的能力上，因为这些因子通常是低表达的。LongSAGE检测到429个TFs，而RNA末端测序检测到799个TFs（Morrissy等，2009）。他们进一步比对了用RNA末端测序与Affymetrix芯片技术共同检测到的基因，在RNA末端测序（测序数据为1000万）检测到的表达活性是Affymetrix芯片的13倍。最后，他们指出RNA末端测序方法能检测反义链的短序（Antisense tags），还可以反映已知基因的未注释的外显子和UTRs。这些都说明RNA末端测序技术比LongSAGE或DNA芯片技术更具有优越性。

RNA末端测序技术另一个优点在于其检测反义转录本的能力。正义与反义基因由同一基因位点上相反的两条链编码，产生的转录本互补。若需研究基因组的两条链准确的转录表型，则RNA末端测序数据优于RNA测序（Morrissy等，2009）。因为利用polyA特性产生不同的3′端序列，RNA末端测序能够识别转录亚型。

RNA测序技术是将一个样本所有的RNAs或有polyA结构的RNAs随机片段化，反转录后添加接头，最后进行测序。

相对于芯片，RNA测序背景信号干扰更小，而且RNA测序检测基因表达的动态范围可以超过5个数量级，远高于芯片的几百倍（Mortazavi等，2008）。RNA测序更利于检测低表达基因或差异基因（Wang等，2009，Marioni等，2008）以及RNA亚型（Isoforms）（Sultan等，2008）。近来，Fu等也指出RNA测序得到的表达值比芯片数据更接近于蛋白表达水平。Bradford等比较了Applied Biosystems SOLiD平台与Affymetrix Exon 1.0ST芯片，发现RNA测序比外显子芯片更易于检测差异表达的外显子，显示了SOLiD 技术平台的极大的动态检测范围。

由于测序深度的优势，RNA测序同样能够用于发现基因中新型转录区域（未注释区域）。Mortazavi等分别对小鼠的脑、肺及表皮肌肉组织进行测序，将得到的数据与小鼠cDNA数据库进行比对，得到1.4亿条顺序，发现了已知基因的约17000候选区域（新外显子，新5′和3′延伸片段），以及596个新的候选转录本（Mortazavi等，2008）。同时

RNA测序还是检测转录过程中所发生变化的理想方法，包括基因剪接。因为多种癌症组织及细胞的存在，最终需要我们在单一细胞水平进行研究（Tang等，2010a；Tang等，2010b），所以RNA测序将发展成为癌症转录组研究的标准平台。

RNA测序还能用于SNp及染色体变异分析。Shah等（2009）利用RNA测序鉴别卵巢颗粒细胞瘤（Granulosa□cell tumors，GCTs）的FOXL2中周期性错义突变位点402CG（C134W）。GCTs是成年女性中最常见的卵巢性索间质恶性肿瘤（SCST）之一，FOXL2基因编码的转录因子对卵巢颗粒细胞的发展起着重要作用。研究证明，在89例成年卵巢颗粒细胞瘤（GCTs）中有86例（97%）发生FOXL2变异，14例卵巢泡膜细胞瘤中存在3例（21%），10例少年卵巢颗粒细胞瘤中有1例（10%）。而在49例其他类型的卵巢性索间质恶性肿瘤及无关联的卵巢癌或乳腺癌中，该变异并不存在（Shah等，2009）。他们通过对4例GCTs进行全转录组测序，发现FOXL2上单一或周期性的突变几乎存在于所有的成年卵巢颗粒细胞瘤中。突变体FOXL2是成年卵巢颗粒细胞瘤发病机制的潜在驱动力。

近来，Morin等利用新兴高通量测序技术鉴别经常发生的体细胞突变EZh2基因，发现EZh2在滤泡性淋巴瘤（Follicular lymphoma）中的突变率为7.2%，而在弥漫性大B细胞淋巴瘤（Diffuse large B cell lymphoma，DLBCL）中的突变率为21.7%。EZh2编码组蛋白转甲基酶，负责h3K27的三甲基（Morin等，2010）。他们指出该变异导致EZh2蛋白（Tyr641）中SET区域的单一酪氨酸被替换。

4.新兴高通量测序技术在全基因组范围转录因子DNA结合位点中的应用染色质免疫沉淀□测序技术（ChIp□Seq），作为取代染色质免疫共沉淀□芯片（ChIp□Chip）的新技术，能很好地在全基因组范围分析DNA结合位点。与ChIp□Chip一样，ChIp□Seq也需要染色体的免疫共沉淀，在这个过程中，通过甲醛将转录因子交联至基因组ＤＮＡ上，随后裂解细胞并超声打断基因组ＤＮＡ，至长度小于１kb。用特异性的抗体沉淀ＤＮＡ□蛋白复合物，然后去除ＤＮＡ和蛋白之间的交联，最后分离出ＤＮＡ片段用于测序分析。染色质免疫沉淀□测序技术是一种基于全基因组范围内高分辨率和无偏好性的DNA□蛋白结合分析方法。其最早用于全基因组范围体内CTCF、NRSF和STAT1结合位点的分析，现在染色质免疫沉淀□测序技术已成为研究DNA和转录因子结合的标准方法。Ji等的最近研究表明，在NRSF转录因子结合位

点的分析上，染色质免疫沉淀□测序技术比染色质免疫共沉淀□芯片技术具有更好的性能。比较这两种方法发现，染色质免疫沉淀□测序技术具有更高的分辨率，并且不依赖于预先定制的探针。ChIp□Seq相对于ChIp□Chip的优点

ChIp□SeqChIp□ChipChIp□Seq的优点材料用量低：最低可达10ng4μg低几百倍的用量可少做很多个Ip反应适用性广：任何已有基因组的物种窄：必须有芯片的物种不受芯片限制分辨率±50bp±500~1000bp位置比对可以提高一个数量级，更加精确敏感性有很强的伸缩性，敏感性随测序量增加而增加低：基于杂交和高低丰度序列的比例只要简单地增加测序量就能得到满意的检查敏感度交叉杂交没有：每条序列都被单独测序非常多即使在复杂的基因组中也能得到高质量的数据

我们还应用染色质免疫沉淀□测序技术研究了雄性激素受体（AR）对细胞生长抑制反应程序的改变（Lin等，2009）。雄性激素受体在很多雄性表型发育中起重要作用，并且在很多疾病中起一定作用，例如前列腺癌。雄性激素受体依据细胞的不同而表现为肿瘤抑制和促进作用。pC3细胞来源于四期前列腺癌骨转移细胞（Kaighn等，1979）。很多研究表明，表达雄性激素受体的pC3细胞表现生长抑制和转移能力的下降。我们通过染色质免疫沉淀□测序分析发现，在pC3细胞内有6629个雄性激素的基因组结合位点中，大约22.4%结合位点位于转录起始位点2kb范围内。应用CisGenome 软件的吉布斯采样算法（Gibbs Motif sampling algorithm）进行模体（Motif）分析，在我们鉴定到的三个新的AR结合模体有两个结合模体具有一致的结合序列：CGAGCTCTTC。27%的雄性激素结合位点包含这两个结合模体。而仅有2.9%的雄性激素结合位点包含TransFac数据库中的雄性激素结合矩阵（Matrix）：M00481，M00447和M00962序列。

三个新的雄性激素结合矩阵的一致性序列（Consensus sequence）。这些雄性激素结合矩阵分别定位在1448，1012，317雄性激素DNA结合区。矩阵的一致性序列标记是用网络标记软件构建的。来自我们发表的文章的图四4.6新兴高通量测序技术在肿瘤表观基因组学研究中的应用表观遗传学包括组蛋白亚基的翻译后修饰，DNA胞嘧啶的甲基化修饰和多种RNA介导的调控，它在肿瘤的发生中起着重要的作用。新兴高通量测序技术已经被广泛地应用于表观遗传学的研究。

第13章 新兴高通量测序技术在组蛋白修饰研究中的应用

在真核生物中，组蛋白结合DNA，组蛋白中的氨基酸残基因甲基化、乙酰化、泛素化、磷酸化等修饰形成组蛋白标记，含有这些特定标记位点的组蛋白可调控他们结合的DNA在特定的细胞类型或状态下的转录状况。组蛋白特定赖氨酸的甲基化，如h3K4、h3K36、h3K79、h3K9、h3K27和h4K20，能调控染色体结构和基因活性。这种甲基化可以表现为单一的、两个的或三个的，每种状态都表示不同的转录状态（Bernstein等，2006）。如：h3K4me3表示启动子在激活状态；h3K27me3表示启动子在非激活状态；h3K4me3与h3K27me3同时存在则表示启动子已经准备好被激活（Bernstein等，2006）；h3K9me3与异染色质和基因沉默有关（Franz等，2009）。

Barski等利用新兴高通量测序技术研究了全基因组范围组蛋白的20个赖氨酸和精氨酸的甲基化状态、变异组蛋白h2A。Z、RNA聚合酶II、转录阻抑物CTCF蛋白等的分布情况。他们在启动子、绝缘子、增强子和转录区域均发现了组蛋白的甲基化现象（Barski等，2007）。另外他们发现h3K4me1，h3K4me2，h3K4me3和h3K36me3（me1，me2，me3指代甲基化程度，一个、两个或三个甲基）与基因激活相关；h3K27me2和h3K27me3与基因沉默相关。图4.8显示了活跃基因（Active gene）区域（STAT1和STAT4）与非活跃基因（Inactive gene）区域（MOB1）的甲基化状态。h3K4甲基化通常意味着基因激活，而h3K27的甲基化通常意味着基因的沉默，RNA聚合酶II标志着转录的起始位点，CTCF标志着组蛋白甲基化区域的边界。

激活基因和非激活基因的甲基化图谱。（a）激活基因和非激活基因的甲基化图谱在UCSC基因组浏览器上典型的视图。（b）1000个激活基因和1000个非激活基因及上下游5kb序列的甲基化图谱。经Elsevier Limited许可复制Zhang等构建了组蛋白修饰数据库，用于存放和整合组蛋白修饰。组蛋白修饰可以用染色质免疫沉淀□测序（ChIp□Seq）、染色质免疫共沉淀□芯片（ChIp□Chip）和定量染色质免疫沉淀（qChIp）检测发现。目前，该数据库中存放了包括甲基化、乙酰化和泛素化的43个位置的组蛋白修饰。该数据库中可以用组

蛋白修饰、基因ID、功能分类、染色体位置或肿瘤名字来进行搜索，它还包含一个可视化工具hisModView。

需要强调的是，由于组蛋白标记可以跨几千个碱基对，所以转录因子染色质免疫沉淀□测序的算法不适合用来分析组蛋白标记。Zhang等开发了SICER软件（Spatial clustering approach for the identification of ChIp□enriched regions），用于检测这种长跨度的峰信号，分析组蛋白标记。

1.新兴高通量测序技术在DNA甲基化研究中的应用

植物和哺乳动物的基因组中，1%~6%胞嘧啶存在甲基化现象，而且这种甲基化在肿瘤中也起着重要的作用（Montero等，1992；Ballestar，Esteller，2008）。在人类基因组中大约有29000万CpG岛，CpG岛的甲基化状态，被称为甲基化组（Methylome）。利用新兴高通量测序技术，人们已经能在单碱基水平上研究DNA甲基化。用于甲基化组分析的方法有以下三种。①甲基化测序（Methyl□Seq），即基于限制性内切酶的方法，通过对非甲基化的片段进行酶切，然后剩下甲基化的片段，并进行测序（Brunner等，2009）。②甲基化免疫共沉淀□测序（Methylated DNA immunoprecipitation□Seq，MeDIp□Seq），通过针对于5□甲基胞嘧啶（5mC）的抗体，免疫沉淀甲基化的片段（Down等，2008），然后测序。Butcher和Beck进一步改进了甲基化免疫共沉淀□测序技术，发明了自动甲基化免疫共沉淀□测序技术（AutoMeDIp□Seq）（Butcher，Beck，2010）。③新兴高通量测序的亚硫酸氢钠测序技术（BS□Seq），即采用亚硫酸氢盐转化结合新兴高通量测序技术来研究DNA的甲基化组。该方法通过亚硫酸氢盐将未甲基化的胞嘧啶C转变为尿嘧啶U，测序结果中未甲基化的胞嘧啶仍为胞嘧啶，而甲基化的胞嘧啶变为胸腺嘧啶。

RRBS样品制备；（b）RRBS序列比对；（c）小鼠胚胎干细胞富集序列区域测序覆盖率。来自Smith等的图一（Smith等，2009），Copyright（2009），经Elsevier许可复制DNA甲基化，主要通过甲基化CpG结合域（MBD）蛋白重构的染色质重塑作用来行使功能，于是人们利用MBD富集分离与甲基化CpG岛，结合新兴高通量DNA测序，即甲基化CpG结合域□测序（MDB□Seq）。Serre等用该方法对三株同源但甲基化程度不同的细胞系进行了甲基化图谱的分析，成功地找到了已知的甲基化区域和几百个新的甲基化区域（Serre等，2010）。不同的甲基化分析技术各有利弊，MeDIp□Seq和MDB□Seq由于已先富集

目的序列，所以在费用上比BS□Seq要低。BS□Seq能在单一碱基水平测定整个基因组的甲基化图谱，但最主要问题是亚硫酸氢盐转化可能不完全，要是没有先富集目的区域，直接对整个基因组进行测序，费用也非常高。为了节约成本，Smith等开发了RRBS法（Reduced representation bisulfite sequencing），即通过酶切富集甲基化的CpG区域，然后用亚硫酸氢盐转化，再进行测序。在同样的测序深度下，RRBS法可以得到更高的CpG岛的覆盖率。

2.新兴高通量测序技术在其他方面的应用

新兴高通量测序技术在生物和医学研究中不断扩展，一些新的应用被不断提出。这些应用包括将新兴高通量测序应用于RNA的研究中，如识别RNA结合蛋白的靶标和识别小RNA（miRNA）的靶标。

CLIp□Seq（Crosslinking immunoprecipitation sequencing，交联免疫共沉淀测序）是通过抗体免疫共沉淀RNA结合蛋白，获得与RNA结合蛋白交联的RNA，并测序来研究RNA结合蛋白所结合的RNA。Stanford通过CLIp□Seq技术，获得了人的肾脏胚胎干细胞转录物的剪切因子SFRS1的23632个结合位点（Sanford等，2009）。他们发现了一个富含嘌呤的包含GAAGAA核心序列的10聚体，GAAGAA核心序列存在于大多数与SFRS1免疫共沉淀的外显子、内含子、非编码RNA和基因间序列中。同时他们发现SFRS1更倾向于结合外显子RNA片段中，离5′或3′剪切位点长度为21~40碱基的区域。通过与人类基因突变数据库（human gene mutation database）比对发现，与人类疾病相关联的许多基因，在与SFRS1相结合的区域存在突变。

CLIp（交联免疫共沉淀）技术采用紫外光254nm进行RNA与蛋白质的交联，效率较低，而且交联的位置难以发现，背景非交联的RNA片段含量高。hafner等开发了pAR□CLIp（photoactivatable□ribonucleoside□enhencedcrosslinking and immunoprecipitation，光敏核糖核苷酸增强型交联和免疫共沉淀）技术以克服CLIp的缺陷（hafner等，2010）。在pAR□CLIp中，光敏核糖核苷酸相似体（如4□硫尿苷和6□硫鸟苷）被加入到活细胞中，整合到活细胞新生的RNA转录物上。用365nm紫外光交联有很高的效率。pAR□CLIp□Seq的优点在于它能够精确找出交联中的cDNA的碱基突变。当使用4□SU时，交联序列中的胸腺嘧啶转变成胞嘧啶；当使用6□SG时，交联序列中的鸟苷转变成腺苷。交联序列中突变的存在可以将其从背景RNA中分离出来（hafner等，2010）。hafner等（2010）成

功 将 pAR□CLIp□Seq 用 于 RNA 结 合 蛋 白 （ 如 UM2 、 QKI 、 IGF2Bp1□3）所结合的mRNA的研究。

小RNA在癌症的研究中起着重要的作用，其作用机制是通过种子顺序（Seed sequence）5′端的6~8个碱基的核酸序列与靶标mRNA进行碱基互补配对（Ruvkun，2006；Tong，Nemunaitis，2008）。如Illumina Genome Analyzerllx提供一套全新独特的适用于各个物种的小RNA深度鉴定与定量分析的研究平台，通过大量的平行测序，可以发掘、鉴定并定量出全基因组水平的小RNA图谱，可以更深入地研究其功能。这种低成本快速测定数百万标签序列的新方法，被广泛应用于小RNA的表达谱研究中（hackenberg等，2009；Willenbrock等，2009），用来找出癌症中差异表达的小RNA（Wyman等，2009；Szczyrba等，2010；Weng等，2010），该技术平台为解密小RNA图谱提供了独一无二的方法。

第14章 未来测序技术面临的挑战

新兴高通量测序在癌症研究方面取得巨大突破，同时有很多需要解决的问题与面临的挑战，其中包括样本制备、不同测序平台的比较与应用、数据分析存储和生物信息学研究等。如下主要探讨在癌症研究中面临的两个主要问题的解决方法。

1.有效区分致癌基因的“驱动突变”和“随从突变”

为充分理解癌症的生物学机制，寻找癌症新的治疗靶点，在癌症样本基因测序的数据结果中，分析庞大的数据量可以发现上万个基因变化。因此，如何有效区分诱发癌症的驱动突变（Driver mutation）和癌症引起的基因组不稳定所造成的随从突变（passenger mutation），是测序研究中的一个重大挑战。COSMIC（Catalogue of somatic mutations in cancer，癌症体细胞突变目录）数据库记录了肿瘤发生过程中的体细胞突变位点，利用此数据库相关资源，可以比对分析已验证基因的突变位点（Bamford等，2004；Forbes等，2008；Forbes等，2010）。数据库记录有141212个突变位点，23907个单一位点突变。整个资源数据库总结了来自于10383篇文献的2760220个实验。同时，此数据库还包括了28例癌症样本的基因组全序列。

Carter等设计了一种新的计算方法——癌症特异性高通量体细胞突变注释算法（简称为ChASM），用于核酸突变筛查。ChASM可以验证并优先挑选改变基因功能和促进肿瘤细胞增长的错义突变，此算法具有高敏感性和特异性（Carter等，2009）。他们将此方法应用到胶质母细胞瘤研究中，鉴定出607个错义突变中有49个是驱动突变（Carter等，2009），频率达到8%，包括了胶质母细胞瘤中IDh1基因的驱动突变（Yan等，2009）。他们将此方法进一步应用到24例胰腺癌研究中，对963个体细胞错义突变进行鉴定筛查，发现了可能的驱动突变（错误发现率小于0.3），所含驱动突变基因包括已报道的三个胰腺癌驱动基因（p53，SMAD4，CDKN2A）和15个新的基因，包括编码激酶pIK3CG，DGKA，STK33，TTK 和pRKCG的基因，细胞周期相关基因NEK8，及细胞黏附相关基因CMAS，pCDhB2等（Carter等，2010）。

2.解决肿瘤异质性问题和充分利用福尔马林固定石蜡包埋组织样品肿瘤是一种复杂疾病。肿瘤临床样本不仅包括肿瘤不同发展阶段的不同分子表达谱的肿瘤细胞，还包括正常细胞（如成纤维细胞和浸润性淋巴细胞）。因此，有必要把肿瘤临床样本中的细胞分成不同的细胞群进行研究。然而，根据细胞的哪一种特征进行分群是我们面临的一个重大挑战。要全面理解肿瘤细胞的异质性，也许必须对单一细胞群进行分析。目前，某些基于单一细胞群测序的新兴高通量测序技术已经在应用，如单细胞转录组测序分析等（Tang等，2010a；Tang等，2010b）。

肿瘤的福尔马林固定石蜡包埋组织样本库，丰富易得，并且可以长期临床随访，因此福尔马林固定石蜡包埋样本的研究具有重要临床意义。Weng等对样本的冷冻组织和福尔马林固定石蜡包埋组织分别进行了miRNA深度测序，测序结果表明miRNA表达谱相似（Szczyrba等，2010）。Weng等应用基因芯片、深度测序和RT□pCR等方法进一步研究发现，与基因芯片所得结果相比，深度测序方法与RT□pCR方法显示了更好的相关性（Szczyrba等，2010）。Wood等应用高通量测序方法对纳克量的福尔马林固定石蜡包埋组织的DNA样本进行测序，高精度地多重分析了基因拷贝数变异（Wood等，2010）。

Beck等研究了基于高通量测序技术的3′端测序表达定量方法（3Seq），应用于肿瘤的福尔马林固定石蜡包埋组织的基因表达定量分析。他们采用基因芯片方法和3Seq方法，研究了23例软组织肿瘤的冷冻样本和福尔马林固定石蜡包埋样本的基因表达谱，软组织肿瘤分别为韧带样型纤维瘤和孤立性纤维性瘤（Beck等，2010）。测序结果表明，应用3Seq方法，在冷冻组织的韧带样型纤维瘤和孤立性纤维性瘤中可发现差异表达基因大约9600个，在福尔马林固定石蜡包埋组织的韧带样型纤维瘤和孤立性纤维性瘤中发现大约8100个差异表达基因；而用基因芯片技术仅发现了这两种肿瘤冷冻组织中大约4640个差异表达基因，在福尔马林固定石蜡包埋组织中仅发现69个差异表达基因（Beck等，2010）。比较结果可知，3Seq方法具有更大的优势，随着测序技术的日益更新和方法的不断改进，其将被更广泛地应用于癌症人群研究中。

总而言之，我们可以看到新兴高通量测序技术将在对癌症基因组和肿瘤生物学的全景式理解中起主要的推导作用。正常人类的遗传变异对鉴定和理解癌症相关的基因突变也非常重要，国际千人基因组计划于

2008年1月启动，目标是全球合作建立全面详尽的人类基因组遗传多态性图谱。近来，美国国家癌症研究院（NCI）宣布将扩展癌症基因组图谱研究计划，将研究的肿瘤类型增加到25种，同时将应用高通量技术研究几种儿童癌症疾病，每种儿童癌症样本例数至少为100例。国际癌症基因组联盟将协调完成50种主要癌症和（或）其亚型的大规模基因组研究，他们提议在基因组水平、表观基因组水平和转录组水平系统地研究25000例癌症样品，找出致癌突变的图谱和功能，确定临床病理亚型，用于预后和治疗方案的管理，并发展新的治疗方案。

第15章 质谱的基本原理

细胞是生命最基本的组成单元。而四大类化合物（即氨基酸、脂肪酸、核酸和糖）则是任何有生命的有机体不可或缺的组成部分（Alberts等，2002）。其中，脂肪酸形成的膜结构（主要由磷脂组成）界定了细胞的物理空间，并介导了细胞与环境的相互作用。遗传信息则储存在由核酸编码成的基因中，而基因的功能则是由占细胞干重一半以上的蛋白质来执行的。细胞从糖中获取能量，并继而以脂肪酸和多糖的形式储存起来。

组学技术的目的是无差别地分析和描述基因（基因组学（Genomics）/转录组学（Transcriptomics））、蛋白质（蛋白组学，proteomics）和代谢产物（代谢物组学，Metabolomics）。系统生物学则运用这些大规模组学的策略来描述基因、基因产物以及与环境相互作用的分子（即代谢产物、脂质、碳水化合物、激素等）之间的相互作用。由于生物体系中器官由多细胞体系构成，这使得生物体复杂程度达到了另一个层次，而利用代谢组学（Metabonomics）分析生物液体和组织可将高通量技术和组织学联系在一起（Nicholson等，2002）。了解这些组成部分是如何在活细胞（或其他生理学实体）中进行装配、交流和行使功能的，需要具备提出并验证假说的能力。分析化学提供了测量工具，而计算模型则为这些数据的整合以及从中获取有用信息提供了必要的框架（Ideker和Lauffenburger，2003）。系统生物学分析整合了实验数据和计算模型，它利用计算模型检测预期假设与实验观察的匹配程度，并最终帮助提出新的假说（Ideker等，2001）。

质谱分析（MS）作为分析化学的常用工具之一，能够对生物学体系中几乎所有分子的结构和数量进行无偏倚的整体分析（Yates，2004）。大规模质谱分析实验为系统生物学中从抽象（大规模）模型到详细（局部）模型任何一种计算模型提供输入数据（Ideker和Lauffenburger，2003）。质谱能够系统地分析系统生物学三个主要子域，即基因组学、蛋白组学和代谢组学（包含其分支脂质组学、糖组学等）中涉及的化合物。到目前为止，蛋白组学（分析和鉴定大量蛋白）是质谱分析技术中最为前沿的领域（Yates，2004）。

本章将简单介绍质谱的基本原理，重点阐明质谱在蛋白组学和脂质组学中的分析作用。质谱已经在蛋白组学中被广泛应用，且正在被引入

到脂质组学的研究领域内。目前，蛋白组学和脂质组学实验所应用的分析方法基本是附加于或基于质谱分析手段的。本章也将重点突出一些整合了蛋白组学大规模质谱分析数据以及从芯片研究得到的mRNA表达谱和脂质组学数据的研究。

1.质谱原理

质谱仪主要由离子源、质量分析器和检测器三个部分组成。这三个部分共同作用使质谱仪能够测量分子的一种基本物理化学特性——分子的气相离子的质荷（m/z）比。

2.离子化

生物样品导入质谱仪后，质谱仪首先将其离子化成挥发性的气相离子。为了能够可靠地并在一定程度上无偏倚地将生物分子离子化，质谱仪主要倚仗两种软电离技术，即电喷雾离子化（Electrospray ionization，ESI）和基质辅助激光诱导解吸离子化（Matrix□assisted laser desorption ionization，MALDI）。由于软电离技术的重要性，ESI和MALDI的发明者均获得了2002年的化学诺贝尔奖（Fenn，2003；Tanaka，2003）。

在ESI中，毛细管的顶端（正极）和质谱仪的进口端（负极）之间的高电场（2～5kV）能够将液相生物分析物转变成带电荷液滴，并随后蒸发形成气相离子；然后气相离子顺电势和气压梯度的降低被引导入质谱。ESI的一个重要优点是使分离技术和质量仪的连接变得简单化。

在MALDI中，生物分子先是与能量吸收剂（基质）共同结晶成混合物（Karas，hillenkamp，1988）。然后基质分子快速地吸收辐射能量（来自激光脉冲）促使基质□分析物混合物汽化，并使分析物离子化。

对于MALDI和ESI来说，离子化过程的效率是测试敏感度的重要决定因素，它决定了仪器对特定浓度分析物的反应强度。

多肽可以在酸性缓冲液中迅速形成离子。依据分子的极性，可以相应地使用正或负离子化（例如阴离子脂质、带有负电荷的磷酸化多肽等）；而非极性化合物可能要通过形成金属簇促进其离子化（例如三酰基甘油的锂加合物）（han，Gross，2003）。在对脂质的高通量分

析中，分析的敏感度依赖于脂质的浓度；脂质在较高浓度时容易形成聚合物，所以仪器线性应答曲线只能在脂质低浓度时获取（han，Gross，2003）。对于极端亲水的化合物，如碳水化合物，化学衍生法或许可以使其更好地被离子化（Zaia，2004）。

3.质谱仪分类

用于分析生物测定物的质量分析器有飞行时间分析器、四极杆分析器、四极离子阱分析器和离子回旋共振分析器等类型。

描述质量分析器性能的特征参数主要有如下三个（Yates，2004）：①分辨率，或生物分析物的m/z值与m/z峰宽的比值，即（m/z）/（Δm/z）；②质量精确度，即分子质量的测量值与实际相对分子质量的接近程度（用百万分率表示，10-6）；③扫描速度，即质谱仪获取数据的速度。

1.飞行时间（TOF）质量分析器

飞行时间质量分析器根据离子在一定距离的无电场区域内测定的飞行时间来计算该离子的m/z，离子化的分子首先从加速电压获得预定量的动能，此动能为Uz，其中U指加速电压，z指电荷。由此可知离子的运动速度是v=（2Uz /m）1/2，在U一定的情况下，由m/z决定其速度v，这里m指离子分子的质量。换而言之，在同一距离中，低m/z的离子将比高m/z的离子运动得快。

利用重复脉冲产生离子的离子源（例如，MALDI）是TOF质谱仪的天然选择。在理想状态下，加速电压将赋予每个m/z相同的离子群相同的起始速度。然而，汽化和离子化过程使得起始动能离子不均匀分配，从而导致相同m/z值的离子具有不同的起始速度。因此有相同m/z值的离子当以不同的起始速度被加速并穿过无场管道到达监测器时会有前后时间偏差（Δt）。此Δt时间偏差终导致峰宽Δm/z的形成，继而影响质谱的分辨率。而利用较长时间的飞行管，可以最大限度地降低初始速度的影响。另外可以通过电场反向镜将飞行路径逆转使得具有相同m/z但不同起始速度的离子聚焦，其原理是初始速度较低的离子在反向镜中被反射得更快，反射式飞行时间质谱仪的分辨率是10000，质量精确度为（10～20）×10-6.鉴于加速电压为+20～+30kV，TOF质量分析器能够在10-5秒检测m/z为500～4000的离子，而且不需要扫描。当然，检测速度同时依赖于生成离子所需的时间和其他因素。

基本由单一质量分析器构成的单极质谱仪示意图。（a）反射器飞行时间：m/z是由通过无场管的离子飞行时间计算生成的，电场反射镜补偿离子能量值的散布。（b）在四极杆滤质器中，由两对极性相反的电极杆形成包含离子运动的电场。通过改变RF和DC，四极能够选择性分析轻离子（低m/z）或重离子（高m/z）

2.四极杆质量分析器

四极杆质量分析器由两组对称的电极组成，相对的两个电极组成一对，其中一对带有射频电压（RF）和直流电压（DC），另一对杆有极性相反的电压。四极杆滤质器通过离子飞行的稳定性来选择特定质荷比的离子：通过变换RF幅度和DC电势，四极杆滤质器仅赋予特定m/z值的离子稳定的飞行轨道来通过滤质器并到达检测器，而其他m/z值离子则因为不具备稳定飞行轨道而被滤质器清除。通常，四极杆质量分析器的质量精确度为300×10-6.值得指出的是，在实际应用中四极杆滤质器常用单位分辨率而不是其真实分辨率。其中单位分辨率是指四极杆滤质器分辨一个分离单位的m/z信号的能力，即分辨两个相邻单位的m/z的能力。近年来，仪器方面的进步已经将四极杆的质量精确度提高到5×10-6，从而达到5000分辨率。

（c）四极杆离子阱将进入的离子锁定在质量分析器内，然后按照m/z值大小射入检测器。离子阱独自就能发挥串联质谱仪的功能，它能依次进行离子富集、破碎和碎片分析（这一过程被称作时间串联）。（d）在离子回旋共振阱中，被锁定的离子按照与质量电荷比值相匹配的特定频率旋转。

四极离子阱是由一对环形电极和两个呈双曲面形的端盖电极组成的，在环形电极上加射频电压或再加直流电压，上下两个端盖电极接地，离子阱的射频电场决定了离子在X、Y轴方向的运动，而端盖电极决定了离子在Z轴方向的运动。通过电场的变换，离子按m/z值递增的顺序从离子阱进入检测器，即利用离子飞行轨道的不稳定性来选择特定质荷比的离子（即保留下来的有稳定飞行轨道的离子m/z是不确定的）。这与四极杆滤质器正好相反，在离子阱中，除了m/z值与特定的挤出电压相匹配的离子被排出外，其他离子均被储存在离子阱中。离子阱的质量精确度和单位分辨率与四极杆滤质器相似，分别为300×10-6和2000.由于生成离子都被高效使用，故离子阱比四极杆滤质器更加灵敏。

4.离子回旋共振质谱仪

质量分析器也能利用静态高磁场来捕获离子。由于在高磁场中捕获的离子表现出回旋加速运动，带有这种分析器的质谱被称为离子回旋共振质谱仪（ICR），在ICR质谱仪中，离子在磁场中的旋转频率与其m/z值是成反比的。在ICR离子阱中，质量分析器与检测器合二为一。简单地说，离子在ICR中以某一特定的频率做回旋运动，并因此产生能被ICR离子阱中检测板（电极）检测到的电流（镜像电流）。由于离子回旋频率与m/z是成反比的，被记录下来的信号时间间隔被转换为频率（傅立叶转换，FT），最终计算得到m/z。傅立叶转换□离子回旋共振质谱（FT□ICR MS）具有高质量精确度（1~2）×10-6和高分辨率（>1×105）。

然而精确地分辨和测定离子的质荷比并不能用来描述生物分子内在的共价结构。现今质谱仪已经可以通过以下方法来逐步分析待测分子的相对分子质量和分子结构：①分子离子的质量测定；②碎裂反应；③分子离子碎片的质量测定。这便是利用串联质谱仪（MS/MS）分析生物分子结构的常用操作步骤。

将碰撞池置于两个四极杆滤质器中间的质谱仪便是串联质谱仪的一种，也被称作三级四极杆系统，待测母离子质量测定在第一个滤质器中完成，随后母离子在碰撞池中被惰性气体（如氩气和氦气）碰撞诱导解离（CAD）释放出碎片离子，而碎片离子则在第二个滤质器中被分析。三级四极杆质谱仪也能通过监测母离子和某一碎片离子的产物（多离子反应监测）对离子进行高度特异和精确的定量。三级四极杆质谱仪常用于代谢物组学的研究。

不同类型质量分析器构建的多级质谱示意图。（a）三级四极杆质谱（QQQ）。（b）四极杆□飞行时间质谱（Q□TOF）系统。

（c）三级四极杆线性离子阱质谱（QTrap）。（d）线性四极杆离子阱（LTQ）。（e）线性离子阱□傅立叶转换□离子回旋共振质谱（LTQ□FTMS）。与三级四极杆质谱类似，与三级四极杆质谱不同的是，Q□TOF第三部分用TOF质量分析器代替四极杆，从而提高了分辨率和质量精确度（Morris等，1997）。而在四极杆线性离子阱质谱，线性离子阱质量分析器替代了TOF质量分析器，其分子结构解析（MS/MS）是在线性离子阱质量分析器中进行的，这种质谱仪结合了

四极杆滤质器的高选择性和线性离子阱质量分析器的高灵敏度分析能力。

单独利用线性离子阱（LIT）也不失为一种简单、功能强大的串联质谱方案，线性离子阱通过端盖电极控制离子的进入，而当离子射出离子阱时，它们从离子阱两边射入检测器。这种设计的好处在于整个检测过程没有损失离子。线性离子阱的扫描速度相当快，每秒钟能产生5张谱图（Schwartz等，2002）。线性离子阱可独自按顺序完成MS/MS分析所有步骤，包括分离目标母离子，对其进行CAD（碰撞诱导解离）并随后测量解离反应产物。一般来说，离子阱类质量分析器能够多（n）级（MSn）进行该循环（分离，解离和质量分析）。图5.3三种不同蛋白酶消化的标准蛋白混合物的数据依赖性质谱分析。10min时间窗的LC□MS/MS分析。（a）由于质谱仪在记录全扫描MS谱图和三个连续的MS/MS谱图之间不断交替，色谱图呈现锯齿状。在保留时间为73.32min（在色谱实验中一分钟被平分为100份）时的全扫描MS谱被记录为2761谱图。（b）这张谱图从m/z 400到1400，测量了1000m/z单位范围内的离子。三个丰度最高的离子被选择进行进一步MS/MS分析。一旦某一个离子被选中用于MS/MS碎裂分析，它将在有限的时间内被列在一个排除清单上以避免重复分析。所以，在进行2761扫描之后，仪器将继续对没有列在排除清单上的离子进行MS/MS测序。因此，m/z 509.2，1050.9和701.3对应的碎裂图谱分别记录在2762（c），2763（d）和2764（e）扫描中。而从2765扫描开始，质谱仪将再次重复上述循环步骤采集数据。随后的Sequest数据库搜索鉴定到下面的肽段序列：2762为来自卵清蛋白的KIKVYLpR（两电荷），2763为来自白蛋白的LKECamCamDKpLLEKShCamIA（两电荷），2764为来自白蛋白的LKECamCamDKpLLEKShCamIA（三电荷）。白蛋白肽段中出现的半胱氨酸是被脲基甲基化的傅立叶转换质谱仪（FTMS）在质量测定的精确度和分辨率方面是无与伦比的，但是在进行多级质谱实验时它并不非常有效。然而，结合线性离子阱作为第一个质量分析仪的混合型FTMS仪器可以用来进行高扫描率（在线性离子阱中）的MS/MS实验，并保证测量的高精确度和分辨率。

质谱仪的运行由仪器控制系统指挥。比如，仪器控制系统监测质量分析器、检测器和一定真空条件下的离子源部分。除了维持仪器参数外，仪器控制系统软件也是设计不同质谱实验的界面。其中一个典型例子是数据依赖性谱图采集实验。图例中包括了仪器程序的详细信息。

目前，大多数大规模多肽分析实验是以电脑控制的数据依赖性谱图采集法为基础的，在这个过程中，质谱会对多个生物分子的m/z进行测定（全扫描质谱），随后对从全扫描质谱中得到的丰度最高离子进行串联质谱分析。当然，除数据依赖性谱图采集法之外，还有更多专门的分析方法（包括母离子扫描，产物离子扫描，中性丢失离子扫描）可用来描述生物分子的精细结构成分。例如，母离子扫描可被用来描述脂质的类别。

大规模质谱实验分析生物样品在仪器采样速度（扫描速度）和所收集数据的信息度之间是相互冲突的。总之，我们仅能在合适的灵敏度（见前章节）、适当的分辨率和质量精确度的条件下尽可能地从实验中提取信息。而且，在分析混合物时，多张解离图谱对于完整的鉴定是非常必要的。为全面描述一生物实体而设计实验、选择仪器时，应考虑上述因素是否会影响预期的结果。本章集中讲述了用大规模质谱分析中最关键的一个方面，即用生物分子的串联质谱分析或称生物分子MS/MS测序。

质谱的生物分析物测序

单独的质荷比测定并不能阐明离子尤其是肽段和蛋白的内在结构，所以对生物分子完整的物理化学定性还需要进一步分析。串联质谱首先自多个生物分子的混合物中逐个选择母离子，并使其碎裂而获得每个母离子对应含有的子离子的特定图像（模式）的MS/MS谱图。而测序图谱和适当的多级MS方法一起能够描述生物分子的结构和图谱改变。以下介绍肽段和脂质的MS/MS（测序）图谱的一些专业术语和突出特征。

1.多肽的MS/MS图谱。肽键碎裂产生两种主要离子，即b离子（N末端带电荷）和y离子（C末端带电荷）。通过多肽的MS/MS可以得到肽段的氨基酸序列。多肽的MS/MS图谱类似于柱形图，条形（碎片离子）通常表示相差一个氨基酸质量的肽段碎片。将数据库中所有肽段的理论图谱或之前采集的MS/MS图谱进行一一匹配，找到匹配的图谱就能得到相应的多肽鉴定结果。蛋白质的大规模分析包括搜索成千上万的MS/MS图谱。因此，数据搜索的结果应经过统计评估，为蛋白鉴定提供过滤参数。

人心肌肌钙蛋白的肽段。肽段的碎裂峰是一个双指纹图谱，肽序列能从氨基末端或者羧基末端读取。解释了组成肽序列的不同氨基酸序

列。图（a）总结了从MS/MS谱中得到的多肽质量梯度信息。（b）MS/MS谱中质量梯度之间的氨基酸简写表示相对分子质量与两碎片峰质量之差等同的某氨基酸残基。

2.多肽是氨基酸单元的聚合体。多肽在质谱中主要是沿肽段的主链碎裂，并产生一个质量梯度的双向指纹图谱，有时还含有从氨基酸支链上碎裂出来的某些化学基团。

3.脂质的MS/MS质谱。脂质MS/MS图谱是由脂质头首基的结构以及脂肪酸长链的XX:Y比例所决定的（XX是指脂肪酸中碳原子的总数，Y指碳碳双键的数目）。脂质的MS/MS质谱一般比多肽的质谱简单。一般只要利用在MS/MS质谱中尤为突出的脂质头首基特征峰以及恰当的扫描方式便足以描述脂质的类别。然而，用MS/MS鉴定脂质有时需要合成同型物来进行完整结构的鉴定。由于脂质代谢产物具有化学多样性，与合成的同型物的MS/MS质谱对比鉴定的方法也用于脂质代谢产物的分析。

阐明生物分析物的精细结构，尤其在比较分子结构差异时，也许会要求比MS/MS碎裂更多的方法。最近改进的质谱技术已经允许通过多级（即MS3和更高）碎裂和数据依赖方式检测中性丢失相结合的方法进行生物分析物结构的常规分析（见上面的质谱部分）。

第16章 生物分析物的分离

不管是分析蛋白质、多肽、核酸、脂质、碳水化合物、药物制剂，还是其他生物有机分子等，对生物分子的全面分析倚仗于生物分子的分离。做质谱之前有效分离生物分子是必需的步骤。分离技术的进步促进了质谱在生物分子分析中的应用，尤其是毛细管色谱与电喷雾质谱联用技术的出现，既降低了样品的流速，又提高了检测的灵敏度。复杂生物混合物的高分辨率分离手段对于降低离子化过程中的离子抑制作用和增加检测动态范围都是非常重要的。对于多肽类，可采用毛细管液相色谱仪（LC）在多肽抵达电离源之前进行混合物的简化。另外，对于相对分子质量较低的复合物（如代谢产物），如果是挥发性或者能够衍生为挥发性化合物的，那么利用毛细管气相色谱仪（GC）也能够对复杂混合物进行简化。气相色谱与质谱联用（GC□MS）是代谢组学研究的主要方法。不过在此，我们将主要介绍液相色谱在蛋白质组学中多肽大规模分析方面的应用。

蛋白质组学是一种生物学分析方法，主要涉及两种蛋白鉴定方法：一种是对通过鸟枪法分析获得的多肽序列信息进行重组的方法，另一种是通过直接测定蛋白相对分子质量，随后碎裂蛋白的方法。

在自下而上的蛋白质组学中，蛋白水解酶消化蛋白质（在溶液或凝胶中）得到的多肽由串联质谱仪测序。样品进入与质谱仪联用的毛细管柱上，并利用高压将体积为1mL的样品加到毛细管柱上。微柱/毛细管色谱法是目前蛋白质组学研究中常用的方法。有兴趣的读者可以参照相关技术指南来构建纳升级流量hpLC柱、微型ESI和毛细管上样。

多维正交色谱（针对多肽的差异化分离原理）可用来改进复合肽混合物的质谱分析，在多维正交色谱分离中，上游柱产生的每个馏分在下游柱中被进一步分离。最常见的组合方式是在上游设置离子交换柱，在下游则利用反相（分析）柱。在线多维色谱和串联质谱的联用则为鸟枪法测序研究整体蛋白组学提供了一个强有力的自动化平台与Sequest数据库搜索算法一起，多维色谱□串联质谱形成了多维蛋白鉴定技术（MudpIT）平台的核心。

一维色谱和高准确度的质量测定（借助于FTICR运行）也被应用于鸟枪法测序，这种方法能为多肽提供精确的质量标签。在接下来的分析

中，精确的质量标签和一维色谱中的保留时间一起可作为大规模的蛋白质组学实验中应用的多肽特异标志符。

针对一些有特殊物理化学性质的多肽亚群的分析则有一些专门设计的蛋白组学方法。这些方法不通过液质在线分析的形式，大多是采用多种亲和方法制备用以检测特殊物理化学特征的样品。最常见的例子是利用固定金属亲和色谱（IMAC）富集磷酸多肽以及利用凝集素色谱分离糖基化多肽。而通过一种对角线色谱分离法对比针对某个氨基酸特异性的化学衍生反应前后两次色谱分离，则可以用来研究含有这个特定氨基酸的多肽。这些方法大多只富集某一类多肽而浪费了其余的样品。如果将亲和技术和多维分离结合，最终要根据其母蛋白对经过特殊修饰的多肽进行整合分析。

液相色谱技术为大规模肽段分析提供了广泛多样的选择，但是对特殊类别的脂质却仅有极少的分析方法。Gross和他的同事利用脂质的电特性引进了离子源内分离方法。在这种方法中，ESI的极性（阳离子/阴离子模式）可被分别用来分析能被优先离子化为阳离子（中性种类）或阴离子（阴性种类）的脂质群体。类似的二维电喷雾质谱也应用于磷酸多肽的多维分析，即在正离子模式中被MS/MS测序的磷酸多肽，可以进一步采用阴离子电离对之前的测定进行补充。

依赖生物测定物的大小和电荷的比值来进行分离的毛细管电泳（CE）是蛋白质组研究中另一种可行的分离方法。Tong等（1999）将Rp和CE串联，用于分析核糖体蛋白。除应用于多维色谱外，由于CE提供了比GC更快的分离速度，可以将CE作为独立的分离手段用于代谢组学研究。

MudpIT平台由以下三部分组成：（i）样品准备；（ii）在线联机液相色谱□串联质谱；（iii）数据分析。（i）蛋白样品用化学消化剂CNBr或特定蛋白酶（Endo□lys C，胰蛋白酶）化学消化（Washburn等，2001）。此外，也可用非特异性蛋白酶和特异性蛋白酶（弹性蛋白酶、枯草杆菌蛋白酶、胰蛋白酶）相结合的方法消化样品并混合进行LC□MS/MS分析（MacCoss等，2002），或仅用非特异性蛋白酶（蛋白酶K）消化样。使用非特异性蛋白酶可以增加蛋白翻译后修饰区域的覆盖序列。（ii）样品随后进入毛细管柱，多肽选择性地从SCX和Rp洗脱下来，之后进入质谱仪，离子化后进行串联质谱分析。（iii）数据分析由一系列软件包完成，之后，肽段信息被集中后用于蛋白的鉴定（Sadygov等，2004）。基于研究目的和标准，使用对结果进行进

一步分析。最后利用Relex软件对稳定同位素标记的样品进行定量，以获取相对蛋白量信息（MacCoss等，2003；Venable等，2004），或者使用数据统计学模型（Liu等，2004）或根据实验参数，在没有稳定同位素标记的情况下凭经验计算蛋白的量无论是用四极杆或者TOF检测器，GC□MS都是分析代谢产物的一种流行的方法。GC的一个显著特点是能够预测生物分析物的相对保留时间。对于预先明确的化学结构，质量测定（m/z）和由GC□MS实验测定的相对保留时间足以确定该化合物。分离技术可以通过增加峰值容量即增加生物分子化合物的分离度来优化MS平台的分析能力。

第17章 定量蛋白质组学

定量蛋白质组学是在总体水平上研究生物体受到刺激（Stimulus）或干扰（perturbations）时蛋白表达水平的改变，是对mRNA水平改变的重要补充。基于质谱技术的肽段鉴定和定量的主要步骤。①蛋白被水解为肽段的混合物，经过高效液相色谱（hpLC）分离进入质谱仪。整个过程包含多个步骤，特别是肽段的离子化、质谱图的获取以及选择合适的母离子进一步打碎、离子化并得到二级质谱图。得到的数据可用于样品间的定量，或者通过与数据库的比对得到肽段的氨基酸序列。使液态样品形成带电液滴并最终离子化，而点在固相基质上的肽段则通过基质辅助激光解吸（Matrix□assisted laser desorption ionization，MALDI）离子化。②通过检测到的肽段离子（母离子，precursor ion）的质核比推测肽段的相对分子质量。③选中的母离子被分离出来。④打碎选中的母离子，常用的方法有碰撞诱导解离（CAD）。⑤碎片子离子（product ion）的质量在子离子谱图中被解析并记录，由离子碎片的质量推测得到肽段的序列。肽段的含量是由母离子的信号强度来判断的，常见的方法是比较目标肽段及用同位素标记的同样序列的参照肽段。除以上步骤外，质谱实验还包括样品准备（蛋白样品的制备、蛋白水解和肽段的分离）以及后续数据的处理与分析。

通常采用以下两种策略实现定量。①在蛋白水平的定量策略：将蛋白复合物进行二维凝胶电泳，比较蛋白在凝胶上染色的深度进行相对定量，再对差异蛋白进行质谱分析定性。②在肽段水平上的定量策略：此类方法通常需引入一种稳定同位素（13C、3h或15N）标记的基团，不同的蛋白混合物被酶（如胰酶）水解成肽段，来源于其中一种蛋白混合物的肽段被轻同位素（12C、2h或14N）基团标记，而另一来源的肽段被重同位素基团标记，通过同位素区分肽段在质谱中所形成的峰高和面积或者报告离子的峰高和面积，对不同样品间的相同蛋白质进行定量.

在定量方式上，定量蛋白质组学分为相对定量（Relative quantification）和绝对定量（Absolute quantification）两种。相对定量主要是确定两种、三种或更多种（如病理状态和正常状态，正常组织、肿瘤转移和非转移等）不同状态下蛋白的相对差异，多用于差异蛋白的鉴定（Identification）；绝对定量则通过已知量的特定多肽的掺

入来确定蛋白混合物中目的蛋白水平的绝对量，多用于差异蛋白的验证（Validation）。

按标记方法，定量蛋白质组学可分为两大类，即标记法和无标记的方法（Label free）。标记法可以分为体内标记法和体外标记法。前者如SILAC等，即利用细胞自身的代谢活动将同位素标记的氨基酸基团掺入蛋白；后者如ICAT、ITRAQ、AQUA、18O等。以下将分别详细介绍各种定量蛋白质组学的方法。

定量蛋白质组学研究方法。标记法需要在样品处理的过程中加入与蛋白共价结合的同位素标记，研究者可以在细胞培养或者蛋白质和肽段的处理过程中选择相应的试剂，根据同位素间相对分子质量的差值，混合后的样品可以被同时检测并计算相对浓度，大大减少了多次质谱检测引起的误差。

1.蛋白水平的定量策略

基于凝胶的实验方法（Gel□based workflow）有二维电泳（Two□dimensional sodium dodecyl polyacrylamide gel electrophoresis，2□DE□SDS pAGE）和差异显示凝胶电泳（Differential in□gel electrophoresis，DIGE）。

二维电泳（Two□dimensional electrophorese）是等电聚胶电泳和SDS□pAGE（Sodium dodecyl sulfate polyacrylamide gel electrophoresis）的组合，根据蛋白质两个独立的特性——等电点和相对分子质量，先按照蛋白的等电点（pI）进行等电聚胶电泳，再根据相对分子质量大小对相同等电点的蛋白进行SDS□pAGE分离后染色得到二维分布的蛋白质电泳图。二维电泳能够有效地分离一个复杂生物混合物中的蛋白质，分辨出成千上万个蛋白质。凝胶上的蛋白点可以通过放射性标记或各种染色方法染色而观察到。染色方法包括银染、考马斯亮蓝或荧光染色等。蛋白点之间染色的差异反映该点蛋白含量的差异，取差异蛋白点进行质谱鉴定以实现差异蛋白的定性和定量。目前，该技术发展成熟、使用较为广泛，实验步骤依次包括样品制备、样品标记、双向电泳分离、图像获取、图像分析、抠点、酶切、点靶和MALDI□TOF蛋白鉴定等。但在实际应用中该方法仍存在一定的缺陷，例如难以有效分离鉴定高相对分子质量、低相对分子质量、极酸性、极碱性和疏水性强的蛋白质。

二维凝胶电泳结果。在第一维凝胶中通过等电聚胶电泳分离蛋白，等电点相同的蛋白再进入第二维SDS□pAGE电泳，按照相对分子质量大小被分离开，电泳后的凝胶经考马斯亮蓝染色后便可以观察到蛋白点的分布和染色情况。染色的差异表明蛋白量的差异，借助图像分析软件可以分析蛋白的含量，并挑选感兴趣的蛋白点进一步做质谱鉴定。

荧光差异双向凝胶电泳在二维凝胶电泳的基础上引入了荧光染色的方法。不同样品经多重荧光分析（Cy2，Cy3和Cy5）标记，混合后加入二维凝胶电泳分离，在不同波长的激发下同一块凝胶上的蛋白点将被检测到带有不同颜色的荧光信号，借助于图形捕获和分析软件可以对不同样品中的同一蛋白点进行定量，克服了二维凝胶电泳重复性差的缺点，避免了胶与胶之间的差异对定量的干扰。DIGE还引入了内标，将实验中的样品等量混合后上样于同一凝胶中，电泳后每个蛋白点都有相应的内标，软件自动根据每个蛋白点的内标对其表达量进行校正，保证所检测到的蛋白丰度变化的真实性，极大地提高了结果的准确性、可靠性和重复性（Unlu等，1997）。

DIGE实验流程。来自动物正常左心室（Normal LV）、心室肥大（LVh）和心力衰竭（LVh/hF）组织的三组蛋白样品分别被标记上Cy2，Cy3和Cy5荧光染料后混合，并进行二维凝胶电泳，通过不同波长的激光激发得到对应三组样品的3张凝胶图片，运用DeCyer软件对图片中的蛋白点定量分析，抠下感兴趣的有差异的蛋白点，酶解后经串联质谱分析，结合蛋白数据库检索鉴定出其中含有的蛋白质。

2.肽段水平的定量策略

二维电泳难以分离高相对分子质量、低相对分子质量、极酸性、极碱性和疏水性强的蛋白，而液相分离的方法克服了此缺点。基于肽段水平的定量蛋白质组学的流程和原理。蛋白混合物先在溶液内或胶内被胰酶水解为肽段，肽段混合物进入液相色谱（离子交换液相色谱或反向液相色谱）被连续梯度分离后离子化进入质谱进行鉴定，再对质谱得到的数据进行生物信息分析，主要包括蛋白的定性和定量分析。

基于液相的定量蛋白质组学的流程。复杂的蛋白样品在溶液内或胶内被胰酶水解得到肽段混合物，经离子交换液相色谱和反向液相色谱梯度分离，分离的肽段直接进入质谱仪进行鉴定，得到的二级和一级质谱图通过软件分析和数据库比对搜索最终得到混合样品中所含的蛋白质信息，并根据是否带有同位素标记选择相应的定量分析方法

第18章 基于串联质谱的蛋白质定性分析

定性即鉴定样品含有何种蛋白。运用质谱对蛋白的鉴定是基于相对分子质量的比对原理。基于串联质谱的蛋白质定性分析中，蛋白经胰酶水解为肽段，肽段离子化后经一级质谱检测得到一级质谱图，反映了每个肽段离子的质核比；二级质谱检测时，一级质谱中高丰度的肽段离子被选取进一步由CID打碎成包括b离子和y离子在内的离子碎片。二级质谱记录离子碎片的质核比，得到相应的二级质谱图。蛋白分子的鉴定依赖于生物信息分析，将实际质谱得到的母离子（肽段）的相对分子质量和子离子的相对分子质量与理论酶切打碎形成的离子相对分子质量比对。基于该原理设计的常用软件有Sequest、Mascot和X！tandem等。

肽链断裂示意图。进入二级质谱前，母离子肽链将被碰撞诱导解离等方式打断，生成带有电荷的离子碎片，其中最常见的裂解位置是连接两个相邻氨基酸残基的酰胺键，产生b离子和y离子。相邻的b离子或y离子的质量差别对应于一个氨基酸残基的质量，由此推断出多肽的序列，也可由此绘制出已知多肽的理论二级图谱。

依赖生物信息分析软件，例如Sequest等，将实际得到的一、二级图谱与数据库比对，得到多肽的氨基酸序列和对应的蛋白质信息

1.蛋白质定量分析

基于液相的定量蛋白质组学，通常采用两种策略，即稳定同位素标记法和非标记法，本节中我们侧重阐述同位素标记法。以相对定量为例，分别以轻同位素（12C、1h或14N）和重同位素（13C、2h或15N）标记两组蛋白样品中的丝氨酸，标记后的肽段在质谱和液相色谱中行为没有发生改变是同位素标记的定量方法的基础。未标记或标记轻同位素的蛋白样品A与标记重同位素的样品B混合后酶解得到带有不同标记的肽段混合物，相同序列的肽段在液相中同时被洗脱并出现在同一时刻的一级质，由于同位素标记产生的相对分子质量的差值，不同来源的相同肽段在谱图上表现为两个有一定相对分子质量差的同

位素标记的肽段峰，根据此峰的高度或者峰下面积可计算出此肽段来源的蛋白的相对量的差异。

基于标记策略的相对定量方法。样品A，黑色线条为未标记或标记轻同位素的样品。样品B，黑点表示重同位素标记基团。两种样品混合，酶解得到肽段混合物，在一级质谱中同时出现，表现为两个有一定相对分子质量差的同位素标记的肽段峰，根据此峰的高度或者峰下面积可计算出此肽段来源的蛋白的相对量的差异

第19章 常用同位素标记定量蛋白组学方法

1.SILAC方法

SILAC（Stable isotope labeling with amino acids in cell culture）方法也称为体内标记方法，为活细胞在增殖和复制过程中通过代谢将蛋白进行同位素标记的方法。细胞在两种培养基中经过至少6代的培养后，细胞内的蛋白分别被普通氨基酸（Unmodified amino acid）和重同位素标记的氨基酸标记（重同位素标记的赖氨酸、精氨酸或两者同时标记）。蛋白等量混合后进行SDS□pAGE凝胶电泳和胶内酶解，经液相色谱分离后进入质谱检测，最后进行生物信息学分析。该方法需使ILAC原理流程。两组细胞分别在含有轻同位素标记（如12C，14N）的必需氨基酸或重同位素标记（13C，15N）的必需氨基酸的培养基中培养，由于细胞本身不能合成生长所需的某种氨基酸（如L□Lysine，L□Arginine），在细胞生长的过程中，这些带有同位素标记的必需氨基酸在细胞体内的代谢便使得细胞内的蛋白带上了同位素的标记，因此细胞每增殖一代就会有一半的蛋白被标记上标签，最终接近100％的标记效率。

用特殊处理的不含游离氨基酸的血清，Invitrogen和Thermo等多家生物公司已将该技术商业化。SALIC操作简单，能够标记活细胞体内的蛋白质，效率高且精确；但是该方法不适用于组织和原代非增殖细胞来源蛋白的标记，同时无游离氨基酸的特殊血清和重同位素氨基酸对细胞的生长和增殖可能有潜在的影响。目前，该方法已经用于蛋白混合物组分的研究、信号通路中酪氨酸激酶底物的研究以及肿瘤标志物的研究等。

2.同位素编码的亲和标签技术

于1999年开发的同位素编码的亲和标签技术（Isotope coded affinity tags，ICAT）可以有效定量比较两种细胞状态下蛋白质表达。ICAT试剂由三部分组成：一个生物素亲和标签；一个与8个轻氢或者8个重氢原子的连接剂（生成轻或重两种标签分子）；一个能与半胱氨酸形成共价键的Sh反应基团，可以用轻试剂标记一种细胞状态（如：早期肝

癌细胞），用重试剂标记另一种细胞状态（如：晚期肝癌细胞），然后混合同等量的两组标记蛋白质。蛋白混合物经胰酶消化成多肽混合物后，通过亲和柱将半胱氨酸位置上被标记的多肽（约90%的蛋白质有一个或多个半胱氨酸）分离。其方法的优点在于通过ICAT的结合以及亲和纯化降低了蛋白样品的复杂性，有利于后续的质谱鉴定；缺点是不适用于无半胱氨酸蛋白的定量研究，同时该方法操作过程复杂、步骤多，因操作带来的误差也相应增多。

ICAT原理。（a）ICAT试剂结构。ICAT试剂由一个生物素亲和标签，一个与8个轻氢或者8个重氢原子的连接剂（生成轻或重两种标签分子），及一个能与半胱氨酸形成共价键的Sh反应基团组成。（b）ICAT 标记策略，两种蛋白样品分别标记上轻或重ICAT试剂后按一定比例混合进行胰酶消化。酶解后，被标记的肽段因为ICAT上的生物素可通过亲和层析与未标记的肽段分离、纯化。被标记轻、重同位素的肽段混合物经过高效液相色谱串联质谱分析，定性和定量。

美国应用生物系统公司（Applied Biosystems Inc）在传统的ICAT技术上又开发了可切的同位素编码的亲和标签技术（Cleavable ICAT）。其优点是用13C同位素进行标记，使得被重标记的多肽与被轻标记的多肽在反相色谱中同时被洗脱（Co□elute）并同时进入质谱仪被定量分析，这样定量的误差比轻氢或重氢同位素标记（它们在反相色谱中不被同时洗脱）要小。另外，该技术引进了可被酸切的基团，生物素亲和标签在进行串联质谱分析之前被切除，减小了标记基团，扩大了可检测的多肽的范围。

3.iTRAQ技术

iTRAQ（Isotope tags for relative and absolute quantitation）技术由美国应用生物系统公司推出，可同时对4个样品进行标记、检测和定量分析，最近该公司又推出可同时标记8个样本的iTRAQ技术。概括了该技术的定量原理。iTRAQ试剂分子由三个部分组成：①报告基团，用于定量，其相对分子质量分别为114，115，116，117；②平衡基团，用于平衡报告基团带来的相对分子质量的差异；③反应基团，与氨基酸的N末端反应结合。肽段离子（母离子）由CID打碎，形成的b离子和y离子进入二级质谱用于蛋白的定性，被打断游离出的报告基团则用于定量。

iTRAQ试剂结构。（a）iTRAQ试剂分子由三个部分组成：对定量起关键作用的报告基团，其相对分子质量分别为114、115、116和117；平衡基团，相对分子质量分别为31，30，29和28，用于平衡报告基团带来的相对分子质量的差异；反应基团，与氨基酸的N末端反应结合。（b）这四个iTRAQ标签可与四个样品结合，同时对它们进行定量分析。来源于Ross等

iTRAQ实例。对已知肽段按照1:1:1:1进行标记，标记后的4个样品等量混合。（a）一级质谱母离子的同位素分布。（b）二级质谱中，出现在低相对分子质量区域的报告基团。（c）和（d）分别显示了肽段产生的b离子和y离子的同位素分布。

该方法已经广泛应用，使用该方法研究了在表皮生长因子（Epidermal growth factor，EGF）刺激下表皮生长因子受体（Epidermal growth factor receptor，EGFR）酪氨酸磷酸化状态的改变。

基于iTRAQ的酪氨酸残基磷酸化定量分析。（a）4组细胞在25mmol/L EGF环境下分别诱导0，5，10，30min后，提取蛋白、酶解、除盐，将4组肽段混合物分别与iTRAQ试剂反应后混合，其中含有磷酸化酪氨酸的肽段通过IMAC亲和色谱富集。（b）在二级质谱中，通过肽段断裂产生的b离子和y离子可以推断出肽段的序列和磷酸化残基的位置。（c）报告基团出现在低质核比值的区域，根据不同样品带有的不同报告基团的相对丰度，可以推测样品间酪氨酸磷酸化的相对水平。磷酸化抗体免疫沉淀后上清液中的肽段被用于对样品浓度和磷酸化水平的定量的均一化处理。来自Zhang等的图一（Zhang等，2005）

4.氧18（18O）标记法

氧18（18O）标记法：蛋白被胰酶消化时，18O与多肽羧基端的两个16O的原子发生交换，进而标记含赖氨酸和精氨酸的所有多肽。该方法相对比较廉价，实验也相对简单，可以应用于大量样本的标记。对这一方法的最新改进，即把胰酶消化后的多肽干燥后，再加h218O和胰酶（Yao等，2003；Bantscheff等，2004）保证了18O和多肽羧基端的两个16O原子交换的完全，大大提高了定量蛋白的准确性。另外，还有不用同位素标记的定量蛋白质组学（Label free quantitative proteomics）方法，此技术还有待进一步完善。

5.绝对定量法

绝对定量法（Absolute quantification，AQUA）用同位素标记已知多肽掺入到所要检测的标本中，这些多肽可在合成时加入含稳定同位素标记氨基酸，比如用13C和15N标记的氨基酸合成多肽。采用掺入同位素标记多肽做定量分析的方法可以降低背景离子干扰，质谱仪一般采用选择性跟踪（Selected reaction monitoring，SRM）和多样性反应跟踪（Multiple reaction monitoring，MRM）来定量分析多肽产生的特定多肽，具有很高的灵敏度和特异性。

美国哈佛大学Gygi教授实验室用该方法进行定量蛋白质组学研究。该方法的发展和应用主要包括两个阶段。

第一阶段，根据目标蛋白的氨基酸序列和水解蛋白酶的种类选择内标肽段。在研究翻译后修饰的试验中，需注意内标肽段包含待研究的修饰位点。这样合成的内标肽段与目标蛋白酶解后得到的对应肽段的氨基酸序列是完全相同的，并且内标肽段上的某个氨基酸将被替换成带有同位素标签的，从而与样品肽段区分，比如在Leu上标记6个13C和1个15N。这样得到的内标肽段与目标蛋白水解后得到的对应肽段具有相同的理化性质，并且相对分子质量相差7Da。接下来，合成的内标肽段经过LC□MS/MS分析其保留时间、片段离子强度，并选择合适的离子进一步做SRM试验。在SRM试验中，三重四极杆质量检测器的第一重四极杆（Q1）用来选择母离子，符合一定质核比（m/z）的离子被选中进入下一重离子碰撞室（Q2）并被打成碎片，得到的碎片离子将进入三重四极杆（Q3）中，在窄质核比（m/z）范围内的单个片段离子将被检测到。

第二阶段是在混合物样品中检测蛋白或被修饰蛋白的含量。细胞裂解液经SDS□pAGE电泳分离并染色后，在目的相对分子质量范围的胶条可被分离，然后与内标肽段混合并进行胶内酶解与LC/MS分析。此时，蛋白水解得到的对应肽段与内标肽段的保留时间和分离谱图是一样的，使用SRM方法可以从复杂的肽段混合物中精确地检测到内标蛋白和待分析蛋白，根据加入的已知浓度的AQUA肽段及其与目标肽段离子谱峰面积的计算，可得知这种蛋白在细胞中的表达或修饰情况。由于内标肽段与样品蛋白一起参与了胶内酶解的过程，因此肽段的提取效率、样品制备（包括真空悬干）的损耗和进入LC/MS分析时的可变因素不会影响样品和内标肽段的浓度比率的确定。

Kirkpatrick等用该方法研究了蛋白酶抑制剂MG132对hEK293细胞中泛素蛋白聚合体（polyubiquitin）表达水平的影响（Kirkpatrick等，

2005），并发现在经MG132处理的细胞中K48链肽段的表达量是未经处理的细胞中表达量的5倍。

美国SIGMA公司现拥有绝对定量法专利并销售相关产品，其他公司也可提供合成同位素标记多肽服务。除检测蛋白的表达水平外，AQUA还可以用来检测细胞中的蛋白翻译后修饰（posttranslational modification，pTM）。Gygi等用AQUA方法分析了细胞周期依赖的人细胞的分离酶（Separase）的1126Ser磷酸化位点（Gerber等，2003）。针对目标蛋白和磷酸化的蛋白合成了不同同位素标记的多肽（EpGpIApSTNSSpVL*K和EpGpIApSTNS（pS）pVL*K。），内标多肽上的Leu被同位素标记，通过LC□SRM方法选择性地分析分离酶在细胞有丝分裂前期、分裂开始以及分裂间期三个阶段的1126Ser位点的磷酸化情况。他们发现分离酶在细胞分裂中-后期过渡时磷酸化并在细胞分裂后期去磷酸化，同时发现hela细胞中有34%分离酶的1126Ser位点被磷酸化。

AQUA方法的缺点在于其定量的精确性或多或少会受同位素标记的多肽的纯度和目标蛋白酶解不完全的影响，并且同位素标记的多肽价格相对昂贵。为了克服这些缺点，Beynon等发明了一种新的方法：QCAT（Beynon等，2005）。对于一组需要定量分析的蛋白，首先为每个蛋白选择一个或多个能够唯一代表该蛋白的内标肽段，将编码这些内标肽段的cDNA序列整合成一条DNA，克隆到载体中，并在大肠杆菌中表达出内标肽段的联合体QconCAT。如果在含有同位素标记氨基酸的培养基中培养细菌，细菌就可以利用培养基中同位素标记的氨基酸合成同位素标记的蛋白质，经纯化、酶解后便可得到多个同位素标记的多肽。编码QCAT的cDNA质粒也可用体外表达系统表达蛋白，在表达系统中加入同位素标记的氨基酸，纯化、酶解后也可得到多个同位素标记的多肽。通过分析质谱中得到的样品肽段和内标肽段的比值，就可以对含有该肽段的蛋白进行定量分析。在QCAT中，每个Q肽段代表一个蛋白，在选择时要注意以下几点：①该肽段不能含有半胱氨酸（Cys）残基，因为该残基会阻碍所表达的蛋白分子内和分子间二硫键的形成，并在QCAT过程中引入一个半胱氨酸残基。②选择的肽段在待分析的蛋白中必须是独一无二的。③根据检测原理的不同，这些肽段的相对分子质量应当满足一定的范围。比如，对于基质辅助激光解离和时间飞行质谱（MALDI□TOF），在1000~2000Da的质量范围内检测的灵敏度较高而干扰较低。最后，试验证明在做MALDI时

75%的信号较强分子是以Arg结尾的胰酶酶解肽段。Johnson等用该方法对鸡的肌蛋白混合物进行了绝对定量。

第20章 无标记定量蛋白质组学

由于标记定量蛋白质组学实验过程繁琐，标记试剂价格昂贵，较难对大规模的样品同时进行比较，于是人们开始寻找无标记的方法，而质谱色谱技术的发展，也使无标记定量蛋白质组学成为可能。无标记的（Label free）研究方法在样品处理过程中不加入任何会改变目标蛋白的肽段相对分子质量的同位素标签，根据比配样品间对应离子的相对强度达到相对定量的目的。该方法适用面广、较为经济，但样品必须逐一检测，对实验的重复性以及后期的数据分析要求较高。目前，无标记定量蛋白组学根据原理的不同分两类：一类是基于质谱峰强度的方法，另一类是基于鉴定蛋白的肽段数的方法。

1.基于质谱峰强度的方法的原理：在液质联用检测中样品的肽段浓度越高，其在质谱中被检测到的峰强度（面积）也越大。通过比较不同LC□MS中相同质核比的峰强度，来确定蛋白的相对表达量。它的基本过程一般由计算机软件来完成，分以下几步：①峰信号的计算；②峰信号处理；③峰面积的计算以及差异计算。质谱峰强度比较常用软件

2.基于鉴定蛋白的肽段数的方法的原理：相对简单，根据一个蛋白在LC□MS中相应肽段所检测到的二级谱图的次数来比较蛋白的相对表达量，同一肽段中鉴定到的二级谱图的次数越多，相应的蛋白表达量也越高。其实现方法是，先将所有的二级谱图与数据库进行比对，确定图谱所对应的肽段，然后将属于同一蛋白的谱图进行计数，从而比较蛋白在不同样品中的表达差异。peptideprophet和proteinprophet会将谱图数作为其结果的一部分呈现出来。

Old等（2005）比较这两种方法，发现基于鉴定蛋白的肽段数的方法在检测蛋白丰度变化时更灵敏，而基于质谱峰强度的方法在评估蛋白比率时更准确。

以串联质谱为基础的蛋白质组学主要用到如下几方面的软件：①格式转化软件；②肽段蛋白鉴定分析软件；③定量比较分析软件；④后期分析软件。

第21章 格式转化软件

在对生物数据进行分析时，我们可能需要用到很多不同的软件工具去完成各种相同或不同的功能，不同软件的输入文件格式往往不同，从一些软件中获得结果的格式常常不便于之后的分析。数据格式时常限制我们对数据的信息挖掘、组合和分析。所以数据格式转化软件在整个蛋白质组学信息分析中必不可少。以串联质谱为基础的蛋白质组学中，最常用的是原始数据与数据库检索相关的转化软件，包括完成质谱原始数据转化工具和数据库格式转化工具.在得到蛋白鉴定或定量结果之后，根据不同的研究目的，后期的分析将更为多样，所需要的格式转化软件也更多样，因此也常需要自己编写脚本或者手工完成特殊的转化目的。

对串联质谱获得的数据，推断肽段序列的方法有两种。一种方法是从头测序，根据二级或者二级以上的质谱图谱直接推断出肽段序列。这种方法对图谱的质量要求较高，往往需要理论上质谱峰能够比较完整连续地被检测到，对于谱峰的缺失、偏移的容忍度不是很高，因此只有比较理想的谱图符合这一要求。同时从头测序的方法在计算序列上一般采用类似寻找最优路径的方法，算法复杂度比较高，且没有比较理想的替代算法，所以该方法不便于大规模的数据分析。

另一种方法是通过数据库比较分析，需要有一个已知的数据库，将需要分析的二级图谱与数据库中所有的图谱依次进行比较打分，然后人为地设定一个阈值，如果得分最高的数据库图谱与实际图谱的匹配优于阈值，那么就认为该实际图谱所代表的就是数据库图谱所代表的序列。这种方法按数据库分又包含两种不同的分析方式。①数据库本身就是质谱图谱数据库（如X！hunter，SpectraST），数据库中的每一个图谱代表一个序列。其图谱数据来源于实际实验，其构建方法通常为在已知的蛋白序列数据库基础上人为地合成每一个肽段序列，然后依次用质谱扫描。而不同的质谱仪器对于同一序列扫描得到的图谱也有一些差异，所以即便同一蛋白序列数据库也常常需要分别建立不同仪器的图谱数据库。即使在建立完成图谱数据库之后，因为序列数据库的更新，研究工作往往也要求图谱数据库能够及时更新。可想而知，构建和维护一个这样的数据库所需的费用和人力将非常庞大。而且这种方法很不灵活，适用的范围不广，如无法完成有特殊修饰要求的扫描。不过该方法在谱图匹配上确实有很优秀的表现。②直接采用序列

数据库，通过软件模拟，对每一个序列生成一张二级质谱图，便产生了理论上的谱图数据库。这种方法的优点是构建谱图数据库非常方便、快捷以及零成本。同时由于谱图是通过软件模拟产生的，可以基于不同的研究目的灵活地构建谱图数据库，比如以研究磷酸化为目的，则可以在构建的理论谱图中引入磷酸化修饰。目前，该方式在串联质谱数据分析中应用最广泛。

6.6.3定量比较分析软件

蛋白质组学的定量分析方法按其样品标记与否分为两大类。由于样品的标记方法也有多种方式，不同的定量软件往往只适用一种或几种标记方法的定量。常见的一些以标记定量为基础的定量软件及其适用的定量方法。

6.6.4后期分析软件

得到蛋白鉴定或定量结果之后，根据不同的研究目的，后期还会用到很多不同的软件。以研究分泌蛋白为例，后续分析中可能会用到一系列分泌蛋白的预测工具（Secretomep，signalp，targetp，tmhmm等），这些工具在不同的研究领域有较强的针对性。后续分析也是整个分析流程中最为灵活的部分，没有固定的框架，但是其重要性丝毫不亚于蛋白的鉴定或定量。比较常见的有对蛋白鉴定或定量结果进行GO和KEGG分析。GO（Gene ontology，www。geneontology。org）是一个注释基因和蛋白标识的数据库，包含细胞组分、分子功能、生物学过程三方面的信息注释。而KEGG（Kyoto encyclopedia of genes and genomes，www。genome。jp/kegg）数据库包含了代谢、遗传信息处理、环境信息处理、细胞过程和人类疾病等信息。网站上有许多以这两个数据库为基础的信息分析软件。

第22章 蛋白质组学数据分析平台

在质谱数据分析过程中需要用到各种不同工具，费时费力，于是出现了整合格式转化、数据检索、统计分析、定量等各种软件的分析平台。这些平台极大简化整个分析流程，常见的平台软件有Tpp，CpAS，GpM，Spire等。其中，Tpp是一款集合格式转化、数据库检索、肽段和蛋白统计分析、肽段蛋白定量和多样本整合等功能的强大软件。相对于其他分析方式，Tpp平台的优势在于采用自己的统计模型对不同的数据库检索结果进行统计分析，重新分配其肽段和蛋白的鉴定可信度。

1.以小波理论为基础的蛋白质组学定量方法

小波算法能够将功能信号或时间上连续的信号分解成不同频率的信号组分，然后在其对应尺度上进行分别处理（Mallat，1989；Meyer，1993）。相对于传统的傅立叶变换算法，小波算法能够更好地处理非连续尖锐峰形信号，并且能够更加精确地进行信号分解和重构（Crandall，1994）。

目前，基于小波算法已开发出多款针对MALDI，SELDI□TOF和LC/MS等不同类型的蛋白质组学数据的软件。例如，针对MALDI质谱数据，Yang 等比较了5种信号滤波算法（Yang等，2009），包括滑动平均滤波、Savitzky□Golay滤波、高斯滤波、Kaiser时窗函数和小波滤波。实验结果显示，基于小波算法的滤波效果最好。Du等研究开发了一种基于连续小波变换（CWT）的信号峰鉴定算法，能够更有效地区分SELDI□TOF质谱图谱中信号与噪音，并且对强弱不同信号峰的鉴定都能保持很低的假阳性率（Du等，2006）。Randolph 和Yasui 应用一种平移不变小波的变换算法对MALDI□TOF质谱图谱进行基于不同尺度的信号分解、信号特征提取以及定量（Randolph，Yasui，2006）。Alexandrov等开发了软件工具MALDIDWT，能够分析血清蛋白图谱并发现生物标志物（Alexandrov等，2009）。Lange等基于小波理论开发了信号峰辨别算法，从而替代质谱仪厂商提供的绑定软件（Lange等，2006）。Schulz□Trieglaff等开发了另一种算法，利用小波基来模拟同位素标记的信号强度分布（Schulz□Trieglaff等，2008）。后两种算法后来都被应用于OpenMS软件中（Sturm等，2008）。Zhang等利用非抽取小波变换算法，在不提供蛋白质量前提下去除prOTOF质谱数据中

的随机噪音（Zhang等，2009）。针对高通量LC/MS数据，以代谢组学数据为例，Tautenhahn等展示了一种新的信号特征辨别算法centWave，其能够利用连续小波变换和自选高斯拟合算法准确辨别图谱信号峰区域（Tautenhahn等，2008）。

小波理论的应用还能降低质谱数据复杂度，并缩减运算时耗。例如，hussong基于自适应小波变换理论，开发了一种信号特征辨别算法，能够大大加快质谱数据分析。利用小波细节系数来描述信号特征值并能降低质谱数据复杂性。

基于小波理论，我们开发了WaveletQuant软件.该算法基于构建肽段离子的单离子图对其丰度进行定量，首先利用小波算法对其峰图进行优化，将信号分解为4层并计算每一层的相关系数，再通过小波滤噪算法计算信号阈值并过滤噪音，最终重构得到优化信号峰图，从而计算肽段丰度。鉴于是一个已广泛应用的蛋白质组学开源分析软件，我们将程序和Tpp进行整合，取代其原有ASApRatio定量工具。

通过实验比较发现，相对于原来的ASApRatio软件，WaveletQuant能通过拟合单离子图并区分相邻峰图来更准确计算峰图区域，从而得到目标肽段丰度。中两组比较所示，ASApRatio利用Savitzky□Golay 滤波法去噪，其往往将两个相邻峰进行融合并一同定量；而WaveletQuant利用小波算法则能更好地区分相邻高低丰度峰，避免了低丰度噪音峰或其他肽段信号峰对目标肽段信号峰的影响，从而更准确地获得目标肽段定量结果。

蛋白质组学的局限和挑战

相对于基因组学、转录组学，蛋白质组学还未达到全景式（即对所有蛋白质）的研究，目前的蛋白质组学研究方法对蛋白质组的覆盖率还很有限（大多物种特别是复杂生物的覆盖率小于50%）。总结了基于液质联用的蛋白组学的进展及其面临的问题和一些解决办法。但是，要全面提高蛋白质组学研究对蛋白质组的覆盖率，还需有革命性的技术突破。

读累了记得休息一会哦~

公众号：古德猫宁李

- 电子书搜索下载
- 书单分享
- 书友学习交流

网站：沉金书屋 https://www.chenjin5.com

- 电子书搜索下载
- 电子书打包资源分享
- 学习资源分享

第23章 亚蛋白质组学

对蛋白质组进行深入系统的研究可以全景式地揭示生命活动的本质。但是生物体内的蛋白质种类繁多，其表达量具有很宽的动态范围，理化性能也存在很大的差异。其中，高丰度蛋白质表达量大，易于被分离鉴定；而中低丰度蛋白质由于数量庞大、组成复杂而不易被检测。同时，高丰度蛋白质也会掩盖、屏蔽、结合大量的低丰度蛋白质，给中低丰度蛋白质的分离鉴定造成极大的困难。尽管蛋白质组学研究的对象是一个生物体中所有蛋白质，但是受分析技术手段的限制，目前还不可能检测到所有蛋白质。对蛋白质组全谱的分析得到的经常是高丰度蛋白质的信息，很难得到在生命活动中起关键作用的低丰度蛋白质的信息。这也是疾病蛋白质组研究难以取得突破的主要原因。

由于目前的技术条件无法对蛋白质组进行完整的分析，较好的策略是分而治之、逐个击破。蛋白质组可以根据蛋白质在细胞中的定位不同分成不同的亚细胞蛋白质组，也可以根据蛋白质物理、化学性质的不同用电泳、色谱的方法分成不同的馏分，还可以根据蛋白质功能的不同分成不同的功能蛋白质组。将蛋白质组分成一个个亚蛋白质组（Subproteome），然后进行分析，由于样品的复杂度降低很多，因此更有可能发现有意义的低丰度蛋白质。

蛋白质翻译后修饰在生命体中具有极其重要的生理作用。蛋白质由于各种不同的翻译后修饰而使其功能更为完善、调节更为精细、作用更为专一。由于蛋白质翻译后修饰并不是直接由基因决定的，而几乎所有的蛋白质行使功能、功能受阻以及蛋白质性质改变都是由蛋白质翻译后修饰所致的。因此蛋白质翻译后修饰的种类、数量、结构鉴定对于蛋白质生理功能的诠释有着极其重要的作用。本章主要介绍根据蛋白质翻译后修饰分类的三种亚蛋白质组的研究，即磷酸化蛋白质组、糖基化蛋白质组和多肽组。①磷酸化蛋白质组：蛋白质的磷酸化修饰是研究最为广泛的一种翻译后修饰方式，据估计，真核生物细胞通常条件下约有30%蛋白质发生磷酸化修饰。蛋白质的磷酸化和去磷酸化在信号转导过程中有着很重要的调控作用，研究蛋白质的磷酸化有助于揭示疾病发生、发展过程中的调控机制。②糖基化蛋白质组：蛋白质的糖基化也是一种非常常见的翻译后修饰。糖蛋白及其糖链结构的改变能够体现特定的生理或病理状态改变，已知的多种疾病标志性分子都是糖蛋白，如对前列腺癌有提示作用的前列腺特异性抗原、肿瘤

抗原CA125等。因此，分析和研究糖基化蛋白质组将有助于发现疾病标志性分子，提高疾病的预警和诊断能力。③多肽组：内源性多肽物质在生命活动中也起重要的调控作用，从酶稳定性、细胞间信息传递到生物体内环境平衡等生理活动都离不开多肽的功能活性。亚蛋白质组的基本质谱分析方法与一般的蛋白质组是相似的，都是通过先将蛋白质酶解成肽段，然后利用质谱分析这些肽段，来实现鉴定蛋白质的目的。对于亚蛋白质组而言，最重要的是如何将亚蛋白质组分离富集出来，因此如下将侧重介绍亚蛋白质组研究的分离富集方法与技术。

1.磷酸化蛋白质组

蛋白质的磷酸化和去磷酸化几乎调节着生命活动的整个过程，包括细胞的增殖、发育和分化、神经活动、肌肉收缩、新陈代谢和肿瘤发生等。据统计，哺乳动物磷酸化的丝氨酸、苏氨酸和酪氨酸

细胞内有1/3以上的蛋白质可以被磷酸化（Zhang等，2002），蛋白质通过磷酸化改变其三维构象来调节其活性。真核生物中，蛋白质磷酸化主要发生在丝氨酸、苏氨酸和酪氨酸残基上，其比例大约为1800∶200∶1。磷酸化蛋白质功能研究的主要任务之一是鉴定磷酸化蛋白以及磷酸化位点。对单一蛋白质进行研究的传统方法远不能满足分析蛋白质多样性和复杂性的需要，因此利用蛋白质组技术和生物信息学，高通量地研究蛋白质的翻译后修饰已成为必然趋势。

磷酸化的一般过程

随着质谱技术和蛋白质组学的发展，串联质谱技术逐渐被广泛地应用于蛋白质磷酸化的分析。利用串联质谱分析蛋白质磷酸化的主要过程如下：①蛋白质酶解，将磷酸化蛋白质酶解成肽段；②磷酸肽的选择性富集，利用金属亲和色谱等技术将酶解液中的磷酸肽选择性富集起来；③液相色谱串联质谱联用（LC□MS/MS）分析磷酸肽，获得磷酸肽的多级质谱图；④利用质谱图进行数据库检索，获得磷酸肽的可能序列和位点；⑤磷酸肽的最终鉴定，主要利用人工筛选数据库检索结果确定磷酸肽的序列和磷酸化位点。蛋白质磷酸化分析的难点主要有两个方面：一是如何高选择性地提取磷酸肽，避免大量非磷酸肽的干扰；二是如何根据质谱数据自动高可信地鉴定磷酸肽。

2.磷酸肽的富集方法

在磷酸肽的高选择性富集方面，近几年国内外都取得了较大的进展。给出了常用的磷酸肽富集的方法及其优缺点。其中，固定金属离子亲和色谱（Immobilized metal ion affinity chromatography，IMAC）是最常用的方法。在IMAC方法中，铁、镓等金属离子通过螯合基团［如亚氨基二乙酸（IDA）］固载在色谱颗粒上，磷酸肽由于其磷酸根与固定的金属离子之间的强相互作用而被选择性捕获。在实际操作中，蛋白质酶解液一般需要酸化后上样，经过冲洗后用氨水等碱性溶液将磷酸肽洗脱。近几年，新材料如纳米材料（pan等，2006）、整体柱材料（Feng等，2007）等也被用于制备IMAC的载体。其中基于铁离子的IMAC（Fe3+□IMAC）最常用，但其缺点是特异性不高，在富集磷酸肽的同时也富集了一些酸性肽和含组胺酸的肽。为了提高富集磷酸肽的特异性，可以在进行IMAC富集前对肽进行酯化，使羧基失去负电从而减小非磷酸肽与金属离子间的相互作用（Ficarro等，2002）。这种方法已经成功地用于酵母、鼠肝、hT□29癌细胞株等磷酸化蛋白质组的规模化分析（Ficarro等，2002；Kim等，2005；Moser，White，2006）。但由于化学反应控制不好会导致反应不完全、产生副产物等问题，从而使样品更为复杂，因此基于化学反应的方法并没有被普遍采用。最近我们发展了一种具有更高特异性的IMAC技术，其化学结构。不同于一般的IMAC，在这种基于锆离子的IMAC（Zr4+□IMAC）中，螯合基团是磷酸

方法优缺点文献IMAC（固定金属亲和色谱）Fe3+/Ga3+/Al3+□IMAC最常用的方法，一般使用亚氨基二乙酸为螯合基团；但特异性差，除磷酸肽以外，酸性肽、含组氨酸肽也被富集Neville等，1997；posewitz，Tempst，1999；pan等，2006；Feng等，

2007IMAC富集前酯化处理有利于提高特异性，但反应条件难控制，经常由于化学反应不完全和副反应使样品更为复杂Ficarro等，2002Zr4+□IMAC以磷酸基团为螯合基团的新一代IMAC，比常规IMAC有更高的特异性Feng等，2007；Zhou等，

2007金属氧化物ZrO2/TiO2/Al（Oh）3比常规IMAC有更高的特异性；由于没有间隔臂，对比较大的磷酸肽可能有空间位阻Kweon，hakansson，2006；Cantin等，2007；Zhou等，2007离子交换色谱SCX（强阳离子交换色谱）能富集并分级磷酸肽；特异性较低；由于弱相互作用，多磷酸肽容易丢失Ballif等，2004；Beausoleil等，2004SAX（强阴离子交换色谱）能富集并分级磷酸肽；特异性较高；

分级磷酸肽时分辨率较低Nuhse等，2003；Dai等，2007化学方法β□消除仅适用于富集丝氨酸、苏氨酸磷酸化肽；由于受O连接糖肽的干扰，特异性差；化学反应难以控制Oda等，2001磷酸酰胺化适用于所有磷酸肽；由于多步化学反应而产率较低Zhou等，2001

用于富集磷酸肽的固定亲和色谱技术（Jiang等，2008）

基团。锆离子由于与磷酸基团之间的强相互作用而被固载在固相载体上，磷酸肽也由于其磷酸根与锆离子之间的强相互作用而被富集。Zr4+□IMAC特异性非常好，能够有效地将磷酸肽从磷酸蛋白（酪蛋白）与非磷酸蛋白（牛血清白蛋白）之比为1:100的蛋白质酶解液中提取出来。除IMAC外，TiO2、ZrO2等金属氧化物也可以用于磷酸肽的富集（Kweon，hakansson，2006；Cantin等，2007；Zhou等，2007）。一般认为这些金属氧化物富集磷酸肽的特异性比Fe3+□IMAC要好。但是由于金属氧化物颗粒没有间隔臂，可能会存在比较大的空间位阻，对相对分子质量大的磷酸肽的富集效果可能不太好。鉴于磷酸肽带有额外的负电基团，因此也可以用离子交换色谱富集磷酸肽。强阳离子交换色谱（SCX）（Ballif等，2004；Beausoleil等，2004）和强阴离子交换色谱（SAX）（Nuhse等，2003；Dai等，2007）都可以用于磷酸肽的富集。在SCX中，由于磷酸肽保留较弱，因此磷酸肽比大多数非磷酸肽早流出从而实现选择性分离；而在SAX中，磷酸肽保留很强而非磷酸肽基本上不保留，因此也可以实现选择性分离。使用离子交换色谱的优点在于：在富集磷酸肽的同时还可以将磷酸肽分成多个组分，因此有利于磷酸化蛋白质组的规模化分析；其缺点是：特异性不高。此外，还可以利用化学的方法将磷酸肽分离出来，但由于化学反应难以控制，因此很少使用。

3.磷酸肽的鉴定

目前，绝大多数二级谱图（Tandem mass spectrometry，MS2）是通过碰撞诱导裂解（Collision□induced dissociation，CID）获得的。由于磷酸化肽段在CID中容易丢失磷酸（98Da）而难以进一步碎裂形成可以运用于数据库匹配的碎片信息，所以所产生的二级谱图（MS2）质量比较差。为了获得额外的碎片信息，可以将失去磷酸基团的中性丢失峰打碎获得三级谱图（MS3）。但是由于离子在MS3中的峰强度很低，所获得的MS3谱图质量也难尽人意。用MS2、MS3进行数据库检索，利用一般的筛选标准进行筛选会导致很多假阳性鉴定，因此磷酸肽的鉴定一般通过人工确认才能最终获得鉴定。人工确认的缺点是显

而易见的，劳动强度大、通量低、结果不客观、鉴定的假阳性率难评估。运用高质量精度的质谱仪结合反库检索针对磷酸化肽段进行大规模的鉴定，可以得出磷酸化鉴定的假阳性率（Beausoleil等，2006；Olsen等，2006）。但是由于高精度质谱仪器价格昂贵，因此该方法的应用范围受到了很大的限制。我们发展了一种通过主动匹配MS2、MS3的数据检索结果来实现磷酸肽的自动、高可信鉴定的方法（Jiang等，2008）。在这种方法中，将MS2、MS3谱图分别进行数据库检索，收集MS2和相应的MS3能共同鉴定同一磷酸肽的数据，然后选择合适的门槛值，将错误的鉴定去除，实现磷酸肽的自动鉴定。该方法采用先匹配后筛选策略，所有二级谱和对应的三级谱的鉴定结果都应用于匹配，从而避免了常规方法中采用高门槛分别筛选MS2和MS3数据检索结果后所造成的信息量丢失的问题，有利于提高磷酸肽的分析灵敏度。如果采用含反库的混合数据库来检索MS2和MS3的数据，鉴定结果的可信度通过假阳性率的评估得到，有效地避免了人工确认过程中的人为因素。

4.新的碎裂技术

为了克服CID中的中性丢失，最近出现了两种新的碎裂技术：电子转移解离裂解（Electron transfer dissociation，ETD）和电子捕获裂解（Electron capture dissociation，ECD）。给出了ECD与CID的碎裂方ECD与CID碎裂方式不同，ECD碎裂成c和z离子，而CID 碎裂成b和y离子。，ECD几乎不存在序列依赖性，可产生广泛的肽骨架裂解，更为重要的是不稳定的翻译后修饰在ECD条件下都保持稳定，因而将生物质谱的“软”电离能力从MS延伸到MS/MS，使得ECD成为一种非常有前途的翻译后修饰研究工具。然而，ECD技术在除了傅立叶转换离子回旋共振质谱仪（Fourier transform ion cyclotron resonance mass spectrometry，FTICR MS）外的其他类型质谱仪中实现还有困难，而且FTICR比较昂贵，维护费用高，这些都限制了ECD技术的广泛应用。与ECD碎裂方式类似，ETD技术可以使翻译后修饰基团在碎片上保持完整，同时得到肽段骨架序列连续的碎片信息以及修饰位点信息。与ECD相比，ETD技术可以应用在相对比较便宜的离子阱质谱仪中，因而具有更加广阔的应用前景。由于在ETD和ECD中磷酸基团不丢失，因此磷酸位点的确定更准。Molina等利用ETD对人胚肾细胞的磷酸化蛋白质组进行分析，鉴定了1435个磷酸化位点（Molina等，2007）。

第24章 糖基化蛋白质组

与磷酸化一样，蛋白质糖基化也是一种重要的翻译后修饰，据统计约有一半以上的蛋白质发生糖基化。糖基化对蛋白质的结构和功能有着重要影响。与蛋白质相连的糖链对蛋白质的折叠、运输起着重要作用，并且参与了一系列生命活动。异常的糖基化蛋白质也是癌症等疾病的标记物。糖基化蛋白质是一种糖缀生物，由一个或多个寡糖链通过不同连接方式与蛋白质部分的多肽骨架共价相连而成。根据连接方式不同，可以将该糖基化分为4种类型：N型连接、O型连接、糖基磷脂酰肌醇锚定和C型连接（Morelle等，2006）。其中，以N型连接和O型连接最为常见，也是糖基化蛋白质组的主要研究内容。在N型连接中，一般由六至数十个糖基组成糖链，N□乙酰葡萄糖胺还原端与蛋白质中天冬酰胺的酰胺氮以β□1，4糖苷键相连，而且被糖基化的天冬酰胺残基必须存在于Asn□Xaa□Ser/Thr（Xaa是除脯氨酸外的任何氨基酸）基序中。在O型连接中，糖链的还原端与蛋白质肽链的丝氨酸、苏氨酸或羟赖氨酸残基中的羟基氧原子相连，糖链没有固定的结构及核心，也没有明确的基序。N型糖蛋白主要存在于血浆蛋白与分泌蛋白中，而O型糖蛋白质广泛存在于黏液中。

糖基化蛋白质组学分析包括两方面内容：蛋白质的鉴定和糖基结构的确定。由于糖基结构非常复杂，对其进行全面分析目前还有一定困难。相对而言，蛋白质的鉴定比较容易，是目前糖基化蛋白质组分析的主要内容。由于蛋白质样品酶解后，糖肽存在于大量的非糖肽之中，因此糖基化蛋白质组学鉴定的关键是如何高特异性地提取糖肽。

1.糖蛋白和糖肽的富集

目前，常用的糖蛋白、糖肽的富集技术主要是外源凝集素亲和色谱和化学方法，下面分别进行介绍。

双亲和色谱技术富集糖肽凝集素亲和色谱法经常用于糖蛋白质和糖肽的选择性富集。在糖基化蛋白质组分析中，为了提高糖肽富集的选择性，一般采用双亲和色谱技术。先利用亲和色谱将糖蛋白从复杂的蛋白质混合物中提取出来；然后将糖蛋白酶解后再次利用该亲和色谱将糖肽提取出来；将糖肽的糖链去除后，利用LC□MS/MS可以鉴定糖肽，从而实现糖蛋白质的鉴定。伴刀豆球蛋白A（ConA）亲和色谱是

最常用的外源凝集素亲和色谱，主要用于富集高甘露糖型和混合型寡糖的糖蛋白质或糖肽。Kaji等（利用双ConA亲和色谱技术对线虫蛋白质组中的N连接糖基化蛋白质进行了规模化分析。在他们的实验中，富集的糖肽用pNGase F（一种能特异性切除N连接糖基的糖酶）切除糖链，所得到的肽段再用LC□MS/MS分析，鉴定了约250个糖蛋白质，同时通过N型连接的基序确定了大约400个糖基化位点。除Con A外，其他外源性凝集素（如麦胚凝集素、橙黄网胞盘菌凝集素等）也被用于糖蛋白质和糖肽的富集。不同的凝集素可以富集具有不同糖结构的糖蛋白质或糖肽，因此利用多种凝集素亲和色谱可以更全面地分析糖蛋白质组。Yang等利用三种凝集素（Con A、麦胚凝集素、木菠萝凝集素）亲和色谱分析血浆中的糖蛋白，鉴定了约150个糖蛋白质。

2.酰肼化学法

除凝集素亲和色谱外，基于酰肼化学的方法来提取糖肽也很有效。在这种方法中，通过糖链上的基团将糖蛋白质或糖肽固定在固相小球上实现糖肽的选择性提取。目前，基于酰肼化学提取糖肽的方法，第一种方法固载糖蛋白质，第二种方法固载糖肽，它们的最终产物是一样的，即切除糖链的糖肽。①固载糖蛋白的方法：最初由Zhang等建立（Zhang等，2003），为了将糖蛋白固载到固相小球上，先将糖链上的顺式二醇基团氧化成醛基，然后糖蛋白质由于醛基与固相小球上的酰肼反应而被固载。糖蛋白固载后，加入胰蛋白酶，将固载于小珠上的蛋白质酶解。酶解产生的非糖肽将被冲洗除去，而糖肽则还固载在小球上。利用pNGase F将N型连接糖肽从小球上酶切下来，用LC□MS/MS进行分析鉴定。②固载糖肽方法：由Sun等建立，它是在固载糖蛋白方法的基础上发展起来的。在这种方法中，糖蛋白质先被酶解成肽段，然后通过酰肼化学将糖肽固载在固相小球上。将蛋白质酶解成肽段后固载，可以解决膜蛋白的溶解性问题，而且糖肽位阻小，因此更容易被固载。根据Sun等的结果，该方法的糖肽捕获效率接近100%，糖肽的选择性高达91%。

3.其他方法

除了凝集素亲和色谱和酰肼化学外，还有些其他的色谱方法。这些方法一般利用糖肽的特殊物理化学特性，如利用糖基的亲水性以亲水相互作用色谱富集糖肽，利用糖肽比一般的肽体积大以分子排阻色谱将其分离。此外，硼酸亲和色谱、石墨色谱也被用于分离糖肽。但是由于特异性不是很好，这些方法一般用于简单糖蛋白质样品的分析。

4.糖蛋白鉴定的难点

由于糖链的微观不均一性，一个糖基化位点上的糖链类型就可能多达几十种，因此利用带有完整糖链结构的糖肽来鉴定糖蛋白质及其位点是很困难的。在实际的糖蛋白质组分析中往往需要将糖肽上的糖链切除后利用液质联用来进行分析。①对于N□连接糖肽，一般使用N□糖苷酶F（pNGase F）将糖链切除（Kaji等，2003；Zhang等，2003）。pNGase F在切除N□糖链的同时将使天冬酰胺转变为天冬氨酸，使相对分子质量增加0.98，从而起到质量标记N□糖基化位点的作用。N□连接糖肽中的糖链也可以使用内切糖苷酶h（Endo□N□acetylglucosaminidase h）切除。与pNGase F不同，内切糖苷酶h在去糖基化时会将N□糖链五糖核心中与天冬酰胺相连的GlcNAc以外的部分切除，而在糖基化位点处留下GlcNAc，从而起到标记糖基化位点的作用。糖基化位点处的GlcNAc使该天冬酰胺残基相对理论质量增加了203，所以对质谱质量准确度要求不高。内切糖苷酶h最大的不足在于其狭窄的专一性，其实际上只能作用于高甘露糖型和杂合型N□糖链。另外，留在糖基化位点处的GlcNAc在CID条件下并不稳定，经常会脱落，从而起不到标记糖基化位点的作用。这在很大程度上限制了内切糖苷酶h在糖蛋白鉴定中的应用。不过，可以利用其专一性与pNGase F配合使用，阐明N□糖基化的类型。②对于O□连接糖肽来说，现在还没有一种与pNGase F相似的酶能用来实现去糖基化/质量标记糖基化位点的作用，所以O□连接糖基化位点的标记多采用化学法，其中报道较多的是β□消除反应法。该方法基于在碱性环境中Ser、Thr上的O□糖基团会发生β□消除形成1个不饱和的双键，这个双键可以被亲核试剂攻击发生加成反应，使Ser或Thr残基的质量发生一个特定的变化，也就是使O□糖基化位点被质量标记，而且这种质量标记在CID的条件下是稳定的，从而可以通过串联质谱测序的方法得到糖基化位点的信息。除用于O□糖基化研究外，β□消除□迈克尔加成反应还更为广泛地应用于蛋白质磷酸化富集、位点鉴定研究，其中许多方法很值得借鉴到O□糖基化研究中。该方法的不足在于非糖基化或磷酸化的Ser、Thr残基的侧链羟基也可能发生β□消除，尤其在较高的温度和较强的碱浓度下，在应用中应该注意。

对于带有糖链的糖肽的分析，目前应用较多的质谱碎裂方法还是碰撞诱导裂解碎裂（CID）。但在CID中，糖苷键常常优先断裂，而多肽骨架保持完整，从而使肽段序列鉴定、糖基和糖基化位点鉴定困难。近年来，电子捕获裂解（ECD）和电子转移解离（ETD）也被用于糖肽

的分。ECD和ETD几乎不存在序列依赖性，可产生广泛的肽骨架裂解，更为重要的是存在许多不稳定的翻译后修饰，因而将生物质谱的“软”电离能力从MS延伸到MS/MS，使其成为一种非常有前途的翻译后修饰研究工具。

第25章 多肽组学

多肽组学是蛋白质组学的一个分支，主要研究相对分子质量较小的蛋白质或多肽。鉴于多肽的重要生物学意义，多肽组学研究已经获得了广泛的关注。多肽组学所研究的低相对分子质量成分主要包括生理功能多肽（如神经多肽、激素、细胞因子、生长因子等）和蛋白质降解产生的多肽分子。生理功能多肽分子是重要的信号分子之一，在体内发挥着重要的信号传导作用，几乎所有细胞都受多肽调节，它涉及激素、神经、细胞生长和生殖等各个领域。蛋白质降解产生的肽段没有生物活性，但也是多肽组的重要组成部分，它们是蛋白受内源性酶剪切的产物，因此可以在一定程度上反映内源性酶的活性。在体液中，这些由蛋白质降解而来的内源性多肽的数量和功能与人体的生理及病理状态有很大的关系，因此这些分子也是潜在的生物标记物。多肽分子与大量的蛋白质共同存在，因此多肽组学分析所面临的一个挑战就是如何有效地将低丰度内源性多肽从高浓度蛋白质混合物中提取出来。

蛋白质和多肽的物理化学性质非常相似，唯一较大的区别在于多肽相对分子质量较小。因此，一种比较有效的方法是利用他们在相对分子质量方面的差别将多肽与蛋白质分离。利用具有不同相对分子质量截留范围的膜可以实现多肽与蛋白质的分离。目前，应用比较广泛的是超滤技术。通过选择不同相对分子质量截留范围和处理体积的超滤膜可以实现对中等和少量体积多肽组学样品的高通量和高选择性处理。将超滤技术与其他分离方法联用可以实现多肽组的规模化分析。如hu等发展了基于超滤管预富集、体积排阻色谱（SEC）预分离和毛细管液相色谱质谱联用的方法来进行鼠肝中多肽组的大规模分析，实现了对1181个特征多肽的成功鉴定（hu等，2007）。由于高相对分子质量的蛋白质容易吸附在超滤膜孔周围而逐渐使孔堵塞，因此超滤的方法在处理含高浓度蛋白质的样品时效果往往不是很理想。有序介孔材料具有孔道大小均匀、排列有序、孔径可在2~50nm连续调节等特性，使其在基于分子排阻的分离提纯方面有一定的应用潜力。选用合适孔径大小的介孔材料，多肽可以进入孔内而保留，而蛋白质由于不能进入而不保留，这样就可以实现多肽与蛋白质的分离。我们合成了具有高度有序孔道结构的MCM□41有序介孔材料（孔径为20□），利用该材料具有的体积排阻机制和反相色谱保留机制实现了对血浆内源性多肽的高效选择性富集。通过化学改性将该材料分别修饰以强阳离子和强

阴离子交换基团，可以进一步提高材料对内源性多肽的富集选择性和富集容量。利用该方法可以实现对人血浆中内源性多肽组分的高选择性富集和大规模鉴定，共鉴定988种人血浆内源性多肽。该方法的优点：不仅鉴定出的血浆多肽数目是常规膜超滤或液□固吸附提取方法所鉴定的3倍以上，而且显著缩短了样品预处理所需的时间。基于有序介孔材料的固定化金属离子亲和色谱基质还可以用于从复杂的血清中富集内源性血清磷酸化多肽。

除利用相对分子质量的区别外，还可以利用蛋白质与多肽在溶解度等性质的区别实现分离。有机溶剂沉淀是一种比较简单的处理方法，它是利用高相对分子质量蛋白在有机溶剂中溶解度低这一性质将其去除。常用的有机溶剂有乙腈、丙酮、三氯乙酸等。但是由于蛋白沉淀的同时会携带与其结合的多肽分子，因此样品损失大，选择性不高。还有一种常用的方法是固相萃取技术。这种方法一般利用键合有不同官能团的磁珠将多肽捕获并用MALDI□TOF质谱分析，具有速度快、通量高的特点。尽管固相萃取技术在一定程度上可以实现复杂生物样品中多肽组分的选择性富集，但同时样品中含有的大量蛋白质也会被富集出来，必然会影响多肽样品在质谱中的检测。

第26章 C□MS/LC□MS 代谢组学分析

代谢组学是随着生命科学发展而发展起来的。与其他组学（如基因组学、蛋白质组学等）不同，代谢组学是通过考察生物体系受刺激或扰动后，其代谢产物的变化或随时间的变化来研究生物体系的一门科学。高灵敏度、高通量且稳定性好的分析方法是进行代谢组学研究的关键。目前的分析手段主要包括核磁共振技术（NMR）及各种色质联用技术。

气相色谱质谱联用技术（GC□MS）是广泛用于代谢组学分析的色质联用技术之一，其在人类先天性代谢异常研究方面已经有相当长的发展历史，目前已成为植物和微生物功能基因组代谢表型研究的常规分析技术。GC□MS 的主要优点是灵敏度高，可检测到大量低含量的小分子代谢物。另外，GC□MS仪器的购置价格较低，在色谱分析重复性、分辨率和电子轰击电离源得到的质谱碎片重复性方面具有明显的优势，且受基体效应影响较小。GC□MS的主要不足是样品中难挥发或极性较大的代谢产物需经过衍生化后才能进行分析。与GC□MS不同，用于代谢组学研究的另一种重要的色质联用技术——液相色谱质谱联用技术（LC□MS）可直接分析体液及组织提取物，无需衍生化操作，适用于热不稳定、不易挥发、不易衍生化和相对分子质量较大的物质。液相色谱卓越的分离能力和质谱多通道监测的功能使LC□MS技术对检测样品的浓度和纯度要求明显低于NMR技术，甚至对含量极低的物质也能通过优化质谱的扫描模式给出可视化响应。同时，LC□MS技术又有较好的选择性和较高的灵敏度，这些优点使LC□MS联用技术得以在代谢组学研究中大显身手。当然，LC□MS技术应用于代谢组学研究的过程中也遇到了一些挑战，如分析方法的偏向性、方法有限的峰容量造成的峰重叠、潜在生物标志物的鉴定以及海量数据的处理策略等。这些挑战也成了基于LC□MS的代谢组学平台技术研究的热点。

本章主要介绍GC□MS、GC×GC□MS（二维气相色谱质谱联用技术）、hpLC□MS（高效液相色谱质谱联用技术）、UpLC□MS（超高效液相色谱质谱联用技术）、hpLC×hpLC□MS（二维高效液相色谱质谱联用技术）及基于上述联用技术的代谢组学平台的基本组成、优势和发展趋势，并以具体实例介绍上述联用技术在代谢组学研究中的应用。

1.GC□MS代谢组学分析

色谱法具有高灵敏度、高分离能力和高分析速度等特点，是分析复杂混合物的主要手段。气相色谱自20世纪中叶问世至今一直是一种广泛使用的分离分析技术（James，Martin，1952）。由于色谱法在进行定性分析时的主要依据是保留值，难以对复杂未知混合物作定性分析；而质谱法可以测定化合物的相对分子质量、分子式以及提供有关分子结构的信息。因此，色谱法与质谱法联用，即分离技术与鉴定技术结合的联用技术已成为分析技术的一个主要发展方向。在GC□MS联用系统中，气相色谱相当于质谱的分离和进样装置，质谱则相当于色谱的检测器。这样既发挥了各自的优势，又弥补了各自的不足。holmes和Morrell（1957）在1957年首次实现了GC□MS的成功联用，为该技术的快速发展奠定了基础。与其他联用技术相比，GC□MS目前拥有更为成熟的技术和更为广泛的应用，也逐渐成为复杂混合物分析的主要定性、定量手段之一。由于GC□MS能够达到较高的分辨率和检测灵敏度，并且有可供参考、比较的标准谱图库，可以方便地得到待分析代谢物的定性结果，因此近年来被广泛地应用于代谢组学分析。

在GC□MS代谢组学分析中，代谢物通常被分为两大类，即不需要化学衍生的挥发性代谢物和需要化学衍生的非挥发性代谢物。挥发性代谢物不需衍生化步骤即可从气相色谱流出，这类对象的采样方法主要包括直接收集和分析顶空样品、用固体吸附剂富集顶空液体样品中的代谢产物、固相微萃取和溶剂萃取等。挥发性代谢物通常不需要进一步样品制备可直接用于仪器分析。而非挥发性代谢物（如血液和尿液中的氨基酸、脂肪酸、胺类、糖类、甾体类物质）极性强，挥发性低，如果能将这些极性强、挥发性低的物质进行适当的化学处理转化成相应的挥发性衍生物，可以扩大气相色谱的测定范围。常用的衍生试剂主要有硅烷化试剂、烷基化试剂（包括酯化试剂）、酰基化试剂、缩合反应试剂和手性衍生试剂等。如人血清经衍生化预处理，在优化的分析条件下检测到951个峰

人血清分析的GC□TOFMS总离子流图。色谱柱：DB□50（30m×0.25mm×0.25μm）；色谱进样量：2μL；进样温度：270℃，以20℃/min升至300℃，保持4.5。经American Chemical Society许可复制

自Tanaka等（Tanaka，West□Dull等，1980）报道将GC□MS技术用于患者尿液分析以筛查有机酸尿症以来，GC□MS在临床代谢组学研究中的应用取得了快速发展。近期的应用包括尿液和血液中的有机酸谱分析，以确定新生儿代谢异常、有机酸尿、脂肪酸氧化、神经代谢紊乱和2型糖尿病等（Kuhara，2002；Carlson，2004；Yuan，Kong等，2007）。

GC□MS还被用于毒理研究，如Lee等（2007）采用小鼠动物模型，通过给小鼠腹腔注射不同剂量的壬基酚来研究毒性作用。研究结果表明，四氢皮质甾酮和5□四氢皮质甾酮是与壬基酚致毒相关的尿中潜在生物标志物。

德国科研人员最早将GC□MS技术用于植物代谢组学研究（Roessner，Wagner 等 ， 2000；Roessner ， Luedemann 等 ， 2001；Roessner ，Willmitzer等，2002），使得近年来GC□MS作为植物代谢组学研究的重要分析技术得到了广泛的应用。植物代谢组学主要通过研究植物细胞中的代谢组在基因变异或环境因素变化后的相应变化，研究基因型和表型的关系及揭示一些沉默基因的功能，进一步了解植物的代谢途径。最具代表性的是Oliver Fiehn 研究组的工作（Fiehn，2002；Taylor，King，2002）。他们利用GC□MS技术，通过对不同表型拟南芥的433种代谢产物进行代谢组学分析，结合化学计量学方法（pCA、ANN和hCA）对这些植物的表型进行分类，找到了4种对分类有重要贡献的代谢物质，即苹果酸、柠檬酸、葡萄糖和果糖。该结果与线粒体和叶绿体中的基因型结果一致。

pCA分析结果：特征代谢物质苹果酸和柠檬酸。经Oxford University press许可复制

Roessner□Tunali等（2003）采用GC□MS分析方法系统研究了西红柿叶片和果实组织的代谢谱，发现由于果糖激酶AthXK1的过度表达，转基因西红柿的磷酸果糖下降。类似的方法还可用于微生物代谢组学的研究，如Strelkov等（2004）考察了不同生长条件对Corynebacterium glutamicum的影响。Klapa等采用质谱同位素法研究了C。glutamicum生化网络中的赖氨酸生物合成。

2.GC×GC□MS代谢组学分析

二维气相色谱（GC×GC）是20世纪90年代发展起来的具有高分辨率、高灵敏度、高峰容量等优势的多维色谱分离技术（Venkatramani，Xu等，1996；phillips，Beens，1999；Dalluge，Beens等，2003），也是迄今为止能够提供最高分辨率的分离技术。就其相应的检测器而言，传统四极杆质谱采集速度慢，通常采集速度为2张全谱图/秒。目前，商品化的四极杆质谱能达到的最快采集速度也只有20张全谱图/秒，而且使用最大采集频率时会牺牲检测灵敏度。飞行时间质谱（TOFMS）有非常高的采集速度，最高采集频率可达500次全扫描/秒，是目前可以与GC×GC很好匹配的质谱技术（Dimandja，Grainger等，2000）。

在代谢组学研究中，GC×GC□TOFMS分析常用的柱系统与其他复杂混合物分析相似，通常第一维采用非极性或弱极性的色谱柱，第二维采用中等极性的色谱柱。如Synovec等采用GC×GC□TOFMS方法研究酵母细胞内的代谢物组成（Mohler，Dombek等，2006；Mohler，Tu等，2008）。胞内代谢物经衍生化处理后，经GC×GC□TOFMS分析所采用的柱系统为第一维色谱柱RTX□5MS（20m×0.25mm×0.5μm），第二维色谱柱RTX□200MS（2m×0.18mm×0.2μm）。对于经衍生化预处理的代谢谱的数据分析，一般不采用总离子流色谱（Total ion chromatogram TIC），而采用特征性更强的质量碎片的二维谱图。由不同特征质量数得到的酵母胞内代谢物的GC×GC□TOFMS谱。胞内代谢物中含（a）酵母细胞内代谢物的GC×GC□TOFMS等高线图（选择特征质量数m/z 73）；（b）选择质量数m/z 73进行“Sratios”计算的结果；（c）选择质量数m/z 128进行“Sratios”计算的结果；（d）选择质量数m/z 217进行“Sratios”计算的结果。图中标记1，2，3，4的化合物分别是指甲基柠檬酸盐、肌醇、葡萄糖磷酸和胱硫醚（Mohler和Tu等，2008）。Copyright（2008），经Elsevier许可复制

有活性官能团的组分，经三甲基硅烷化（TMS）反应生成含有1个或多个TMS基团的衍生化产物。由于TMS基团（Si（Ch3）3）具有m/z 73的特征离子，因此衍生化后含有TMS基团的胞内代谢物也含有m/z 73的特征质量碎片，通过选择特征质量数可以得到与样品组成相关的信息。m/z 73特征质量数得到的二维谱图由于m/z 73质量数存在的重叠峰引起的二维图干扰比较大，因此作者采用了一种称为“Sratios”的计算方法以消除上述重叠峰的干扰。选择不同质量数，如m/z 73、m/z 128和m/z 217，进行“Sratios”计算后，700个重叠峰分别被简化至85个、37个和23。标记1，2，3，4的化合物分别是指甲基柠檬酸盐、肌醇、葡萄糖磷酸和胱硫醚。

与传统的GC□TOFMS技术相比，GC×GC□TOFMS技术可以提供更高的分辨率和灵敏度，也可以检测到更多用一维方法无法分离的色谱峰。因此，近年来采用GC×GC□TOFMS技术进行代谢组学研究的报道逐渐增多。Welthagen和Fiehn等（2005）最先将GC×GC□TOFMS技术应用于代谢组学研究。他们以哺乳动物生物学研究为例，分析了小鼠脾组织提取物的复杂代谢产物轮廓，证实了GC×GC□TOFMS技术用于代谢组学研究的可行性。之所以选择脾组织，一是因为目前还没有关于脾组织提取物的代谢轮廓分析的相关报道，二是肥胖和过度喂养已经被证实与人类和啮齿目动物免疫反应负相关，通过代谢组学研究有可能帮助揭示其潜在的机制。Welthagen和Fiehn等的分析结果也表明，采用GC×GC□TOFMS方法经重叠峰解析和干扰峰去除后共检测到1220种化合物，大大优于用GC□TOFMS方法检测到538种化合物的结果，而且峰的纯度也得到了很大的改进。研究同时证实了GC×GC□TOFMS方法得到的代谢轮廓可以用于生物标志物的检测。我们应用GC×GC□TOFMS方法分析了青蒿挥发油的成分，并与一维GC□MS分析结果进行比较，定性出303种化合物，其中大部分是萜类化合物。在选定的条件下，青蒿挥发油成分可达到族分离的效果，分为烷烃、单萜、单萜含氧衍生物、倍半萜、倍半萜含氧衍生物五部分。与青蒿素合成密切相关的青蒿酸也被初步鉴定。

青蒿挥发油的全二维气相色谱飞行时间质谱图。横坐标为一维保留时间，纵坐标为二维保留时间。（a）族分离效果图：A为烷烃，B为单萜，C1/C2单萜含氧衍生物，D为倍半萜，E为倍半萜含氧衍生物；（b）被标注的色谱峰为其他文献中报道过的化合物，如青蒿酸（Arteannuic acid）和蒿酮（Artemisia ketone）（Ma，Wang等，2007）。Copyright，经Elsevier许可复制

（a）青蒿酸第一维色谱图；（b）青蒿酸第二维色谱图；（c）青蒿酸二维效果图；（d）青蒿酸结构图；（e）E1：样品中青蒿酸质谱图；E2：NIST库中的标准质谱图（Ma，Wang等，2007）。Copyright（2007），经Elsevier许可复制

上述研究结果表明，GC×GC□TOFMS可以用于代谢组学研究。尽管这方面的工作刚刚起步，GC×GC□TOFMS数据处理方法还不完善，文献报道的二维峰匹配、峰解析等方法也不够成熟，但随着代谢组学研究的不断深入和分析技术的不断创新，可以预计GC×GC□TOFMS技术在代谢组学研究方面将发挥越来越大的作用。

3.hpLC□MS代谢组学分析

近年来，hpLC□MS联用技术以其分离效能高、分析速度快、灵敏度高及应用范围广的优势在代谢组学研究中占据了较大比重。采用hpLC□MS技术，可以使用更为简便的样品预处理步骤，在较短的时间内对选定的靶标化合物进行检测。hpLC□MS，特别是hpLC□MSn能够实现对复杂基质中结构相似的化合物的同时分析，非常适合于代谢产物的代谢轮廓分析及极端复杂基质中靶标代谢产物的分析和鉴定。hpLC□MS还能反映预处理阶段较难分离和不稳定化合物的信息。因此，当对待分析化合物的结构不具备先验知识时，hpLC□MS可以同时测定生物样品中的已知和未知化合物，即可以进行代谢组学分析。

当然，目前基于hpLC□MS的代谢组学研究仍存在许多挑战，这些挑战也正是基于液相色谱及液相色谱质谱联用技术的代谢组学方法中的热点研究问题。首先是分析方法上对化合物的偏向性，如极性化合物在反相液相色谱（RpLC）柱上保留较弱，常常由于离子抑制现象得不到较好的检测结果，这种偏向性主要是由选择的样品预处理方法和分析手段所导致的。在样品的预处理过程中，通常需要使用某种提取方法将代谢物从基质中提取出来进行hpLC及hpLC□MS分析，对于丰度较低的代谢物来讲，往往还需要对其进行浓缩才能进行后续的分析，在这个步骤中不可避免地会因提取方法和样品性质不同而产生偏向性。而在其后的分析方法中，目前使用最多的hpLC模式是RpLC，在这种模式下，极性化合物往往在柱上不能得到保留，在死时间时就被冲出柱外，即反相柱对这些化合物没有分离能力。这些化合物经常会导致离子抑制，从而导致分析方法上的偏向性。其次，与GC□MS联用技术相比，hpLC□MS技术缺乏可供定性参考的规模数据库，在代谢产物特别是结构信息较少的化合物的解析上存在较大难度。近几年，这方面的研究工作正迅速开展（Dunn，Bailey等，现在已有少量液相色谱质谱数据库，并且已有多个组织及公司正在构建代谢产物的谱图库。另外，较受欢迎的两个解决办法是采用高分辨的质谱数据（TOF□MS，ICR□FT□MS）和采用其他分析方法辅助定性。但无论采用上述哪种方法都不能实现简单快捷确定代谢物结构的目的，往往需要综合各种方法及研究者的专业知识来甄别判断。此外，由hpLC□MS的高分辨率所带来的海量数据，需要合适的数据挖掘技术和数据处理方法来提取其中的有用信息。到目前为止，已有多种用于液相色谱和液相色谱质谱联用方法的数据处理方法和软件（。

一个完整的代谢组学分析流程包括样品采集、样品制备、样品分析及数据处理等步骤。根据不同的分析目的，代谢组学又可以分为不同的层次，其对样品采集、样品制备、样品分析及数据处理的要求亦有不同。以代谢轮廓分析为例：代谢轮廓分析的对象是某一类化合物或某个代谢途径上的代谢产物，因此要建立基于色谱及其联用技术的代谢轮廓分析方法，需要选择合适的样品制备方法、分析方法及数据处理方法，从而实现对特定目标化合物的代谢轮廓分析。在这里介绍我们建立的基于hpLC技术的代谢轮廓分析策略。实际上，如果在样品的预处理过程中能够保留尽可能多的代谢物组分，这个策略特别是其中的数据处理部分也适于建立代谢组学研究层次的技术方法。

采用代谢轮廓分析策略，我们（Xu，Di Stefano等，1999；Yang，Xu等，2002）对不同阶段肝病患者的实际尿样进行了hpLC□MS分析，并与健康对照组进行比较，数据解析结果表明，具有顺式二羟基结构的化合物主要是核酸的代谢产物——核苷）是一类与癌症诊断有密切关系的代谢物。在随后的研究中也发现基于液相色谱的代谢轮廓分析不但能够用于癌症的早期诊断，还有效地减少了炎症对诊断的干扰，大大降低了诊断的假阳性率。

尿样中的顺式二羟基代谢产物的反相高效液相色谱分析的典型色谱图

代谢组学方法鉴定的生物标志物和传统生物标志物之间的相关性网络。

另外，我们采用代谢组学分析策略，以尽可能多地保留内源性代谢产物信息为基础，建立了样品分析的技术平台及数据处理流程。并将上述技术用于慢性乙型肝炎的急性发作疾病研究，对慢性乙型肝炎肝功能急性恶化患者及健康对照组的血清样本进行了hpLC□MS分析及数据解析，鉴定了新的肝病生物标志物。结果表明，肝癌诊断中肝炎和肝硬化患者的假阳性率仅为7.4%；应用于慢性乙型肝炎的急性发作样本，诊断正确率为100%，鉴定出1个传统标志物和4个新的标志物——4种溶血磷脂，弥补了现有标志物的不足。实验中采用化学计量学方法，将新标记物与病历中的标记物做相关性分析，得到了这些标记物在代谢网络上的关联。Wilson等也采用hpLC□MS技术对肥胖大鼠的尿液进行了代谢组学研究，结合化学计量学方法比较了不同种属、性别、年龄等肥胖大鼠的代谢谱差异，各组之间得到了很好的区分结果，是三类雄性小鼠晨尿样品的hpLC□MS数据经pCA分析后的得分图。在微生物代谢组学方面，Dalluge等（2004）采用hpLC□MS技术

对发酵过程中的氨基酸实施了监测，通过分析认为其中的一个子集可反映发酵的状态。上述研究结果充分证明了hpLC□MS联用技术在代谢组学研究方面的应用潜力。

三类雄性小鼠晨尿样品的hpLC□MS数据经pCA分析后的得分图。绿色代表裸鼠，黄色代表小白鼠，红色代表小黑鼠，经Elsevier许可复制

UpLC□MS代谢组学分析

随着现代社会与科学技术的不断发展，液相色谱分析的对象也逐渐复杂化，这就对高效液相色谱提出了更高的要求。尤其代谢组学研究针对的是更加复杂的生物样品基质，如血、尿、组织等，并且样品数量巨大，要求液相色谱具有更加高效、快速、灵敏的性能。采用细粒径填料和细内径柱子而获得高柱效（100000~300000）的液相色谱技术，简称超高压液相色谱（Ultra performance liquid chromatography，UpLC）。UpLC系统是利用创新技术进行整体设计，从而大幅度改善色谱分离度、样品通量和灵敏度的最新液相色谱技术。相较于目前分析速度最快的高效液相色谱（hpLC），UpLC的分析速度提高了9倍，分辨率提高了2倍，灵敏度提高了3倍，一次分析所得到的信息量大大超过了高效液相色谱。

许多科研工作者也对UpLC与hpLC的性能进行了比较。Cristiana等应用UpLC技术检测婴儿食品中农药残留，对各类相关的婴儿食品中16种农药残留物进行测定，检测限可达1μg/kg，其检测速度为hpLC的215倍，平均回收率范围是85%~119%，相对标准偏差小于17%。plumb等应用（2.1mm×100mm，3.5μm）C18hpLC系统和（2.1mm×100mm，1.7μm）C18UpLC系统对老鼠胆汁样品进行分析比较（plumb和Castro□perez等，2004），UpLC色谱峰更加尖锐，分辨率更高，可获得更多的色谱信息。Novakova等用标样比较了UpLC、hpLC和使用整体柱的分离效率以及分析期间的系统维持费用（Novakova和Solichova等，2006），得出结论，UpLC比hpLC和使用整体柱时得到理论塔板数、分辨率更高，不对称因子也得到更优的结果，而所使用的溶剂量远远少于hpLC和整体柱。Tomas等应用UpLC技术分析食品中农药残留物（Kovalczuk和Je，检测了苹果样品提取物中17种半极性农药，与hpLC比较，UpLC技术提高了样品的通量，降低了溶剂的消耗量。

老鼠胆汁样品hpLC和UpLC色谱图比较。（a）hpLC；（b）UpLC。

众多的研究结果表明，UpLC较hpLC有更好的分离效率、峰容量以及灵敏度，能提供更适合与质谱联用的接口，这将有助于检出更多代谢物，提高方法通量、灵敏度，改善与质谱联用的定性定量结果。随着代谢组学的迅猛发展，UpLC□MS联用为代谢组学研究提供了更为高效、快捷、灵敏的方法平台。我们应用UpLC□QTOF MS检测了40个肠癆患者及17个健康人的血清样品，ESI+和ESI-的数据都用于pLS□DA分析，肠癆患者与健康人可以得到很好的区分，并发现了9种新的代谢标志物，为重要生物标志物的网络相关图。研究结果还表明，UpLC□MS联用技术在发现疾病可能的生物标志物及揭示病理改变方面是强有力的工具。我们使用UpLC□QTOF MS对癌症患者尿中顺二醇结构的代谢物进行分析，用代谢组学方法寻找可能的临床诊断标志物，同样在更短的时间内得到了更多的代谢物信息，癌症患者与正常人得到了更好区分，与此同时也发现了更多的生物标志物。在使用UpLC□MS联用时发现的对分类贡献最大的15种可能的生物标志物中，有5种在hpLC□MS联用时并没有被检测出来。这一研究结果也进一步证明了UpLC□MS联用技术在代谢组学研究方面的巨大潜力。

肠癆患者血液中可能的生物标志物及其相互关系（Yin和Zhao等，2006）。Copyright（2006），经the American Chemical Society许可复制

在药物代谢组学研究方面，Li等应用UpLC□MS联用技术进行中药试剂淫羊藿提取物对老鼠模型的药效研究，检出并且定性了淫羊藿提取物的几种主要活性代谢产物。对内源性代谢产物进行pCA分析，对照组、模型组和给药组得到了显著的分离。我们（汪江山和赵欣捷等，2006）将超高效液相色谱与飞行时间质谱联用，用于人参皂甙Rg3作用后大鼠尿液代谢物指纹图谱分析及标志物的鉴定。该研究以大鼠尿液及人参皂甙Rg3静脉给药后的大鼠尿液作为测试样品，考察UpLC□QTOF MS这一平台对复杂体系进行分离分析的能力，探讨了如何利用化合物的精确质量和MS/MS数据对给药大鼠尿液中显著变化的未知内源性代谢物进行结构鉴定。是大鼠尿液中代谢物的UpLC□QTOF测定图。对比给药组以及对照组大鼠尿液的代谢物指纹图谱，给药后大鼠尿液中众多代谢物的相对浓度发生了较大的变化。后续研究发现，给药后0~24小时，大鼠生化代谢发生了显著性的扰动；给药48~72小时后，大鼠的生化代谢显示回归平衡状态的趋势。Ma等应用UpLC□MS联用技术对人体血浆中的氨氯地平进行药代动力

学研究，获得了较好的结果，定量下限为0.15ng/mL，精确度RSD15%。

给药后大鼠尿液UpLC□TOF MS质谱。

综上所述，与传统的hpLC相比，UpLC提供的高效、高速、高灵敏性能已经在代谢组学研究中显示了优势，使液相色谱在更高水平上实现了突破。目前，Agilent、Thermal、Shimadzu等公司也已先后推出类似的高效、高速、高灵敏仪器。这些新技术在分离科学表现出新奇而强大的能力，大大拓宽了液相色谱的应用范围，提升其在分离分析科学中的重要地位，也为代谢组学研究的进一步发展提供了重要的技术支持。

5.hpLC×hpLC□MS代谢组学分析

二维液相色谱（hpLC×hpLC）是20世纪90年代初开始发展起来的色谱分离技术。它主要是利用两种不同分离机制的色谱模式，或使用相同模式具有一定正交性的不同色谱柱进行样品的分离。20世纪90年代中后期，hpLC×hpLC的应用研究主要集中在蛋白质和多肽的分离上。自21世纪以来，新型色谱柱填料类型尤其是整体柱的出现和发展，极大地促进了二维液相色谱技术的发展。如利用整体柱代替常规的硅胶基质微粒色谱柱，可以采用更长的第二维分析柱和更快的分离速度，从而在更短的分析时间里获得更高的峰容量。有研究预测，如果加上质谱的分离能力，这样的分离体系可以提供高达几千的峰容量。由于样品在液相色谱的分离是样品在色谱固定相和流动相中的分配过程，流动相的选择对液相色谱的分离起着至关重要的作用，因此可以通过流动相改性剂的作用，影响化合物的保留，产生多种保留机制，为多维色谱分离提供丰富的组合方式。如利用新型色谱填料，通过改变两维的流动相条件，实现酸、碱、中性物质的二维分离。此外，利用某些具有特殊分离机制的色谱填料与反相色谱联用，可以实现聚合物和同分异构体的二维分离。

自1984年Giddings等提出多维分离的概念以来，在理论研究方面，增加峰容量和多维柱系统的正交性始终是人们关注的热点。1987年，Giddings指出，全二维色谱的理论峰容量是两维各自峰容量的乘积（Giddings，1987）。在20世纪90年代中期，又提出了“样品维数”的概念，将样品的复杂性和分离体系的峰容量结合起来（Giddings，1995）。此后，很多针对不同样品，采用不同填料类型和流动相条件

的柱系统的正交性得到考察。Gilar等系统研究了多种色谱柱（包括hILIC Si、C18、苯基、SEC、SCX和五氟苯基（pFp））在不同ph值条件下选择性的正交性（Gilar和Olivova等，2005）。结果表明，通过改变柱系统配置和流动相条件，某些色谱柱之间可以表现出很强的正交性，尤其是SCX/RpLC和hILIC/RpLC，但是所有的实验数据都是在单柱上完成的，并没有考虑到在线构建二维系统时流动相的匹配问题。即使如此，其实验结果对了解柱系统的正交性仍然有指导作用。

2004年以来，一个新的应用研究热点集中在分别用正相色谱（NpLC）和反相色谱作两维，以分离分析脂类等强疏水性化合物。通过增加两维间的流速比，或降低第二维的上样量，辅以在两维分别采用互溶性较好的流动相，规避了正相色谱和反相色谱流动相不兼容的问题。通过这些方法，一些典型的强疏水性化合物得到了分离。Murahashi等（2003）设计了全二维NpLC/RpLC联用系统，第一维采用常规柱，第二维采用平行的整体柱实现快速分离。

不同色谱柱对磷酸化胰蛋白酶酶解片段的LC□MS色谱分离图

到目前为止，二维液相色谱已经用于分离蛋白质、多肽、聚合物、同分异构体、药物和天然产物中的疏水性物质，而几乎所有的hpLC×hpLC柱系统都采用RpLC作为其中一维，利用组分间疏水性的差异对样品进行分离。但对于代谢组学研究中的复杂生物样品，尤其是对于含有大量强极性和亲水性组分的样品来说，如尿样，RpLC会因无法有效保留其中的强亲水性组分而损失很多代谢物的信息。hILIC作为一种新的液相色谱分离模式，已经显示出其在代谢组学分析领域的适用性。

根据我们的研究，基于不同功能基团的hILIC色谱柱的选择性也不尽相同，通过优化两维间的流动相组成，hILIC色谱柱之间也能体现出适中的正交性，因此可以构建用于分离分析强亲水性复杂体系样品的柱系统。我们建立了一种新型的hILIC×hILIC□QTOF MS分离方法，用于分析结构类似的强亲水性样品皂树皂甙提取物中的皂树皂甙单体。皂树皂甙具有多种不同糖苷配体和糖环残基，可能形成多种结构，各单体组分间又具有类似的结构和色谱保留特征，一维色谱很难为其提供足够的峰容量进行分离，峰重叠无法避免。目前报道的绝大多数hpLC（包括SpE）方法采用RpLC填料作为分离介质。考虑到其结构中含有丰富的亲水性基团，以及不同皂树皂甙组分间的区别通常来自糖环残基的不同，采用选择性独特的hILIC色谱柱应能为这种混合物提供

一种新的分离方法。但实验表明，一维hILIC□MS方法不能为这种结构复杂的天然产物提取物提供足够的峰容量。

在上述实验中，样品经hILIC×hILIC分离后的总离子流数据经Transform软件（ver 3.4，Noesys Software package，Research Systems International，Crowthorne）处理为等高线图。从图中可以看出，经过第二维pOh□Et A柱（2.1mm×35mm，5μm polyhYDROXYEThYL ATM）的再次分离，某些在第一维Amide□80柱（2mm×150mm，3μm）上的共流出组分得到了完全分离；而某些在pOh□Et A柱上难以分开的组分也因在Amide□80柱上表现出不同的保留行为而得到分离。在此实验条件下，二维分离空间得到了较为充分的利用，尤其是在50~100min的时段内。

皂树皂甙混合物的hILIC×hILIC□QTOF MS总离子流等高线图。

在以上hILIC×hILIC流路中，两种功能基团不同的hILIC色谱柱构成柱系统，通过优化柱参数和两维流动相条件，为样品提供了正交性较好的二维分离空间，从而显著改善了系统的峰容量和分离能力。两维间的流动相互溶且强度易于控制，其组成又利于与质谱联用。质谱可以为样品的分离提供额外的分辨率，多级质谱还可以提供样品组分的结构信息，进一步提高了该系统对亲水性复杂样品的分离和定性能力。对于该研究中的皂树皂甙混合物，虽然正离子模式可以提供互补的组分信息，但负离子模式更适合于进行结构鉴定。结合二维色谱保留行为和二级质谱特征，46种皂树皂甙单体得到检出，其中8种此前未见报道。此外还排除了44种非皂树皂甙组分的干扰。通过二维分离，多对皂树皂甙同分异构体组分也得到了良好的分离，体现出hILIC×hILIC□QTOF MS系统强大的分离和鉴定能力。

第27章 发展与展望

综上所述，GC□MS/LC□MS技术已经对代谢组学领域的研究产生了很大的推动作用，但其本身仍然有很大的发展空间，并且面临着理论、技术和应用等多方面的挑战。

就GC□MS来讲，该分析技术本身还需要不断完善。如GC□MS分析生物样本中代谢物普遍需要衍生化预处理，因此可发展与多个官能团发生衍生反应且重复性好的高通量预处理方法，不断满足组学分析对通量和信息量的要求。另外，经衍生化反应的代谢物仍有较宽的沸程，有些衍生化的代谢物在较高温度下才能从色谱柱中流出，发展适合复杂代谢物分析的且可在较高温度条件下使用的高效气相色谱柱无疑会对改善分辨率和提高分析方法的精度十分有益。此外，尽管GC□MS有商品化的质谱谱图库，但由于还有相当数量的代谢物或其衍生化产物的质谱图未包含在其中，对这些代谢物的鉴定仍很困难。

同样，LC□MS技术本身也有许多方法学上的问题需要解决。首先，生物体系的复杂性决定了生物体液以及生物组织组成的复杂性，即使LC□MS技术有较高的检测灵敏度，痕量物质的归属和精确定量也还存在不少困难。其次，应用LC□MS技术确证化合物结构的数据库有限，不能满足复杂多样的代谢物研究的需要。虽然科研工作者已经总结了一些与病理和生理变化相关的标志性代谢产物，但是要建立完整的诊断专家系统，实现代谢组学诊断的常规化还需要做大量的工作。

在数据处理方面，用于GC□MS和LC□MS，特别是GC×GC□MS及LC×LC□MS的数据处理方法还需进一步开发。目前还没有用于GC×GC□MS及LC×LC□MS数据处理的商品化代谢组学软件，这在很大程度上影响了GC×GC□MS及LC×LC□MS在代谢组学研究中的广泛使用。

围绕着上述亟须解决的问题，未来代谢组学的重要研究方向之一是合理有效地结合GC□MS、LC□MS技术各自的优势，尽量全面准确地对生物体中代谢物的分布、变化进行定性及定量分析。这些技术的不断发展、成熟及在生命科学中的广泛应用，也将使我们在基因组学、转录组学和蛋白质组学的基础上更好地描述和评估生命系统。

第28章 基因调控网络

1.基因调控网络

DNA 微阵列的广泛应用提供了海量的基因表达谱数据，如人们可以在同一时刻观察到细胞内成千上万mRNA 的相对或绝对数量，这为在分子水平研究基因之间的相互关系及作用提供了技术基础。在转录过程中，一个转录因子（蛋白质）与DNA绑定激活另一个基因的转录，其表达产物有可能是转录因子，后者又激活或抑制其他基因的转录，这样就形成了基因调控路径（Gene regulatory pathway）。基因调控网络（Gene regulatory network）是一组调控因子如何调控另一套基因表达的过程。参与该过程的主要生物大分子包括DNA、mRNA、蛋白质、其他小分子等。基因调控网络的特点：①基因数量大，如人体中共有3万～4万左右的基因；②分子种类多，如DNA、mRNA、蛋白质、分子、大分子等；③基因表达具有时空性，同一个基因在人和动物的细胞周期中可能实现不同的功能；④真核生物中大多数的基因同时受两个及两个以上的基因调控，它们之间的相互作用具有非线性特性。基因间的相互作用是通过细胞内的各种分子、蛋白质、基因之间的相互作用来实现的。

细胞中分子之间相互作用的几种主要方式。果蝇体节极性基因相互作用网络的实例，4个模块分别对应4种作用方式。研究表明该网络非常保守，其他研究也有相似的结论，如细菌对化学毒性的自应性。

基因表达调控机制的研究具有非常重要的理论和应用价值，有助于解决如下问题：在特定的细胞状态下，哪些基因发生了表达？它们的调控情况是什么？它们的表达量是多少？这些基因的产物对细胞的生理活动会产生什么影响？诸如这些问题的答案将揭示生命奥秘，为疾病的预防和治疗提供理论基础。因此，基因表达调控是后基因组时代系统生物学的一个主要研究内容。

细胞内的几种相互作用方式。1.信号传导；2.基因的转录；3.剪接修饰；4.蛋白质相互作用形成复合体

果蝇体节极性的基因调控网络。四边形代表蛋白质，圆圈代表基因，六边形代表蛋白质复合体

2.基因调控网络模型概述

要建立基因调控网络模型，了解细胞的基因表达调控过程，首先必须全面和系统地测量细胞内的各种分子，然后根据这些数据建立调控网络，但目前的测量技术尚不能得到所有生物大分子和相关物质的数据。而要研究这种复杂的动态系统，最理想的数据是时序数据。但目前大部分的数据是细胞在特定条件下的稳定状态数据，即使有时序数据，其数量也很少，而且其时序性不一定与真实的生物过程的时间尺度相一致。再次，基于DNA微阵列或芯片技术的数据获得过程及生物系统本身都存在各种不确定性。这些因素都增加了建模的难度，因此在实际研究中往往采取折中的策略，如根据数据的类型、研究问题的目的选择不同类型的模型。多参数的精细模型（如微分方程动力模型）能够给出系统的详细状况，如蛋白质的浓度、生化反应的动力学等，为防止模型的过拟合，对数据的精度及数量要求比较高。相反，粗粒度的模型（如各种聚类算法）能够揭示系统的一些宏观行为或现象，如哪些基因的表达具有一定的相关性或执行相同的功能，对数据的精度及数量要求较低。因此，精细模型适用于相对较小的独立系统，而粗粒度模型则适于整个基因组范围。

目前，基因调控网络的建模方法主要有线性模型、贝叶斯网络、布尔网络、神经网络、微分方程及其他随机模型等。每个模型都建立在一定的基本假设上，并对数据有不同的要求。

3.布尔网络模型

1969年，Kauffman提出布尔网络（Boolean networks）模型，用于研究基因调控网络。布尔网络模型中，基因的表达状态被离散化为布尔变量，即开（1）或关（0），它们之间的相互作用通过布尔函数描述。其优点是：能以简单的方式反映网络运行过程中复杂的动态行为。模型的重点是研究系统的基本原理而不是其中生化反应的具体细节。虽然真实的生物系统是一个连续的过程，但使用二进制的逻辑语言来描述基因的开和关、上游调控和下游调控、对外界的变化有无反应等，仍然可以揭示基因间相互作用的逻辑关系，如使用布尔模型能够正确区分同一肿瘤的不同亚型。总之，布尔网络模型可用于：①确定基因之间相互作用的一种定性关系，从而有助于发现药物作用靶点；②研究网络的动态行为及其与生物现象（如细胞的状态）之间的关系；③

研究网络的干预，如采取特定的干预措施，避免细胞从常态转向病态或将病态细胞转向细胞凋亡状态等。

一个有5个基因的布尔函数真值表，ji（1≤i≤3）表示每个布尔函数的调控基因，“-”表示空9.3.1布尔网络

布尔网络G（V，F）由节点集合V={x1，…，xn}和布尔函数集合F={f1，…，fn}组成。每个节点xi是一个布尔变量，0和1分别表示基因表达状态的“关”和“开”；函数集合F表示基因之间的相互作用，基因xi∈{0，1}在t+1时刻的状态是由基因xj1（i），xj2（i），…，xjk（i）在t时刻的状态决定的：

xi（t+1）=fi（xj1（i）（t），xj2（i）（t），…，xjk（i）（t））

网络在t时刻的状态就是所有基因的表达状态向量x·t=x1（t）x2…xn（t），状态采用同步更新（Synchronous updating）方式。显然，对于有n个基因的网络，其状态空间由00…0到11…1的2n个状态组成。

布尔网络函数关系通常用真值表的形式表示。给出了一个有5个基因的布尔网络真值表及其预测基因，除第四个基因仅有一个预测基因外，其他每个基因都有三个预测基因。在布尔网络模型中，给定一个初始状态（State），系统就在布尔函数的作用下从一个状态变迁到另一个状态。系统的状态通常分为两类，即暂态（Transient state）和吸引子（Attractor）。暂态是指在整个运行过程中系统仅经过一次的状态；吸引子指经过一段时间后系统不断重复经历的状态。吸引子有两种：一种是由多个状态组成的有限环，11110和11010；另一种是仅有一个状态的单吸引子，0000、00100、10011和11111.到达同一吸引子的所有暂态组成该吸引子的吸引域（Basin）。布尔网络中的吸引子描述了系统的长期行为。在细胞网络中，一个吸引子通常对应特定的生物意义，如细胞的某种显型或细胞的某个分化阶段、正常细胞或肿瘤细胞等。

一个有5个基因的布尔网络的状态转化图

布尔网络是一种确定性模型（Deterministic model），给定一个初始状态，系统会唯一地达到一个吸引子。由于生物系统本身的复杂性和随机性，一个基因可能在不同的条件下采用不同的布尔函数。同时，实际的数据测量过程也带有各种不确定性。为了引入随机性，2002年

Shumulevich提出了概率布尔网络（probabilistic Boolean networks，pBN），其基本思想如下。

（1）每个基因xi有一组布尔函数Fi={f（i）1，…，f（i）l（i）}，下一时刻的状态由其中一个布尔函数f（i）j确定，每个布尔函数的选择概率c（i）j满足∑l（i）j=1c（i）j=1.因此，pBN实际上由N=∏ni=1l（i）个布尔网络构成，每个网络的选择概率是其对应布尔函数的选择概率的乘积。

一个有3个基因的布尔网络的函数真值表

（2）当选择概率参数λ=1时，每经过一时间，系统都要重新选择一个新的网络来决定下一状态，这种网络模型称为暂态pBN（Transient pBN）；当λ1时，系统以当前网络运行，直到某种外部条件改变再重新选择一个新的网络，这种模型称为上下文相关pBN（Context□sensitive pBN）。

（3）实际生物系统中，由于某种原因，一个基因的状态可能突然发生改变，为了表示这种不确定性，引入一个扰动向量γ∈{0，1}n。当第i个分量为1时，表示该基因的值翻转，否则保持不变。假定扰动向量γ的各分量独立同分布，即pr{γi=1}=E［γi］=p。则下一时间步的状态为

x（t+1）=x（t）□γ1-（1-p）n

fk（x（t））（1-p）n

□表示二进制向量的异或，fk表示pBN的一个基础网络，k=1，…，N。

一个包括3个基因的pBN在无扰动情况下的状态转移概率分布。在pBN中，由于加入了随机扰动，任意两个状态之间都可以一个特定概率直接到达，即网络是各态遍历的（Ergodic）。通常使用马尔可夫（Markov）理论研究系统的长期稳定状态分布（Long□run steady state），它类似于布尔网络中的吸引子，代表了细胞的特定状态。当扰动概率p=0.01时，从000到111的网络稳态概率分布为［0.075，0.0028，0.371，0.076，0.0367，0.0424，0.0672，0.731］。pBN模型不仅更加真实地描述了生物系统，而且也为网络的动态干预研究奠定了基础。

pBN的状态转移概率图，γ={0}n

4.网络的动态行为

在布尔网络或概率布尔网络理论下，给定一个初始状态，系统将按照相应的规则，最终演化到一个吸引子或稳态分布。研究系统在小扰动下的演化行为，即扰动后系统是否会到达同一吸引子或稳态分布，对于揭示细胞网络的稳定性和适应性具有重要的意义，这就是网络的动态行为。通常，将系统的演化方式分为有序（Order）、临界（Criticality）和无序（即混沌，Chaos）状态。稳定性（Stability）是系统有序的一个重要指标，稳定性越高，系统对外界变化的反应越迟钝，缺乏灵活性。适应性（Adaptability）是系统无序的标志，适应性越高，系统对外界变化的反应越敏感，缺乏稳定性。临界状态介于有序和无序之间，它同时具有一定的稳定性和适应性。在变化的环境下，各种生物大分子相互作用的细胞网络既能保持一定的稳定性，又能不断适应环境并作适当调整。越来越多的研究表明：生物系统可能具有临界性。因此，临界性可能是生物系统在复杂的环境中能够基本保持不变，并能协调各种复杂行为能力的一个主要原因。如果这一假设正确，那么研究网络在临界状态下处理信息的机制将有助于揭示细胞网络的奥秘。同时，我们猜想临界状态是否有利于网络的调控。Shumlevich提出网络灵敏度（Sensitivity）的概念，0≤S≤1，S=1和S>1分别对应有序、临界和无序状态（Shmulevich和Kauffman，2004）。利用这一参数可以对网络的机制做进一步的研究。

5.基因调控网络的预测

目前，基于pBN模型的网络预测方法主要有以下几种。

（1）确定系数法（Coefficient of determination，CoD）：是Dougherty和Kim等最早提出的一种预测pBN结构的统计方法。CoD主要评价一个基因对提高目标基因表达水平预测的程度。

（2）最佳子集法（Best□fit extensions）：也是一种常见的方法，其主要思想是对布尔函数类型或变量个数加以限制，然后从这些函数集合中找出与观测数据最一致的一个或多个作为目标基因（Target gene）的预测函数。最近，Liu等提出了一种对网络灵敏度正则化的方法，可以提高最佳子集方法预测精度。

（3）基于贝叶斯网络的方法：贝叶斯网络是另一种广泛应用于基因调控网络研究的概率模型，Zhou等提出了一种基于贝叶斯网络的pBN设计方法。其主要思想是首先根据一定的先验知识，构建网络的拓扑空间，从中选择具有高贝叶斯可信度的网络作为pBN的基础网络。

（4）种子基因生成子法（Seed gene generating method）：由hashimoto等提出。其主要思想是：多尺度复杂网络中有一种普遍现象，即在一个大规模的网络中存在大量由少数节点组成的簇，这些簇中的节点紧密相互作用并执行特定的功能，然后这些簇与簇之间再进一步连接形成更高一级的网络。

6.网络的扰动或干预

基于pBN模型的网络干预方法的研究，为疾病的预防和治疗提供理论基础。对网络扰动或干预（perturbation or intervention）通常可以归结为各种优化问题，主要有以下三种干预方法。

（1）基因状态的扰动：改变网络当前状态中一个或多个基因的状态，使网络从一个新的状态演化。扰动的目的是使网络跳出一个不期望的稳定状态或吸引域，通常分为暂态扰动和恒定扰动，前者指仅在当前时刻翻转一个或多个基因的状态，后者指翻转某些基因的状态并保持这些基因的状态不变。基因状态的扰动并没有改变网络本身，所以不影响网络的长期稳定状态。优化目标是寻找能在有限时间以最大概率到达目标状态的一个或多个调控基因。

（2）网络局部结构的干预：从基因调控网络的角度，肿瘤可能是由于某些基因状态之间某种不平衡引起的，原因可能是某种变异引起基因间作用关系的改变，即网络结构的改变。对于这种情况，必须通过改变网络结构（布尔函数）来改变其长期稳定状态分布。优化目标是寻找一种能够到达期望稳态分布的最少改变布尔函数的方法。

（3）外部控制变量干预：在癌症的治疗中，通过放射治疗、化学药物治疗等使网络状态分布远离失控的增生或凋亡状态。Datta等提出了一种基于外部变量的有限域调控方法，通过动态调整外部变量的状态序列（即环境），逐步引导系统从一种稳定状态分布到达另一种稳定状态分布，网络状态的演化可以写为：

w（t+1）=w（t）A（u（t））

w（t）为当前网络的状态分布向量，u（t）表示当前的外部控制向量，A（u（t））是一个与w（t）和u（t）有关的状态变迁矩阵。优化目标是给定一个有限时间域M，寻求到达特定状态分布的最小代价调控序列w（t），即调控策略。Datta等给出的目标函数为：

Jπ（w（0））=E∑Mk=1Ck（z（k），v（k）+CM（z（M））|z（0）

s。t。p［z（k+1）=j|z（k）=i］=aij

k=0，1，…，M为时间步，z（k），v（k）分别为系统当前状态和当前外部控制向量的十进制表示，Ck（*，*）为任一干预步的相应代价，CM为系统演化的终态代价。当z（M）为目标状态时其代价为0，距离期望的状态越远，其代价越大。由于pBN模型的各态遍历性，目标函数实际上是一种期望值。需要指出的是外部变量也可以是诸如网络中所有基因的一个主调控基因。

7.pBN的应用实例

Kim等利用文献中的基因表达数据，其中共有595个基因，用确定系数法（CoD）预测了神经胶质瘤细胞的pBN网络。Shmulevich等利用种子基因生成法构建了一个包括15个基因的子网络，并研究了其中3个基因Tie□2、TGFB3和NF□κB的稳态概率分布。图中水平坐标表示三个基因的状态，表示基因Tie□2关闭、基因TGFB3打开、基因NF□κB关闭。011和111的稳定状态分布明显高于其他状态，因此，基因TGFB3和NF□κB极有可能存在某种调控关系。Strauch等在随后的试验中证实了基因NF□κB 调控基因TGFB3的表达和细胞的迁移。

三个基因Tie□2、TGFB3和NF□κB的稳态分布

Bittner等在研究黑素瘤的转移时发现，在高竞争性转移细胞和低竞争性转移细胞中，基因Wnt5a的RNA丰度具有明显的差异。Weeraratna等随后进一步发现：①通过基因工程方法增加Wnt5a蛋白可以直接改变细胞的竞争转移能力；②通过一种抗体和Wnt5a蛋白结合来阻止其激活受体，能够明显地降低Wnt5a蛋白诱导细胞转移性。pal等利用这方面的数据构建了一个包括基因Wnt5a的含7个基因的调控网络，并采取两种干预方法，即直接使用Wnt5a的抑制蛋白和另外一种基因pIRIN。结果表明，与没有任何干预相比，在特定干预步长，从任一起始状态

开始，两种干预方法最终都以很高的概率保证基因Wnt5a是下游调控的。

一个包含基因Wnt5a的含7个基因的子网络

从任一初始状态开始到达目标状态的代价分布。

8.贝叶斯网络的概述

贝叶斯网络（Bayesian network）是一种概率图模型，可从不完全或不确定的知识或信息中推理，适用于表示和分析不确定的事物。贝叶斯网络B=（G，θ）由两部分组成：①以随机变量X1，…，Xn为顶点的有向无环图G（Directed acyclic graph，DAG）；②以一组局部条件概率分布p（Xi|pa（Xi））（Conditional probability table）表示节点间的相关关系的强度，其中pa（Xi）表示变量Xi的父节点。贝叶斯网络模型的基本假设是：给定变量Xi的父节点pa（Xi），其概率分布与其他非父节点无关。因此，所有变量的联合分布可以简化为：

p（X1，…，Xn）∏ni=1p（Xi|pa（Xi））

一个包含4个变量的贝叶斯网络，给定条件概率表后就可以直接计算网络的联合概率分布，如

p（a=U，b=U，C=D，d=u）

=p（a=u）p（b=u|a=U）p（c=D|a=U）p（d=U|b=U，c=D）

=0.7×0.8×0.4×0.7=0.16

一个包括4个变量的贝叶斯网络

根据贝叶斯网络可以进行由原因到结果的推理或从症状诊断可能的原因。从知识表示的角度，贝叶斯网络提供了一种紧凑的知识表示方法。所有可能的联合概率分布，需要24个列表；而在贝叶斯网络模型中，只需9个条件列表。随着网络规模的增大，这种存储效果更加明显。目前，贝叶斯网络已被广泛应用于基因调控网络的研究，其中的变量或节点通常表示基因。

9.贝叶斯网络的学习

贝叶斯网络实际上蕴涵的是一组不相关的规则集合。显然，两个不同的网络可能蕴涵相同的不相关规则，当两个网络蕴涵相同的不相关规则时称它们等价。因此，从一组数据学习往往是一个贝叶斯网络等价类，而不是一个特定的网络。从一组等价网络类中进一步确定特定的网络是一个非常困难的问题，可行的方法之一是通过目标干预试验来缩小候选范围。贝叶斯网络的学习是指给定一个数据集D和某些特定的先验概率，学习最符合数据的候选贝叶斯网络，主要包括以下两部分。

（1）模型选择：从给定数据集D选择最可能的相关关系，即图G。

（2）参数拟合：给定数据集D和图G，求Xi和pa（Xi）最可能的条件概率p（Xi|pa（Xi））。

其中参数拟合相对比较简单，可以使用最大似然估计（Maximum likelihood estimation）或期望最大（Expectation maximization）算法求解。相反，模型选择则要困难得多，通常通过一个贝叶斯记分函数来评价模型G拟合数据集D的程度：

BayesianScore（G|D）=logp（G|D）

=logp（D|G）+logp（G）+c

logp（D|G）表示边缘似然度，即对任一模型G假定数据集D在所有参数θ上等概率出现；logp（G）表示模型G的先验概率，用来约束模型的复杂度；模型越复杂其出现的概率越小，从而可以防止过学习现象的发生；c是一个与模型无关的常数。

搜索最佳模型是一个非常困难的Np□完全问题，为了降低记分函数的计算复杂度，通常假定BayesianScore（G|D）可以分解为单个节点Xi的记分之和。这样可以根据一条边的增加、删除或反向对BayesianScore（G|D）的影响来决定其可能的状态。实际的应用中还经常使用一些启发式的概率搜索策略，如爬山法、模拟退火算法、马尔可夫蒙托卡罗采样方法等，进一步降低计算复杂度。此外，对每个节点根据其相关性限制父节点范围，从而减小搜索空间。

10.动态贝叶斯网络

贝叶斯网络是一个有无环向图，而在很多实际的生物系统特别是基因调控网络中，反馈回路是其中一种重要的调控机制。为了克服这一局限，人们在传统贝叶斯网络模型的基础上又提出了动态贝叶斯网络。动态贝叶斯网络（Dynamical bayesian network，DBN）是贝叶斯网络在时间变轴上的扩展，反映了一系列变量随时间变化的情况。

动态贝叶斯网络包括两部分：初始的贝叶斯网络B0=（G0，θ0）和转移贝叶斯网络B1=（G1，θ1）。为了处理方便，动态贝叶斯网络通常满足以下两个假设条件。①网络拓扑结构不随时间发生变化，当t=0时，变量间的相关关系为pa（Xi（0））□X（0）；当t>0时，变量间的相关关系为pa（Xi（t））□X（t-1）。②满足一阶马尔可夫条件，p（X（t）|X（0），…，X（t-1））=p（X（t）|X（t-1））。满足上述条件的动态贝叶斯网络是贝叶斯网络在时间序列上的展开，它反映了变量X在所有可能时间域上的状态的联合概率分布：

p（X（0），X（1），…，X（T））=p（X（0））∏Tt=1p（X（t）|X（t-1））

DBN（B0，B1）的例子，B0和B1分别表示分布时初始网络和转移网络。

DBN（B0，B1）的例子9.4.4贝叶斯网络应用

目前，贝叶斯网络已经被广泛应用于基因调控网络的预测（Murphy和Mian，1999；Kim等，2003；perrin等，2003；Zou和Conzen，2005）。1998年，Spellman等使用微阵列杂交方法研究酵母细胞在不同周期调控基因表达数据，确定了其中800个基因的表达随细胞周期的变化有明显变化，并使用聚类算法将其中的250个基因归为8个类别。2000年，Friedman等将这些表达数据离散化为－1，0，1，分别表示基因表达明显很低、没变化和显著表达；用互信息（Mutual information）描述两个基因间关系的强弱，最后用Bootstrap方法确定所得模型的显著性。

基因SVS1的一个局部关系贝叶斯网络。有向边的粗细表示作用关系的强弱

贝叶斯网络的特点之一是具有融合不同数据的能力。2003年，Imoto等将基因表达数据、蛋白质□蛋白质作用数据、蛋白质□DNA作用数

据、转录因子结合位点及从其他参考文献得到的相关信息等集成起来，用以提高贝叶斯网络的学习效果（Imoto等，2004）。该方法的缺点是要求其他信息必须为对称类型。2005年，Bernard等在此基础提出了一种融合转录因子结合位点信息的动态贝叶斯网络的学习方法。仿真结果表明，该方法能够明显提高学习的效果。利用文献中的数据，他们得到有25个基因的局部网络。

酵母细胞周期的局部基因调控网络。实心椭圆表示有转录因子结合位点信息，空心圆圈表示没有转录因子结合位点信息，实线表示实验或相关文献验证的关系，虚线表示模型预测待验证的关系

基因调控网络是细胞网络中的一个重要组成部分，应用基因表达数据研究基因调控网络的方法大概可以分为三类，即以微分方程为代表的精细模型、以聚类方法为代表的粗粒度模型，以及介于两者之间的布尔网络和贝叶斯网络模型。本章主要介绍了布尔网络和贝叶斯网络模型及其在基因调控网络研究中的应用。harri等证明了概率布尔网络（pBN）和动态贝叶斯网络的模型等价性，这一研究可以将这两种模型的优缺点结合起来，取长补短。总的说来，这两种模型具有以下特点：

（1）都能在某种程度上揭示基因之间的作用关系，布尔网络使用的是布尔方程，贝叶斯网络采用的是条件概率。

（2）都能处理数据获得过程及生物系统本身所具有的不确定性。

（3）概率布尔网络模型能够揭示生物过程的动态行为，适合于网络干预、药物作用位点等方面的研究；贝叶斯网络适合于推理和诊断，并表示基因之间作用关系的强弱。

（4）在目前的小样本条件下，贝叶斯网络的学习要比布尔网络更复杂，动态贝叶斯网络的学习则更难。

基因调控路径和网络的研究对于生物信息学来说是一个重大挑战，它不仅需要有效的数据挖掘方法来整合和充分利用海量的异质和异源数据，还需要对基因调控的生物学知识有深层次的理解。基因调控网络反映的是基因之间的相互关系，这些关系还可以反映在DNA序列、转录因子与顺式调控元件相互作用、蛋白质□蛋白质相互作用、蛋白质在细胞中的定位等层面。因此，除需要完善和开发新的模型之外，将

更多的生物学背景知识与表达数据相融合起来研究基因调控网络是一个必然的选择。随着生物信息学和分子生物信息检测技术的发展，我们对基因调控网络的认识会日益丰富和全面。

第29章 因特网上的数据库和工具

近年来，高通量技术发展迅速，积累了很多数据，所以网络数据库越来越多，包括全基因组数据库、蛋白质相互作用数据库以及很多生物代谢途径数据库等。另外，随着系统生物学的发展，分析软件，包括介绍分析软件的算法、模拟工具也越来越多。因特网上主要的数据库和系统生物学所用到的工具，我们不做一一介绍，有兴趣的读者可以自行网络查询，了解各个数据库和工具的功能。另外，Nucleic Acids Research杂志每年1月份有关于生物数据库的专刊，7月份有关于网络数据分析服务的专刊，因特网上的数据库和工具

种类名称功能下载网址算法Clover寻找转录因子结合位点MEGN从大规模基因表达图谱中推导出等同的基因网络Gossip基因组的自动化功能性翻译netOptGene基于计算机（进化规划算法）的代谢工程平台

metabolitcs寻找表达中与大部分变换相关的代谢组dkTFGossip利用Gossip将转录因子结合位点与因子结合net数据格式mzDatamzData的目的是将当前大量的质谱数据组合成一个整体XML SchemamzXML质谱的标准信息输出公式XML Schema数据库BIND生物分子相互作用网络数据库，包括经过整理的一系列相互作用定量生物模型的整理

种类名称功能下载网址数据库DBTSS转录起始位点的数据库DIp相互作用蛋白数据库，包括经过整理的一系列蛋白□蛋白相互作用EnsEMBL有注解的基因组和perl接口

orgEpD真核生物启动子chFANTOM有注解的小鼠转录组jpGpMDB串联质谱数据库，酵母、苍蝇等模式生物遗传与蛋白相互作用的综合数据库基因组的存储、定位和比较hpRD人蛋白参考数据库，包含蛋白结构域，翻译后修饰，蛋白□蛋白相互作用网络和疾病关联的信息整合数据库文本挖掘，蛋白质上的转录因子结合位点的图谱seKarma来自复杂array平台和生物体的不同标识符的定位plOpen proteomics DatabaseOpD是一个公开的数据库，可以存储和传播质谱的蛋白质组学数据pEDRoDB存储、搜索和传播实验蛋白质组学数据uk/pRIDE对蛋白质组学数据来说，pRIDE是一个集中的、标准规范的、公开的数据库，两个阵列平台之间的定位brainarrayResourcerer人类、小鼠、大鼠微阵列基因标识符的定位

magic/rl。plSGDSaccharomyces基因组数据库。S。cerevisiae非常完整的基因组orgSTRING已知的和预测的蛋白□蛋白相互作用数据库2D□pAGE二维聚丙烯凝胶电泳数据库Transfac转录因子结合位点的图谱

第30章 组学数据库

1.美国国家生物技术中心（NCBI）数据库

美国国家生物技术中心（NCBI）数据库提供了广泛应用于生物学研究的一些数据库。其中最重要的是分子数据库，提供了核酸序列、蛋白、基因、分子结构和基因表达的信息。除此之外，还有一些由科学文献组成的数据库。其数据涵盖超过57万物种（2011年为止），拥有1000种以上生物（已经测序好的和正被测序的）的基因组序列和相应的基因图谱。它还与欧洲的EBML和日本的DNA数据库每天进行数据交换，以保证数据的全面性和世界覆盖率。这些数据都可以通过NCBI主页的Entrez搜索引擎进行查询。Entrez将DNA和蛋白的数据与物种分类、基因组定位、蛋白结构、蛋白区域信息和文献整合在一起。

在NCBI的核酸序列数据库中，重要的有基因序列数据库（GenBank）、参考序列数据库（RefSeq）、Unigene数据库和Entrez基因数据库。GenBank（版本182，2011年2月），每两个月发布一次更新（Burks等，1985），含有1.3亿条已知核苷酸序列，1200亿个碱基对。参考序列数据库（RefSeq）是一个二级的非冗余数据库，包括重要模式生物的核苷酸和蛋白产物的序列。Unigene数据库则是把表达序列标签（EST）和全长mRNA序列组织成簇，每一簇代表一个特殊生物个体的一个独特的已知或推断的基因。Entrez基因数据库中的每一个Entrez基因的记录包含了某一个特定基因的很多信息：①包括用图像来显示总结该基因在基因组中的位置，包括其内含子、外显子结构以及邻近基因；②提供mRNA序列的图像显示，并且这些mRNA序列可以显示其他生物特征，如编码区、单核苷酸多态性等；③提供它的表现型的一些信息；④提供蛋白序列的连接以及其保守区域的功能信息；⑤提供与其他相关数据库的连接，如突变数据库的连接。Entrez基因是LocusLink 的延续。

NCBI还有文献数据库pubMed和OMIM（在线的人类孟德尔遗传）。pubMed是由生物医学文献的引用和摘要组成的数据库，引用来源于MEDLINE（http：//medline。cos。com）和其他生命科学杂志。OMIM是由人类基因和遗传病组成的目录，包含了大量的文本信息和与科学文献的链接。

另外，NCBI还有一个重要数据库——基因表达数据库（Gene expression omnibus，GEO）。GEO提供了一个高通量定量实验数据的存储网址，其数据包括mRNA表达水平的数据、基因组DNA的数据（ArrayCGh、ChIp□Chip和SNp），甚至包括蛋白表达数据库，比如质谱的多肽表达谱数据库。另外，它还包括一些不基于DNA芯片技术的表达数据库，比如SAGE表达数据库（Velculescu，1994）和MpSS数据库（Brenner，2000）。随着第二代测序技术的发展，这种不基于DNA芯片的定量数据将会越来越多。GEO的重组结构包括平台（即定量分析方法）、样品名称和研究系列。至2011年3月为止，GEO含有8490种平台、539339个样品以及21830个研究系列。

除以上数据库外，NCBI同时也提供了许多用于操纵序列数据的工具，包括在给定的序列数据库中进行单序列或多序列比对的工具。

第31章 GEO基因表达数据的存储结构

1.欧洲生物信息学研究所的数据库

欧洲生物信息学研究所（EMBL□EBI）提供了很多生物相关的数据库，包括核酸序列、基因和基因组方面的数据库，蛋白数据库，蛋白相互作用数据库，基因表达数据库。它还提供了biomart数据管理和提取系统，对系统的、大规模的数据提取很有帮助。

（1）EMBL核酸序列数据库

EMBL核酸序列数据库整合、组织和分配公共资源中的核酸序列，并且每天与日本DNA数据库（DDBJ）和GenBank中的数据保持同步。EMBL与DDBJ、GenBank组成了全球最重要的三大核酸序列数据库。

（2）Ensembl

Ensembl工程正在开发和维护一个用于管理基因组的序列和真核生物基因组的注释的系统。注释就是指用计算和实验的方法对基因组的特性加以描述。首先，基于组装好的基因组序列对基因组特性进行描述，如重复序列基序、保守区域或单核苷酸多态性（SNp），并对基因进行预测，得到如内含子和外显子等结构元件。SNp是常见的个体间的DNA序列差异，仅仅是单个核苷酸的改变。再者，注释也包括基因编码的蛋白的功能域信息和基因产物在生物体中的作用。

Ensembl的核心组件是一个关系数据库，其存储了组装的基因组序列和由Ensembl自动序列注释软件所产生的注释，这些序列和注释来源于外部的基因组序列和数据。至2004年9月，Ensembl已提供许多物种的基因组注释信息，包括脊椎动物（人类、猩猩、大鼠、小鼠、河豚、斑马鱼和鸡）、节肢动物（蚊子、蜜蜂和果蝇）以及线虫。

（3）Interpro

是一个有关蛋白质信息的数据库，其由蛋白家族、功能域和功能集团的信息组成（Biswas等，2002；Mulder等，2003）。它整合了许多常

用的蛋白信息数据库，并提供了非常强大的工具对通过测序得到的新的蛋白或预测的蛋白进行自动或手工注释。除此之外，Interpro中蛋白信息记录与Gene Ontology（GO）相关联，同时也与Uniprot中蛋白记录相链接。

2.Swiss□prot，TrEMBL和Uniprot

组学数据库中，除核酸数据库外，还有各种各样的蛋白序列数据库，从简单的序列数据库到通用的次级数据库，提供了许多物种的非常有用的信息。Swiss□prot就是最主要的蛋白数据库之一，由瑞士生物信息学研究所（SIB）和欧洲生物信息学研究所（EBI）共同维护，提供了大量的注释信息，至2011年3月，其已经包含525997条蛋白记录，包括蛋白来源（基因名字和物种）、氨基酸序列、蛋白功能和定位、蛋白的功能域、四级结构、文献相关信息、蛋白相关疾病和其他详细信息；此外，Swiss□prot也提供了许多外部数据库的信息，例如核酸序列数据库（DDBJ/EMBL/GenBank）、蛋白结构数据库、蛋白功能和家族数据库、疾病相关数据库等。

TrEMBL使用自动注释工具进行注释，包括了将核酸数据库中的蛋白编码序列理论翻译后得到的注释信息。TrEMBL中的记录与Swiss□prot中的记录并不重复。

Uniprot是提供蛋白相关信息最全面的数据库。Uniprot由Swiss□prot、TrEMBL和pIR中的信息所组成。Uniparc，作为Uniprot的一部分，是最全面的、可以被公共访问的非冗余蛋白序列数据库。

第32章 生物途径网络数据库

1.京都基因和基因组百科全书（KEGG）

京都基因和基因组百科全书是一个关于基因、蛋白质、生化分子、反应和代谢途径的知识库。KEGG由四大部分构成：KEGG基因数据库（KEGG genes），由GENES、SSDB和KO数据库组成；KEGG 配体数据库（KEGG ligand），由COMpOUND、GLYCAN、REACTION和ENZYME数据库融合成LIGAND数据库；KEGG生物途径数据库（KEGG pathway，由pAThWAY数据库组成的蛋白质相互作用的网络）；KEGG BRITE 数据库，生物网络中的有机联系和层次关系。此外，KEGG数据库在四个水平上层次性地分成了许多范畴（Categories）和亚范畴。最高层次的五大范畴是代谢、遗传信息加工、环境信息加工、细胞过程和人类疾病；而亚范畴，如代谢的亚范畴，有碳水化合物、能量、脂肪、核酸和氨基酸的代谢；亚范畴又可以被细分成不同的代谢途径，如糖酵解、三羧酸循环、嘌呤代谢等；最后，第四个水平与KO 同源关系（KEGG orthology）记录相关联。KEGG实体（KEGG objects）：KEGG给予每一个生物实体即数据库中的每一条记录一个特定的编号。

KEGG基因数据库（KEGG genes）提供了国际基因组测序项目所产生的基因和蛋白质的信息。单个基因的信息存储在GENES数据库中，GENES数据库由GenBank、NCBI RefSeq数据库、EMBL数据库和其他公共的物种特异的数据库半自动化生成。然后，GENES数据库中的每条记录分别被赋予了特定的K号码。SSDB数据库包含编码蛋白基因的氨基酸序列的同源性信息，包括直系同源和旁系同源基因的同源性信息。KEGG 释放版本含有15524KO同源关系簇，7505856基因，包含了154真核生物物种，1456种细菌及117种古生菌（Archaea）。

KEGG 配体数据库提供了与细胞过程相关的化合物和化学反应的信息（Kanehisa，2002）。它包括16520种以上化合物（在内部以C number表示，如C00001代表水），单独的碳水化合物的数据库（约11000条记录；以G开头的数字表示，如G10481代表纤维素），8880种以上化学反应（用R number表示，如R00275代表了由激发态的超氧化物变成过氧化氢的反应），16520种小分子和代谢物，10979种糖链和5708种

酶（以EC number表示）。同时它还提供了13342对配对反应物（Reactant pairs）。

KEGG 生物途径数据库提供了由生物途径和蛋白复合体所组成的蛋白质相互作用的信息，提供了416条参考生物途径衍生的156647条生物途径。KEGG生物途径可以通过网页界面，用基因、蛋白质、化合物名称等来查询；也可以通过下载整个数据库，或者通过SOAp服务器来获取数据（SOAp是基于XML的计算机软件间信息交换软件）。

2.其他生物途径数据库

其他生物途径数据库包括：Biopa，panther，Sigpath。BioCarta的特点是以优美的图来显示经典的生物途径，即为p53的生物通路。

p53生物通路。图片来自BioCarta网站，经Biocarta许可复制商业生物途径数据库如下。

（1）Ingenuity 生物途径数据库：其生物途径和网络是基于Ingenuity知识库的。Ingenuity知识库是从文献全文中提取出来的，包括基因、药物、化合物、疾病、信号转导和代谢途径等。

（2）GeneGo：包括500多个经典的信号传导和代谢途径，且内容已被人工验证。

（3）pathway Studio：除了常见的生物途径分析功能外，其MedScan软件还可以直接从生物医学文献中提取生物相互作用的信息。

3.蛋白相互作用数据库

目前，通过各种生物实验手段验证的蛋白相互作用的数据越来越多，加上各种蛋白相互作用预测算法的发展，由此发展了一些常用的蛋白相互作用数据库。数据库的网址及简要说明。

名称说明网址DIp生物实验验证过的蛋白间相互作用，人工验证dipBIND收集文献报道过的蛋白间相互作用偏重蛋白物理及基因关联的相互作用MINT偏重收集来自哺乳动物的蛋白间相互作用hRpD只针对人类蛋白的相互作用STRING包含已知的及预测的蛋白间相互作用

DIp收集了生物手段验证过的蛋白□蛋白相互作用数据。至2010年6月，DIp 包含了274个物种的21889个蛋白，这些蛋白形成了69171个相互作用。

BIND（the Biomolecular interaction network database）主要从文献中收集蛋白□蛋白相互作用数据，目前已有188571个相互作用数据。

BioGrid收集了22个物种的30244个蛋白的238339个相互作用数据。

MINT（the Molecular interactions databases）主要偏重收集来自哺乳动物的蛋白□蛋白相互作用数据，现已收集30563个蛋白形成的83553个相互作用数据。

hRpD（human proteins database reference）是专门针对人类蛋白（39194个相互作用）的数据库（peri等，2004）。

STRING数据库是由欧洲生物分子中心开发的蛋白□蛋白相互作用数据库，它不仅收集已知的相互作用，还包含了预测出来的蛋白□蛋白相互作用。

这些蛋白□蛋白相互作用数据库都有很友好的网站页面（Web interface），可以通过这些网站搜索蛋白的关键词或做序列比对（Sequence blast），以找到能与该蛋白相互作用的其他蛋白，高级用户还可以下载整个数据库并本地化使用。

第33章 基因本体注释和分类

科学知识的积累是一个分散的、平行的过程。在不同的生物体中，功能相同的基因往往被赋予了不同的名字，人们对基因的定位和功能的描述也可能存在着很大差别，如蛋白降解和酶解。对新基因命名和功能的描述也没有系统化和标准化，这样就很难在不同的数据库和生物体中查找某个基因的相关信息。因此，在科学发展，特别是在系统生物学水平上，对基因进行系统的本体化分类和功能注释是必不可少的。为了解决这个问题，从1998年开始，酵母基因组数据库（SGD）、小鼠基因组数据库（MGD）和果蝇基因组数据库（FlyBase）的科学家联合开展了基因本体（Gene ontology，GO）的项目。GO联盟（Consortium）还与其他数据库进行合作，开发和维护了本体自身，注释基因产物，建立合作数据库中基因及基因产物与基因本体的联系，并开发了用于创建、维护和使用基因本体的工具。下面就简单介绍基因本体的一些基本概念。

基因本体包含了基因产物参与的生物过程（Biological process），所属的细胞组件（Cellular component），所发挥的分子功能（Molecular function）三方面注释信息，并将不同的功能概念组织成有向无环图（Directed acyclic graphs，DAG）。可以这么说，基因本体是使用有控制的词汇表和严格定义的一个概念，以有向无环图的形式统一表示各物种的基因功能分类体系，从而较全面地概括了基因的功能信息，纠正了传统功能分类体系中常见的维度混淆问题。在基因表达谱分析中，GO常用于提供基因功能分类标签和基因功能研究的背景知识。利用GO的知识体系和结构特点，可以发掘与基因差异表达现象关联的单个特征基因功能类或多个特征功能类的组合。

基因本体中最基本的概念是术语（Term），至今为止总共有3万多条GO术语。GO里面的每一个条目都有一个唯一的数字标记，如GO：1234567，还有一个术语名来表示这个条目，比如“细胞”、“成纤维细胞生长因子受体结合”或者“信号转导”。每个术语分别属于三大范畴本体（生物过程、细胞组件和分子功能）中的一个。一个基因产物可能出现在不止一个细胞组件中，也可能会在很多生物过程中起作用，并且在其中发挥不同的分子功能。比如，基因产物“细胞色素c”用分子功能术语描述是“氧化还原酶活性”，而用生物过程术语描述就是“氧化磷酸化作用”和“细胞死亡的诱导”，最后它的细胞组件术语是“线粒体基

质”和“线粒体内膜”。本体中的各个术语之间有两种相互关系，它们分别是“is_a”关系和“part_of”关系。“is_a”关系是一种简单的包含关系，比如“A is_a B”表示A是B的一个子集，又如“nuclear chromosome is_a chromosome”也是“is_a关系”。“part_of”关系要稍微复杂些，“C part_of_D”意味着如果C出现，那么它就肯定是D的一部分，但C不一定会出现。比如“nucleus part_of cell”，核肯定是细胞的一部分，但有的细胞没有核。前面说过，本体的结构是一个有向无环图，有点类似于分类树，不同点在于本体的结构中一个术语可以有不止一个父节点。比如生物过程术语“己糖生物合成”有两个父节点，它们分别是“己糖代谢”和“单糖生物合成”，这是因为生物合成是代谢的一种，而己糖又是单糖的一种。我们再来看一个实际的例子，来体会这个概念是怎么实现的。例如，FlyBase（果蝇数据库）中的超氧化物歧化酶，在细胞组件的范畴属于GO术语“细胞质”，在生物过程的范畴属于GO术语“防御反应”和“个体生命周期的决定因子”，在分子功能的范畴上属于GO术语“抗氧化活性”和“铜锌超氧化物歧化酶活性”。GO术语“细胞质”有一个父节点“细胞内”，“细胞内”又有一个父节点“细胞”，“细胞”最终与细胞组件范畴相连。超氧化物歧化酶的其他基因术语与三大范畴相连有着类似的层次结构。

通常来说，单个基因的表达情况的改变不足以反映特定功能/通路的整体变化情况。因为类似于人类社会的组织结构，生物体的功能的实现绝不仅仅是依靠一两个基因功能的改变来实现的。因此过分强调单个基因表达变化，会在后期结果处理中严重干扰对于结果的合理分析，导致偏倚性加大，而且是无法避免的。因此利用GO的结构体系，把参与同样功能/通路的基因进行“功能类”层面的抽象和整合，提供比基因更高一层次的抽象结论，对理解疾病的发病机制或药物作用机制等更有帮助。但是该方法也存在一定的不足，由于生物体内部的调控网络可能具有“无尺度网络”的特点，个别功能重要的基因（主效基因）具有“hub节点”的重要特性，它的功能改变可能对于整个网络来说是至关重要的，在这点上，这些主效基因又具有一定的“自私独裁”特点。而“功能类”之观点模糊了这种差别特性，过于强调“共性”，而忽视了“个性”，这也是“功能类”的一个不足之处，这就需要结合相关的生物学知识来实现。

为了有效地使用基因本体，人们已经开发出许多不同的生物信息学工具，这些工具可以在GO网站上找到，其中包括基于网页的和独立运行的GO浏览器、编辑器、微阵列相关工具及很多用于特殊工作的工具。

接下来我们分别简单介绍一下AmiGO、GoFish和GOstat三个常用的工具。AmiGO是由GO联盟维护的基于网站的GO浏览器，其界面。首先，AmiGO可用于浏览GO术语。GO术语后圆括号中的数字代表了当前选定的数据库中与该术语关联的基因产物的数量。GO后面的七位数字是GO□ID，将GO术语与一个特定的标识符相关联。可以选定一个或多个物种，及一个或多个数据来源来限制搜索的结果。点击GO层次结构的叶子结点（例如未知的生物过程）会弹出一个窗口，显示选定的数据库中与该术语关联的基因。在AmiGO中，人们除了浏览GO术语树，也可以搜索特殊的GO术语以获得与之关联的基因产物或搜索基因产物以找到与之关联的GO术语。最后，还可以得到一个图形化的视图，显示选定的GO术语在本体树上的定位。但是AmiGO的搜索选项非常有限，仅仅能够搜索由OR函数（高级查询下）进行连接的一些术语。为了弥补这个不足之处，相应开发了Java应用程序GoFish。用户可以在GO树上选择不同的GO术语，然后组成复杂的布尔表达式进行高级搜索。例如，我们可以在FlyBase中查找抗氧化的或者在防御免疫中起作用但又不在细胞编程性死亡中起作用的基因产物。当用户选定了特定的数据库后，GoFish的Java应用程序可以下载针对选定的数据库的GO术语和联系，从而往往使GoFish的反应时间比AmiGO快很多。另外一个工具是GOstat，它是利用基因本体系统来检测感兴趣的一组基因中特定的GO术语在统计学上是否出现频率增加。GOstat能够将一列基因与用户提供的一列参考基因或选定的一个数据库中所有的基因进行比较。然后利用Fisher检验或卡方检验鉴定观察的GO术语频率的不同是否具有显著性。程序的输出结果是一列p值，显示输入的该列基因的GO术语的特异性，结果可通过文本或html文件形式获取。用户也可以开发自己的程序，因为所有与GO相关的文件如GO定义或是数据库注释的文件都可以从GO网站上下载得到。

AmiGO网站，一个由GO联盟开发的基于Web界面的GO浏览器。通过它可以浏览GO的层次结构和在数据库中搜索特异性的GO术语或基因产物。GO术语后面括号里的数字代表选定的数据库中与该术语关联的基因产物的数量

读累了记得休息一会哦~

公众号：古德猫宁李

- 电子书搜索下载
- 书单分享
- 书友学习交流

网站：沉金书屋 https://www.chenjin5.com

- 电子书搜索下载
- 电子书打包资源分享
- 学习资源分享

第34章 蛋白结构数据库pDB

世界蛋白数据库是由美国RCSB pDB、欧洲pDBe和日本pDBj（Nakamura等，2002）组成的，旨在成为全世界大分子结构的存储、数据处理和分发中心。生物大分子（如蛋白质和核酸）是折叠成特殊的三维结构来执行功能的。三维结构可以用X射线晶体衍射法或核磁共振（NMR）方法来阐明，以助于理解分子的生物活动及其间可能的相互作用。至2010年6月，pDB存储了5万多种分子的结构。RCSB pDB的主服务器和世界各地的镜像服务器可提供数据库的检索和下载服务，以及关于pDB数据文件格式和其他文档的说明，pDB数据还可以通过发行的光盘获得。世界蛋白数据库可以通过许多不同的方法浏览和查询，如输入雄性激素受体（Androgen receptor），即可找到雄性激素受体的结构，包括它与不同的配体结合时的结构。

pDB网站上显示的雄性激素受体与雄性激素结合时的结构

对分子结构互动，可以通过不需下载插件的方式显示（如用KiNG、JMol、WebMol、pyMol），也可通过下载RasMol或Swiss pDB Viewer插件来显示。NCBI也提供了一个分子模型数据库（the Molecular modeling database，MMDB），该数据库含有4万多个蛋白和多核苷酸的三维结构，并且这些蛋白和多核苷酸与NCBI的其他数据（如顺序、文献、物种的系统分类等）联系在一起。

序列相似的蛋白可能具有相似的三维空间结构，结构相似的蛋白质可能有共同的祖先。这种关系对于了解蛋白质的进化和发展是非常关键的，同样对于分析基因组序列数据也是非常重要的。为了分析蛋白质序列与结构之间的关系，认识不同折叠结构的进化过程，人们需要对蛋白质结构进行分类并建立结构分类数据库。因此，我们将两个最重要的蛋白质结构分类数据库SCOp（Murzin等，1995）和CATh（Orengo等，1997）介绍如下。

（1）SCOp数据库

SCOp数据库（Structural classification of proteins）的目标是提供结构已知的蛋白质之间结构和进化关系的信息，所涉及的蛋白质为蛋白质结构数据库pDB中的所有蛋白质，每年更新一次，目前版本是1.75。

SCOp的结构分类主要是人工完成的，通过图形显示器观察和比较蛋白质结构，并借助于一些软件工具进行分析，如同源序列搜索工具等。SCOp数据库从不同层次将蛋白质结构分成许多层次，以反映它们结构和进化的相关性，自下而上通常将它们分成家族、超家族和折叠类型。SCOp数据库的第一个分类层次为家族，其依据为蛋白质之间的序列相似性。通常将相似性在30%以上的蛋白质归入同一家族，即它们之间有比较明确的进化关系。如果序列相似性较低，但其结构和功能特性表明它们有共同的进化起源，则将其归为超家族。无论有无共同的进化起源，只要二级结构单元具有相同的排列和拓扑结构，即认为这些蛋白质具有相同的折叠方式。结构的相似性主要依赖于二级结构单元的排列方式或拓扑结构。自上而下，SCOp首先从总体上将蛋白质分类，例如全α型、全β型、以平行折叠为主的α/β型、以反平行折叠为主的α+β型；然后将属于同一结构类型的蛋白质按照折叠、超家族、家族层次组织。例如，SCOp 1.75版本有46456个全α型蛋白质，该结构类型下有284个折叠类。在284个折叠类中，第一个超家族是类球蛋白；类球蛋白又包含4个家族，其中第一个家族包含6个结构域；每个结构域下有很多不同物种的蛋白质成员。

（2）CATh数据库

CATh数据库也是一个重要的蛋白质结构分类数据库，其含义为类型（Class）、构架（Architecture）、拓扑结构（Topology）和同源性（homology）。它由英国伦敦大学UCL Orengo教授小组开发和维护。与SCOp数据库一样，CATh数据库的构建既使用计算机程序进行自动分类，也进行人工检查。CATh数据库的第一层分类依据是蛋白质结构域。与SCOp不同的是，CATh将蛋白质分为4类，即α主类、β主类、α□β类（α/β型和α+β型）和低二级结构类。低二级结构类是指二级结构成分含量很低的蛋白质分子。CATh数据库的第二层分类依据为由α螺旋和β折叠形成的超二级结构排列方式，而不考虑它们之间的连接关系，形象地说就是蛋白质分子的构架，如同建筑物的立柱、横梁等主要部件，这一层次的分类主要依靠人工方法。第三层分类依据为拓扑结构，即二级结构的形状和二级结构间的联系。第四个层次为结构的同源性，它是通过序列比较再用结构比较来确定的。CATh数据库的最后一个层次为序列（Sequence）层次，在此层次上，只要结构域中的序列同源性大于35%，就被认为具有高度的结构和功能相似性；对于较大的结构域，则至少要有60%与小的结构域相同。

第35章 TRANSFAC和EpD转录子数据库

基因的转录活性及生物学功能，受DNA启动子区域转录因子的相互作用所控制。因而，更好地理解这些相互作用对于系统地描述细胞的生物学过程非常有用。启动子数据库，如TRANSFAC（Wingender等，1996）或EpD（Cavin perier等，1998），为相关的科研人员提供了非常有用的信息。

（1）TRANSFAC

TRANSFAC包括真核转录因子（Transcription factor）及其DNA结合位点的信息。该数据库需要注册才能使用，但学术用户只需使用非商用邮箱进行注册就可以免费使用。TRANSFAC由SITE、GENE、FACTOR、CELL、CLASS和MATRIX六种不同的ASCII文本文件所组成。SITE包括DNA上转录因子结合位点的信息，GENE给出了位点所属基因的简要描述，FACTOR描述了与位点结合的转录因子，CELL给出了结合于一给定位点的转录因子的细胞类型的信息，CLASS提供了转录因子所属类的信息，MATRIX给出了转录因子结合位点的核苷酸分布矩阵。SITE表中的字段名字以及简短描述。

字段字段描述AC登录号ID标识符DT数据与作者TY序列类型DE基因描述RE基因区域SQ调控元件序列EL调控元件名称SF转录因子结合位点起始位置ST转录因子结合位点终止位置S1转录起始位点外的结合起始位置BF转录因子信息MX矩阵号OS物种OC物种分类SO转录因子来源MM方法CC注释DR外部数据库链接RXMedline IDRN参考文献号码RA参考文献作者RT参考文献标题RL参考文献数据每一组不同的表可以被全局地搜索或者局限于某一特殊的字段。可以使用通配符“*”和“？”进行文本搜索，如使用“*”可以搜索出所选定表中的所有记录。当前，TRANSFAC公开的版本包括14490条FACTOR记录（包括miRNAs）、31039条SITE记录、41859条Factor□Site相互作用记录、67947条GENE记录及1422条MATRIX记录。MATRIX可能对高级用户如从事生物信息学工具开发的人员较有用，因为其包含了转录因子结合位点不同核苷酸的频率信息。

（2）真核生物启动子数据库（EpD）

真核生物启动子数据库（EpD）由瑞士生物信息学研究所（SIB）维护，包含了大量的真核生物聚合酶II的启动子，这些启动子的转录起始位点已被实验验证，而序列可以通过EMBL数据库（http：//epd。vital□it。ch/）得到。至2010年10月，EpD已经是基于EMBL版本号105的数据库，包括了4806个启动子。从EMBL版本72开始，启动子根据一种有限的层次性系统进行分类，与用EC命名法对酶进行分类并无不同。这个层次性的根部是：①植物启动子；②线虫启动子；③节肢动物启动子；④软体动物启动子；⑤棘皮类动物启动子；⑥脊椎动物启动子。这种层次性系统目前已经被放弃，原因是通常很难决定一个新的启动子的分类，特别当基因产物是一个多功能蛋白时。跟TRANSFAC相似，EpD也基于ASCII的文本文件，它包含不同字段，可由前两列字段标识符进行唯一标识；且查询功能也相似，可以对所有字段或字段子集进行查询。例如，在所有字段中查找“超氧化物歧化酶”，能得到三条记录。对于每一条记录的更多信息，可以通过文本文件内容中链接或通过FASTA或EMBL格式保存的启动子序列（加上下游核酸序列）得到；也可以在这种基于层次结构的分类模式中选取一组启动子，并可以下载得到启动子序列。这样可以很容易地获取小鼠甚至所有脊椎动物的启动子序列。此外，通过EpD的FTp服务器（ftp：//ftp。epd。unil。ch/epd/），可以下载其完整数据库。

最近，EpD增加了一个新的数据库EpDnew，它包含了用新的高通量的转录起始位点的定位方法得到的数据，如CAGE（Cap analysis gene expression，基因表达的帽端分析）和TSS□seq（TSS，Transcription start sites，转录起始位点□测序）分析得到的数据。组蛋白h2AZ和h3K4me3，DNA聚合酶II（polII）和DNA甲基化的ChIp□seq（染色质共沉淀□测序）的大规模数据被利用。目前EpDnew含有9716人基因的启动子和9773小鼠基因的启动子。

第36章 BRENDA数据库

对于作为计算系统生物学核心的动力学模型来说，蛋白酶的动力学数据是必需的。因为每一个蛋白酶通常需要单独的纯化和反应条件，故其动力学实验数据很难得到并且非常耗时。此外，相关的文献往往发表在不同领域的各种杂志上，难以整合在一起。因此，BRENDA作为一个全面的蛋白酶信息系统被开发出来。首先，BRENDA是一个二级数据库，提供了大量与蛋白酶相关的动力学数据。这些数据是通过文本搜索技术从大量的生物医学文献中收集得到的，并且可以通过Web界面进行访问和查询。给出了BRENDA收集的动力学相关数据的类型和各种不同类型的记录数目。例如数据库中的酶包括了4379不同的EC值和超过10万个的Km值。BRENDA的一个优势就是可以用多种方法进行搜索，如很容易在C。elegans中找出超过一个特定相对分子质量或最适在30℃以上反应的所有酶。搜索结果还可以空格分隔的文本文件形式下载以用于更进一步的研究。在使用高级查询特性时，用户更可以用表中字段信息构建复杂的查询。如用户在不知道相应EC 值的情况下想要找出所有参与糖基化的酶，或在不知道精确学名的情况下找出相关的蛋白酶，BRENDA系统中的ECTree浏览器和TaxTree搜索就能够给用户提供非常有用的帮助。它们提供了类似浏览器的一个接口，可以用EC 值或物种名字在层次结构中进行搜索。BRENDA还提供了一个用于基因本体术语搜索的接口，与酶关联的GO术语可以直接链接到AmiGO浏览器。BRENDA同时与其他数据库链接，能够提供与蛋白酶相关的更多信息，比如催化反应的底物和产物可以用化合物结构来显示；生物物种信息可以链接到物种数据库NEWT；序列信息可从Swiss□prot中得到；如果相应三维结构已被解析，则可以与pDB数据库相关条目链接。最后，BRENDA还提供了用于产生实验数据的相应文献信息（如pubMed ID和摘要等）。

序列和注释数据库，如RefSeq、GenBank和NCBI上的各种数据库，只提供了基因组中单个基因的信息。这些数据库中的每条目录记录了一个基因和它所表达的蛋白结构和功能的信息。然而，对于生物学，特别是系统生物学来说，最重要的就是蛋白、核酸和其他分子间复杂的相互作用，例如，由复杂生物反应过程组成的代谢网络。然而，这些关于生物过程和代谢途径的数据并没有包含在上述以基因为中心的数据库中，但是常常可以在相关的科学文献中找到。Reactome的开发就为了填补这一空缺，它是一个开放的在线的由人类基本的代谢过程组

成的数据库，由冷泉港实验室、欧洲生物信息学研究所（EBI）和Gene Ontology联盟共同管理。显示了Reactome主页的一个截图。整个Reactome数据库被分成一些基本的生物过程的模块，其每一个模块都由一个或多个相关领域的专家人工审核过。每一模块同样可以以它们的更新时间作为参考，因此可以像科学文献一样被引用。一方面，Reactome数据库可以给用户提供许多有用的信息，例如，用户不熟悉的某个特异基因的产物。另一方面，Reactome数据库还可以被计算生物学家用来进行大量的数据挖掘以期能从中得出某种生物学意义上的结论，例如，由cDNA芯片实验获得的表达数据。Reactome中的数据模型由三种基本数据类型组成：物理实体，反应和代谢途径。物理实体可以是参与生物过程的任何一种分子或复合物。反应以一个或多个物理实体作为输入，以产生一个或多个物理实体作为输出。代谢途径由多个反应组成。因此，一个生物反应网络的拓扑结构可以通过这三种数据类型进行描述。同时，Reactome的基因产物可以链接到其他数据库（如Swissprot，Ensembl和RefSeq等）。然而，Reactome数据库的不足之处在于没有包括真实的生理模型中所需要的动力学数据（如比例常数或浓度梯度等）。

Reactome是一个由人类的基本生物过程组成的数据库。主页提供了人类反应网络的一个图形化显示。图中的每一个箭头代表了一个事件，亦即一个生物过程。在这个生物过程中，输入记录在一步或几步之后被转化成输出记录。图表下面列举了Reactome数据库包括的各种模块。用鼠标对准模块的名字将会显示与之关联的生物过程。同时，对于每一个模块，可以显示其他生物体的相关生物过程。

与此同时，Reactome提供的“pathfinder”工具，可以帮助用户找出两个物理实体之间的最短路径，例如，代谢物D□果糖和丙酮酸之间的最短路径，或者由初级mRNA到加工后的形式间的反应步骤。计算好的路径以图形化的形式显示。pathfinder同时也提供了排除特殊实体的可能性，例如，代谢物ATp或NADh，它们具有较强的连接性，它们的输入可能导致用户并不想要的路径的生成。

第37章 基因组微阵列

对功能基因组信息进行系统性的整合是生物信息学和计算生物学的一个主要挑战。与局限于单种数据类型的序列数据库不同，功能基因组实验所产生的数据相当复杂，只有同时提供大量的实验信息、使用的材料和精确的实验条件等才有生物学意义。此外，很多实验在技术上只对一个物种的基因有效可行。因此，基于基因在不同物种间的直系关系，将很多不同实验来源和不同生物体的功能基因组信息进行整合，显得非常重要。近几年，已经开发出很多相关的数据库。例如，关于表达数据，有很多遵循MIAME标准的数据库，包括：EBI的ArrayExpress，NCBI的Gene Expression Omnibus（GEO），斯坦福微阵列数据库SMD。与此同时，还有很多蛋白质相互作用数据库，蛋白质结构pDB，基因诱捕（和其他功能基因组方面的数据库。此外，还有很多收集一个特定生物体所有基因各种相关信息的辅助性数据库，如小鼠，大鼠，线虫，果蝇，斑马鱼（Danio rerio），或者酵母菌。对于这种数据库的访问，通常是基于单个基因的查询，因而对于一次显示成千上万个基因来说并非是一个好的选择。

这种情况驱动了基因组微阵列的开发。基因组微阵列（Genome matrix）是由马克斯普朗克分子遗传研究所和德国基因组研究中心（RZpD）开发的一个数据库/数据库接口系统。这个系统以带颜色的正方形形成的微阵列的形式显示不同生物体的不同基因，每一列代表了一个基因和它在其他物种的同源基因，而每一行则代表了这些基因的生物功能的相关信息。因此，对于每个基因（或一组基因），该系统以可理解的方式显示了尽可能多的信息，这样既能够输出一组基因（例如，都含有关键词“激酶”），又可以查看染色体上与某个基因相邻的基因（例如，DNA序列上下游的基因）。此外，基因的特殊数据信息以尽可能的方式通过方块的颜色来表示。例如，方块的颜色可以标识基因的功能（GO注释）、基因的表达水平（以红-绿或黄-蓝范围显示出基因的表达水平）或者仅仅表明存在基因的特殊类型信息（如蛋白质的X射线晶体结构，显示大鼠基因敲除后发育表型的图像，线虫中RNAi表型，等等）。点击这些方块能够显示存储在本机、数据库或网站上可能的相关信息。由于这种系统的要求并不高，它可以很容易地适应拥有重要信息的数据库的变化。

基因组微阵列是一个数据库/数据库接口系统，以微阵列的形式提供了基因组的信息。每一列代表一个生物体的一个特殊基因（或者它的同源基因），每一行代表了基因功能相关的特殊类型的信息。这种特殊的数据类型的信息以尽可能的方式通过方块的颜色来表示。点击某一个方块，会弹出一个菜单，显示与方块相联信息的简短描述

第38章 生物系统建模工具

如何利用已有的组学数据（包括基因组学、转录组学和蛋白质组学以及蛋白质相互作用数据、通路信息）对生物系统建模是一个严峻而又非常必要的科学问题。尽管各类注释数据未尽完善，但是根据已有的数据，借助计算机预测算法，可以构建相对完整的生物系统网络，从而追踪代谢物在生物体内的代谢途径，发现调控过程等。这类建模工具中比较著名的包括斯坦福研究所（SRI International）生物信息学课题组开发维护的pathway Tools，欧洲生物信息学研究所开发的BioModels。net在线网络服务，及奥克兰生物工程研究所开发的CellML项目。

pathway Tools是一个用于分析通路、基因组和系统生物学的整合工具，主要包括以下3个交互式的组件：pathoLogic、pathway/Genome Navigator和pathway/Genome Editors。用户可以通过pathLogic模块构建物种特异的通路/基因组数据库pGDB（Organism□specific pathway/genome database）；用pathway/Genome Navigator查询、可视化和分析pGDB的基因组信息、代谢通路和调控网络；用pathway/Genome Editors编辑pGDB。所谓pGDB是指一种生物信息数据库，整合目标物种的基因组序列、基因组注释数据、代谢通路、信号转导通路和调控网络等信息，例如EcoCyc。pGDB可帮助分析基因表达、蛋白质表达和代谢反应，是获取、整合和交流关于该物种最新科学发现的中心资源。已有的pGDB列表可通过访问BioCyc主页获得。当然，用户可以通过pathoLogic输入Genbank格式或pathoLogic格式的注释数据以及FASTA格式的序列，直接创建自定义的pGDB。pathway Tools不仅可用于创建、查询和编辑pGDB，还可以用于预测代谢通路、操纵子和转运反应，在完整的基因组图表上绘制自定义的基因组范围的数据，完成代谢网络和调控网络图表。除此之外，pathway Tools还包括操纵子的比较分析工具，代谢网络的可达性和终止代谢物的分析工具，以及代谢网络中代谢物的追踪工具。目前，至少已有75个研究小组将pathway Tools应用于200多个不同类别的生物的基因组。

除了用户友好的图形交互式界面外，pathway Tools还提供了perl，Java和Lisp的应用程序接口（ApI），及Oracle或MySQL格式的数据库系统接口。另外为了与其他应用工具对接，pathway Tools还可以将pGDB的子集数据导成SBML，BiopAX，Genebank和FASTA等格式。

BioModel。net的在线工具则致力于建立一个标准，供研究者交流共享生物系统研究成果。BioModel。net项目组定义了统一的建模标准，提出了注释模型信息最小化准则MIRIAM（Minimal information required in the annotation of models）和模拟实验信息最小化准则MIASE（Minimal information about a simulation experiment）；创建了可控制的词汇表，如系统生物学本体SBO（Systems biology ontology），动力学模拟算法本体KiSAO（Kinetic simulation algorithm ontology），动力学描述术语TEDDY（Terminology for the description of dynamics）；提供了数据描述语言，如模拟实验描述标记语言SED□ML（Simulation experiment description markup language）；建立了一个注释好的生物模型数据库BioModels Database。利用BioModel。net资源，用户可以在BioModels数据库存储、检索计算模型，也可以通过接口自己编程远程查询及获得信息。

用户可以直接在BioModels数据库查询、模拟已有的生物模型，也可以递交自己的模型。显示的是数据库已注释好的Wnt_ERK_Crosstal模型。用户递交的模型必须是SBML或CellML格式的文件。要创建自定义的模型，可以参考BioModels。net设定的MIRIAM标准，用SBO描述。

BioModels数据库显示界面。最上方有3个下拉式菜单，分别可以导出SBML及其他格式的文件，查看各个格式的反应图并进行在线模拟。页面上还显示模型的总体概况、模型组件、模型反应及参数、各个反应的数学公式等

CellML是一个与BioModels。net类似的项目，目标在于存储和交换以计算机为基础的生物数学模型。CellML项目已拥有500多个在线生物模型，覆盖了细胞周期、基因调控、信号转导、代谢、神经生物学、免疫学、血液循环、钙动力学、内分泌、合成生物学等大类，同时开发了相关的工具，包括建模环境、可视化和注释工具、模拟工具、验证工具、转换工具以及处理CellML文档的应用程序接口。利用这些工具我们可以创建自己的生物模型，并可进行模拟仿真实验，从而帮助我们观测生物体的反应过程。我们也可以直接对这些项目已提供的模型进行分析，完成研究。

第39章 蛋白质与其他分子相互作用

生物信息方法介绍蛋白质与其他分子间的相互作用与识别是21世纪生命科学研究的前沿和热点，细胞内各种重要的生理过程，如基因的复制、转录、翻译以及细胞周期调控、信号转导、免疫反应、细胞的增殖、分化和死亡等，都是以蛋白质与其他分子间相互作用为纽带进行的（Jones等，1996）。本章主要介绍研究蛋白质与蛋白质、小分子化合物及DNA相互作用的常用生物信息工具、算法、软件、数据库，及提供在线服务的网站。常用的生物信息方法中，我们着重介绍蛋白质□蛋白质作用结合位点的预测及分子对接，蛋白质□小分子结合位点的预测及基于分子对接的虚拟筛选，常用小分子化合物数据库，及蛋白质□DNA结合位点的预测。

1.蛋白质□蛋白质相互作用

最近几年来，蛋白质□蛋白质相互作用一直是研究的热点和难点之一。许多相关的生物信息算法、软件和数据库已经被开发出来，第10章就介绍了相关蛋白质□蛋白质相互作用数据库。本节主要介绍相关蛋白质□蛋白质作用结合位点预测方法，及蛋白质□蛋白质对接（Docking）算法和软件。

2.蛋白质□蛋白质作用位点预测

一般来说，人们通过总结蛋白质□蛋白质复合体形成的规则，并根据这些规则进行经验学习和机器学习，从而预测蛋白质□蛋白质作用结合位点。经验学习一般是通过对已知的蛋白质□蛋白质复合体做序列和结构上的相关属性分析，从中找出规则并推导出评分函数（Scoring function），进而应用评分函数进行结合位点预测。机器学习也是通过计算蛋白质□蛋白质作用界面与其他表面的序列、物理化学属性的不同之处，通过神经网络与支持向量机器等人工机能机器学习的方法来预测作用位点。总结了网站服务器所用的蛋白质□蛋白质结合位点预测方法的特征属性。这些特征属性有蛋白质序列上的，如进化保守性（Conservation）和氨基酸在作用界面的倾向性（Residue interface propensity），及几何上的平面凹凸性，还有物理化学属性，如静电性、亲水性、疏水性及蛋白质亲水表面积（Solvent accessible surface area）等。

一般来说，预测的流程是：首先根据已知的蛋白质□蛋白质复合物构建一个训练集（Training dataset），该训练集通常是从蛋白质结构数据库pDB（protein data bank）中筛选得到的。对训练集中所有蛋白质□蛋白质复合体做上述特征属性的计算分析，将这些属性参数输入机器学习算法进行训练，再与已知复合体的结合位点情况进行对比，进而改进算法。近些年来，人们已经设计和开发出许多基于结构预测蛋白质□蛋白质结合位点的方法。常见的几种蛋白质结合位点预测方法的名称、网址和所用的机器学习的方法，常见蛋白质结合位点预测方法的名称、所用的学习方法、属性及网址

预测方法所用学习方法所用特征属性web网址pINUp经验打分方程序列保守性，位点倾向性支持向量机序列保守性，位点倾向性，几何凹凸性，疏水性，表面溶剂可及性神经网络序列保守性，位点倾向性经验打分方程序列保守性，位点倾向性，几何凹凸性支持向量机和神经网络表面溶剂可及性就现有的各种蛋白质结合位点预测方法而言，其所使用的蛋白质□蛋白质作用界面与其表面的序列、物理化学等属性情况不尽相同。有的考虑较为全面，使用了全部的属性；有的则只使用了其中的部分属性。当然并不能单单靠使用属性的多少来判断一种算法的优劣性。因为使用的属性越多，计算所花费的时间也就越多，而且其效果也不一定就比使用属性少的算法好。所以具体怎么使用，用户应该根据具体情况，特别是预测对象来具体确定，在运算速度可以接受的前提下，应尽可能多考虑一些特征属性。

MetappI是我们基于上述五种位点预测方法开发的一套蛋白质结合位点预测方法，用户可以在网站上提交蛋白质结构或者给出pDB ID和链ID，metappI能够自动把用户提交的结构转交到其他五个网站，把它们所预测的位点信息收集回来并根据metappI算法处理后，将所预测的位点信息以文本的形式及基于JMOL可视化形式返回给用户。

①未做处理的蛋白质结构数据被直接分别提交给上述五种方法的服务器；根据五个服务器的预测情况分别对每个表面氨基酸进行可信度（Confidence score）打分，被越多的服务器预测为结合位点的氨基酸得分越高；具有最高可信度分值的蛋白质表面区域（Surface patch）则被选出来。②再使用MeSu算法预测出最终的结合位点。MeSu是一种基于蛋白质溶剂可及表面（Solvent accessible surface）的蛋白质表面连续区域（Continuous patch）扩增方法。首先，MeSu将蛋白质溶剂可及表面表示成由许多点组成的网状区域；接着MeSu从距离上一步选中的

初步结合位点质心最近的一点出发，将离其最近的几个点加进本区域集合，然后将离新加点最近的点加入该集合，将该过程一直递增直到该区域扩增到所需要的大小。这种扩增方法保证了最后得到的蛋白质表面区域在三维空间上是连续的。③与这些点相关联的蛋白质氨基酸就是metappI最终预测出来的蛋白质□蛋白质结合位点。验证结果表明，metappI算法的预测准确度比单个预测方法提高了很多。

MeSu蛋白质表面区域连续扩增方法示意图，中心点为始发点，箭头表示扩增的方向。来自huang等的图一利用metappI对pDB ID 1A6W的L链（蓝色）预测出来的蛋白质结合位点（红色区域），可以看出，预测的位点与真实的位点很接近，紫色为与L链相互作用的h链蛋白质□蛋白质对接

如何将两个相互作用的单个蛋白质结构构造成复合体结构一直以来是结构生物信息学的一个研究热点。人们可以利用蛋白质□蛋白质对接（protein□protein docking）的方法模拟构建出可能蛋白质

复合物结构。蛋白质□蛋白质对接的一般流程：已知蛋白质A与蛋白质B相互作用，对接的第一步是尽可能地构建大量的复合物构象（1000～10000个左右）；第二步用一种评分函数对这些构象进行打分排序，以期从中找出与天然复合物相近的构象；最后一步是对找出的少量构象进行能量优化，使之更加接近天然复合物的结构。

目前，国际上所开发的蛋白质□蛋白质对接程序大多能够提供大量的待选构象，但其中仅含有少量的正确构象。因此，对接的主要工作在于如何从大量构象中挑选出与天然结构尽可能相近的构象，这就需要一个很好的打分函数（Scoring function）来对这些构象进行排序。一般来说，评分函数基于自由结合能（Free binding energy），但是自由结合能的计算量太大。首先，人们常常用快速的基于蛋白质表面几何互补的方法构建出很多复合物构象；然后，才对排在前面的构象计算自由结合能（打分函数）以挑选出正确的构象。近些年来，许多蛋白质对接的算法被设计和开发出来。从大方向上，我们可以将其分类为刚性对接（Rigid□body docking）和柔性对接（Flexible docking）两种。

刚性对接（Rigid□body docking），在整个对接过程中蛋白质结构被认为是不会发生变化的。刚性对接过程可以分为以下几步：①从已知的两个蛋白质结构数据出发，并假设在整个对接过程中这两个蛋白质的

结构是不会发生变化的，通过旋转和移动蛋白质三维结构的方式，对其进行表面几何匹配；②使用设计好的专门的打分函数对每次匹配结果进行打分；③滤除不可能形成结构的匹配构象；④对剩下的匹配构象进行得分排序；⑤使用已知的柔性参数，将匹配构象进一步优化。

在刚性对接中如何准确表示蛋白质的三维结构，是后期预测的关键。快速傅立叶变换（Fast fourier transform，FFT）算法就是运用比较多的方法之一。FFT的主要思想是将蛋白质的三维结构投影到一个三维的栅格中，然后进行对接，很大限度地提高了对接速度和准确率。下面着重介绍基于快速傅立叶变换的蛋白质□蛋白质对接算法。

1.基于快速傅立叶变换的蛋白质□蛋白质对接算法

基于快速傅立叶变换对接的算法最初由Katchalski□Katzir在20世纪90年代初提出（。整个蛋白质□蛋白质对接流程，蛋白A为大蛋白，通常也称作受体（Receptor），映射到固定的三维网格A中；蛋白B为小一点的蛋白，通常也称作配体（Ligand），映射到动态的网格B中。网格的步长（Grid spacing）可以设为1埃（angstrom，10-9米），网格的大小（x，y，z）由蛋白结构大小决定。在网格里的每一个节点被赋予一个整数，赋值函数如：

protein A ap，q，r=1Surface cell

-15Interior cell

0Elsewhere，protein B bp，q，r=1Interior cell

0Elsewhere

由公式可以算出这两个网格的几何匹配程度。Cα，β，γ=∑Nl=1∑Nm=1∑Nn=1al，m，n·bl-α，m-β，n-γ（11□2）

由公式算出每个网格节点（Grid points）的数值。网格B在空间上（x轴、y轴和z轴）围绕着网格A做移动搜索，当匹配值最大时停止搜索。为了加快运算速度，对网格所有节点的数字做快速傅立叶转换运算，以找到网格B（蛋白质B）的移动坐标（Translation）。然后蛋白质B围绕x轴、y轴和z轴旋转一个角度（alpha，phi，psi 角）得到另外一个构象，重复以上映射及移动搜索流程。最后可以得到蛋白质B相

对于蛋白质A的空间几何位置，这个相对空间几何位置由6个参数（旋转角alpha、phi、psi角及移动坐标α、β、γ）决定，根据这6个参数可构建出蛋白质AB复合体的结构。每一个相对空间几何位置都有一个蛋白质B和蛋白质A的表面几何匹配程度得分Cα，β，γ（Geometric complementary score）。蛋白质□蛋白质对接之后所输出结果就是这6个参数加上这个得分。

然而，如果从公式直接计算Cα，β，γ，对于每一次的α，β，γ坐标转换要计算N的3次方。为了降低运算复杂度，Katchalski□Katzir等（1992）开创性地引入离散傅立叶变换。首先，对网格A和B的所有离散值都做一次离散傅立叶变换，变换可以表示为DFT（A）和DFT（B），然后DFT（A）复数耦变换可以表示为DFT1（A）。根据傅立叶变换原理，C的离散傅立叶变换的值DFT（C）为DFT1（A）·DFT（B），对DFT（C）做反向傅立叶变换就可以直接得出C的值。通过傅立叶变换，整个流程的计算复杂度可以由O（N6）降到O（N3·logN）。

自Katchalski□Katzir等提出FFT算法降低蛋白质□蛋白质对接运算速度之后，全世界很多研究小组参与到开发基于FFT的蛋白质对接软件中来。目前，著名的基于FFT算法的对接软件有FTDOCK（Sternberg等，1998），patchDock，ZDOCK及BDOCK。这些软件的下载网址和总结。其中FTDOCK和patchDock对位于蛋白质A表面的网格节点赋予相同的值；ZDOCK和BDOCK对蛋白质表面的弯曲程度做了一定的描述，位于表面凹面的网格节点会被赋予更大的整数值。BDOCK通过扫描7个方向（x，y，z轴及四个立体对角线）计算出蛋白表面的弯曲度数，有关该度数的计算详见LIGSITEcs的算法。

2.常用蛋白质□蛋白质对接软件

对接软件说明网址BDOCK基于FFT，蛋白表面几何互补打分，源代码公开基于FFT，提供网站对接服务基于FFT，蛋白表面几何互补及静电互补，源代码公开

对接软件说明网址hAARDOCK实验位点数据可以使用ICM集成软件的一个模块，商业化基于FFT，提供网站对接服务基于FFT，python代码公开DOCK基于蒙特卡罗动力学模拟，源代码公开，侧链优化，速度慢基于FFT，蛋白表面几何互补及能量方程，提供Linux下可执行文件

可以通过比较对接出的构象与真实的天然态结构，来评价一个分子对接算法的好坏，算出均方根偏差（Root mean square deviation，RMSD）。一般来说，蛋白质B的RMSD值小于10A可以视为相似结构（Near□native complex）。根据起始的输入蛋白质结构，对接可以分为复合态（Bound）对接与自由态（Unbound）对接两种。复合态对接就是从已知的蛋白质复合物结构出发，先把两个蛋白质分开再通过对接构建出复合物的结构。自由态对接就是从以单体结构存在的两个单个蛋白质结构出发，构建出复合物的结构。由于相互作用的影响，两个蛋白质形成复合物后有着一定构象变化，所以自由态对接相比复合态对接要困难许多。还有一种情况是，当蛋白质单体结构未知时，我们可以通过同源建模（homology modeling）的方法先模拟预测出单体结构，再通过对接的方法构造出复合体结构，然而，通过同源建模模拟出的单体结构存在着很大的不确定性，则对接出的复合体结构就更加不准了。

复合态对接与自由态蛋白质□蛋白质对接方法比较

美国Ziping Weng教授课题组除开发ZDOCK（Chen等，2003a；Chen等，2003b）、RDOCK（Li 等，2003）和ZRANK（pierce等，2007）之外，还从pDB数据库里筛选出来Benchmark数据集。后者一直被从事分子对接算法开发的科学家用来验证对接算法的好坏（huang等，2008）。如今Benchmark数据集已包含124个已知复合体结构及相应的单体结构，根据对接的难易程度分类，88个复合体为容易对接的，19个对接难度为中等，剩下17个为高难度对接（构象变化极大）对Benchmark数据集其中的四个蛋白质复合物进行复合态对接的结果，蓝色与绿色是已知结构，红颜色是通过BDOCK对接出来的最接近于天然结构的构象。

BDOCK对Benchmark数据集中四个蛋白质复合物进行复合态对接的结果，对接出来的配体构象（绿色）与天然态结构（红色）非常接近（RMSD值2.0）

2.预测作用位点及实验数据在蛋白质□蛋白质对接中的应用

上节介绍的分子对接通过盲目搜索（Blind search）来构造可能复合物的结构，再通过打分方程挑出可能与真实接近的构象。如果我们能通过实验方法或位点预测方法（如metappI方法等）得知相应蛋白质结合位点的信息，即哪些氨基酸参与了蛋白质之间的相互作用，蛋白质□

蛋白质对接问题就变得稍微简单一些。这些位点信息既可以应用在对接的过程中，也可以应用在后期结果过滤处理中。对于前者来说，我们可以只搜索蛋白质表面位于结合位点的区域，其他表面区域可以忽略不计。ZDOCK软件（Chen 等，2003a；2003b）就提供了这个功能。在后期对接结果过滤中，我们可以设计一个过滤方程，只考虑结合在预测位点附近的复合物构象。结果是对接出的复合体结构与位点信息有着紧密的关系。如果作用位点信息错误，则据此对接出的复合体构象也就与真实结构相差甚远。

3.柔性对接（Flexible docking）

在基于傅立叶变换分子对接的算法中，蛋白质被简化处理为刚性（Rigid）分子。然而，天然蛋白质是动态存在的，从X射线得出的晶体结构只是其能量最低时的一个构象。且在两个蛋白质相互接近对方并发生相互作用形成复合体的过程中，由于各种作用力的影响，单体结构往往会产生构象变化（Conformational change）。这些构象变化包括侧链（Side chain）原子的运动、主链（Backbone）及铰链（Loop）的空间构象改变等。因此，我们需要对单体结构做必要的柔性处理才能真实反映这些构象变化。在对接过程中考虑蛋白质结构变化的对接方法就叫做柔性对接，但这需要搜索更多的自由度，需要更多的计算时间。随着计算机处理能力的增强，软对接（Soft docking）及基于分子动力学模拟（Molecular dynamic simulation）的分子对接方法等柔性对接算法得到了开发和发展。

4.蛋白质□小分子相互作用

基于药物靶标结构进行药物设计及虚拟筛选是生物信息学中极为重要的研究领域，其前提条件是事先知道小分子药物在靶标蛋白质表面的作用结合位点，这些结合位点往往位于蛋白质表面的凹处（Cleft）或口袋（pocket）。一般来说，运用几何的方法就可以快速准确地查找到这些凹处和口袋。近几年来，许多算法和软件已经被成功开发，用以在蛋白质表面上搜索这些口袋以预测小分子结合位点（Levitt 等，1992；Laskowski，1995；hendlich等，1997；Laurie等，2005；huang 等，2006；Kawabata 等，2007；Weisel 等，2007；Capra等，2009；Le Guilloux等，2009；Kawabata，2010；Tripathi 等，2010）。研究表明，一些计算机模型与算法，通过分析蛋白质结构的几何属性、物理化学属性，分析、统计已知的蛋白质与小分子复合物的结合体，通过机器学习的方法，在蛋白质小分子结合位点的预测上及对已知位点再

进行蛋白质□小分子对接取得了不错的成果（Laskowski，1995；Liang等，1998；Binkowski等，2003；Laurie等，2005；Glaser等，2006；huang等，2006）。本节主要介绍相关的蛋白质□小分子结合位点的预测方法、蛋白质□小分子对接算法和软件及基于分子对接的虚拟筛选和常用的小分子化合物数据库。

第40章 蛋白质□小分子结合位点的预测

1.单一的蛋白质□小分子结合位点的预测算法

已开发并投入使用的蛋白质□小分子结合位点预测方法的精度和速度存在着很大的差别，彼此之间也有很深的渊源。本节主要介绍常用的蛋白质□小分子结合位点预测算法。总结了常见的蛋白质□小分子结合位点预测算法，列出了它们的类型、名称、出现的时间、基本方法以及基于该算法开发出来的网站服务器网址。本节将介绍常用的蛋白质□小分子结合位点预测算法。常用的蛋白质□小分子结合位点预测方法总结

分类名称时间基于几何基于能量Web网址基于网格pOCKET

α□shapeCAST目前常用的蛋白质□小分子结合位点预测算法可以分为三类：基于网格的、基于球体的和基于α□shape的。基于网格的预测算法首先是将蛋白质映射到一个三维网格（3D grid）中（Levitt等，1992），然后对网格的每一个节点进行相关的操作运算，如果一个节点满足一定的几何或者能量条件，就可以判定它位于结合位点。基于球体的方法一般是首先在蛋白质的表面或者里面初始化球体用以填充空白区域，满足一定几何或者能量条件的球体所在的区域就是结合位点。α□shape是一种新型的计算几何理论，用以解决一些复杂的空间几何问题。基于α□shape的预测算法比较少，这类算法首先使用α□shape理论对蛋白质原子进行划分并定义结合位点，然后按照结合位点的定义进行寻找，另一种分类方式是按照基于纯几何信息的还是基于分子间的作用能量的来进行分类。下面将详细介绍重要的蛋白质□小分子结合位点的算法。

pOCKET（Levitt 等，1992）算法是最早的基于几何信息的蛋白质□小分子结合位点预测方法之一。该方法诞生于1992年，是基于三维网格的一种预测算法。在该方法中，作者首次提出了使用“蛋白质□溶剂□蛋白质”事件（protein□solvent□protein events，pSp events）来定义结合位点的思想。所谓的蛋白质□溶剂□蛋白质事件是对一条扫描线上的节点进行的描述。首先将整个蛋白质映射到一个三维网格中，网格的

节点将会按照其与蛋白质的关系被赋予不同的属性，有“蛋白质节点”、“溶剂节点”。网格的节点被认为是“蛋白质节点”的依据是，这个节点到最近的蛋白质原子的距离小于3埃，否则这个节点就被判定为“溶剂节点”。然后对于网格中的每个节点，从x，y，z轴三个方向上对齐进行扫描，直到扫描线遇到不同性质的节点或者在某个方向上扫描完所有的网格节点而停止，这期间，每条扫描线都经过了一系列的网格节点。如果这条扫描线的两端是“蛋白质节点”，中间是一系列的“溶剂节点”，那么可以说明这个节点在蛋白质的外面，即可能是结合位点的所在部位，这就是一次“蛋白质□溶剂□蛋白质”事件。如果这个节点的三条扫描线中，“蛋白质□溶剂□蛋白质”事件的个数大于某个阈值，那么就判定这个节点位于蛋白质表面的口袋中，即可能为蛋白质□小分子结合位点。

LIGSITE（hendlich 等，1997）算法诞生于1997年，是对pOCKET算法的改进和扩充，该算法也被后人进行了多次改进。由于在pOCKET中，结合位点的定义过于依赖蛋白质和坐标系之间的旋转角度，所以LIGSITE除了扫描网格节点的三条坐标轴方向的“邻居”节点以外，还扫描了节点所在的立方体的四条对角线，以此来降低结合位点的定义对网格中蛋白质的构象的依赖。因此，在LIGSITE算法中，对每个网格节点一共进行了7个方向的扫描。最初，LIGSITE由于条件限制只进行了10个蛋白□配体复合物的测试，发现有7个小分子结合在最大的结合位点上，2个结合在第二大的结合位点上，有1个结合在第三大的结合位点上。

pocket□Finder算法是对LIGSITE算法的一次改进。与LIGSITE一样，pocket□Finder也是计算每个网格节点的pSp事件的个数，不同的是它使用比LIGSITE算法更好的参数来进行7个方向的扫描。在pocket□Finder中网格之间的间隔为0.9埃，探针半径1.6埃，pSp事件的阈值个数为5，这些参数可以降低第一次预测到的结合位点的体积。

LIGSITECS是我们对LIGSITE算法的第一次改进。该算法中，蛋白质表面采用了康诺利表面（Connolly surface）（Connolly，1983）定义法，结合位点的判断使用较为精确的“表面□溶剂□表面”事件（Surface□solvent□surface events，SSS event）来代替“蛋白质□溶剂□蛋白质”事件。该方法与前面介绍的几种方法一样首先将蛋白质映射到一个3D网格中，网格间隔为1.0埃。每个网格节点按照性质的不同被标记成“蛋白质节点”、“表面节点”和“溶剂节点”。LIGSITECS算法对每

个节点也同时进行了7个方向的扫描（包括x、y、z三个坐标轴和节点所在的立方体的四个对角线）。如果一个“溶剂节点”至少含有5个“表面□溶剂□表面”事件，那么它就可能是结合位点，会被标记为“口袋（pocket）节点”。最后，对所有的“口袋节点”根据他们的空间相似性做聚类分析，形成最终的结合位点区域。

LIGSITECSC方法是我们对LIGSITE算法的第二次改进，由于引入了氨基酸的保守性问题，因此该方法已经不是一个纯粹的基于几何的方法。我们发现结合位点周围的氨基酸具有一定的序列保守性，因此根据氨基酸的序列保守性信息来修正LIGSITECS算法的预测结果，这就是LIGSITECSC算法。该算法对LIGSITECS计算得到的结合位点，再次计算表面氨基酸的保守性度（Degree of conservation），保守性信息从ConSurf□hSSp数据库中得到，然后根据此保守性计算结果对预测出的结合位点进行重新排序，即进行结果修正，实验表明该算法取得了很好的效果，最大预测准确率达到了86%。

pocketpicker是Weisel等对LIGSITE算法的进一步改进，该算法仍然将蛋白质映射到一个3D网格，然后计算每个网格节点的埋藏程度（Buriedness），以此来判断一个网格节点是不是最终的结合位点。pocketpicker算法示意图与LIGSITE的前两种改进相比，pocketpicker在对网格节点的过滤以及扫描节点方向上有着明显的差别。在pocketpicker算法中，首先对网格节点进行过滤，过滤掉离蛋白质表面稍远的节点以及在蛋白质内部的节点，即节点过滤的结果是只保留了离蛋白质表面的距离小于某一程度的并且在蛋白质外部的网格节点，这一步确保了预测结合位点的有效性，因为真正的结合位点也是在这些被保留的区域内。然后，对保留的节点进行扫描，对每个网格节点计算“表面□溶剂□表面”事件的个数。与LIGSITE方法相比，该算法对每个节点一共进行了30个方向的扫描。这30个扫描方向构成如下：首先，将每个节点看做是一个正八面体的中心，正八面体的每个面都是一个正三角形，连接这个三角形每条边的中点在每个面上再构建一个正三角形，然后取这个新构建的正三角形的每条边的中点，该节点与每个中点的连线就是一个扫描方向，这样，八个面就可以构建出24个扫描方向，再加上八面体的六个顶点构成的扫描方向，因此一共是30个扫描方向。pocketpicker的30个扫描方向示。pocketpicker增加了对每个节点的扫描方向，虽增加了准确度却降低了整体的运算速度。

pocketpicker算法示意图。（a）网格点过滤示意图，过滤掉离蛋白质比较远的网格点（i区域），过滤掉在蛋白质内部的网格节点（ii区域），只保留蛋白质表面周围的外部网格点（iii区域）。（b）找到的蛋白质表面凹陷处的网格点，并把它们聚类在一起（黑色节点）。（c）pocketpicker节点扫描的方向示意图，扫描以正八面体为中心的30个方向。来自Weisel等的图六

VICE也是一种基于网格和节点扫描的几何预测算法，且它的扫描算法非常有趣而迅速。将蛋白质映射到一个3D网格中后，就开始了对每个网格节点的扫描。VICE的节点扫描有三重向量，向量的大小从小到大，保证了扫描的密度。并且在VICE中，扫描线的方向是根据网格的位置来确定的，而不是根据传统的罗盘方向（比如对角线之类的），所以在该算法中，确定方向用的永远都是整数而没有浮点数，由于在计算机中整数运算比浮点数运算要快很多，并且节点扫描又是程序中相对耗时的步骤，因此VICE方法有较快的节点扫描速度，也是一种效率较高的算法。每条扫描线可能有三种情况：一是碰到了蛋白质表面（内部向量），二是伸向了蛋白质的外部（外部向量），三则是由于向量的长度不够而停止扫描。对于第三种情况，首先可作为第二种情况来考虑，然后使用大长度的向量继续进行确定。扫描之后，每个网格节点根据内部向量数占总体向量数的百分比进行分类标记，大于50%则认为该节点是内部的（Inside），会被保留，否则归为外部的（Outside）。此外，对于较难判断的节点（百分比为45%～55%），VICE还会对它们进一步使用不同的向量集进行判断，直到可以确定为止。扫描完成之后，就会在蛋白质的表面形成很多个互相独立的节点簇，这些簇就代表了可能的结合位点。VICE还计算了每个簇的大小和体积，以及每个簇中节点的邻居节点数目，小于某个阈值（一般为4）的节点会被删除，于是每个簇被继续精化。最后按照簇的大小进行排序以确定最可能的结合位点位置。VICE在第一位点的预测精确度达到83%，而在前三个位点的预测精确度达到了90%。

VICE 算法示意图。（a）VICE的三重扫描向量示意图。（b）VICE扫描向量经过的路径，从绿色的节点开始，遇到红色的蛋白质停止，中间橘黄色的都是此向量路径上面的节点。（c）VICE对网格点的扫描示意图。绿色的向量为外部向量，黑色的向量为内部向量，粉红色的为停顿（Stalled）向量，由于向量长度的限制无法判断其结果，暂时作为外部向量处理，须用长度更大的向量重新扫描计算。（d）VICE扫描之后确定的结合位点区域，上面的数字标记了内部向量占总向量

数目的百分比。（e）结合位点中的每个节点都有一个最小的邻居节点值，节点上的数字代表了与它们相邻的节点数目。黄色标记的1和3因为邻居节点数太小将会被删除。也是一种基于3D网格的蛋白质□小分子结合位点预测算法，并且考虑了蛋白序列的进化序列保守性（Evolutionary sequence conservation）来提高预测精度。ConCavity算法包括三个步骤，分别是创建网格（Grid creation）、位点提取（pocket extraction）和氨基酸映射（Residue mapping）。首先，根据蛋白质的几何结构和序列的进化保守性将蛋白质映射到一个三维网格中；然后对每个网格节点计算一个代表与结合的配体重叠的相似度的分值，每个网格节点分值的计算都遵循一定的形态学标准；然后对这些网格节点进行聚类分析，将连续的、高分值的网格节点归类在一起，这个过程结束的时候可以在蛋白质的表面得到一系列的相互独立的网格节点聚集成的节点簇（Cluster），这些簇代表了最终的结合位点；最后，对这些结合位点进行高斯模糊（Gaussian blur）计算（s=4□），将结合位点映射到蛋白质表面的氨基酸，然后给每个氨基酸分配一个最高的重合的网格节点的分值，距离高分值网格节点空间区域最近的氨基酸的最后分值最高，也就是最有可能直接和小分子作用的表面氨基酸。

ConCavity算法示意图。（a）创建网格，计算每个网格节点的配体重叠相似度分值。（b）位点提取，找到高分值的网格点聚集区。（c）氨基酸映射，计算出构成结合位点区域的氨基酸组成。（a）初始化两个大的探针球体示意图，标示为“1”为p1，标示为“2”为p2.（b）初始化一个小的探针球体S。（c）使用黑点标示的蛋白质受体结构。（d）多标量扩展灰色的框代表p1的扩展结果，白色的框代表p2的扩展结果，数字“1”代表p1的闭操作结果，数字“2”代表p2的闭操作结果。（e）多标量闭（Closing）操作示意图。数字“1”代表p1的闭操作结果，数字“2”代表p2的闭操作结果。（f）多标量探测定义出来的结合位点示意图，数字“1”代表p1的结合位点定义结果，数字“2”代表p2的定义结果。经John Wiley and Sons许可复制GhECOM与下文将要介绍的phECOM都是由同一个研究小组开发的，不同的是GhECOM是基于网格的，而phECOM是基于球体的。GhECOM中结合位点区域的定义采用数学形态学（Mathematical morphology）方法，通过不同半径的探针球体与蛋白质表面进行扩展（Dilation）、侵蚀（Erosion）、开（Opening）和闭（Closing）四种不同的数学形态学操作，可以很准确地定义节点位点区域。然后通过使用不同半径的探针球体计算多标量分子体积（Multiscale molecular volume）来计算结合位点的位置。

在结合位点的定义中，小的探针球体可以预测到很深的结合位点，大的探针球体可以预测到相对大的结合位点。因此，通过使用不同大小的探针球体进行预测，为定义结合位点的深度和大小提供了很大的帮助，可以很方便地确定结合位点的大小和深度，并可以进行结合位点的比较等工作。需要注意的是，GhECOM算法是基于网格的，因此定义中采用的不同大小的探针球体实际上在计算的时候都是包含不同数目网格节点的正n面立方体。

Laskowski小组在1995年提出了SURFNET算法，该算法与如上所述的算法的差别是没有使用三维网格方法，而使用了基于三维球体的方法。在该方法中，结合位点的定义完全通过这些球体，如果一个球体把两个原子分开而且又没有包含其他的原子，那么这个球体就是一个小分子结合位点。首先放置一个初始球体，保证两个给定的原子在这个球体相反方向的表面上，球体与这两个原子的范德华表面（van der Waals surface）相切。如果这个球体与其他原子的范德华表面有交集，那么不断地缩小球体的半径，最终使球体与任何原子的范德华表面不相交。最后体积太小的球体将会被舍去，只保留半径在1~4埃的球体。该过程执行完毕后，会出现很多相交的球体组成的互相独立的集合，这些集合叫做空隙区域（Gap region），其在蛋白质的内部和表面，分别称作空洞（Cavity）和裂缝（Cleft），就是该方法预测出来的蛋白质□小分子结合位点。SURFNET方法曾经对67个酶□配体（Enzyme□ligand）复合物进行分析，发现小分子结合在最大位点上的概率达到83%。

SURFNET算法示意图。二维化描述的位点探测过程。（a）蓝色的球体为初始化后的探针球体，位于两个原子范德华表面（Van der Waals surface）的中间位置，然后减小它的体积，使其不与任何原子的范德华表面重叠，最终的探针球体显示为红色。（b）最后许多的探针球体形成了许多分布在蛋白质表面的簇（Cluster），这些簇定义了结合位点的大小和形状。来自Weisel等的图一

SURFNET□Consurf算法是对SURFNET算法的改进，不仅可以找到蛋白质表面结合位点的位置，而且可以确定结合位点的形状。它对每个蛋白质分子采用两个阶段的预测。第一个阶段，使用SURFNET程序来预测蛋白质表面的结合位点，以作为潜在的结合位点（potential binding sites）。第二阶段对这些潜在的结合位点按照大小进行“修剪”，去掉离ConSurf□hSSp数据库中定义的高保守性氨基酸很远的区

域。最后剩下的那些最大的潜在结合位点就是蛋白质与小分子结合的位置。该方法使用pDB数据库的244个非冗余的蛋白质□小分子复合物的结构作为测试数据集进行测试，发现其预测精度可以达到75%。对于配体结合在大的高保守性的位置，该方法可以给出更加匹配的结合位点的位置和形状，且对酶和非酶的蛋白质都可以达到不错的预测精度。

pASS（Brady 等，2000）也是一种基于球体的预测方法。与SURFNET不同的是，pASS中的球体是探针球体（Sphere probe），所以这些球体都是初始化在蛋白质外部的，其实pASS是用球体一层一层地填充蛋白质表面的空洞和裂缝的。pASS算法的第一步是在蛋白质表面初始化一层表层探针球体，对每一个探针球体都计算其埋藏分数（Burial number），即探针球体球心8埃距离之内的原子的数目。然后对这些球体进行“剪枝”，只保留埋藏分数大于某个阈值的球体。剪枝完成之后每个探针球体都会被分配一个权值，这个权值与该球体周围的被保留的球体的个数以及球体本身的埋藏分数有关。然后pASS算法会重复此过程。最后会在蛋白质表面形成很多独立区域，它们包含了许多埋藏分数值很高的探针球体，通过识别这些球体可以筛选出活性位点（Active site point，ASp）。然后对这些球体按权值大小进行降序排列，保留权值大于某个阈值的球体，而且它们的距离大于8埃。最后，pASS方法再对筛选出来的探针球体按权值进行降序排序，并作为最后的输出，第一个结合位点就是最大的小分子结合位点。

（a）在蛋白质表面初始化一层表层探针球体。（b）通过不断地添加表层探针球体和剪枝操作，蛋白质的表面会形成很多独立的包含高埋藏分数（burial number）的结合位点区域。

phECOM与GhECOM算法一样，都使用了不同大小的探针球体，不过phECOM是真正的基于球体的预测算法。在该算法中，首先在蛋白质的范德华表面（van der Waals surface）初始化一层小的探针球体；然后初始化一层大的探针球体；最后，与大探针球体重叠的小探针球体将会被删除，剩余的小探针球体将会被保留并确定为最终的结合位点区域。实际上也可以这样理解phECOM的结合位点定义方式，即小的探针球体可以进入但是大的探针球体不能进入的区域，必然是蛋白质表面的凹陷、口袋或裂缝处，即配体小分子结合的地方。

（a）初始化小的探针球体。（b）初始化大的探针球体。（c）删除与大探针球体重合的小探针球体，最终剩下的小探针球体所在的区域就

是结合位点区域，标示为灰色。来自Kawabata等的图二是一种精确度较高的蛋白质□小分子结合位点预测算法，其核心思想是α□sphere理论。α□sphere依赖于沃罗诺伊轮盘（Voronoi tessellation），简单地说，α□sphere是表面有四个原子但内部没有原子的一个球体。Fpocket算法的第一步是计算蛋白质分子的所有α□sphere，Fpocket使用另一个软件包Qhull来计算得到所有的α□sphere数据，然后对所有的α□sphere按照它们接触到的原子的类型进行标记，得到预过滤（pre□filtered）的α□sphere集合。第二步是对这些α□sphere进行聚类分析，聚类分为三步，首先是基于沃罗诺伊邻接顶点（Voronoi vertice neighbours）的聚类操作，然后对其进行精化，再计算出每个α□sphere簇（Cluster）的重心（Mass center）。第三步是对空间上重心距离相近的簇进行合并，于是小的α□sphere簇特别是表面上的α□sphere簇将会被合并到一个大的α□sphere簇中。最后一步是执行多次连接聚类（Multiple linkage clustering），即将一个簇中的所有顶点和另一个簇中的顶点比较计算距离，如果这个簇中有特定个数的α□sphere和另一个簇中的α□sphere相近，那么将这两个簇合并成为一个簇。经过这样的聚类分析，小的和极性的（polar）α□sphere都会从蛋白质的表面去除。这些经过聚类分析产生的簇就是潜在的小分子结合位点，它们还将按照它们结合小分子的能力排序。Fpocket前三个预测到的结合位点的准确率最高可以达到94%。

CAST（Liang等，1998；Binkowski等，2003）采用另一种基于α□shape的方法来预测蛋白质表面的结合位点，计算蛋白质表面原子构成的三角形。蛋白质表面可以使用三维α□shape理论描述成由许多三角形构成的网。CAST就是利用这种描述来定义和寻找蛋白质的小分子结合位点。它首先得到以蛋白质表面的原子为节点的三角形集合，然后对这些三角形进行合并操作。所谓合并是将一个小的三角形合并到它周围大的三角形，小分子结合位点就是最后得到的空三角形的集合。CAST对SURFNET方法测试过的67个酶□配体中的51个酶□配体做过测试，准确率达到74%。

CAST算法示意图。（a）使用泰森多边形法对原子进行划分。（b）α□shape（灰色的三角形和黑色的线）和形成的3个至少包含一条黑线的三角形，这3个三角形就是需要合并的对象。（c）三角形合并示意图，两个小的三角形1和3合并到三角形

以上介绍的方法都基于蛋白质表面的纯几何信息进行预测，是纯几何算法，而不需要知道任何关于小分子以及分子间作用关系的信息。此外，还有基于能量的计算方法，其一般原理是计算蛋白质表面结合能量最高即分子间作用力最大的位置，这些位置可能就是小分子结合的位置。比如Q□SiteFinder就是一种基于能量的计算方法，它只是简单地计算了蛋白表面结合位点的范德华作用力（van der Waals interaction energies）。另一类基于能量的计算方法考虑的因素就比较多了，不仅考虑到与小分子化合物相关的理化属性，还考虑到不同的分子或者基团以及与溶剂的相互作用情况，是真正的基于能量的计算方法，如GRID（Goodford，1985）和FLApsite。下面就简单介绍一下Q□SiteFinder和GRID算法。

Q□SiteFinder方法是在蛋白质的表面初始化一层甲基（□Ch3）球形探针（Sphere probe），并计算蛋白质和每个探针之间的范德华作用势能，势能强的探针将被保留，因为小分子将会结合在分子间作用力强的区域，这样就会形成很多个相互独立的探针聚集区域，可以计算每个区域内部的探针数量，然后按照探针的数量大小对这些区域进行排序。拥有探针数量最多的区域，即结合能最大的区域将会排在首位，作为潜在的小分子结合位点。Morita等优化了Q□SiteFinder计算相互作用能量的方法，使用了更好的散射探针技术和更密集的势能参数。优化后的Q□SiteFinder对35个互相结合的蛋白□小分子复合物结构进行测试，准确率达到了80%。

GRID（Goodford，1985）计算了分子探针和结合区域中重叠的三维网格受体的作用能量。这个相互作用能量是里那□琼斯位能（Lennard□jones）、库伦力（Coulombic）和方向性氢键（Directional hydrogen□bond）的能量总和。探针分子可以是很多种不同的化学基团，而且这些基团可以自由转动达到最佳的结合状态来计算结合能（Goodford，1985）。

2.metapocket算法

metapocket方法是我们开发的寻找蛋白质与小分子结合位点的算法工具。我们发现，如果只是单一使用某一个预测算法来寻找蛋白质小分子的结合位点，总是不能达到很高的精度，也没有可比性，而且很多算法的代码是开源公开的或具有web服务器，于是我们考虑综合很多种现有的优秀的预测算法来提高算法精确性，这就是meta思想的起源。事实证明，使用meta方法可以使算法的整体预测精度明显提高。

210bound：Top1：81%Top3：95%程序结构4个预测方法线性执行进行；容错能力弱8个预测方法并行进行，速度更快；容错能力强*预测精度：精度测试数据集与LIGSITECSC算法中使用的数据集（huang等，2006）一致。

对每一个蛋白质结构，第一步，使用LIGSITECS，SURFNET，Q□SiteFinder，pASS，Fpocket，ConCavity，GhECOM和pOCASA这8个基本的预测算法进行蛋白质□小分子结合位点的先行预测。第二步，实际上每个预测算法给出的结果都是经过聚类分析的，存在一个个簇，因此这一步就是计算每个簇的重心作为这个簇的结合位点的坐标，然后以标准的pDB格式文件输出预测的结果数据。为了配合程序的容错性机制，需要对每个预测算法的最终结果进行检验，验证是否给出了正确的结果，如果没有给出正确结果，则该算法将不参与后期的计算。由于每个算法在打分函数和聚类分子中采用的策略不一样，所以它们给出的预测结果不能直接地用来比较，因此为了使每个算法的预测结果有比较性，我们对每个算法的结果都计算一个Z□score值。然后对算法结果中的节点按照Z□score值的大小进行降序排列，取每个算法的前3个数据进行后期计算。这个过程的结果是得到了N×3个独立的结合位点坐标数据（N代表了成功执行的算法个数），接下来将对这些坐标进行两轮的聚类分析。metapocket使用简单层次聚类方法进行聚类分析。第一轮聚类分析将空间相似性（Spatial similarity）相近的结合位点归类到一起。相似性距离阈值（Distance threshold）为8埃，这个过程的结果是在蛋白质的表面形成了一个个包含不定个数结合位点的簇（Cluster）。第二轮聚类分析针对第一次聚类分析生成的簇进行，考察这些簇的空间相似性，然后将两个或者多个簇进行合并，结果是形成了最终的包含不定个数结合位点的簇。然后，对每个簇计算一个Z□score值，这个值是簇内所有结合位点的Z□score值的总和，metapocket将按照Z□score值对簇进行降序排序，拥有最大的Z□score值的簇就是最大的结合位点，也是小分子最容易结合的地方。metapocket同样使用了LIGSITECSC中用到的测试数据集。

metapocket的界面用户可以输入要预测的蛋白质pDB的ID和对应的蛋白质链的ID，或者上传自己的pDB格式的蛋白质数据，通过简单的参数设置即可进行预测计算。metapocket网站使用了JMol插件，给用户提供了在线查看预测结果的功能，在线查看预测结果的界面。

metapocket web服务器中预测结果的JMol可视化界面

展示了两张metapocket的预测结果，在这两个蛋白质中，metapocket准确地在第一位预测出小分子的结合位点。其中，蛋白质表面被标记成绿色，小分子呈棍状表示，并标记成柠檬色，metapocket预测的结合位点为红色标记，ConCavity预测的结合位点为黄色标记，Fpocket预测的结合位点为蓝紫色标记，GhECOM预测的结合位点为蓝色标记，LIGSITECS预测的结合位点为粉红色标记，pASS预测的结合位点为火红色标记，Q□SiteFinder预测的结合位点为海洋蓝色标记，SURFNET预测的结合位点为橘黄色标记。此外，蛋白质表面的结合位点区域也标记成黄色。

3.蛋白质□小分子对接与虚拟筛选

虚拟筛选的定义：针对药物靶标蛋白质的三维结构或定量构效关系（Quantitative structure□activity relationship，QSAR）模型，从现有小分子数据库中，搜寻与靶标蛋白质结合或符合QSAR模型的化合物。虚拟筛选的目的是从几十到上百万个分子中发现有潜力的化合物，集中目标，大大降低实验筛选化合物数量，缩短研究周期，节约研究经费。蛋白质□小分子对接（protein□ligand docking）是依据配体（小分子）与受体（蛋白质）作用的“锁□钥原理”（lock and key principle），模拟小分子配体与受体蛋白质的相互作用。小分子与蛋白质相互作用是分子相互识别的过程，主要包括静电作用、氢键作用、疏水作用、范德华作用等。通过计算，可以预测两者间的结合模式及亲和力，从而进行药物的大规模虚拟筛选。蛋白质□小分子对接首先在蛋白质表面产生一个填充小分子表面的口袋或凹槽的球集，然后生成一系列假定的结合位点。这些结合位点信息可以由前一节介绍的位点预测算法得到或根据实验数据得到。依据蛋白质表面的这些结合点与小分子的距离匹配原则，将小分子投映到蛋白质分子表面，来计算其结合模式及亲和力，并对计算结果进行打分，评判小分子与蛋白质的结合程度。

给出了基于蛋白质□小分子对接的虚拟筛选流程，整个流程都可利用软件自动完成，例如AutoDock（Morris，1998；Morris等，2009）、Dock（Kuntz，1982）、GOLD（Verdonk等，2003）、FlexX（Kramer等，1999）等，但是这些软件的对接算法、打分方程、网格计算等都不尽相同。

常用蛋白质□小分子对接软件的引用比例，来自Sousa等的图二（Sousa等，2006），Copyright（2006），经John Wiley and Sons许可复制

4.常用蛋白质□小分子对接软件介绍

（1）AutoDock

AutoDock是由Scripps的Olson科研小组开发的分子对接软件包，用于预测小分子（如底物或药物候选化合物）与已知3D结构的蛋白质受体的最佳结合方式（Morris，1998）。AutoDock采用模拟退火和遗传算法来寻找小分子和蛋白质最佳的结合位置，用半经验的自由能计算方法来评价小分子与蛋白质之间的匹配情况。整个对接程序由AutoDock和AutoGrid组成。AutoDock执行配体和描述靶标蛋白的网格之间的对接；而AutoGrid预先计算这些网格。AutoDock已广泛应用于X□ray晶体学、基于结构的药物设计、先导化合物的优化、虚拟筛选、组合库设计、蛋白质□蛋白质对接及化学机制研究等。参与AutoDock研发的小组非常多，版本经常更新，最新的版本为AutoDock。AutoDock 4是一个自由软件，以GpL协议发行，可免费下载。与以前的版本相比，AutoDock 4主要包括以下改进：更加精确与可信的对接结果，可指定靶标蛋白质的柔性，可用于蛋白质□蛋白质相互作用的评估。AutoDock开发小组也开发了一个图形用户接口程序AutoDockTools（简称ADT），可帮助用户方便地完成小分子对接计算以及结果分析。

（2）AutoDock Vina

AutoDock Vina是用于药物发现、分子对接和虚拟筛选的新的开源程序，有多核运行、高性能、高准确度和易于使用等特点。AutoDock Vina 是由Scripps研究机构的分子图形实验室的Oleg Trott博士设计并应用的。在Vina中，用户不需要理解其实现细节及微调搜索参数和对结果进行聚类，只需要提供对接的分子结构和结合位点的空间描述，网格图的计算、原子电荷的分配及对接过程都在后台自动化实现。相对于AutoDock4，Vina显著提高了结合模式预测的平均准确率，并且简化了分子对接的前期准备过程，更便于不熟悉编程的科研人员使用。由AutoDock Vina把小分子Imatinib对接到c□Abl激酶上的一个构象。

AutoDock Vina对接结果。这是对接结果中排名第一的小分子构象（红）与真实小分子结构（绿）的对比。可以看出经过分子对接所产生的结果与真实的结构比较接近。小分子为Imatinib，蛋白质C□Abl kinase结构未显示（pDB ID：1IEp）

（3）DOCK

DOCK由Kuntz小组于1982年开发，是第一款用于分子对接的软件，最新版本为DOCK 6.4（Kuntz，1982）。DOCK的开发经历了一个由简单到复杂的过程：DOCK1.0考虑的是配体与受体间的刚性形状对接；DOCK2.0引入了“分而治之”算法，提高了计算速度；DOCK 3.0采用分子力场势能函数作为评价函数；DOCK 3.5引入了打分函数优化以及化学性质匹配等；DOCK4.0开始考虑小分子配体的柔性；DOCK 5.0在前面版本基础上，采用C++语言重新编程实现，并进一步引入GB/SA打分。DOCK程序现已成功地应用于药物分子设计领域。DOCK 6.0对5.0进行了扩展，在包括原有特性的基础上，增加了最小化过程中额外的打分选项、pB/SA溶解打分等。

（4）FlexX

FlexX是德国国家信息技术研究中心生物信息学算法和科学计算研究室开发的分子对接软件，目前已经作为分子设计软件包Sybyl的一个模块实现商业化（Kramer等，1999；Cross，2005）。其使用碎片扩增的方法寻找最佳构象，根据对接自由能的数值选择最佳构象，对接速度快，效率高，可以用于小分子数据库的大规模虚拟筛选。总体来说，FlexX是一套非常全面的药物设计软件包，包含各种对接方法，如基于药效团限制条件的对接、考虑受体柔性的对接等。而它的柔性对接不仅考虑到活性位点蛋白的构象，更重要的是它能给出整个受体蛋白随配体的变化得到的蛋白构象，其快速对接的特点令用户完全可以在组合化学合成之前将虚拟的化学库进行对接和评价。

（5）GOLD

GOLD是一款用来计算小分子结合到蛋白质结合位点的对接程序，是GOLD软件包（GOLD软件包中hermes用来结构可视化和操作，GOLD用来完成蛋白质□小分子对接，GoldMine用来作对接结果的后期处理和可视化）的一部分。在分子对接领域，GOLD以其准确性和可靠性而著称。GOLD应用遗传算法实现蛋白质和小分子对接，通过hermes实现简单的用户交互界面。在GOLD算法中，用户可以对整个小分子和蛋白质的部分区域进行柔性处理，还可以进一步自定义10个氨基酸的蛋白质侧链和主链柔韧性。GOLG的打分函数有GoldScore，ChemScore。GOLD还提供应用编程接口（ApI），允许用户修改GOLD打分函数以便实现自己的打分函数或者提高现有的打分函数。

这些常用的蛋白质□小分子对接算法和软件中，在准确率上，没有哪个算法占有绝对的优势，每个对接软件都在对某个对接对象或集合进行对接时比较有优势。

虚拟筛选过程中，需要对大规模的小分子化合物进行前期处理。这些小分子的来源主要是网络上可用的公共数据库，下面介绍网络上三个比较常用的小分子数据库，这些数据库都可进行在线查询得到想要的小分子化合物，或者可以进行全部下载，在本地进行查询和对接。

（1）DrugBank

DrugBank数据库是一个独一无二的生物信息学和化学信息学资源库，包含详细的药物信息（比如化学、药理学和制药学信息）和综合的药物靶点信息（比如序列、结构和生物途径信息）。该数据库包括近4800个药物条目，其中有经美国食品药物管理局（FDA）批准的1350种小分子药物和123种生物大分子药物（蛋白质、肽段）；71种营养药品；3243种正处于临床测试阶段的药物。此外，将经FDA批准的药物条目链接到2500条非冗余蛋白质（比如药物靶标）序列上。每个药物条目包含100多个数据域，其中一半是关于药物或化学物属性的信息，一半是关于药物靶标蛋白质属性的信息，包括商品名、化学结构、蛋白和DNA序列、互联网上的相关链接、特征描述及详细的病理信息等。DrugBank支持全面而复杂的搜索方式，包括文本搜索、序列搜索、化学式搜索，而且提供很好的专用搜索工具，便于不熟悉数据库搜索的科研人员使用，来完成其研究需要，比如，检索新的药物靶标、比较药物结构、研究药物机制以及探索新型药物等。

（2）ZINC

ZINC是一个可用于化合物虚拟筛选的免费数据库（Irwin，2005）。该数据库包含了130多亿个可购买到的小分子化合物，这些化合物都可用于蛋白质□小分子对接，并且都具有3D结构信息。ZINC由美国加利福尼亚旧金山大学药物化学部的Shoichet实验室建立，其设计初衷是用于基于结构的虚拟筛选，因此ZINC团队致力于分子的存储表达方式。它提供了多种分子数据格式（比如，MOL2，SDF，SMILES），并且提供这些分子的来源及可购买信息的连接。

（3）pubChem

pubChem分子生物活性数据库是NIh分子库计划的一个组成部分（Wang，2009），由NCBI的Entrez信息检索系统的三大数据库组成，包括：pubChem Substance，包含6900万多条记录；pubchem Compound，多达2700万个唯一的结构；pubChem BioAssay，包含43.4万条生物测定，每条生物测定包含很多个信息点。这三大数据库之间包含相互的链接。pubChem也提供了一个快速的化学结构相似性搜索工具。pubChem的化学结构记录与其他Entrez数据库的链接可以提供更多相关生物信息，包括与pubMed科学文献和NCBI的蛋白质3D结构资源的链接。

5.蛋白质□DNA相互作用

蛋白质和DNA相互作用后形成的复合物在许多细胞活动中扮演着极为重要的角色。比如最为常见的转录因子，它在基因表达中结合到某基因上游的特异核苷酸序列上，从而调控基因的转录；又比如DNA复制、DNA修复、病毒转染、DNA折叠和修饰等，这些过程中都有DNA结合蛋白的参与。但是，目前有关蛋白质和DNA结合的生物机制及过程还不是很清楚，想要通过X射线的实验方法得到他们的结晶复合物结构还是非常困难的。因此，有关蛋白质和DNA相互作用的研究备受关注，许多研究小组尝试用生物信息学预测的办法得到蛋白质和DNA结合的相关信息，如结合位点等（Ahmad等，2004；Wang等，2006；Yan等，2006；hwang等，2007；Ofran等，2007；Tjong等，2007；Andrabi等，2009）。

在蛋白质结构数据库pDB中，目前已有超过6万个实验测定的蛋白质结构（Berman等，2000），其中有关蛋白质□DNA复合物的结构还不到1000个，而真实存在的蛋白质□DNA复合物的数量远远超过这个数量。许多科研人员尝试通过蛋白质的序列信息或结构信息来预测蛋白质的DNA结合区域。因结构数据有限，基于蛋白质序列信息预测的方法更为广泛和有效。目前，主流的基于序列信息的DNA结合位点的预测服务器有DISIS（Ofran等，2007），DNABindR（Yan等，2006），BindN（Wang等，2006），BindN□rf（Wang等，2009），Dp□Bind（hwang等，2007）和DBS□pRED（Ahmad等，2004）等。这些预测算法使用了多种多样的特征属性，比如：氨基酸的频率、进化信息、序列保守型、二级结构信息、溶剂可溶性信息、静电势、疏水性、BLOSUM62矩阵、位点特异性打分矩阵等。预测方法涵盖了常见的机器学习方法：支持向量机（Lv等，2010）、贝叶斯网络、人工

神经元网络（De Roach，1989）、随机森林（Calle 等，2011）等。总结了几种最重要的预测方法所用的特征属性、预测方法以及Web网址。另外，还有其他一些DNA结合位点预测的方法，因未能提供在线服务器等原因，这里不再赘述

第41章 蛋白质□DNA 结合位点预测算法

1.DISIS

DISIS的在线网址是这种方法发表在2007年的生物信息学杂志上（Ofran等，2007）。DISIS构建了独特的氨基酸残基数据库，其中包含23862个与DNA结合的氨基酸残基和103202个不与DNA结合的氨基酸残基。DISIS的特征属性包括了进化信息（Evolutionary profile）、序列保守型、预测的二级结构和预测的溶剂可及性（Solvent accessibility）。预测方法上，DISIS采用了支持向量机和神经元网络结合的办法。因为引进了进化信息，因此其运算速度慢，计算一个蛋白约需要5~10分钟。

2.DNABindR

DNABindR是美国犹他州立大学的研究者于2006年开发的蛋白质的DNA结合位点的预测服务器。其构建的数据库pISCES包含从pDB数据库中筛选出来的171个蛋白质□DNA复合物。筛选原则是蛋白质的序列同源性≤30%；蛋白质结构的分辨率3.0□□序列长度至少40个氨基酸。此方法运用的机器学习方法是贝叶斯网络方法。预测用到的特征属性包括：相对溶剂可及性，序列熵（Sequence entropy），二级结构，静电势和疏水性。这种方法识别DNA结合位点的总体准确率（Accuracy）达到71%，对应的敏感性（Sensitivity）为53%，特异性（Specificity）为35%。

3.BindN

BindN 是一种在线预测DNA/RNA和蛋白质结合位点的工具，传输服务，预测结果可以直接在页面上显示。机器学习方法是支持向量机方法。此方法用到的特征属性有：侧链pKa值，疏水性及氨基酸的分子质量。BindN用到两个数据集：①pDNA□62，包含62个蛋白质□DNA复合物，该数据集在之前的研究工作中被用过多次；②pRINR25，包含174个蛋白质□DNA复合物，是当时从pDB数据库中按照蛋白质序列相似性≤25%，结构分辨率3.5□的原则筛选出来的。

4.BindN□rf

BindN□rf 与BindN由同一研究小组开发，BindN□rf 运用了新的机器学习方法——随机森林法，又增加了几项特征属性，预测精度亦得到了提高。BindN□rf增加了一个新的数据集——pDC25t，该数据集的数据不包含在pDNA□62中，两两之间的序列相似性也≤25%。除了BindN中用到的序列特征属性外，新增加了基于BLAST（Altschul等，1990）的序列保守型、生物化学属性和位点特异打分矩阵pSSM。

5.Dp□Bind

Dp□Bind是一个基于蛋白质序列信息的在线预测方法。Dp□Bind使用了三种机器学习方法，分别是支持向量机方法、内核逻辑回归（Kernel logistic regression）和分逻辑回归（penalized logistic regression）。网站还允许用户自己选择特征属性，可选基于序列的BLOSUM62矩阵或者基于位点特异性的pSSM打分矩阵（position specific scoring matrix）。

6.DBS□pRED

DBS□pRED由日本Shandar Ahmad开发。这个课题组做了很多有关DNA和蛋白质结合位点的工作，还开发了另外一种DNA结合位点预测的工具——DBS□pSSM，其跟DBS□pRED的区别是引入了位点特异性打分矩阵（pSSM）的信息。DBS□pRED使用了三个数据库测试：①被很多科研人员使用过的pDNA□62数据集；②NRTF□915，是从SWISS□pROT数据库里收集的无冗余的转录因子蛋白质；③CNTR□3332，是从SWISS□pROT数据库搜集的非转录因子的对照数据集。其使用的机器学习方法为神经网络，使用的特征属性有蛋白质序列信息、溶剂可及性和二级结构信息。DBS□pRED通过email向用户提供预测结果，计算速度比较快。

7.metaDBSite 算法介绍

目前，尽管DNA结合位点预测的方法有很多种，但每种方法都有它本身的特点，特征属性不同，预测方法不同，所用的数据集不同，预测的精度也各有千秋。在这种情况下，用户在使用时就面临如何选择的难题。为了解决这个难题，我们课题组开发出了一种统一整合各种服务器预测结果的DNA结合位点的预测系统——metaDBSite。

metaDBSite用户只需提供目标蛋白质序列和email地址，metaDBSite可以在十几分钟内完metaDBSite web服务器预测主界面

成计算，并将预测结果自动发送到用户提供的email邮箱。用户提交的蛋白质序列被自动提交到DISIS，DNABindR，BindN，BindN□rf，Dp□Bind和DBS□pRED六个在线服务器上进行预测。预测完成后结果被自动下载到后台并进行预处理，之后作为支持向量机的输入值进行预测。为了与其他单个算法比较及验证metaDBSite算法的准确率，我们从pDB数据库中搜集整理出一个序列冗余性30%的DNA□蛋白质复合物的数据库（包含316个蛋白质□DNA复合物）作为我们的训练数据集。结果表明，metaDBSite在预测DNA结合位点上的表现优于各种单个方法，能够成为DNA结合位点预测的一个有用工具。两个metaDBSite预测DNA结合位点的例子，pDB ID是1CMA_A和206M_A。其中蓝色区域是metaDBSite预测的DNA结合区域，红色区域是真正的DNA结合区域。从图中可以看到，metaDBSite比较好地预测出DNA结合区域。但是，有一些非DNA结合位点被预测成了结合位点。与其他方法相比，metaDBSite已经降低了假阳性，但是仍未能根除假阳性。这也对DNA结合位点预测算法的改进提出了更高的要求。

metaDBSite 预测流程示意图，“+”号表示DNA结合位点，“-”号表示非DNA结合位点两个metaDBSite预测蛋白质（色氨酸操纵子，pDB编号1CMA_A；多头绒泡菌，pDB编号206M_A）DNA结合位点的例子。蓝色是metaDBSite预测的DNA结合区域（1CMA：26个氨基酸；206M：54个氨基酸）。红色是真正的DNA结合区域（1CMA：7个氨基酸；206M：17个氨基酸）。所有真正的DNA结合区域都被正确预测到了（精确度值为100%）11.4小言

由于蛋白质与其他分子之间的相互作用一直以来是一个研究热点，许多相关的生物信息学工具已经被开发出来，本章介绍的算法、软件及数据库只是其中的一部分，难免有些遗漏，再加上版本的更新，不足之处希望读者见谅。为了阐述方便，本章用到的一些算法及流程示意图标注了引用文献及出处。本章介绍到的一些生物信息工具是由本课题组最近几年开发的，如蛋白质□蛋白质作用位点预测软件metappI、蛋白质□蛋白质对接软件BDOCK、蛋白质□小分子作用位点预测软件LIGSITEcsc和metapocket及蛋白质□DNA作用位点预测软件metaDBSite。今后本课题组会进一步提高和完善相关预测算法，开发

出更好的生物信息学工具，供广大的相关生物科研人员使用，提高对蛋白质与其他分子相互作用机制的了解。

第42章 系统生物学模拟工具

目前，系统生物学已有BioSpice、DBSolve、E□Cell、Gepasi、Jarnac、StochSim和Virtual Cell等模型模拟分析语言和软件。本章主要介绍SBML、CellDesigner和Cytoscape三个软件，它们功能相对综合，具有较好的应用前景。

1.系统生物学标记语言

系统生物学标记语言（Systems biology markup language，SBML）是美国加州理工学院教授Erato Kitano所提出的一种基于XML与UML来描述和分析系统生物学的模拟软件，用来制作计算机可读的生物网络模型，包括代谢网络、细胞信号传导网络、基因调控网络和其他网络（Finney和hucka，2003；Gillespie，Wilkinson等，2006；Schmidt和Jirstrand，2006；Schulz，Uhlendorf 等，2006；Rodriguez，Donizelli等，2007）。XML是一种在万维网（Internet）上广泛使用的传输和交换数据的格式。

开发SBML的目的是使其不依赖于特殊工具和软件就能进行生物学模型描述，并且用SBML描述的细胞生物网络模型是可以被携带的（Model transportability），即一个用SBML描述的模型可以在不同用户与平台间相互交流，无论对方使用什么软件、硬件或者操作系统，只要用户使用与SBML兼容的软件，就可以对同一模型进行模拟，而无需针对用户操作系统进行特定修改，能够更方便直接地运用模型。此外，模型的互通性（Interoperability）还有很多优势，一方面，用户可以共享模型，另一方面，他们可以使用不同模拟软件对同一模型进行模拟，从而比较整合各软件的结果。

再者，以SBML格式作为网络模型描述的标准，有利于模型的发表和共享。以往发表的模型都用微分方程、代数方程、反应方程式或者网络图谱等来描述。如果以SBML格式描述模型，并通过网站共享，则能够让读者及文章审稿人下载并独立进行测试，从而观察该模型是否能够得到所描述的效果。目前，有些杂志已要求将SBML作为网络模型描述标准，比如，“Molecular Systems Biology”杂志等，他们要求作者提供SBML格式的数据和模型。

最后，用SBML格式描述的模型将有更长的使用寿命。因为以往的模型都有其独特的数据输入格式，并且必须使用自己编写的软件来进行模拟，所以这样的模型常常只能使用特定的软件进行模拟。因此，假如一位研究生或者博士后编写了一个模型，当他离开后，该模型也会随着其模拟软件的丢失而丢失。另外，商业软件也常常要求特定的输入格式和特定的数据格式，因此，当其软件更新或者不再支持原有软件或硬件时，其对应模型也同样会丢失。因此，将SBML格式推广成生物网络模型描述的标准化将大大促进生物网络的研究。

验证软件用于验证SBML编码的模型是否符合SBML、XML格式。可视化软件能够结合Graphviz dot 和XSLT脚本，并利用网络浏览器来显示格式为gif的图像。用户可以直接在sbml。org网站上进行模型模拟，但可视化功能只局限于含有100个以下反应式的模型；如果下载XSLT脚本到本地，并利用相应软件，就可以对含有任何数量反应式的模型进行模拟。

2.LibSBML

LibSBML是一种针对C语言或者C++语言执行的函数库。其提供了应用程序的编程接口——ApI，可以读取、存储、验证以及操作用SBML表示的文件和数据。而且LibSBML还包括Java、python、perl、Lisp、MATLAB等语言，基本上兼容了常用计算机语言，并且具有携带性，可以在Linux、Windows、苹果系统等不同平台上使用。

3.基于MATLAB的SBML工具箱

SBML工具箱（Systems biology toolbox）是一个基于MATLAB语言和LibSBML函数库开发的软件工具包，在Windows、Unix、Cygwin和苹果系统上都能运行。它可以把SBML模型转换成MATLAB可读的格式，然后利用标准的MATLAB求解器Solvers和模拟器对模型进行模拟。目前，该工具箱包括将SBML模型转换成MATLAB数据结构的函数，以及读取和操作这些结构的函数等。MATLAB的Symbolic格式已经用MATLAB的常微分方程求解器（ODE solver）进行模拟。最新版本SBML工具箱（SBTOOLBOX2）（www。sbtoolbox2.org）设计图

SBTOOLBOX2主要功能如下：①展示模型、测量数据及实验描述；②SBML格式读入和输出；③特定模拟和随机模拟（Deterministic and stochastic simulation）；④可视化（Visualization）；⑤稳定态和稳定

分析（Steady□state and stability analysis）；⑥代谢控制分析；⑦化学计量分析（Stoichiometric analysis）；⑧参数敏感度分析（Local，Global，Customizable metric，Oscillating systems）；⑨部分保守性（Moiety conservations）分析；⑩分叉点分析（Bifurcation analysis）；决定复杂行为的局部机制分析；统计方程；信号处理方程；15个全局和局部优化算法；线性规划和二次方程式规划求解器（Linear and quadratic programming solvers）。

最近，针对SBML工具箱（SBTOOLBOX2）还开发了SBpD扩展包（the SBpD extension package），为SBML提供更多支持，包括高速模拟、整合项目中的模型、试验和方法数据；同时增加了新功能支持完整的模型构建过程，包括模型构建、模型模拟、模型分析、模型缩减以及参数估计、验证等。其主要功能如下：①综合模型、模型试验和测量数据形成课题；②自动生成C语言编码的模拟模型，可比MATLAB integrator 提高模拟速度30～200倍；③参数估计（多实验、多参数）；④参数吻合分析（parameter fit analysis）；⑤残差分析（Analysis of residuals）；⑥可辨识性分析（Identifiability analysis）；⑦模型降阶分析（Model reduction）；⑧37个预先安装的动态速率定律（Kinetic rate laws）；⑨动态灵敏度分析（Dynamic sensitivity analysis）；⑩模型参数的手动调节。

4.基于Mathematica的MathSBML

数据模拟软件Mathematica为用户提供符号运算环境，目前运用非常广泛。符号计算，又称为计算代数系统，可以对方程式进行符号化处理，可以将反应式、能量以及作用方法用符号表示等，这对生物学家研究生物分子网络互作是很方便的。开发MathSBML的主要目的是使Mathematica可以读取处理SBML模型，并将其转换成基于Mathematica的模拟及绘图等回归方程式，从而用Mathematica计算并显示。因此，Mathematica所有标准功能都针对SBML模型运行。另外，MathSBML利用微分代数方程及应用程序编程接口对SBML模型进行操作及绘图，并可以html网页格式显示结果，使用户更容易和直观地理解结果。

第43章 MathSBML软件具体应用

1.SBML 格式转换

SBML格式转换程序能把其他语言描述的模型转换成SBML格式的模型。目前，针对两个模型转换的程序有KEGG2SML和CELLML2SBML。在SBML网站上也有一些软件，可以将SBML1模型转换成SBML2模型。

KEGG2SML是perl语言程序，可以将KEGG生物代谢途径数据库文件转换成SBML格式文件。它基于LIGAND 数据库来完成此转换功能，并且兼容SBML所有版本，同时还提供支持CellDesigner的标签（）。KEGG生物代谢途径数据库拥有非常完整的数据库及应用软件，包括蛋白相互作用数据库、生物代谢途径数据库、化学反应数据库（LIGAND）、序列相似性数据库（SSDB）、基因功能分层分类数据库（BRITE）等。KEGG2SML 兼容FreeBSD、Linux以及微软Windows的Cygwin平台，其运行需要先安装perl、Expat、perl函数库XML parser和libxml□perl，以及KEGG数据库。KEGG数据库可以从KEGG官方网站下载。

CELLML2SBML可以将CELLML格式模型转换成SBML格式模型。CELLML也是一种基于XML的模拟语言，用来储存和模拟生物学模型。虽然这两种语言功能相似，但是其目的不同。CELLML主要针对解剖学和细胞成分的模拟，它有AnatML和FieldML等语言分类。目前，CELLML开发者也在开发SBML2CELLML工具，以将SBML格式模型转换成CELLML格式模型。辅以XSLT处理器，CELLML2SBML可以在Linux和Windows系统平台运行。

2.应用程序编程接口

MathSBML中的应用程序编程接口（Application programming interface，ApI）可用ApI控制的SBML的特征；除ApI以外的MAThSBML的指令模块。

3.MathSBML的主要指令

MathSBML 的主要指令如下所示：

SBMLRead读取SBML的模型到Mathematica软件；SBMLNDSolve 提供微分代数等方程解；SBMLplot为SBMLNDSolve的求解提供画图功能；SBMLWrite可把SBML 模型转换成其他模型；SBMLCopy把机器生成的SBML 文本变成易被人读的格式；Model Builder 提供构建模型所需的一系列功能，包括增加或减少模型中的客体（Objects），即反应式、物种等。4.SBML模型读取（SBMLRead）

生物学模型数据库（http：//www。ebi。ac。uk/biomodels/）是用来储存、寻找和共享已发表生物学模型的数据库，由一个包括美国、英国、日本、印度等多国科学家国际合作团队发展起来的。其中的生物学模型都有注释，及与其对应的数据来源，如发表文章、功能途径、化合物数据库和功能词汇（Controlled vocabularies）等。它还可以通过各种关键字对生物学模型进行搜寻，如模型ID、模型构建者、模型对应发表文章。生物模型数据库中的模型可以直接下载用于计算和模拟（Le Novere，Bornstein等，2006）。

通过生物学模型数据库Biomodels（http：//www。ebi。ac。uk/biomodels/）可以搜索并下载到很多以SBML格式储存的模型，包括细胞模型、细胞周期模型等。例如，对细胞模型感兴趣，我们可以在生物模型（Biomodels）数据库中搜索酵母菌DNA复制模型（BIOMD0000000056.xml），并将其下载到本地计算机。然后利用SBML模型读取指令模块（SBMLRead）就能够将该模型读入Mathematica的计算环境中，从而用户即可进行下一步分析工作。

第44章 CellDesigner 的主要特征

在生物学研究中，很多生物调控网络（包括基因调控网络和生化反应网络）用图形来表示。如能开发一个能够对这些生物网络图形进行编辑及模拟的软件，将大大方便生物学研究工作。因此，科学家开发了软件CellDesigner，它基于系统生物学图形标注（Systems biology graphical notation，SBGN），与SBML兼容，并且能利用基于系统生物学工作台（Systems biology workbench，SBW）的软件工具进行模拟和分析。

CellDesigner主要功能特征有以下几点：

●与SBML相兼容，包括SBML 1和SBML 2；

●与基于系统生物学工作台（SBW）的模拟分析软件相兼容；

●与SBMLOEDsolver，即常规微分方程求解器相兼容，进行模拟分析；

●能够直接链接数据库，进行数据交流；

●基于Java编写，通用于不同操作系统平台。

CellDesigner的最终目标是发展成一个标准化的生物网络途径、图形编辑器。其标准化主要归结为三方面：①图形标注（Graphical notation）；②模型描述（Model description），使模型能够在不同工具软件中相互交流及运用；③应用软件的整合环境（Application integration environment），CellDesigner设计出的模型可以用其他应用软件进行模拟和分析。

1.系统生物学图形化标注

由于人类大脑对图形有更直观感受，因此使用图形来表示生物学模型更直观有效。虽然SBML是一种机器可读的软件模型，并可以用来在不同软件之间交流生物学模型，但对于普通人来说，它是不可读的，所以需要开发一个生物学模型的可视化标准。针对这一需求，日本东京Kitano实验室最先提出了系统生物学图形化标注（Systems biology

graphical notation，SBGN），并开发了可应用的系统生物学模型的标准，能够支持SBML格式的图形化。在计算机科学及工科领域，都已有各种图形化标注标准。因此，提出生物学模型图形化标注标准，将对系统生物学的发展起到促进作用，能够让更多用户使用图形化标注软件构建生物学网络模型，并对其进行模拟分析。

系统生物学图形化标注的主要目标有：①用图形标注来代表不同种类的生物分子及其之间的相互作用；②使图形标注在语意和图形上明确而不模糊；③可以附加其他注释；④可以用其他模拟软件工具将图形表示的模型转换成数学公式，从而对它进行分析和模拟；⑤可以用软件来支持和帮助图形化过程；⑥其注释和标示可以灵活变换。Kitano实验室提出的系统生物学图形标注标准版本1中的流程图标示（process diagram）。如，蛋白由非直角长方形表示，并且蛋白具有两种状态，即非激活状态和激图系统生物学图形标注的标准（SBGN Level 1）活状态，激活状态的蛋白在图标外面加上虚线。同时，在该长方形图标角落还可以用圆圈来注释该蛋白的修饰状态，如磷酸化、甲基化、酰基化等。相比于一些传统方法，Kitano提出的系统生物学图形化标注能更精确地描述生物学模型。例如，传统方法的生物学模型一般用箭头表示蛋白激活状态，如蛋白A与蛋白B结合并激活蛋白B，就用箭头来表示激活。但生物体常常有另一种状况，当蛋白A和蛋白B结合之后，并不能激活蛋白B，可能是由于B蛋白的某单个位点被磷酸化就不能被激活，而必须同时有多个位点被磷酸化才能被激活。因此，应用传统方法无法精确表示该状态信息，但使用Kitano提出的图形化标注就能更精确地表示这类状态详细信息。

SBGN构建的糖酵解过程。系统生物学图形化标注语言可用来标示代谢网络和代谢途径，包括描述简单化学分子、大分子以及它们之间的流程图、催化反应等。

SBGN构建的类胰岛素生长因子（IGF）的信号传导过程。系统生物学图形化标注语言可用来描述蛋白状态的变化、蛋白复合体以及细胞结构的分区；另外，流程图中利用亚图（Submap）可简化IGF信号传导的描述。

CellDesigner所绘图形可以用基于系统生物学工作台（SBW）的软件工具进行模拟分析。CellDesigner本身就是由SBW驱动的应用软件，安装SBW之后，在CellDesigners上就能够直接驱动模拟器，如从SBW菜单中选择模拟服务或者JARNAC模拟服务，就可以实时模拟。由SBW驱

动的模拟软件有许多，包括ODE（即常规微分方程模拟器）、Stochastic（随机模拟器）、MATLAB模拟器以及Bifurcation analysis（分歧分析软件）等。

第45章 Cytoscape数据整合及网络显示分析平台软件

1.Cytoscape简介

计算机辅助的生物学网络建模是系统生物学的基础。生物学网络包括生化作用网络、基因动态表达网络、细胞内分子作用网络、代谢调节网络等（Mendes，1997；Endy和Brent，2001；Loew和Schaff，2001）。通过生物学网络建模可以系统地整合信息，理清脉络，预测功能网络，为下一步的实验验证提供线索，最终揭示系统的调控机制及发现全新调控网络。

随着现代生物学技术（如基因芯片、蛋白质组学技术、高通量测序等）的迅猛发展，实验数据越来越多，产生的速度也越来越快，因此，海量数据的分析现已成为科研工作者的研究瓶颈。使用以往的方法已不能很好地对现代生物学数据进行分析，如何从海量的枯燥数字中快速、准确地寻找出有意义的数据对每一个科研工作者来说都至关重要。

通常，网络建模以细胞内生物途径或疾病发生发展相关信号转导通路为引导，利用大量已报道文献进行信息收集和挖掘。对于不同实验方法的数据，由于其格式不同，经常需要用不同软件来打开，并进行分析。pajek，Graphlet和daVinci可以用来显示基本分子相互作用关系。Osprey和pIMrider（pim。hybrigenics。com）不仅可以显示分子之间的相互作用，还可以将构建的模型和已有分子相互作用及功能数据库等连接比较，比如BIND（Bader，Betel等，2003）、DIp（Xenarios和Eisenberg，2001）和TRANSFAC（Wingender，Chen等，2001）等。对于表达谱等其他信息，也可用GeneCluster、Treeview和GeneSpring等软件进行分类和可视化。因此，如能把多种数据信息整合在一起，以一种多层次的方式显示，并整理归纳，则能大大促进生物学网络调控关键位点的预测。

Cytoscape软件（www。cytoscape。org）就是这样一个可以整合并图形化生物分子相互作用网络及各种高通量实验数据的免费开源生物信息平台软件（Shannon，Markiel等，2003）。该软件可同时导入蛋白□

蛋白相互作用、蛋白□DNA相互作用和基因相互作用等信息数据，并以图形化形式将各种相互作用关系呈现出来。Cytoscape基于Java平台，具有很好的扩展性，可以通过编写各种插件来增加其功能。它最初是由美国系统生物学研究（Institute of Systems Biology）、美国加州大学圣地亚哥分校（University of California，San Diego）、斯隆□凯瑟琳癌症研究所（Memorial Sloan□Kettering Cancer Center）、巴斯德研究所（Institute pasteur）和美国安捷伦公司（Agilent Technologies）多方合作开发的免费开源软件。如今美国加州大学旧金山分校（University of California San Francisco）、加拿大多伦多大学（University of Toronto）以及美国国家医学生物综合信息中心（National Center for Integrative Biomedical Informatics）等科研单位相继参与了Cytoscape的合作开发，全世界很多顶级生物信息实验室都为其开发了许多不同功能的插件，可以实现导入、输出各种格式数据并进行各种功能分析等。

Cytoscape的主要功能列举如下。

1.数据输入

（1）输入、建立分子相互关系网络文件（SIF格式），此文件包含了蛋白□蛋白或蛋白□核酸相互作用信息。对于酵母和其他模式生物来说，可从BIND和TRANSFAC等数据库下载得到网络文件；

（2）以GML格式载入和保存已经建立的网络文件；

（3）以XGMML（eXtensible graph markup and modeling language）格式载入和保存网络文件，并且保存每一个节点和相互作用的属性；

（4）导入信使RNA表达数据；

（5）导入Gene Ontology和KEGG数据。

2.可视化

（1）通过强大的可视化效果编辑自定义网络数据显示模式；

（2）通过不同颜色或不同视觉效果表示基因表达数据，如显示p值等；

（3）以二维方式显示数据；

（4）放大和缩小显示网络结构；

（5）通过高效渲染机制可以轻易显示超过10万个节点的大型网络。

3.插件

（1）网络分析，通过现有数据的过滤来选择子集及相互作用；

（2）找出关键的子集和信号通路；

（3）找出节点聚集簇。

Cytoscape的操作界面：区域1为软件控制区域，区域2为软件显示区域，区域3为属性显示区域。

2.插件安装

Cytoscape是一个开放的软件，全世界许多实验室为其开发了功能繁多的插件，因此正确安装和使用插件对全面提升Cytoscape的功能非常重要。

1.分析插件（Analysis plugins）：用于分析当前网络；

2.网络属性插件（Network and attribute I/O plugins）：用于导入不同格式的网络及属性数据；

3.网络推理插件（Network inference plugins）：用于根据实验数据预测出网络结构；

4.工程增强插件（Functional enrichment plugins）：用于增强网络功能；

5.通讯/脚本插件（Communication/scripting plugins）：用于通讯或编写Cytoscape脚本。

3.BiNGO插件的安装和使用

BiNGO插件能够在Cytoscape导入Gene Ontology数据，并为每一个基因添加基因注释，构建以目的基因功能为基础的结构图。

1.BiNGO 下 载 地 址 为 http ： //chianti 。 ucsd 。 edu/cyto_web/plugins/pluginjardownload 。 php ？ id=199 ， 下 载 得 到 BiNGO。jar，并将其复制到Cytoscape安装目录中的plugins文件夹中，重启Cytoscape就能在菜单栏plugin下找到“BiNGO 2.0”。

2.启动BiNGO后，Cluster name中输入新建功能结构图的文件名；paste Genes from Text中输入基因的名字；Select ontology中选择生物功能分类，包括分子功能（Molecular function）、生物过程（Biological process）和细胞组分（Cellular component）三类；Select organism中选择相应的种属。

BiNGO启动界面。Copyright（2001），经the American Association for the Advancement of Science 许可复制

酵母半乳糖代谢过程（Galactose utilization）中的半乳糖转换酶gal家族具有非常丰富的生物学功能，想了解gal家族中gal1，gal2，gal3，gal4，gal5，gal6，gal7在细胞中参与的具体生物功能，“Cluster name中”输入“gal”，勾选“paste Genes from Text”，在其下方输入框中输入“gal1gal2gal3gal4gal5gal6gal7”，中间以空格分开。选择细胞功能分类，并确认选中“Saccharomyces cerevisiae”。点击运行开始比对GO数据库，BiNGO的输出结果包括一个有众多细胞功能组成的网络图，显示了这7个半乳糖转换酶参加的所有生物功能，下面有一个详细输出窗口，显示了每个生物途径所包含的基因，此外还有p值颜色提示框。

酵母半乳糖代谢途径。半乳糖转换酶（gal）是一个复杂的转换酶家族（Ideker，Thorsson等，2001）

BiNGO运行结果视图，包括功能结构图、详细输出窗口以及p值颜色提示框

4.Cerebral插件的安装和使用

Cerebral插件可以根据已有数据自动生成一个以“分子在亚细胞单位内空间分布以及其分子互作”为基础的高度可视化的网络结构图。不同的

分子根据其所在细胞结构位置，在结构图中被分配至不同的结构层面，同时它们的功能与相互作用关系也被清晰表现出来。

1.Cerebral插件所需下载文件有两个：cerebral□v1.2.jar和prefuse。jar。下载文件存入Cytoscape安装目录的plugins文件夹中，重启Cytoscape，在菜单栏plugin的下拉菜单中就会出现“Create Cerebral view”、“Restore previous Cerebral views”、“Export Cerebral view”三个选项。同时左侧“Cytopanel 1”栏中出现新的“Cerebral”选项。

2.导入数据

①点击菜单栏Import→Network（multiple file types）导入tlr4_sif。sif文件；

②点击菜单栏Import→Node Attributes分别导入tlr4_localization。noa和tlr4_function。noa文件。

简单介绍一下noa格式文件：

用记事本打开tlr4_localization。noa文件，显示如下：

Localization

ACTB=cytoplasm

ACTC=cytoplasm

ApOE=extracellular

ATF2=nucleus

BIRC3=cytoplasm

BTK=cytoplasm

……

第一行数据为“Localization”，从第二行数据开始，第一列是分子名，第二列是“=”，第三列是分子在细胞中所在的位置。

3.点击菜单栏plugins→Create Cerebral view，在新建网络图的设置中，首先在Localization选择网络图的细胞结构排列顺序，再在Downstream nodes中选择interaction并设置网络图所要表达的分子相互作用种类，然后选择进一步分类的原则（例如Function），最后点击Create Layout，建立网络图。

4.以分子在亚细胞单位内空间分布以及其分子相互作用为基础的高度可视化的网络结构，其中绿色的区域代表所有具有相同功能的分子。当鼠标指针移到某一节点时，该节点所涉及的所有其他路径与节点都会被清晰表示出来。

以“分子在亚细胞单位内空间分布以及其分子互作”为基础的高度可视化的网络结构图。其中绿色的区域代表所有具有相同功能的分子
Agilent Literature Search插件安装和使用

Agilent Literature Search文献检索插件，利用最新pubMed文献信息的检索建立目标基因的网络相互作用模型。该插件也可以在Cytoscape官方网站下载得到，将下载文件存入Cytoscape安装目录的plugins文件夹中。

当鼠标指针移到某一节点时，该节点所涉及的所有其他路径与节点都会被清晰表示出来

点击菜单栏plugins→Agilent Literature Search，出现运行界面，在Terms输入框中输入所研究基因的名字，可以输入多个，以换行符分隔；Context输入框中输入感兴趣的研究方向的关键词；Use Aliases，可以选择在检索的时候是否同样检索基因的别名，比如rac也可以是prkba，一般作勾选；Use Context，是否使用Context输入框的内容进行检索；还可以选择种属，提高检索的特异性。完成以上设置后，程序会自动在Query Editor中生成检索条件语句，点击运行按钮即可运行。

Agilent Literature Search插件运行界面

如希望对皮肤黑色素瘤（Melanoma）中beta□catenin，p53，wnt5a，ifnb，nfatc，il6等6个基因的相互作用网络进行分析，则在插件窗口中输入。点击运行后，可以找到55篇最新的文献，同时自动生成如中的网络结构图。

建立皮肤黑色素瘤（Melanoma）中beta□catenin，p53，wnt5a，ifnb，nfatc，il6等6个基因的相互作用网络模型的检索方法

检索到55篇最新文献，同时自动生成相互作用结构图

插件可以保存和载入检索结果，同时可以在View菜单中选择更多的数据库，比如Online Mendelian Inheritance in Man（OMIM）和United States patent and Trademark Office（USpTO），如果勾选全部的数据库，用以上步骤同样的检索条件可以检索到146篇文献和专利等信息，同时生成一个由688个节点形成的网络。

勾选全部数据库后，对以上 6 个基因再做检索，自动生成由688个节点构成的网络

5.MiMI plugin插件的安装和使用

目前，存储蛋白分子相互作用和信号通路的数据库有许多，如BIND，IntAct，GO，KEGG，DIp，hpRD，Interpro和BioGRID等。而每一种数据库都有其特定数据格式、分子命名及附加信息，这种情况将我们难以完整找出目标信息。因此，密歇根大学医学生物计算中心开发了MiMI（Michigan Molecular Interactions）数据库，它深度整合（排除冗余）了多个著名的蛋白分子相互作用和信号通路数据库，使用户能一次得到多个数据库的结果，同时还提供相应文献等相关附加信息的组成部分之一，针对Cytoscape，MiMI plugin插件可以完成上述工作，并生成可视化网络结构图。

1.下载MiMI pulgin后，复制下载文件到Cytoscape安装目录的plugins文件夹中，重启Cytoscape即可激活插件。

2.点击菜单栏plugins→MiMI plugin→Query，启动插件。MiMI plugin插件启动界面

3.选择From File选项卡，并下载演试数据，Iron和Sphingolipid相关基因列表，以及这些基因在GO数据库中的所属生物途径列。

MiMI plugin插件From File选项卡

4.Load Gene File→选中hemeANDSphingolipidGenes。txt（上一步下载的目标基因列表文件）。选好种属人类（homo sapiens），选择分子类型（如所有分子），选择需要提取信息的数据库（如选择所有数据源），同时选择搜索方法为“1.Query genes+nearest neighbors”，即只提取列表中基因及与其有直接关系的基因。

5.点击“Search”按钮，插件会根据导入的目标基因列表，从相应的数据库提取目标基因相关的分子相互作用关系网络。左边为图形说明，右边为网络图。

目标基因相关的分子相互作用关系网络搜索结果

6.点“Network”选项卡，则会出现导航窗口，通过拖动蓝色阴影部分可以移动右边网络图，菜单栏中还有放大缩小按钮能控制网络图大小变化。可以用鼠标点击进行选取（在鼠标点击同时按住Shift键可以连续选择多个点）；也可以用Ctrl+A，全选所有节点，被选择的节点会呈现黄色高亮。窗口下方“Data panel”部分用来显示节点、边及网络属性信息。如要查看节点信息，点击第一个按钮出现下拉菜单，可以选择“Description”，“Function”，“Gene Name”和“pathway”这四项，此时在Data panel中就显示出所选节点的四项信息。用户可以通过点击每一列列名进行排序，并找出自己最感兴趣的节点。右键点击节点之间连线，还可以链接到NCBI的数据库，查看相关的文献。

7.如对ARSD这个基因感兴趣，则可以在ESp搜索中输入“ARSD”栏（需安装插件EnhancedSearch，将下载文件复制到Cytoscape安装目录的plugins文件夹），软件会找到ARSD节点，并将其显示为绿色高亮。

8.但由于交叉边非常多，并且点过于分散，所以这种显示方式使用户很难查看哪些基因与ARSD直接相关。于是我们可以改变显示方式，点击菜单栏Layout→Cytoscape Layout→Edge□weighted Spring Embedded→All nodes→（unweighted），得到新的网络显示图。可以放大ARSD周边部分，并进一步取出与ARSD直接相关的基因，点击菜单栏Select→Nodes→First neighbors of selected nodes，所有与ARSD直接相关的节点都在ARSD附近，并加上黄色高亮。

所有与ARSD直接相关的节点都在ARSD附近黄色高亮显示

9.可见除最右下角的两个节点外，所有节点都在一个网络中，这提示我们最初载入的Iron相关基因与Sphingolipid相关基因之间有关联。于是我们引入GO数据库信息对这些基因进行分类。点击File→Import→Attribute from Table（Text/MS Excel），导入已下载的CytoihemeAndSphingNohdgReadInList。xls。用鼠标右键修改每一列名称，在2号红框处选择“Show Mapping Options”，在3号红框中选择“Gene Name”作为网络关键属性，从而将GO信息导入网络中。点击Import，完成导入。

10.选择VizMapper选项卡，在Node Color选项中选择以属性Go term的值进行颜色区分，自定义以不同颜色对应不同的值，如Cholesterol homeostatis为蓝色，Sphingolipid Metabolism为红色，iron homeostatis为黄色。

11.有密集的基因簇和松散的基因簇，并且来自不同生物途径的基因之间没有直接关系。当然也可以手动选取一批节点，然后用工具SAGA在KEGG数据库中寻找所属的生物途径。如选取hTR1A，GNA13，ARhGEF12，GNAI1，GNAQ，pRKACA，ARhGEF11，EDG3和ABCA1这些假设在生物途径中的基因，可以在ESp搜索栏中输入“hTR1A GNA13ARhGEF12GNAI1GNAQ pRKACA ARhGEF11EDG3ABCA1”（以空格符分隔），然后选取这些节点的所有相邻边，点击菜单栏Select→Edges→Select Adjacent Edges。鼠标右击节点EDG3→MiMI plugin→SAGA→Do SAGA。

用SAGA在KEGG数据库中寻找生物途径结果

12.在弹出的新窗口中，保留默认选项，点Query按钮，将弹出SAGA结果网页，并以表格形式表示我们选取的基因和KEGG数据库中的生物途径比对情况。

13.如点击Match2，可以得到所提交基因的具体比对情况。

14.点击Link to KEGG picture，显示详细生物途径示意。

15.对于整个网络，也可以进行聚类分析。应用插件MCODE，MCODE能分离出网络结构中关系紧密的簇，不一定有生物学意义，但这些分子往往具有共同生物途径。点击菜单栏plugins→MCODE→Start

MCODE，在Cytoscape窗口左边出现MCODE的面板。保留默认选项，点击Analyze按钮，结果将显示在窗口右边，一共分析出10个簇。

16.对于每个簇，都可以创建子网络进行详细研究，如将排名第一的簇创建子网络，并在VizMapper选项卡中选择使用MiMI的显示风格。

读累了记得休息一会哦~

公众号：古德猫宁李

- 电子书搜索下载
- 书单分享
- 书友学习交流

网站：沉金书屋 https://www.chenjin5.com

- 电子书搜索下载
- 电子书打包资源分享
- 学习资源分享

第46章 最佳子集法1925.5MS定量生物分析

定量生物分析是系统生物学研究中的一个关键性实验。传统上用MS定量化合物是利用稳定同位素内部标准来校正因离子化效率、色谱的重复性、样品操作误差或基质效应造成的离子流图的差异。如果稳定同位素同型物的浓度已知，就能够进行绝对定量而不是相对定量。如下主要介绍定量蛋白质组学的研究。

定量蛋白质组学研究的是蛋白的相对表达量或者绝对量。相对表达水平的定量蛋白质组学分析在差异显示二维电泳（2DE）中具有一定基础。然而，2DE在定性（如低丰度蛋白、膜蛋白等）和定量（如翻译后修饰造成相同蛋白呈现出复斑）实验中都有缺点。LC□MS/MS的液相蛋白质组学方法比传统的2DE分离更加灵敏和特异，但是相对定量的完成离不开某种形式的稳定同位素标记或控制良好的色谱（见下文）。

自下而上的蛋白质组学，作为一种分析蛋白相对表达水平的方法，能够将多肽水平的定量信息弥补起来重新组合成其母蛋白的信息。同位素标记和无稳定同位素两种方法都能提供质谱在多肽水平进行高通量定量的能力。

同位素标记方法是由日益增加的变异体标记组成的。如果可以对生物体系统进行操作，代谢标记的方法可以为多肽的相对定量提供精确的测量方法，如下为几种同位素代谢标记方法：①15N掺入所有氨基酸（Oda等，1999；Washburn等，2003）。15N掺入通过C。elegans（Krijgsveld等，2003）和大鼠（Wu等，2004）的代谢标记已应用于生物水平。②富集带有13C的氨基酸（Stocklin等，2000）。③在细胞培养过程中仅掺入选定数目的同位素富集的氨基酸，SILAC（Ong等，2003）。

当生物样品不能进行同位素标记时，可以采用化学标记的方法。多肽/蛋白的稳定同位素标记有多种可能的化学方法（Lill，2003）。Aebersold等发明的ICAT（同位素编码的亲和标签）是化学标记多肽最成功的方法之一（Gygi等，1999）。ICAT试剂包括三个部分：一个亲

和标记物（通常是生物素），一个交联剂（同位素编码的）和一个巯基特异性反应基团。这些年已发明了许多ICAT试剂（Zhang等，2004）。ICAT方法的特殊优点是：它通过多肽亚群的亲和力富集而简化蛋白消化混合物。然而，由于大部分ICAT试剂的靶点仅仅是半胱氨酸残基，致使它仅适用于包含半胱氨酸的多肽。也正因为如此，蛋白翻译后修饰区域的鉴定范围也缩小为仅包含半胱氨酸的多肽（MacCoss，Yates，2001）。

定量分析最严格的要求之一是标准化合物以非常精确的起始点加入。所以，需要对生物样品的整体类型（尤其人类组织）进行定量分析时，就强调了无稳定同位素标记的定量蛋白质组学的必要性。绘制蛋白复合物图谱（Bondarenko等，2002）和对特异磷酸化的天然参考肽段（来源于目的蛋白）的相对定量（Ruse等，2002）都已进行实验的验证。消化蛋白质的印迹数量与天然参考肽段特异点消化得到的数量相似（Willard等，2003）。可以通过绘制外部标准反应曲线得到绝对量（修饰的物质的量/蛋白物质的量）。另外，在蛋白溶液中加入天然肽段（放射性标记的）并同时进行消化，也可以在计算消化效率的同时得到蛋白的绝对量（Barnidge等，2003）。加入同位素标记内部标准的目的肽段（经蛋白酶解处理），随后用单反应监测实验（最灵敏的MS常规定量实验之一）监测分析物及其标记标准（被称作AQUA的方法）也可以得到绝对量。Liu等考虑到蛋白图谱的实际问题，在用MudpIT进行多肽的大规模分析时（Liu等，2004）在统计学上评估了利用数据依赖采集方法对肽段离子的采样情况。他们的模型显示，图谱取样对相对蛋白量评估精确度超过两个数量级。

定量蛋白组学实验为系统生物学产生输入数据。从这个预期来看，为了有效地描述一个复杂体系，蛋白表达水平在统计学上至少应有两倍的差异（prudhomme等，2004）。最近，通过使用代谢标签的有机物（Venable等，2004；Wu等，2004）和定量蛋白组学相关的先进软件（MacCoss等，2003；Venable等，2004），MudpIT为核心的技术方案被改进后，蛋白表达水平很小的差别也能精确定量。

细胞脂质的整体量化采用了不同于稳定同位素标记的方法。由于离子化效率主要依赖于脂质的极性，不同的脂质需采用单独的内部标准进行定量（han，Gross，2005）。为成千上万种脂质合成稳定同位素的内部标准几乎是不可能，所以这一方法很不实用。虽然脂质组学研究最近有所进展，但还不完善，方法学的发展是在广泛的技术中不断成

熟起来的。我们将在下一部分看到，蛋白质组学技术已经从液相色谱和质谱技术及液相色谱质谱联用技术充分获得益处。

5.6蛋白质组学平台

LC和MS联机可以生成多维数据的色谱图。普通的色谱图形中插入了MS和（或）MS/MS谱的信息。色谱图形和m/z强度的结合可用于无稳定同位素标记的情况下提取定量信息（Radulovic等，2004）。

LC□MS实验中有三个参数能完整描述生物分子：连续的m/z值（在MS和MS/MS图谱中出现），保留时间（RT）和峰面积（即与数量成比例的曲线下面积）。实验生成的数据能以其中任一参数为基础再次检查验证，这为假设驱动质谱分析创造了条件。举个例子，两组样品的比较通常只强调数量上有变化的离子和继而能用串联质谱分析鉴定其性质的离子，于是我们将这些离子作为有价值的信息。这种方法一般采用MALDI和四极离子阱或Q□TOF。而MALDI允许不断重复地分析直至样品用完。

假设驱动质谱的分析过程分为两步：①高通量方式的样品调查产生一个假设（通常使用MALDI□MSn或LC□MS/MS）；②用不同的策略验证假设（可使用MALDI□MSn或LC□MS/MS）。MALDI四极离子阱，以预测少量母离子的m/z值为基础，对蛋白样品进行深入研究（Kalkum等，2003）。另外，MALDI□MS/MS能够鉴定中性丢失（例如，源自丝氨酸/苏氨酸磷酸化多肽的h3pO4中性损失）的存在。然后，利用定向MS3阐明中性丢失的多肽结构（Chang等，2004）。在LC□MS/MS实验中，ESI线性离子阱的数据依赖性中性丢失采用相同实验方法进行无假设分析对比。多肽的差量评估、MS/MS质谱或中性丢失扫描都可以作为假设驱动质谱的标准。Graber等（2004）采用MALDI提出一套依赖于以结果驱动质谱所需数据的关系数据库的整体工作流程并决定接下来的实验使用ESI。

最近，Venable等（2004）发明了一种非数据依赖性方法以获取串联质谱数据，获取这些数据的目的是利用类似于线性离子阱的快速扫描仪器进行大规模定量蛋白质组分析。非数据依赖的串联质谱方法不像数据依赖性获取方法需要数据的质谱审查，而是不加甄别地选择MS/MS图谱。

最终，任何蛋白质组学平台的基础作用（假设驱动或无假设）是阐明多肽的结构，及作为细胞内事件的功能和翻译后修饰的精确氨基酸位点。这些数据通过描述蛋白的序列有效区、修饰及相对表达水平，重组蛋白框架，最终描述生物系统的参数。