第六章　计算蛋白质组学

计算蛋白质组学是从上世纪 90 年代中期开始发展起来的一个新的研究领域。在目前为止，其主要的研究手段是通过对蛋白质质谱数据的分析来检测蛋白质及其变异和修饰的情况。在这个领域的发展过程中，生物信息学的发展和实验手段的成熟是环环相扣、互相促进的。在本章，我们就此领域中当前研究的一些主要问题做一介绍，并给出一些参考文献。需要指出的是，迄今为止，新的仪器、实验手段、生物信息学的分析方法仍然在不断的涌现。虽然本章的内容可以起到入门的作用，对这个研究领域感兴趣的读者，仍然需要追踪最新的文献，以了解可以入手的最新的科研问题。

§6.1　基础知识

§6.1.1　氨基酸和肽的质量

一个肽的结构是由肽键连接而成的氨基酸序列 (图 6.1)。一个氨基酸的性质主要由它的侧链（图中的 R 和 R_i, $i = 1, 2, 3$）决定。当两个氨基酸形成肽键时，会脱掉一个水。所以在肽链中的实际上是氨基酸脱掉一个水形成的残基。肽链的一端（左端）是一个氨基 (NH_2)，一般也称为 N 端；另一端（右端）是一个羧基 (COOH)，一般也称为 C 端。

表 6.1 给出了构成蛋白质的 20 种主要氨基酸的残基的质量。其中质量的单位是碳 12 原子质量的十二分之一，被记为 1u 或者 1Da（道尔顿）。一个肽序列在计算机中就表示成为这 20 个氨基酸的字符串。从表 6.1 可以看出，除了 I 和 L 两种氨基酸有着相同质量之外，其它的氨基酸的质量互不相同。一个肽的质量，就等于它包含的每一个残基的质量和，加上两端多出来的 H_2O 的质量。这样，通过测量蛋白质以及它的序列的子序列片段的质量，就有可能区分不同的蛋白质。这也是利用质谱来对蛋白质序列进行分析的一个基本的出发点。

§6.1.2　质谱仪和质谱

质谱仪 (mass spectrometer) 是测量分子质量的一个基本仪器。质谱仪本身已经有很多年的历史了。它的基本原理是通过电场或磁场对电荷的作用对不同质荷比（质量除以电荷数）的带电离子进行分离和检测，并将检测到的电信号通过计算转换成为带电离子的质荷比。但是传统的将分子电离的手段会破坏蛋白质这样的大分子。一直到上世纪 80 年代之后，随着 MALDI（基质辅助激光解析电离）和 ESI（电喷雾）这两种电离方法的发明，研究人员才可以比较容易的将蛋白质和肽这样的大分子进行可控的电离。MALDI 和 ESI 的广泛使用，使得利用质谱仪进行蛋白质组学的研究成为可能。在目前，蛋白质组学中使用的离子源 (ion source) 一般都是通过增加一个或多个质子的

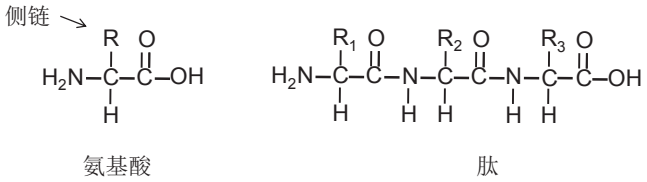

图 6.1: 一个肽的分子结构示意图。

方式将肽带正电，但是实验上带负电也是可能的。

现在蛋白质组学中使用的质谱仪的种类有很多，主要可以从离子源和质量分析器（mass analyzer）的不同来区分。从离子源上，分为 MALDI 和 ESI 两种。MALDI 形成的离子一般会带一个电荷，而 ESI 形成的离子会带一个或多个电荷。质量分析器用于分离不同质荷比的离子，在蛋白质组学中常用的

表 6.1: 20 种常见氨基酸残基的质量和分子式。

名称	三字母代码	单字母代码	单一同位素质量	平均质量	分子式
丙氨酸	Ala	A	71.03711	71.08	C_3H_5NO
精氨酸	Arg	R	156.10111	156.2	$C_6H_{12}N_4O$
天冬酰胺	Asn	N	114.04293	114.1	$C_4H_6N_2O_2$
天冬氨酸	Asp	D	115.02694	115.1	$C_4H_5NO_3$
半胱氨酸	Cys	C	103.00919	103.1	C_3H_5NOS
谷氨酸	Glu	E	129.04259	129.1	$C_5H_7NO_3$
谷氨酰胺	Gln	Q	128.05858	128.1	$C_5H_8N_2O_2$
甘氨酸	Gly	G	57.02146	57.05	C_2H_3NO
组氨酸	His	H	137.05891	137.1	$C_6H_7N_3O$
异亮氨酸	Ile	I	113.08406	113.2	$C_6H_{11}NO$
亮氨酸	Leu	L	113.08406	113.2	$C_6H_{11}NO$
赖氨酸	Lys	K	128.09496	128.2	$C_6H_{12}N_2O$
甲硫氨酸	Met	M	131.04049	131.2	C_5H_9NOS
苯丙氨酸	Phe	F	147.06841	147.2	C_9H_9NO
脯氨酸	Pro	P	97.05276	97.12	C_5H_7NO
丝氨酸	Ser	S	87.03203	87.08	$C_3H_5NO_2$
苏氨酸	Thr	T	101.04768	101.1	$C_4H_7NO_2$
色氨酸	Trp	W	186.07931	186.2	$C_{11}H_{10}N_2O$
酪氨酸	Tyr	Y	163.06333	163.2	$C_9H_9NO_2$
缬氨酸	Val	V	99.06841	99.13	C_5H_9NO

有离子阱（Ion Trap）、四连杆（Quadrupole）、飞行时间（Time of Flight）、傅里叶分析（Fourier Transform）、轨道阱（Orbitrap）这几种。读者需要意识到不同类型的质谱仪的工作原理大相径庭，有着不同的优缺点，其数据特点（譬如质量的精度）也会不同。对于需要从事质谱数据分析的生物信息学研究人员，需要对各种仪器的性能和特点有更为深入的了解，并参阅专门介绍质谱的书籍和文献。在这里我们只是在图 6.2 以飞行时间质谱仪为例来介绍一下质量分析器，以便给读者一个比较直观的概念。对于理解本章的内容，这个简单的概念就足够了。图 6.3 给出一个具体的质谱的例子。注意，质谱仪本身并不是一个为定量分析设计的仪器，所以一个谱中的带电粒子的质荷比可以测量的很准确，但信号峰的强度却只有大概上的比较意义，只是在某些特殊的实验配置下才能被利用起来。

§6.1.3 同位素峰、误差和噪音

虽然理论上讲，质谱中的每个信号峰对应了待测混合物中的一个质荷比下的离子，但是由于各种各样的原因，实验中产生的质谱比理想情况下要复杂很多。这些实验中的复杂性也是质谱数据分析中的主要困难所在，所以在阐述质谱的生物信息学分析之前，有必要对实验产生的质谱的复杂性作进一步的了解。

同位素峰和电荷数计算

自然界中存在的很多元素都有同位素。同位素是同一种元素的不同质量的原子，它们含有相同的质子数，却因为中子数目的不同而造成质量上的差异。表 6.2 中列出了蛋白质中的几种常见元素的天然同位素质量及其出现比例。

因为存在同位素的原因，同一种蛋白质或者肽的质量会有所不同。如果仅仅考虑碳 12 和碳 13 这两个同位素造成的影响，相对于没有碳 13 的肽，

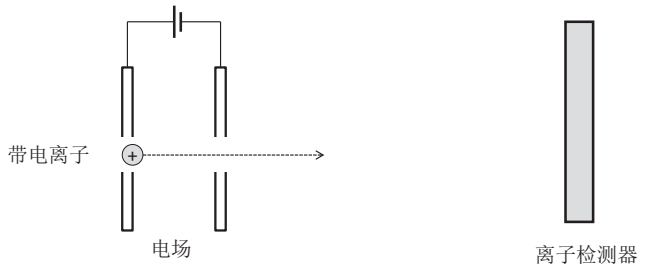

图 6.2: 时间飞行质量分析器的原理示意图。一个带电的离子在电场的作用下获得动能，其获得的飞行速度跟离子的质荷比的平方根成反比。所以通过测量离子从建立电场到到达检测器的飞行时间，就可以计算出待测离子的质荷比。

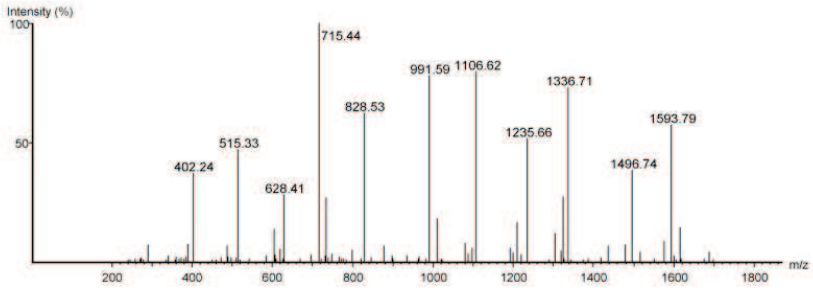

图 6.3: 质谱例图。每一个信号峰代表了质谱仪检出的一个质荷比下的信号。峰的横坐标（即峰上标识的数字）代表质荷比，峰的强度代表了检出的同样质荷比下的带电离子的多少。

包含一个碳 13 的肽的质量就多了约 1Da。同样，包含两个碳 13 的肽就多了 2Da。这样，在质谱中同一个肽就会给出多个同位素峰。而且由于一个肽中含有很多个碳原子，其中包含多个碳 13 的可能性随着肽的变长而变大，所以随着肽质量的变大，其同位素峰也会变得更强。图 6.4 展现了不同质量的同位素峰的分布情况的例子。

同位素峰的存在一方面给数据分析增加了复杂性，另一方面也带来了一定的好处。特别是对于电喷雾质谱，同一个分子可能会形成不同电荷数的离

表 6.2: 蛋白质中常见的同位素质量及其在自然界中的比例。

名称	符号	质量	比例 (%)
氢	^{1}H	1.007825	99.9885
	^{2}H	2.014102	0.0115
碳	^{12}C	12.000000	98.93
	^{13}C	13.003355	1.07
氮	^{14}N	14.003074	99.632
	^{15}N	15.000109	0.368
氧	^{16}O	15.994915	99.757
	^{17}O	16.999132	0.038
	^{18}O	17.999160	0.205
磷	^{31}P	30.973762	100
硫	^{32}S	31.972071	94.93
	^{33}S	32.971458	0.76
	^{34}S	33.967867	4.29
	^{36}S	35.967081	0.02

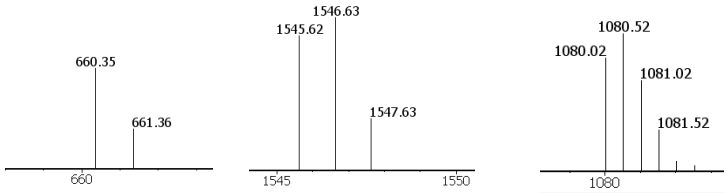

图 6.4: 左图和中图展示了同位素峰的强度分布会随着肽质量的变化而有所不同。右图展示了一个二价的相邻同位素峰距离为 1/2 Da。

子，从而在质谱中形成多个不同质荷比的峰。这个时候就需要根据两个相邻同位素峰之间的距离判断每种离子的带电荷数。设电荷数为 z，由于两个相邻同位素质量差 1Da，所以质荷比就差 1/z Da （如图 6.4）。对于同位素峰的信息学处理，一般来言是先把所有的多电荷的峰转换为对应的单电荷的峰，并把同一离子的多个同位素峰叠加在一起作为单一同位素（monoisotope）的峰进行处理。这样的预处理并不总是必要的，但会降低后续的信息处理复杂度。也正因为一个信号峰的电荷数通常可以通过同位素峰来计算，所以虽然质谱仪检测到的只是质荷比，我们通常可以认为我们已经得到了质量。

质量误差和分辨率

另一个在质谱处理中无法回避的问题是质量上的误差。任何一个质谱仪都有一定的测量精度。质量的误差一般是由 ppm (part per million) 来指定的。譬如一个质谱仪的误差是 15ppm，指的是在测量质荷比等于 m 的离子时，偏差可能会有 $\pm 15 \times 10^{-6} m$ 。不同的质谱仪类型的质量误差会有较大的差异。衡量质谱仪优劣的其它指标之一是分辨率。分辨率指的是当两个信号峰的质荷比接近到什么程度时，仪器仍然能够区分出这是两个峰，而不会错误地认为它们是同一个峰。譬如说一个分辨率是 10000 的仪器，能够在质荷比 m 附近区分两个距离为 $\frac{m}{10000}$ 的信号峰。

噪音峰和假阳性

一张质谱中往往会包含很多的噪音峰。噪音峰的形成原因是多种多样的，譬如待测样本中的杂质和质谱仪中的信号处理造成的假峰都可能会形成噪音峰。另外一种情况就是在分析连续质谱的时候，如果对肽断裂机制的模型考虑不够周全，也会造成把一些本来正常的碎片离子形成的峰当做噪音峰处理的情形。噪音峰的反面就是信号峰的缺失，指的是理论模型中预测到的一个峰未在实验谱中观察到。

无论是噪音峰还是信号峰缺失，都给质谱的生物信息学分析造成了极大的困扰。在本书书写阶段，实验室中常见的实验条件下产生的谱中有很大比例（接近或超过 50%）并不能提供足够的信息来对蛋白质和肽进行准确的鉴

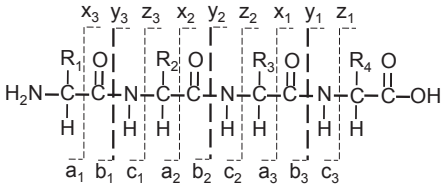

图 6.5: 一个肽在连续质谱中的断裂可以发生在各种位置，形成多种碎片离子。

定。如果生物信息学的算法仅仅满足于对每一个谱提供一个 “最优解”，那么在这些质量比较差的谱上鉴定错误率就会很高。大量的假阳性结果就会将正确结果污染，造成结果完全不可用。

一个实用的生物信息学算法，必须充分考虑到实验谱是不完美的。在实际的质谱分析中，一般的解决方式是提供一个结果验证的统计方法，用来对不可靠的分析结果进行过滤，以保证最终输出的结果中具有较低的错误率。

§6.1.4 连续质谱

连续质谱 (tandem mass spectrum) 是指在质谱仪中先对混合的肽离子通过质荷比进行过滤，选择一个特定质荷比肽离子之后，将其通过特定机制进行断裂，并测量碎片离子形成的质谱。选中的那个肽离子称为母离子 (precursor ion)，而断裂后形成的离子称为碎片离子 (fragment ion)。有时候连续质谱也被称为二级质谱 (MS/MS)。连续质谱中碎片离子的质量提供了肽的结构信息，使得肽的鉴定成为可能。

在连续质谱中使用的断裂机制有许多种，目前比较常用的包括 CID (Collision Induced Dissociation) 和 ETD (Electron Transfer Dissociation)。本章的介绍主要以 CID 为例，但是会在适当的时候指出 ETD 的不同。从理论上讲断裂可以发生在肽链的多个位置，从而形成 a、b、c、x、y、z 等各种碎片离子 (图 6.5)。在 CID 中，肽离子通过跟质谱仪中加入的其它气体分子撞击获得能量而形成断裂。在 CID 谱中强度最高的碎片离子类型是 y 离子和 b 离子。在 ETD 中，带多个正电荷的肽离子通过获取一个电子的方式获得能量而形成断裂。在 ETD 谱中最常见的碎片离子类型包括 c、z 和 $z^\bullet$ 离子。$z^\bullet$ 离子和 z 离子差不多，但它的质量比相应的 z 离子质量高出 1Da。不同碎片离子的质量可以通过这个离子所含残基的总质量加上一个质量偏差来进行计算。表 6.3 给出了每一种常见碎片离子类型的质量偏差。

这样，给定一个肽的序列和这个肽的实验谱，就可以计算各种碎片离子的理论质荷比，并由此对实验谱中的信号峰进行匹配。图 6.6 给出了一个肽和它的 CID 谱的匹配情况。当实验谱的质量较为理想时，正确的肽一般会满足两个条件：

1. 大多数高强度的峰会被碎片离子解释；

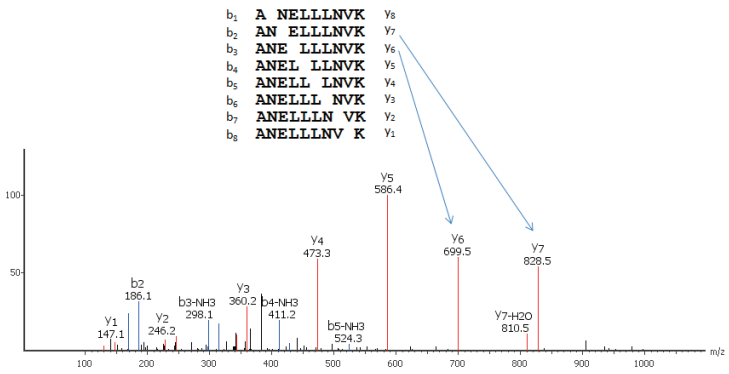

图 6.6: 一个肽和它的 CID 谱的匹配情况。

2. 大多数理论碎片离子都会在谱中找到相应的峰。

所以，这两个条件就成为判断一个肽是否对应某个实验谱的重要依据。图 6.6 所示的肽与谱的匹配就属于非常理想的。

然而很多情况下实验谱的质量并不理想。另外由于肽的断裂机制较为复杂，采用简单模型计算出来的理论上的碎片离子也不一定会在谱中出现，而且实验谱中也会出现一些无法通过简单模型预测的峰。所以在实际的肽鉴定过程中需要使用较为复杂的打分函数来判断肽谱匹配（peptide-spectrum match），并通过统计来对肽鉴定结果进行质量控制。这些算法和统计模型会在稍后的章节中再做详细介绍。

§6.1.5 复杂蛋白样本的处理

我们后面几节中介绍的算法往往假定算法的输入是单张的连续质谱，但是实验产生的数据往往包含上万甚至上百万张连续质谱。所以在开始介绍算法之前，有必要对实验做最后一点说明。

首先，到本书成文为止，现有的连续质谱仪和数据分析的发展水平仍然无法满足直接对整个蛋白进行流程化分析，而是处于一个刚刚能够流程化地

离子类型	质量偏差	取整	离子类型	质量偏差	取整
a	-26.9871	-27	x	44.9977	45
b	1.0078	1	y	19.0184	19
c	18.0344	18	z	1.9918	2
			$z^\bullet$	2.9997	3

表 6.3: 在连续质谱中常见的碎片离子类型。质量偏差等于离子质量减去其包含的所有氨基酸残基质量得到的质量差。

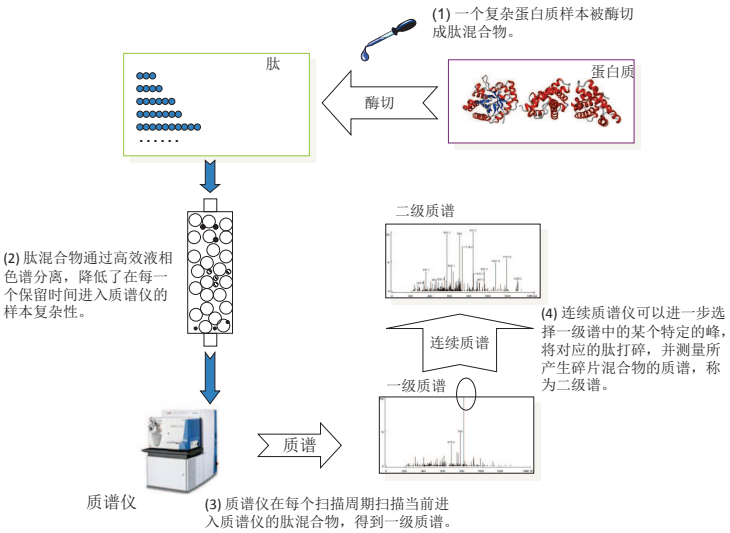

图 6.7: 一个常见的从下而上的蛋白质组学液相色谱 - 连续质谱试验流程。

鉴定一个较短（一般小于 50 个氨基酸）的肽的水平。当蛋白不是很大的时候，也有个别前沿的研究组直接通过整个蛋白质的连续质谱鉴定整个蛋白序列（称为“从上到下”或者 top-down）。但是业界最为流行的办法仍然是“从下到上”（bottom-up）的办法，也就是说把蛋白质通过水解酶进行酶切成为较短的肽，然后对每个肽进行鉴定，得到肽结果之后再回过头来鉴定整个蛋白质。

实验中的另外一个复杂性来自于蛋白质样本。蛋白质组学在多数情况下研究的是复杂的蛋白质样本，甚至经常包含上千种蛋白质，酶切之后就会形成数以万计的肽。直接对这么复杂的样本进行连续质谱的分析超出了现有质谱仪的能力。所以一般的实验流程中会包括一个或者多个分离的步骤，将蛋白质以及水解之后的肽进行分离之后再输入质谱仪。这样在每一时刻进入质谱仪的肽混合物都不太复杂，从而降低谱的复杂度。

较为通用的分离方法是通过高效液相色谱（HPLC）进行的。由于不同的肽具有不同的疏水性，它们在液相色谱中的保留时间（Retention Time）也就不同。通过液相色谱和连续质谱串联（LC-MS/MS），质谱仪在不同时间测量出来的谱也就对应了不同保留时间的肽或者肽混合物。图 6.7 展示了一个较常见的 LC-MS/MS 实验流程。

在 LC-MS/MS 的实验中，由于同一保留时间出现的肽混合物仍然会很复杂，仪器往往会做基于数据的采集（Data Dependent Acquisition 或者简称 DDA），具体流程如下：

1. 肽混合物从 HPLC 进入质谱仪，质谱仪会首先测量这个混合物的质谱，

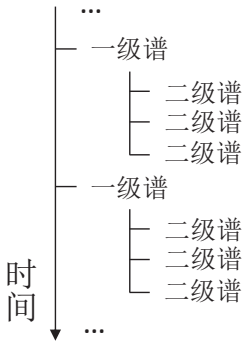

图 6.8: 一个基于数据获取的液相色谱 - 连续质谱试验的数据形态。

这个谱被称为一级谱或者 survey scan，其中每一个信号峰理论上应该对应一个肽离子。

2. 仪器会从一级谱中按一定策略选择几个信号峰作为母离子，对它们中的每一个分别进行断裂并测量二级谱（连续质谱）。

3. 因为 HPLC 输出的肽混合物随着时间会逐渐改变，所以仪器在测量了几个二级谱之后，会再次回到第一步并不断重复步骤 1 和 2 。

这样一来，一个 LC-MS/MS 的数据就具有如图 6.8 所示的形态。

§6.1.6 肽鉴定的基本方式

质谱数据分析最根本的一个问题就是如何通过一个二级谱鉴定出产生这个谱的肽序列。肽鉴定的算法研究大体上分为两类。一类假定肽的序列在一个蛋白质数据库中出现，这样，就可以通过搜库的办法来找到一个与质谱最匹配的肽。另一类则不使用蛋白质数据库，直接从谱来构造出一个肽的序列，这种方法叫做从头测序 (de novo sequencing)。这两种算法各有其优缺点。相比于搜库的办法，从头测序放弃了对数据库的要求，所以适用面会更广，但是同时也增加了算法的难度，对谱的质量也有着更高的要求。

在蛋白质组学的早期研究中，搜库和从头测序被认为是适应于有数据库和无数据库两种不同情形下使用的两种独立的算法。但是近期来研究者意识到数据库的有无并不是绝对的。即便是在有数据库的情况下，实验样本中也往往包含一些数据库中尚未收录的肽。反过来说，即便正在研究的物种不存在一个完整的蛋白质数据库，一个相似物种的蛋白质数据库中也可以提供大量相同或者相似的肽供鉴定的算法参考。所以，一个更为合适的做法是对于一个样本，通过多个算法的结合使用来鉴定出尽可能多的肽序列。