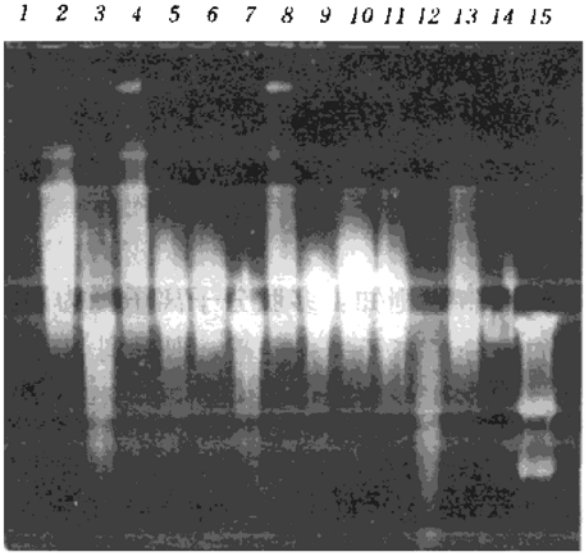


**图 5 高分子量 DNA 用不同限制酶消化的情况**

*1*：酵母染色体；*2*：未消化处理的胶块 DNA；*3*～*13*：用 11 种不同的内切酶消化的 DNA；*14*：λDNA；*15*：λDNA/Hind III 分子量标记。每 *1/6* 胶块，*37*℃消化 *3* h.

## 参 考 文 献


1 Hamilton R H, Kunsch U, Temperli A. Anal Biochem, 1972; **49**: 48

2 Willmitzer L. In: Vasil I K ed. Cell culture and somatic genetics of plants. Orlando: Academic Press, 1984; **1**: 454

3 钟 翎，李文哲，刘良式. 生物化学与生物物理进展，1992; **19** (5): 352

4 Rizzo P J, Pederson K, Cheng J H. Plant Sci Lett, 1978; **12**: 133

5 Slater R J, Venis M A, Grierson D. Planta, 1978; **144**: 89

6 孙敬三，钱迎倩. 植物细胞学研究方法. 北京：科学出版社，1987: 369～371

7 Kuethl L Z. Naturforsch, 1964; **196**: 525

8 Luthe D S, Quatrano R S. Plant Physiol, 1980; **65**: 305

9 张自立，俞新大. 植物细胞和体细胞遗传学技术与原理. 北京：高等教育出版社，1990: 320

10 Yamagachi J, Lim P Y, Aratani K *et al*. Cell Structure and Function (Japan), 1992; **17**: 87


**Isolation of Megabase DNA from Rice Nuclei.**

Wang Chunxin, Xie Yiwu, Liu Liangshi (*School of Life Science, Zhongshan University, Guangzhou* 510275, China).

**Abstract** A convenient and efficient procedure for isolation of rice HMW DNA had been developed. Etiolated seedings were ground into the fine powder with a pestle and mortar under liquid nitrogen, then extracted with a protective medium. Filtrate was layered on discontinuous sucrose gradient, then the purified nuclei were embeded in agarose gel. After proteinase digestion, the HMW DNA was released. Pulsed-field gel electrophoresis showed that the sample size ranges from 200 kb to 3 Mb and in the majority of 2.2 Mb. Rice YAC library was constructed successfully, using these easily digestable DNA preparation.

**Key words** rice, Mb DNA, pulsed field gel electrophoresis (PFGE)

# 序列搜索算法在多肽质谱解析中的应用*

方慧生 相秉仁 安登魁

（中国药科大学分析计算中心，南京 210009）

**摘要** 序列搜索算法由三部分组成：搜索过程、搜索得到多肽的各氨基酸残基的评分及两端（N 端、C 端）搜索得到的多肽的合并过程. 通过若干实际多肽质谱的解析，结果表明，该算法对多种序列专一性

*江苏省自然科学基金资助课题. 收稿日期：1994-08-30，修回日期：1995-02-15

离子并存的未知多肽质谱的解析，可获得较满意结果. 尤其是它的评分方式及标准，比较适合多肽质谱图的实际情况，可最大限度地判断解析结果的准确度. 为从事用质谱测定多肽一级结构的分析工作者提供了一比较简便且可靠的手段. 也为质谱法快速测定蛋白质或多肽序列及其在生物学中的普及提供了一条方便之路.

**关键词** 序列搜索法，质谱，多肽，序列专一性碎片离子

分子生物学的迅猛发展对微量蛋白质或多肽序列测定的要求日益提高，质谱技术的发展适应了这一要求并逐渐成为最有前途的方法之一[1,2]，它包括两个过程：多肽或蛋白质质谱图的获得及其解析，即从获得的质谱图中提取相应的结构信息. 随着待测样品分子量的不断增加，相应的质谱图亦愈趋复杂，如何快速地从这复杂的质谱图中最大限度地提取相应的结构信息以满足准确、快速的要求，已成为质谱法测定蛋白质或多肽序列整个过程中最困难、最关键的步骤之一. 因此，计算机辅助多肽质谱的解析已成为质谱测定蛋白质序列的研究中最为活跃的领域之一[3~6]. 我们在现有发现的多肽碎片离子的基础上，通过对 100 余篇有关质谱法测定多肽序列的文献中近 600 张多肽或蛋白质质谱图的分析研究发现：蛋白质或多肽质谱图中以 a、b（N 端）和 x、y（C 端）型离子为主且按一定规律分布；高质荷比处的杂质离子峰较少；不同氨基酸侧链对其骨架裂解的影响程度不一. 据此，我们建立一序列搜索算法，按碎片离子质量数依次由大到小进行搜索，可比较快速、准确地解析未知多肽质谱，得到可信度较高的多肽序列，为质谱法快速测定蛋白质或多肽序列及其在生物学中的普及提供了一条比较方便的途径.

## 1 基本原理

一般地，在多肽离子裂解得到的碎片离子中，含末端（如 N 端）的碎片离子被称为序列专一化离子(sequence-specific ion，本文所指序列专一化碎片离子不包含侧链裂解得到的碎片离子)，其裂解规则见文献 [9] 的图 1 (Scheme 1) 和文献 [11] 的图 1，得到的序列离子参见文献 [3] 的图 1 (Scheme 1).

由裂解规则知：多肽分子离子峰为 y 型离子；如果分子离子失去羟基，则得到 N 端 b 型离子. 对衍生化多肽，可分如下三类讨论：

第一类，仅 N 端衍生化：若作为 y 型离子，其质量数为 $[M+H]-group+1$，其中 $group$ 为衍生化基团质量数；N 端碎片离子不受其影响；

第二类，仅 C 端衍生化：相应的 b 型离子质量数为 $[M+H]-group$，C 端碎片离子不受影响；

第三类，N 端、C 端均衍生化：分别同上两类.

本算法基本思路为除了分子离子峰以外，其余离子峰若作为 N 端碎片离子，则认为它是 a、b、a+1、c 中的一种；如果作为 C 端碎片离子，则认为是 x、y、y−2、z、z+1 中一种. 然后根据两峰（指同一碎片离子裂解前后所形成的离子峰）质量数之差及相应离子种类之间质量数的关系寻找与之相匹配的氨基酸，如此不断搜索直至所有的碎片离子峰均搜索完毕. 最后根据多肽裂解规则对多肽中各氨基酸评分，以分数之高低判断解析之可信度.

据以上基本思路，本算法之描述分三方面：搜索过程、多肽中各氨基酸的评分及两端（N 端、C 端）搜索得到的多肽的合并过程.

### 1.1 搜索过程

搜索过程依被搜索的碎片离子所属种类分为含 C 端离子搜索和含 N 端离子搜索. 由于二者过程相同，为便于描述，以含 C 端离子搜索为例介绍.

设 n ($i$) 为搜索起点峰号 ($i=1, 2, 3, 4, 5$ 代表离子类型 x、y、y−2、z、z+1)，由于每搜索一次均涉及到两个离子峰，我们称质荷比高的离子为高离子峰，峰号记为 h ($i$)，另一个

为低离子峰，峰号记为l（$i$），其质荷比分别记为 $M$［h（$i$）］、$M$［l（$i$）］，以 $k$ 代表第 $k$ 个氨基酸残基，见表1. 质量数记为 $M$（$k$），$N$ 为

**表1 氨基酸残基名称缩写代码及质量数**

| 序号 | 代码 | 质量数 | 序号 | 代码 | 质量数 |
|---|---|---|---|---|---|
| 1 | G | 57.04 | 10 | D | 115.03 |
| 2 | A | 71.04 | 11 | Q/K* | 128.06 |
| 3 | S | 87.03 | 12 | E | 129.04 |
| 4 | P | 97.05 | 13 | M | 131.04 |
| 5 | V | 99.07 | 14 | H* | 137.06 |
| 6 | T | 101.05 | 15 | F | 147.07 |
| 7 | C | 103.01 | 16 | R* | 156:10 |
| 8 | I/L | 113.08 | 17 | Y | 163.06 |
| 9 | N | 114.04 | 18 | W | 186.08 |

*碱性氨基酸残基（不包括Q）.

质谱图所含碎片离子峰数. $H$（$i$，$j$）代表第 $i$ 种离子与第 $j$ 种离子的质量数之差，于是搜索过程算法可如下表述：

（1）置n（$i$）=1，$p$=1；

（2）若n（$i$）=1，则置 $i$=2（即分子离子峰为y型离子）；
否则置 $i$=1；若n（$i$）>1且ion（$p$，$i$）=0转步（9）

（3）置h（$i$）=n（$i$），$tos$=0，$tok$=0，$p$=1；

（4）置 $j$=1；ion（$p$，$j$）=0；

（5）置l（$j$）=h（$i$）+1，计算 $\delta$m：
$\delta m = M[h(i)] - M[l(j)] + H(i,j)$
若 $\delta m > 231$ 转步(8)

（6）置 $k=1$，判断 $|\delta m - M(k)| < \varepsilon$？
成立：则第 $k$ 个氨基酸为一可行解，记
$No[p+1,j] = l(j)$，$tok = 1$，
$tos = 1$，
$ion(p+1,j) = 1$；转步(8)
否：若 $k = 18$ 转步(7)
若 $k < 18$ 置 $k = k+1$ 转步(6)

（7）置 $l(j) = l(j) + 1$ 转步(5)

（8）若 $j = 5$ 转步(9)
若 $j < 5$ 置 $j = j + 1$ 转步(4)

（9）若 $n(i) = 1$，置 $p = p + 1$，$n(i) = n(i) + 1$，转步(2)
否：若 $i < 5$ 且 $tok = 1$ 置 $p = p + 1$ 转步(4)；
若 $i < 5$ 且 $tok = 0$ 置 $i = i + 1$ 转步(2)；
若 $i = 5$ 且 $tos = 0$ 置 $p = p - 1$ 转步(2)；

（10）若 $n(i) = N$ 结束，否则置 $n(i) = n(i) + 1$ 转步(1)

### 1.2 评分

多肽中的各肽键，由于其所在位置（指离开N端或C端的距离）不同及其相应氨基酸残基侧链的影响，它们的裂解程度不等，导致处在不同位置的氨基酸形成的肽键会裂解成不同类型的碎片离子，如脯氨酸的亚氨基形成的肽键就不易生成z（或c）型离子. 因此，质谱对同一多肽不同位置的氨基酸残基所提供的信息不均等，致使解析得到各氨基酸的可信度有差异. 所以，若仅给出每个可行解的总得分数（即各氨基酸残基得分数之和）并以此作为它可信度的唯一评判标准，则在某些可行解中部分可信度较低（得分较低）的序列会掩盖可信度较高的序列. 为此，我们分别给N端离子搜索得到的和C端离子搜索得到的肽的各氨基酸残基评分. 最后判断结果的可信度以各氨基酸残基得分为主要标准：得分越高，可信度也越高；反之，则越低. 这可得到两种可信度：肽的氨基酸序列的总可信度及各氨基酸残基位置的相对可信度. 此外，不同种类离子的形成尚与碱性氨基酸残基所处位置有关：若位于N端附近则易形成N端离子（a、b、c）；反之，易形成C端离子（x、y、z）. 据此，具体的评分规则如下确定：

C端离子搜索得到的多肽中各氨基酸得分规则如下：

规则1：高、低离子峰所属离子种类相同且为x、y中一种，得分数5；对z型离子（除脯氨酸外）得4分，否则得0分；

规则 2：高、低离子峰所属离子种类不同，且其中之一为 y 型，另一为 x 型，得分数为 4；

规则 3：高、低离子峰所属离子种类不同，且其中之一为 z 型离子：若该氨基酸为脯氨酸，则得 0 分；否则得 3 分；

规则 4：对同一种高质量峰，若氨基酸残基相同但其低质量峰不同（所属离子种类也不同），其得分为这二者之和；

N 端离子搜索得到的多肽中各氨基酸残基得分规则如下：

规则 1：若其高、低离子峰均为 b 或 a 型离子，得 5 分；

规则 2：若高、低离子峰均一个为 a 型，另一为 b 型且该氨基酸不是 R、H 则得 4 分；若为 R、H 中一种，则得 0 分；

规则 3：存在 a-a（即高离子峰为 a 型，低离子峰亦为 a 型，下同）或 b-a 型氨基酸，同时又为 a-c 或 b-c 型氨基酸，则加 4 分；

规则 4：同一氨基酸存在两种型号离子，则得分为其二者之和.

### 1.3 合并

根据被解析的多肽序列的唯一性，C 端离子和 N 端离子的峰号应不同，若有相同情况，依次按如下原则选择：

a. 碱性氨基酸存在位置：若碱性氨基酸位于 N 端，则为 N 端离子；若碱性氨基酸位于 C 端，则为 C 端离子；

b. 各氨基酸残基得分数：优先考虑氨基酸残基得分数较高的离子.

## 2 结 果

现以 β-Casmorphin-5 等六种多肽为例，分别用本算法进行序列搜索，所得结果分别列于表 2、3、4.

**表 2 β-Casomorphin-5 解析结果**

| N 端* | C 端* |
|---|---|
| G(5:b)P(5:b)F(5:b) | Y(5:y)P(5:y)F(5:y)R(4:z+1) T(4:x)A(4:y)N(4:x) |
| G(5:b)P(5:b)F(4:a) | Y(5:y)P(5:y)F(5:y)G,P(5:y) T(4:x)A(4:y)N(4:x) |
| G(5:b)P(5:a)F(5:a) | Y(5:y)P(5:y)F(5:y)Gls(4:0,0:z) |
| G(5:b)P(4:a)F(4:b) | Y(5:y)P(5:y)XLe(4:x) Y(5:y)P(5:y)G(4:y−2)G(3:x) |
| G(5:b)P(4:a)F(4:b) | Y(5:y)P(5:y)A(3:z+1)C(3:x)GLs(3:z) |
| W(4:a+1)F(5:a+1) | Y(5:y)P(5:y)A(3:z+1)C(3:x)R(3:z+1) |
| W(4:a+1)C(4:c) | Y(5:y)P(5:y)F(5:y)R(4:z+1) T(4:x)A(4:y)F(5:y)N(4:x)E(4:z)<br>T(4:x)P(5:x)F(6:x) T(4:x)P(5:x)F(4:y−2)<br>T(4:x)W(4:x) T(4:x)A(4:y)F(5:y)N(4:x)E(3:z)<br>T(4:x)A(4:y)F(5:y)A(4:x) T(4:x)A(4:y)F(5:y)A(4:z) |
| G(5) P(9) P(9) | Y(5) P(5) F(5) (G,P)(5) |

*N 端指用含 N 端碎片离子解析得到的序列，其 N 端在右，C 端在左；C 端指用含 C 端碎片离子解析得到的序列，其 N 端在左，C 端在右.

本文解析的质谱数据源于文献[8～10]. 其中，表 4 中的 1、2、3、4 肽源于文献 [8]，5、6 分别源于文献 [9，10]，其共同特点是多种序列专一性离子并存.

所有结果（见表 2、3、4）均用 C＋＋语言编程并在长城 286 微机上计算得到.

**表 3 结果比较**

| | |
|---|---|
| 解析结果 | $H_2N$—YPFPG—OH |
| 实际序列 | $H_2N$—YPFPG—OH |

表4 6个多肽的解析结果

| 多肽名称 | 解析序列 | 实际序列 |
|---|---|---|
| HGH 41～44 | K(5)Y(5)S(5)F(5) | KYSF |
| β-Casomorphin-5 | Y(5)P(5)F(9)P(9)G(5) | YPFPG |
| HGH 70～76 | K(8)S(5)N(5)L(9)K(5) | KSNLQ |
| | L(9)L(9) | LL |
| HGH 31～37 | F(5)E(5)E(9)A(5)Y(8) | FEEAY |
| | L(5)P(5) | SF |
| Substance P(1) | (P,R)(5)K(9)P(9)K(5) | RPKPQ |
| | K(8)F(9)F(9) | QFF |
| | G(5)L(5)M(5) | GLM |
| Val-gramicidin A | (G,V)(5)A(5)L(5)A(5) | VGALA |
| | V(5)V(5)V(5)W(5)L(5) | VVVWL |
| | W(5)L(5)W(5)L(5)W(5) | WLWLW |

*各序列均为N端在左，C端在右.

表2的结果是多肽β-Casomorphin-5质谱数据的解析结果（该肽的解析过程比较典型），以此为例，可了解本算法的具体解析步骤. 其结果列于表3. 表2中所列各种可行解的氨基酸残基表达形式为：A（B：C），各字符代表含义为：A代表20种氨基酸残基的名称，其中L代表L和I，K代表Q和K（参见表1）；B是该氨基酸残基得分数；C为低离子峰所属离子类型.

对峰号为0（即质荷比最小的离子峰），根据搜索过程中得到的离子种类及其本身所含的质量数，用分支定界法[11]求解其所含的氨基酸种类.

表4中各肽氨基酸残基的A、B同表2，所列结果是根据合并原则得到的最终结果.

## 3 讨 论

由表2和3可知：

在含N端离子搜索得到各可行解（即多肽）中，可行解1、2、3、4、5具有相同的氨基酸序列，其中，脯氨酸P残基的低离子峰不仅有b型离子，也有a型离子；相应地F残基也有a、b两种离子，根据打分规则，它们的得分数均为：4+5=9分，高于可行解6、7. 因此，可认为含N端离子搜索得到的可行解为GPF.

与上类似，在含N端离子搜索得到的各可行解中，子序列YPF（P、G）中各氨基酸残基得分均为5，其中（P、G）是由0号峰（质核比最小的离子峰）作为y型离子时，由分支定界法计算得到. 它们的得分数均高于其它多肽，故其可靠性最强.

根据合并原则，得到其最终结果为：

$$H_2N—YPFPG—OH$$

由表4可知：

多肽1、2、3解析得到的序列与实际序列基本相符，不同的是尚未区分出Q和K，L和I，其原因是二者残基质量分别相同，仅用序列专一性离子不能分辨.

多肽4中C端两个氨基酸残基L-P的得分数低于该肽中其它氨基酸，依据本算法的评分标准：其可信度低于该肽中的其它氨基酸. 事实表明（参考多肽4的实际序列）：多肽4 C端两氨基酸残基应为S和F，不是L和P，而其它序列完全相符，这说明本文评分标准及方式比较合理：最大限度地判断解析结果的准确度. 同时，也反映了“某些多肽，单纯用质谱技术无法测定出其一级结构的全部序列”这一事实.

多肽5、6（即substance P和Val-gramicidin A）解析得到的氨基酸序列与实际序列基本吻合. 未能解析出的序列是N端的两个氨基酸残基. 其主要原因是多肽裂解不完全. 出现这种情况，我们用分支定界法求解其可能的氨基酸组成：多肽5 N端的两个氨基酸残基为P、R，而多肽6为G和V. 与实际序列相比：组成相同，但具体的氨基酸序列不确定.

由以上结果及其讨论可知：运用序列搜索算法同时结合分支定界法，可比较准确地提取多肽质谱图所含的一级结构信息，尤其是它的评分方式及标准，能为人们提供同一可行解中各氨基酸残基的相对可信度. 显然，它可比较准确、快速地解析某一多肽的氨基酸序列，使质谱法在多肽及蛋白质序列测定中应有的优点得以充分地显示，并为它在生物学中的应用普及开辟了一条比较可行途径.

## 参 考 文 献


1 Hill R E. Clinica Chimica Acta, 1990; **194**: 1
2 方慧生，相秉仁，安登魁. 药学进展，1993; **4**: 196
3 Johnson R S, Biemann K. Biomed Environ Mass Spectrom, 1989; **18**: 945
4 Scoble H A, Biller J E, Biemann *et al*. Anal Chem, 1987; **239**: 327
5 Rubino F M. Spectroscopy Letters, 1992; **25**: 811
6 Grey C A, Steven C, Rivier J E *et al*. International Journal of Mass Spectrometry and Ion Process, 1993; **126**: 137
7 Roepsstorff P, Fohlman J. Biomed Mass Spectrom, 1984; **11**: 601
8 Rubino F M. Biological Mass Spectrometry, 1992; **21**: 451
9 Martin S A, Biemann K. International Journal of Mass Spectrometry and Ion Process, 1987; **78**: 213
10 Johnson R S, Martin S A, Biemann K. International Journal of Mass Spectrometry and Ion Process, 1988; **86**: 137
11 方慧生，相秉仁，安登魁. 中国药科大学学报，1994; **25**: 177


**The Application of Searching Sequence-Specific Ions Algorithm in the Interpretation of Polypeptide Mass Spectra**. Fang Huisheng, Xiang Bingren, An Dengkui (*Analysis & Computer Center, China Pharmaceutical University, Nanjing* 210009, China).

**Abstract** An algorithm, called searching sequence-specific ions, is proposed here for interpretation of mass spectra of unknown polypeptide. This program is composed of three parts: searching, scoring and merging. The method successfully interpreted some unknown polypeptides' mass spectra. One of the major advantages of this program over algorithms described earlier is its scoring ability which can rank the confidence of every amino acid residues in the interpreted polypeptide. It greatly facilitates the determination of the amino acid sequence and provides a pathway for the application of mass spectra to biology.

**Key words** searching sequcnce-specific ions algorithm, mass spectra, polypeptide, sequence-specific ions

# 超临界 $CO_2$ 技术萃取蛋黄磷脂*

赖炳森** 毛中兴 沈晓京** 孙树秦

（解放军北京医学高等专科学校生物化学教研室，北京 100071）

**摘要** 采用新型物理分离技术——超临界 $CO_2$ 萃取法，提取天然蛋黄粉中的磷脂. 在 40 MPa，先去除蛋黄粉中甘油三酯和胆固醇，再萃取磷脂. 结果显示，磷脂纯度为95%，$N/P$ 比值为1.003，$\lambda_{max}$ = 214 nm，薄层层析显示磷脂着色点清晰，并去除了绝大部分甘油三酯和胆固醇. 此法操作简单、产品质量高、安全和不污染环境. 还可得到天然纯蛋黄油和蛋白.

**关键词** 超临界 $CO_2$ 萃取，磷脂，蛋黄

磷脂是一类含磷酸根的脂类，是生物膜的基本组成成分，又是脂蛋白的载体，具有重要生物学活性. 磷脂中含高度不饱和脂肪酸，是人体不饱和脂肪酸的重要来源. 磷脂分子又是

* 全军“八五”医药卫生科研基金资助项目.
** 联系人.
收稿日期：1994-09-19，修回日期：1994-12-26