生物信息学概论

第十二章 计算蛋白质组

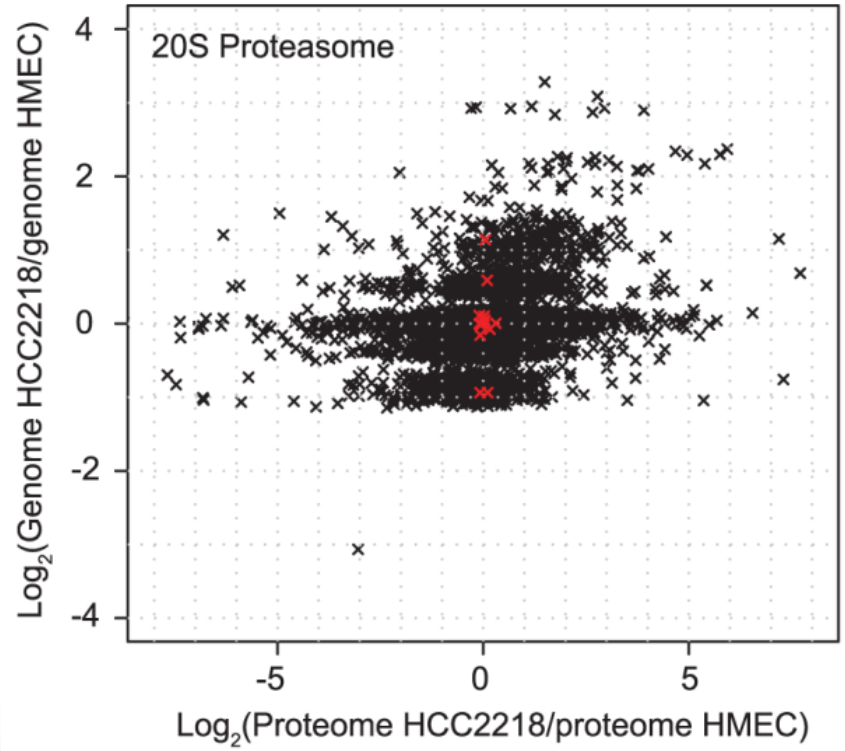
蛋白质组

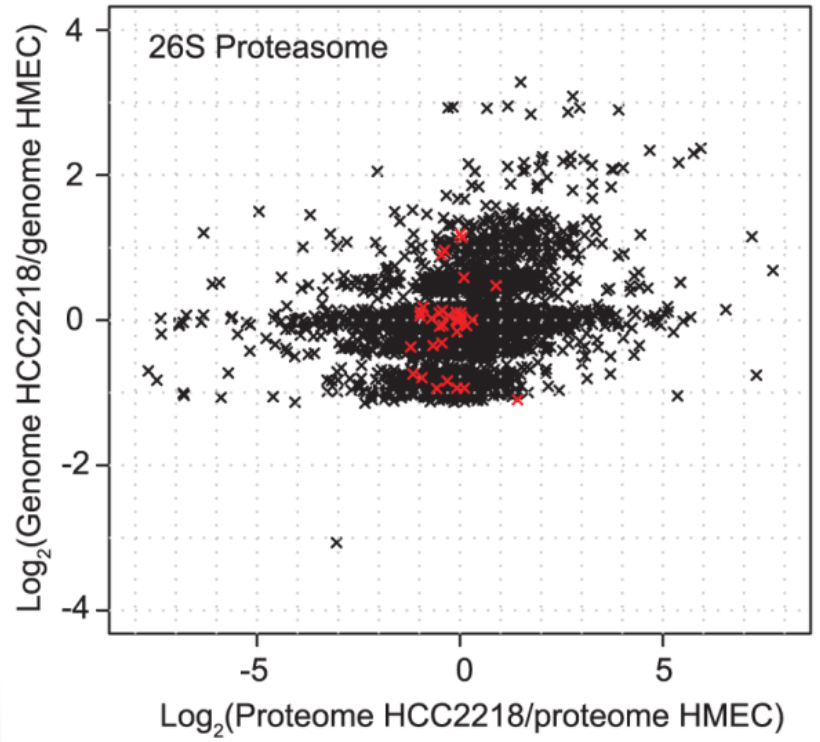
- 相同基因组，不同蛋白质组

黑凤蝶 (Black Swallowtail) – 幼虫 & 蝴蝶

蛋白质调控

- Human mammary epithelial cell (HMEC)
- HCC2218: breast cancer cell line





Geiger et al., “Proteomic changes resulting from gene copy number variations in cancer cells”, PLoS Genet. 2010, 6, e1001090.

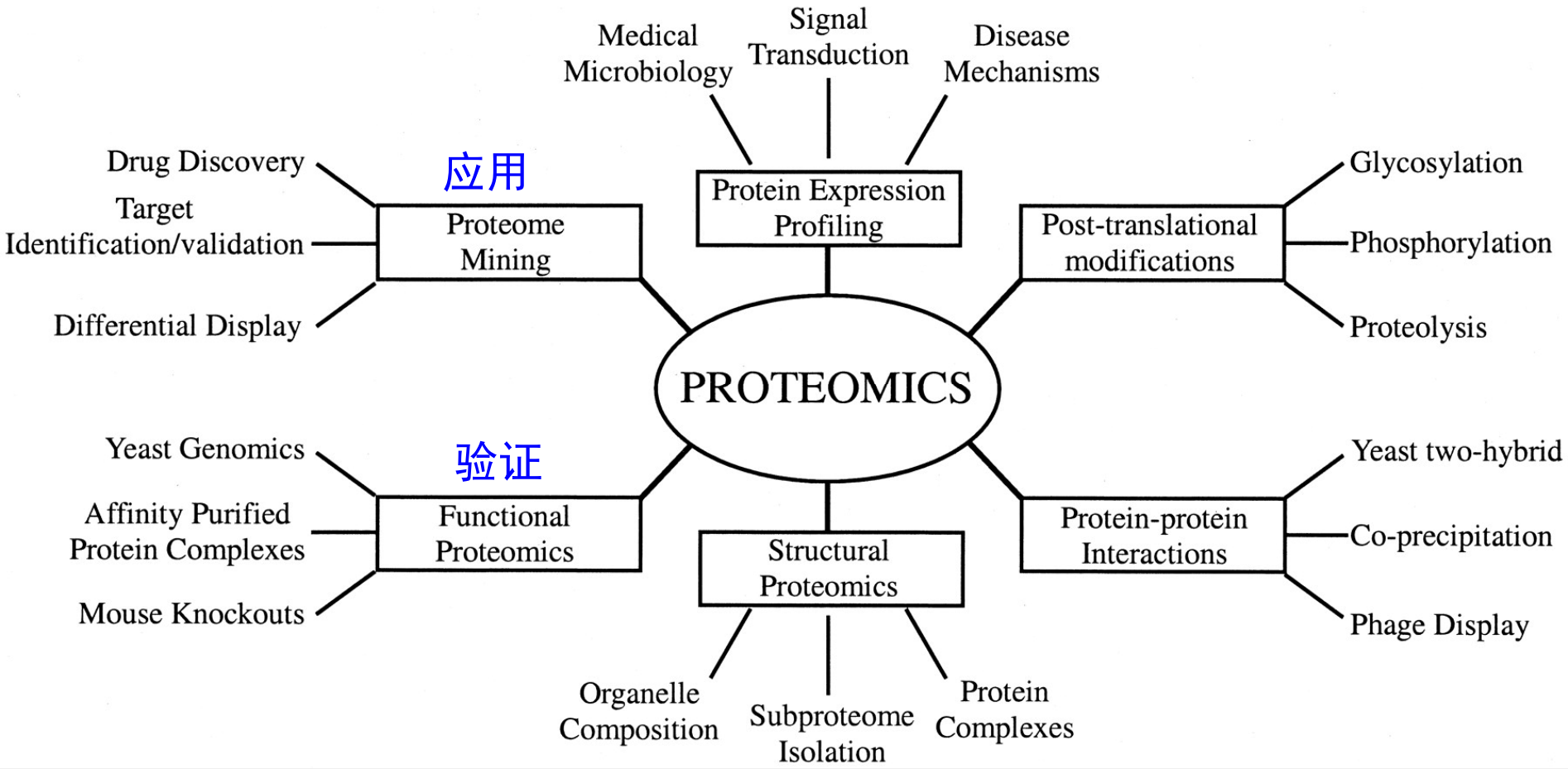
蛋白质组

- Proteomics, 1995年
  - 基因组对应的蛋白质产物的合集
  - 鉴定和定量所有的蛋白质
  - 特定信号通路、细胞器、细胞、组织、器官、物种
  - 为生物系统提供准确和详尽的蛋白质信息
- 主要研究的问题
  - 一个生物学样本里包含多少种蛋白质？
  - 这些蛋白质的浓度/丰度是多少？
  - 在不同的样本里有多少蛋白质表达水平改变？
  - 有哪些翻译后修饰？
  - 这些蛋白质在细胞或器官中定位在哪里？
  - 蛋白质如何与其他蛋白质或分子相互作用？

蛋白质组 *vs.* 转录组

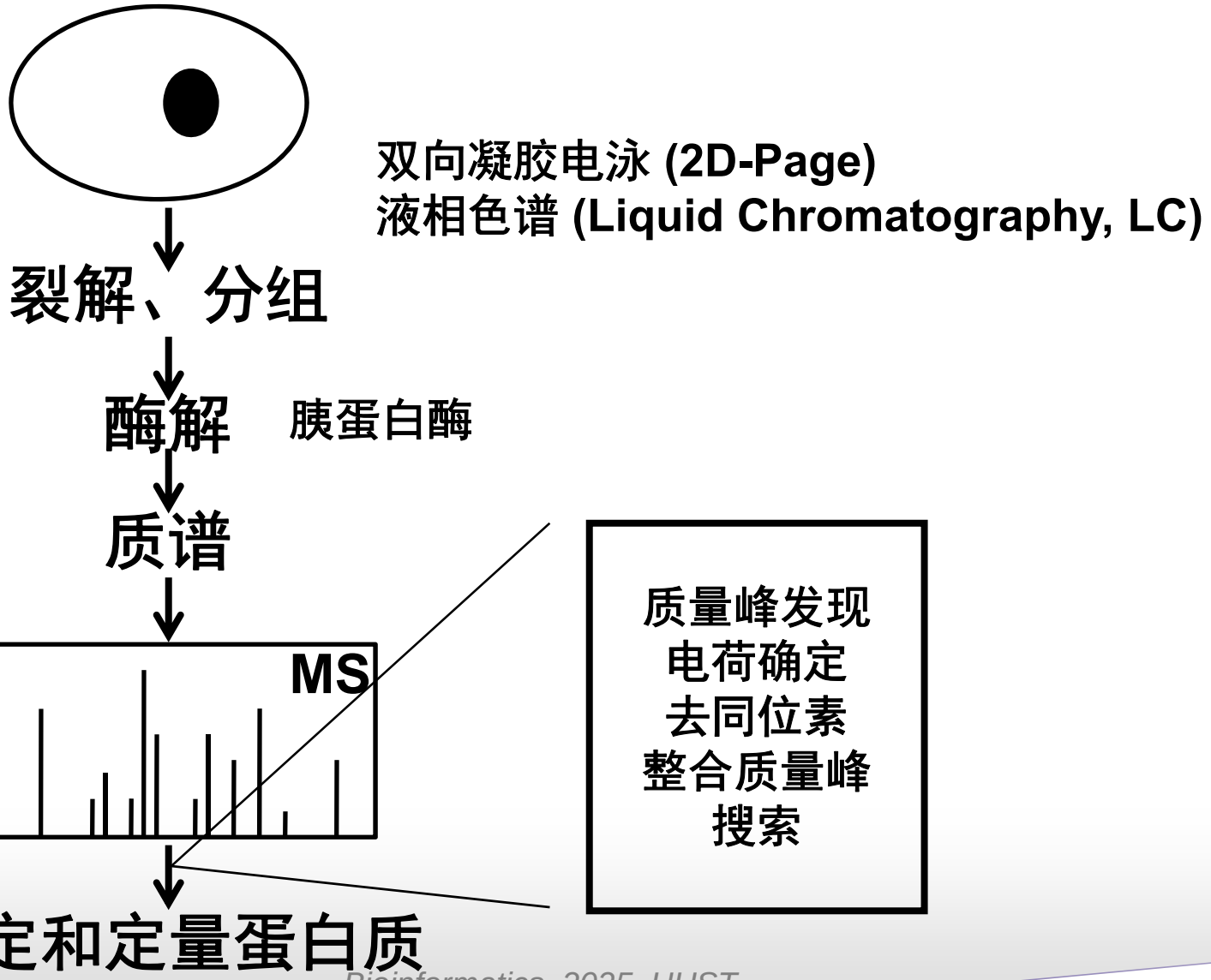
- 不是所有的mRNA都会翻译成蛋白质
- 编码蛋白质的RNA的转录水平，与表达水平并不总是相关
- 蛋白质的活性：
  - mRNA，RNA剪接，翻译后修饰

蛋白质组研究的类型及应用



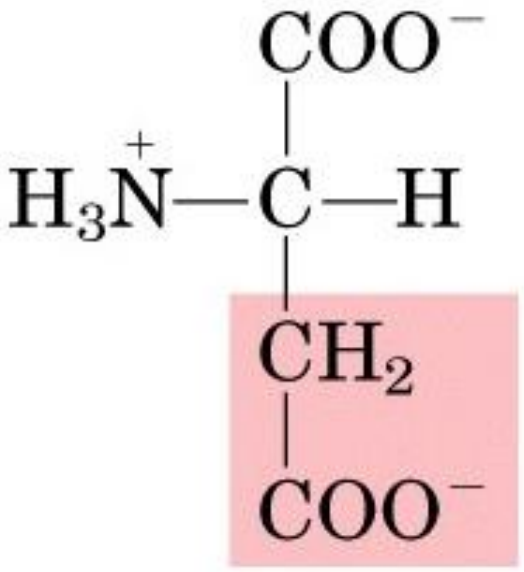
Graves and Haystead (2002) Microbiol & Molec. Biol. Rev. 66, 39-63

基于质谱的蛋白质组

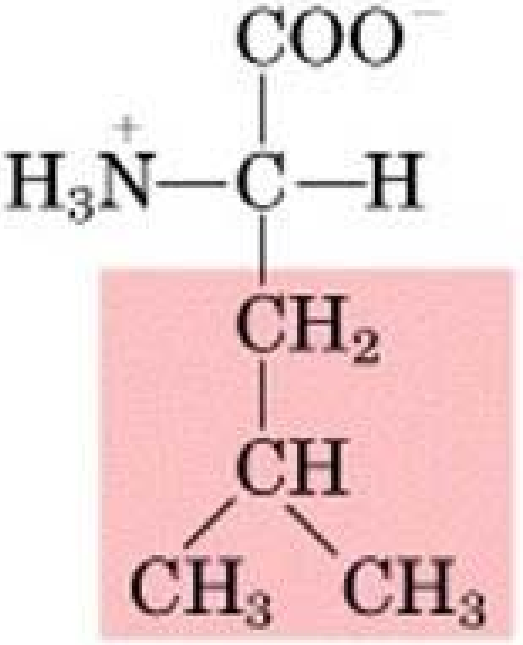

双向凝胶电泳 (2D-Page)
液相色谱 (Liquid Chromatography, LC)
裂解、分组
酶解
胰蛋白酶
质谱
MS
质量峰发现
电荷确定
去同位素
整合质量峰
搜索
鉴定和定量蛋白质

氨基酸残基

- 天冬氨酸 *vs.* 亮氨酸

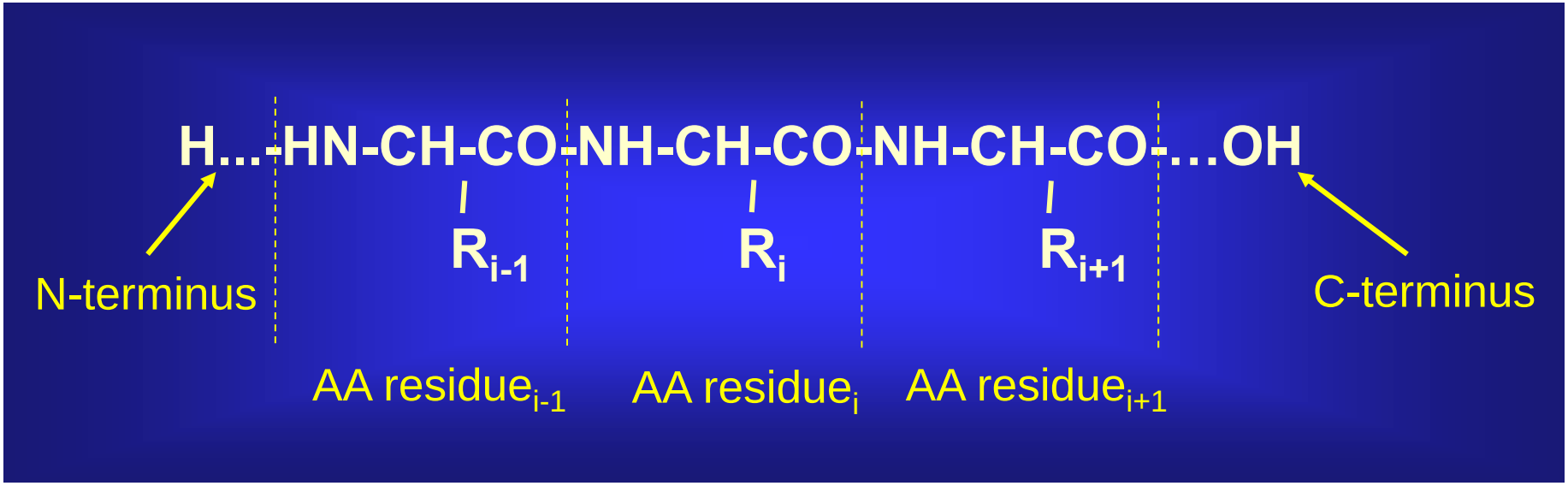


Aspartate

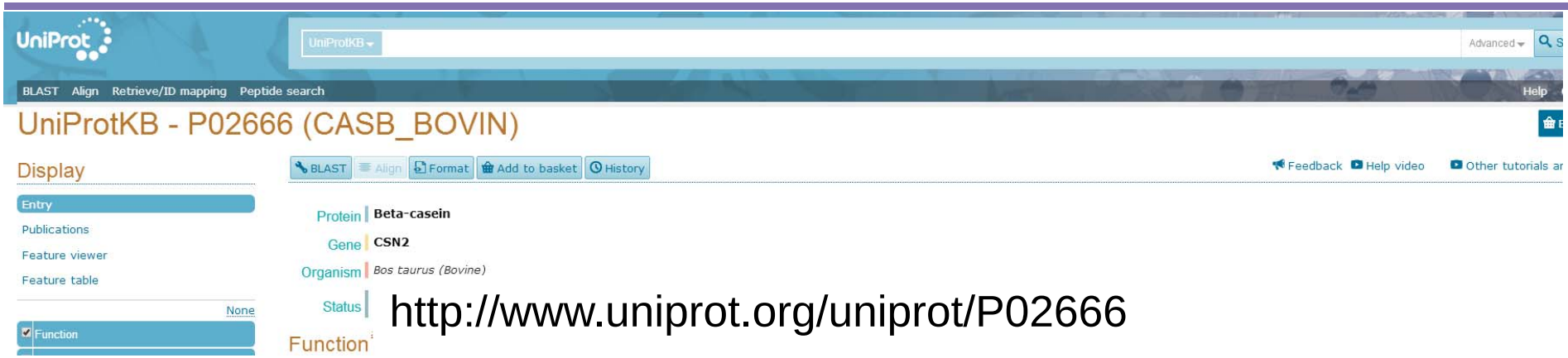


Leucine

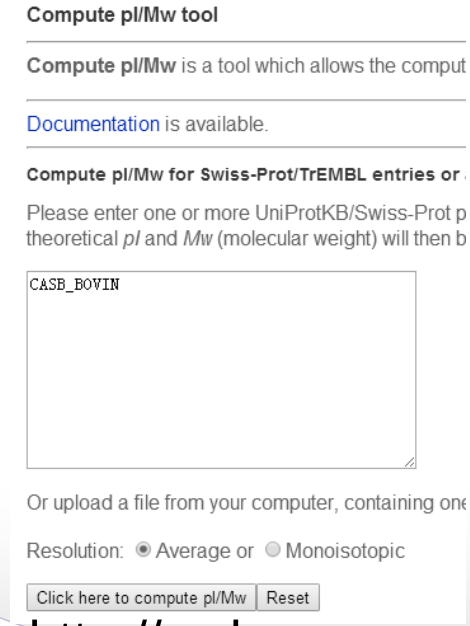
蛋白质序列

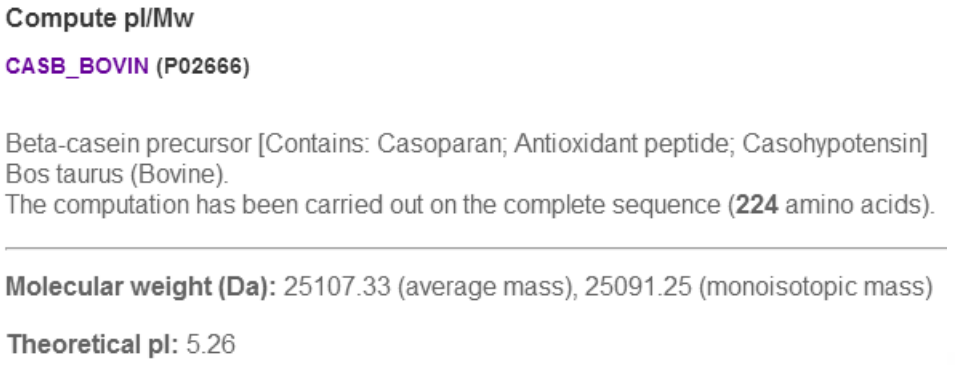

H…-HN-CH-CO-NH-CH-CO-NH-CH-CO-…OH
R_{i-1}
R_i
R_{i+1}
N-terminus
C-terminus
AA residue_{i-1}
AA residue_i
AA residue_{i+1}

等电点/分子量



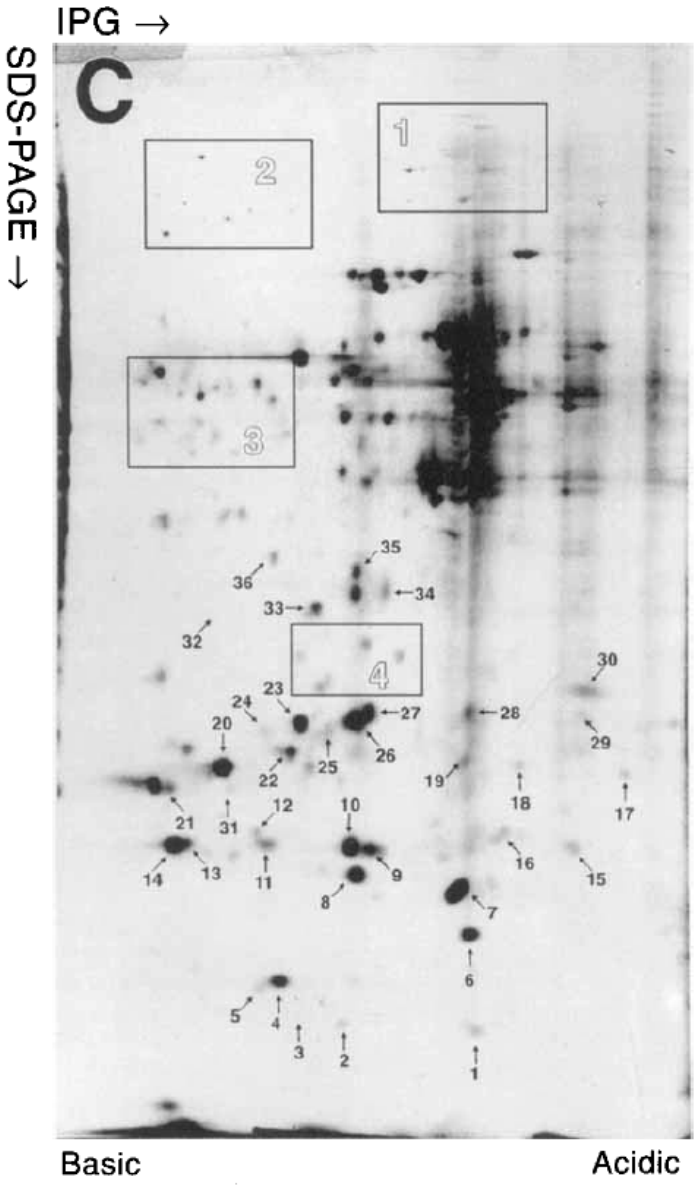
http://www.uniprot.org/uniprot/P02666

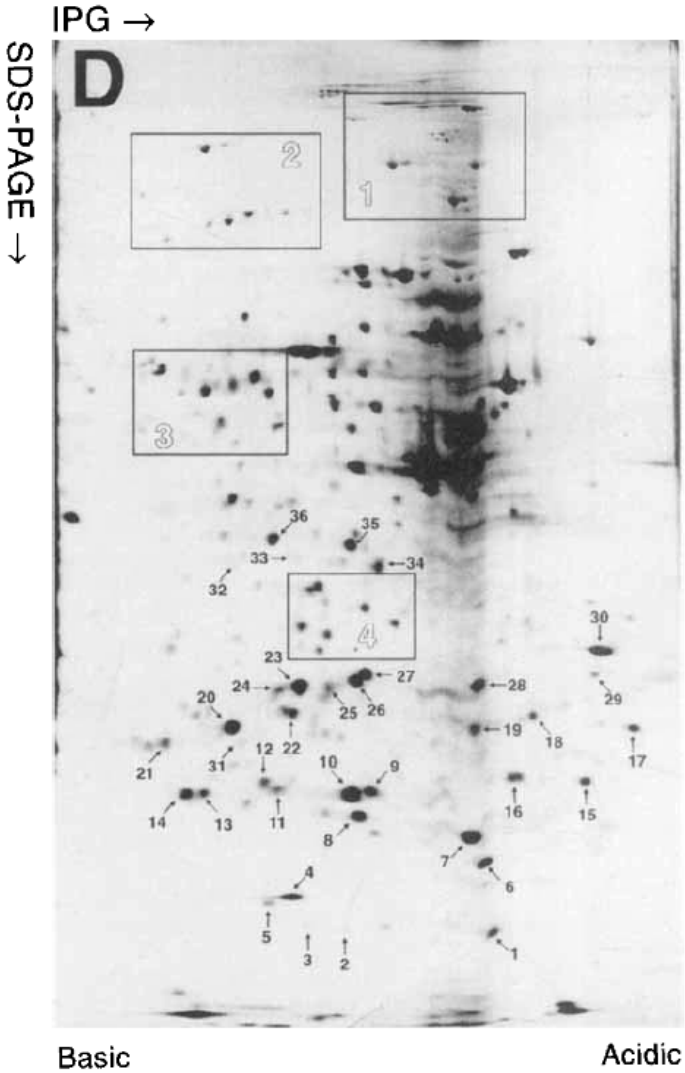




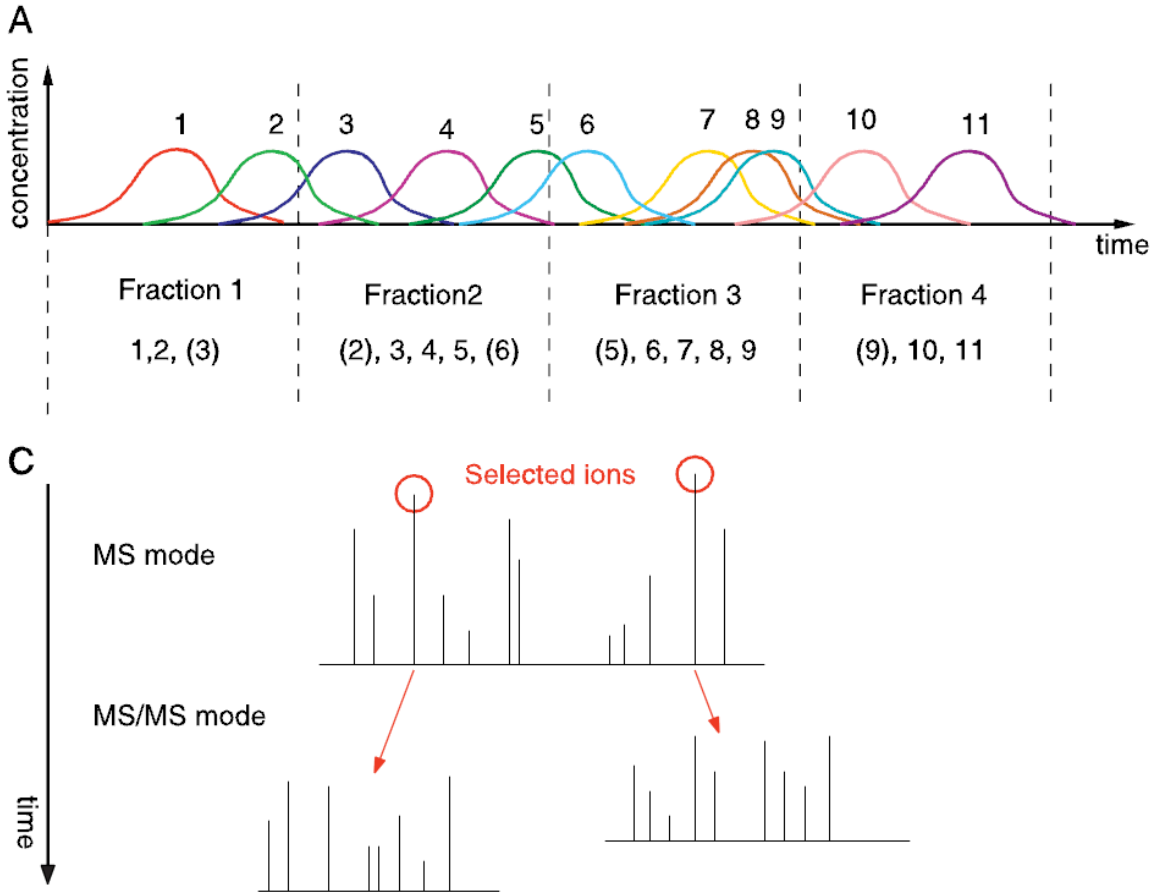
http://web.expasy.org/compute_pi/

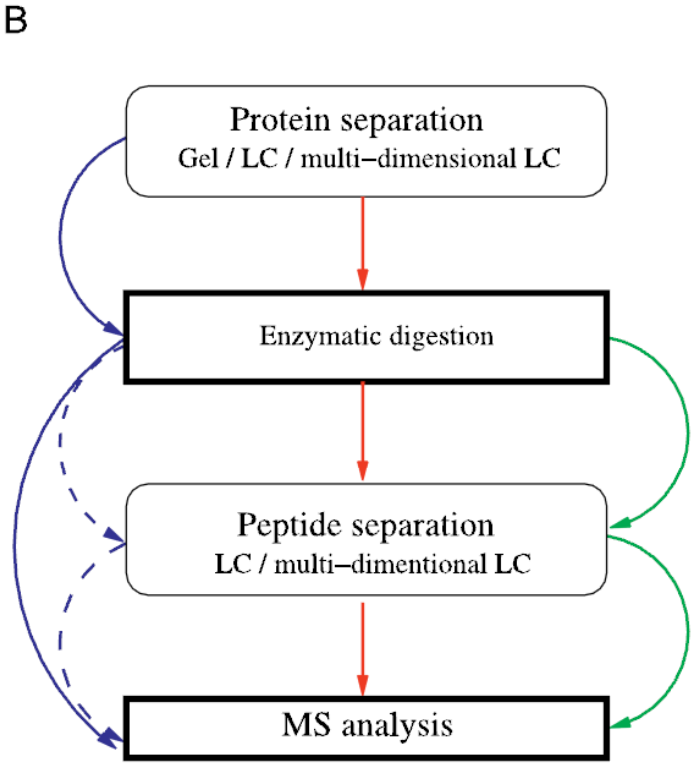
二维凝胶电泳 (2D-Page)





色谱分离





doi:10.1371/journal.pcbi.0030114.g001

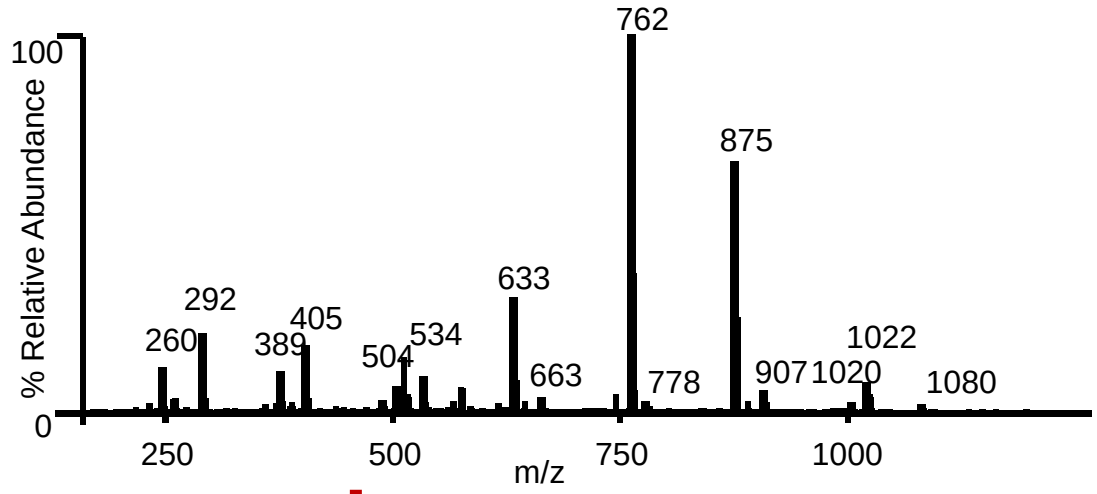
肽段酶解

- 肽酶 (Peptidase)
  - 胰蛋白酶 (Trypsin)：K/R↓, not with P
  - 糜蛋白酶 (Chymotrypsin)： 疏水性氨基酸
  - Arg-C：R ↓
  - Glu-C：E↓
  - Lys-C：K↓
  - V8-protease：E↓
  - Pepsin – 酸性环境，随机切割

谱图匹配

氨基酸质量

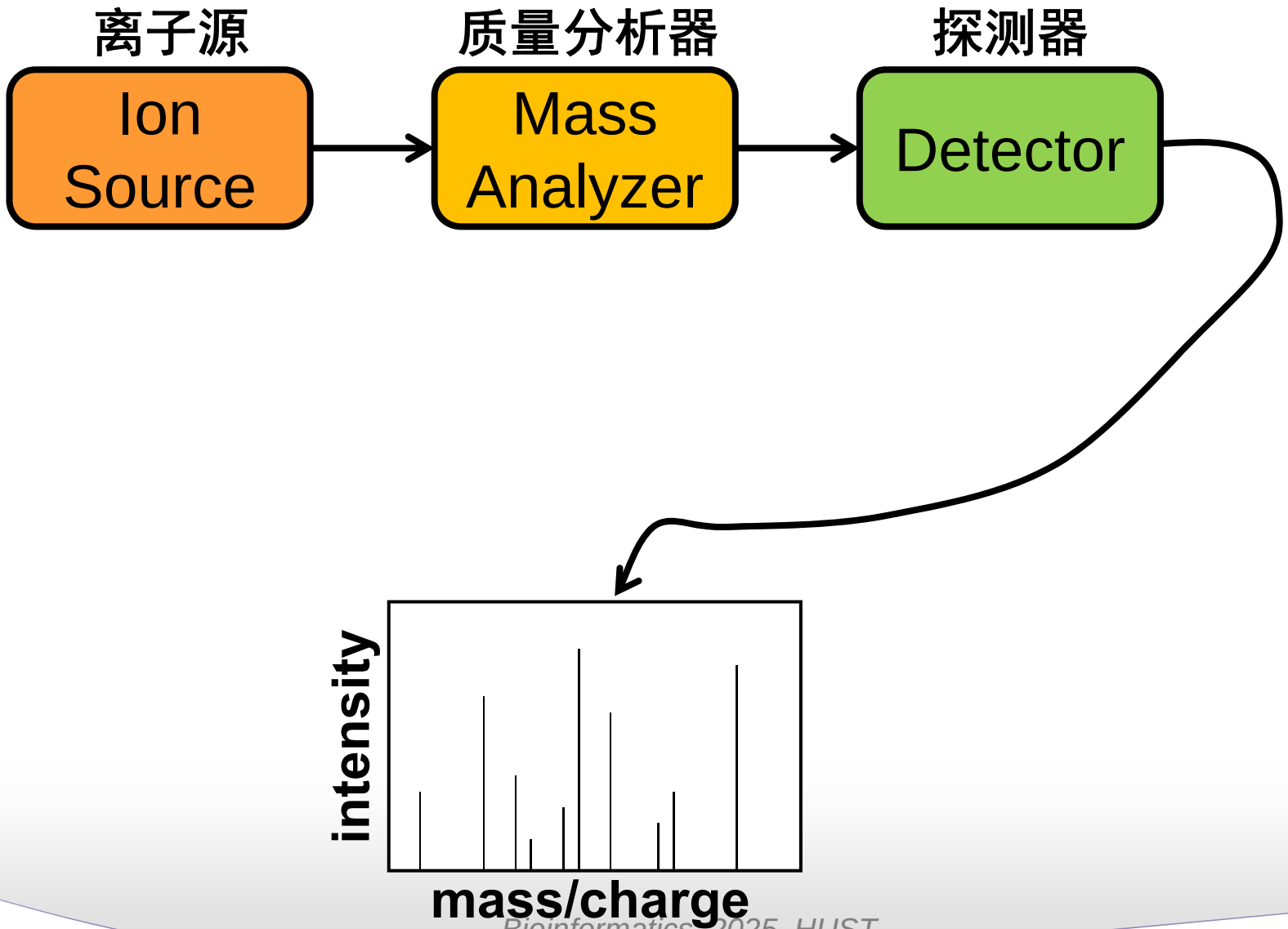
1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	71.0371	71.0788
R	Arg	$C_6H_{12}ON_4$	156.101	156.188
N	Asn	$C_4H_6O_2N_2$	114.043	114.104
D	Asp	$C_4H_5O_3N$	115.027	115.089
C	Cys	C_3H_5ONS	103.009	103.139
E	Glu	$C_5H_7O_3N$	129.043	129.116
Q	Gln	$C_5H_8O_2N_2$	128.059	128.131
G	Gly	C_2H_3ON	57.0215	57.0519
H	His	$C_6H_7ON_3$	137.059	137.141
I	Ile	$C_6H_{11}ON$	113.084	113.159
L	Leu	$C_6H_{11}ON$	113.084	113.159
K	Lys	$C_6H_{12}ON_2$	128.095	128.174
M	Met	C_5H_9ONS	131.04	131.193
F	Phe	C_9H_9ON	147.068	147.177
P	Pro	C_5H_7ON	97.0528	97.1167
S	Ser	$C_3H_5O_2N$	87.032	87.0782
T	Thr	$C_4H_7O_2N$	101.048	101.105
W	Trp	$C_{11}H_{10}ON_2$	186.079	186.213
Y	Tyr	$C_9H_9O_2N$	163.063	163.176
V	Val	C_5H_9ON	99.0684	99.1326



质量差异

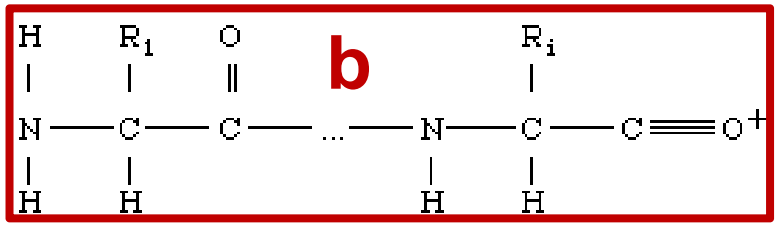
与质谱图一致的蛋白质序列

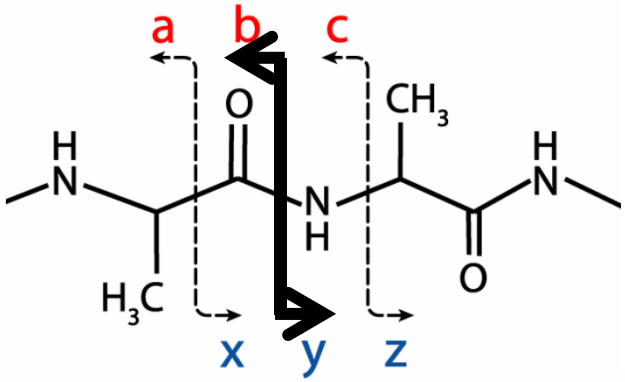
质谱

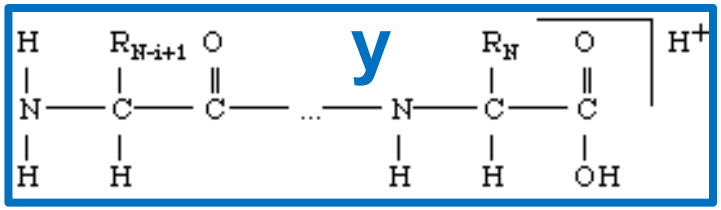

离子源
质量分析器
探测器
Ion Source
Mass Analyzer
Detector
intensity
mass/charge

肽段碎裂：二级质谱

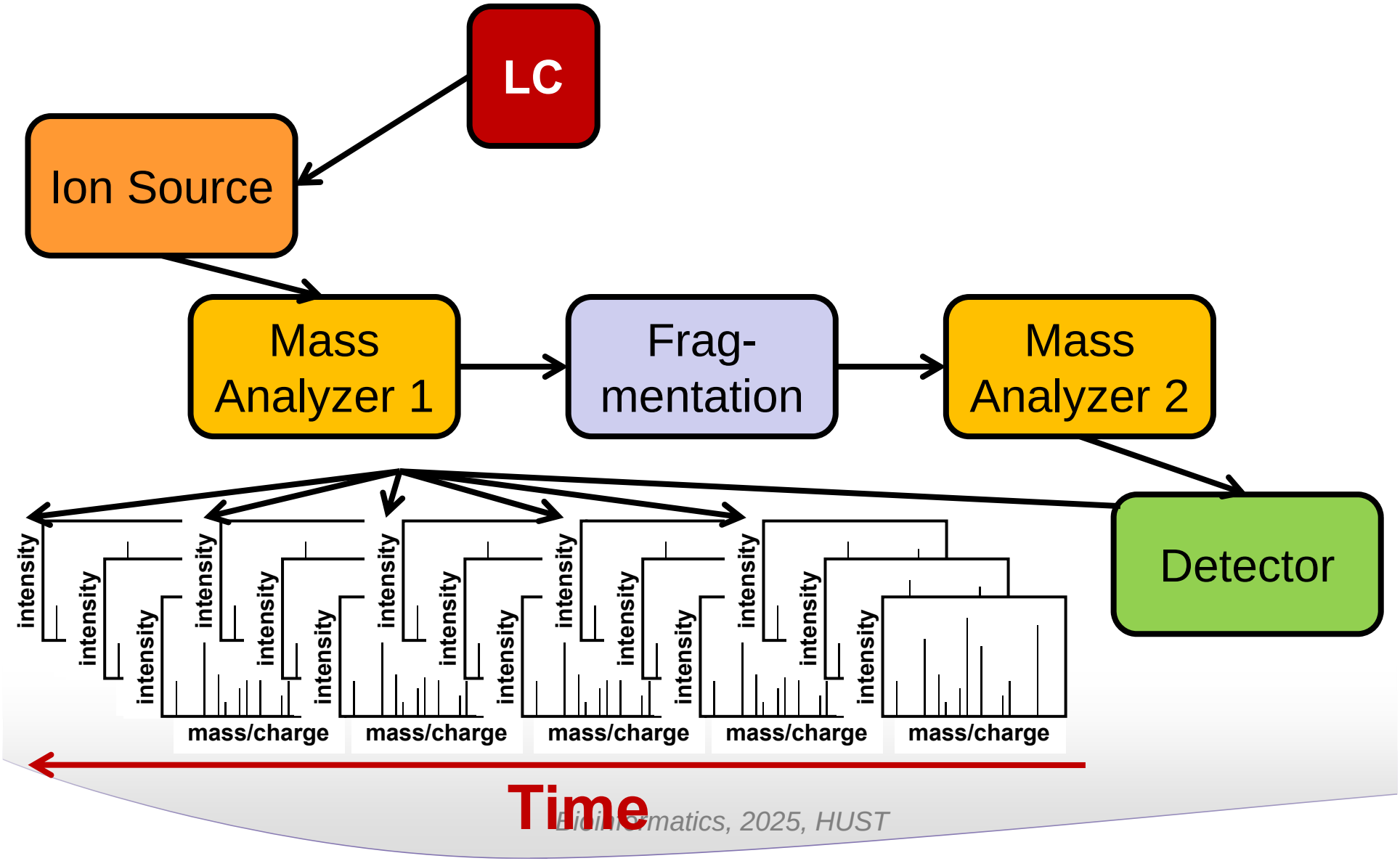
Tandem MS, MS/MS



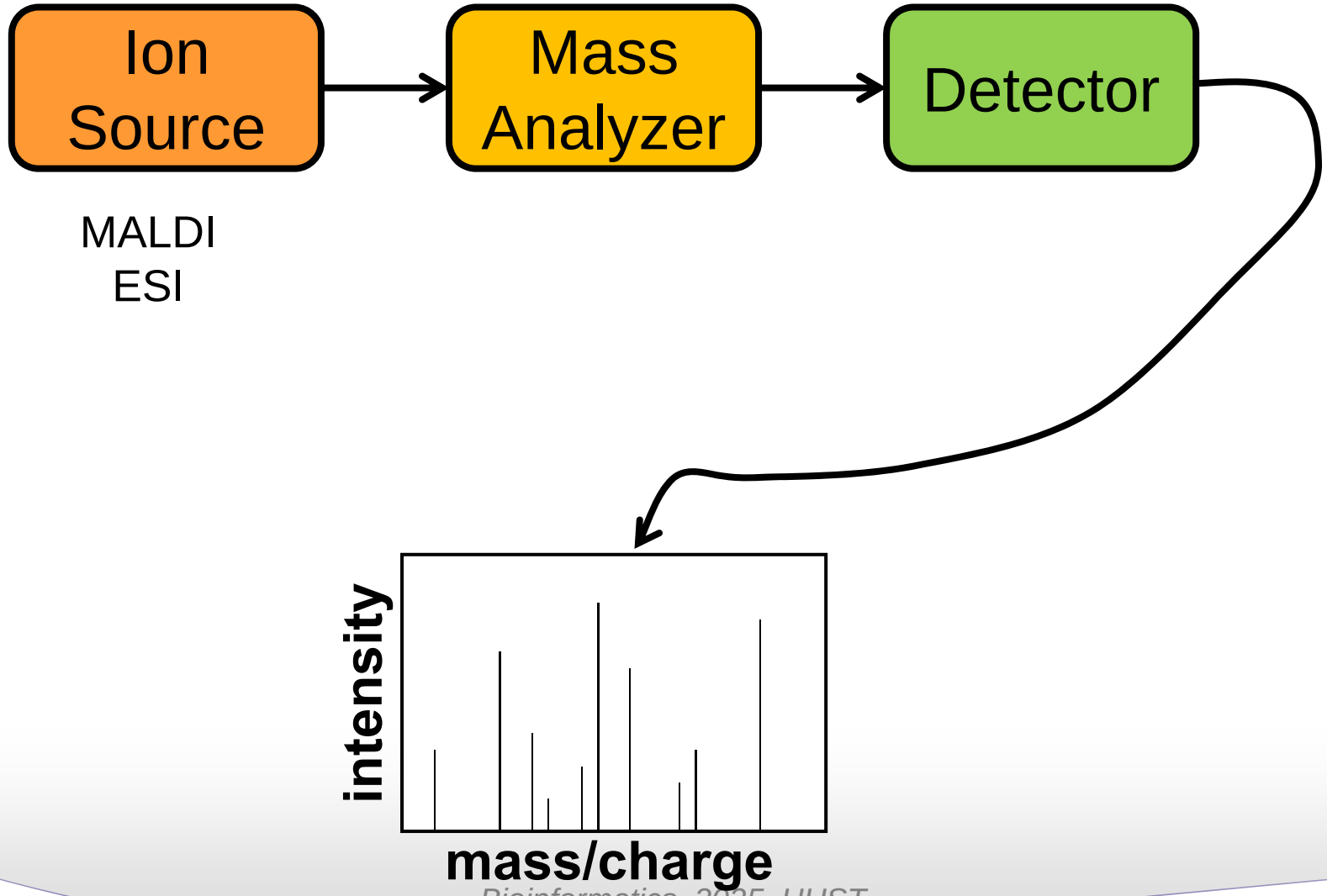




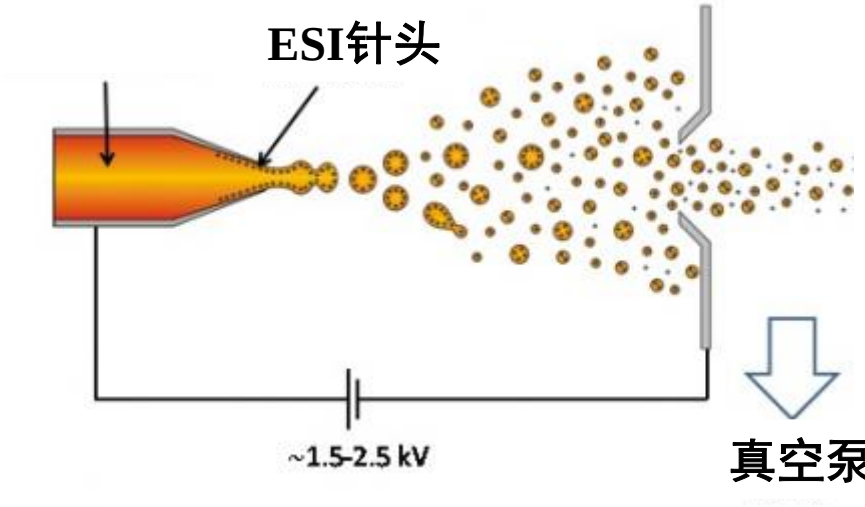
LC-MS/MS


LC
Ion Source
Mass Analyzer 1
Frag-mentation
Mass Analyzer 2
Detector
intensity
mass/charge
Time

离子源


Ion Source
Mass Analyzer
Detector
MALDI
ESI
intensity
mass/charge

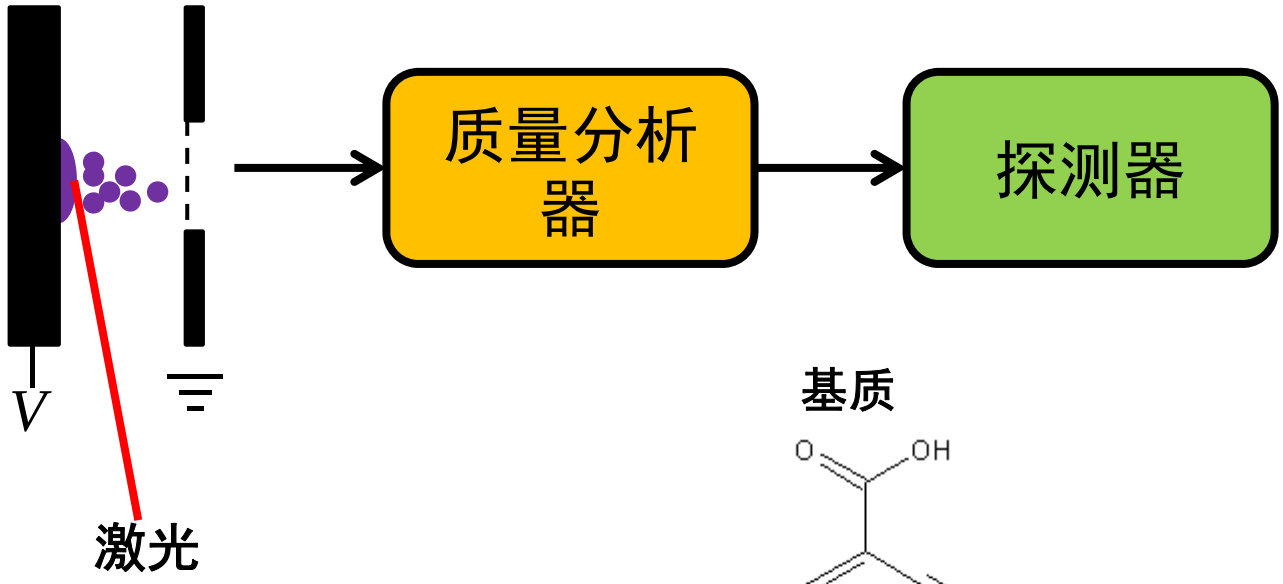
电喷雾



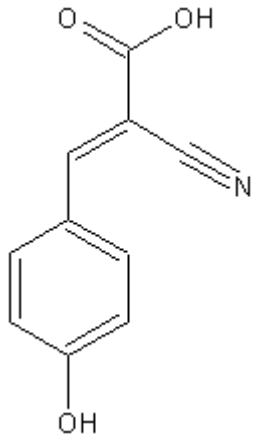
质量分析器

探测器

基质辅助激光解吸电离

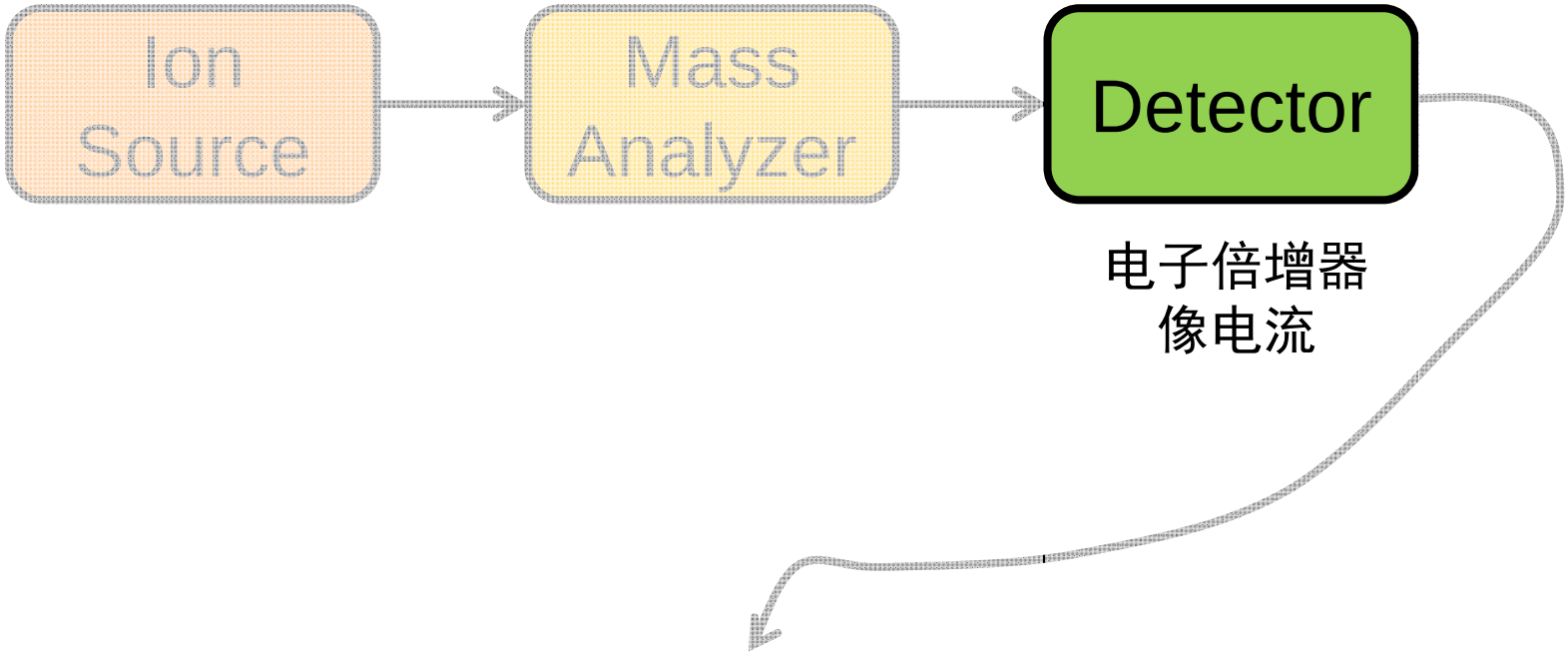


基质



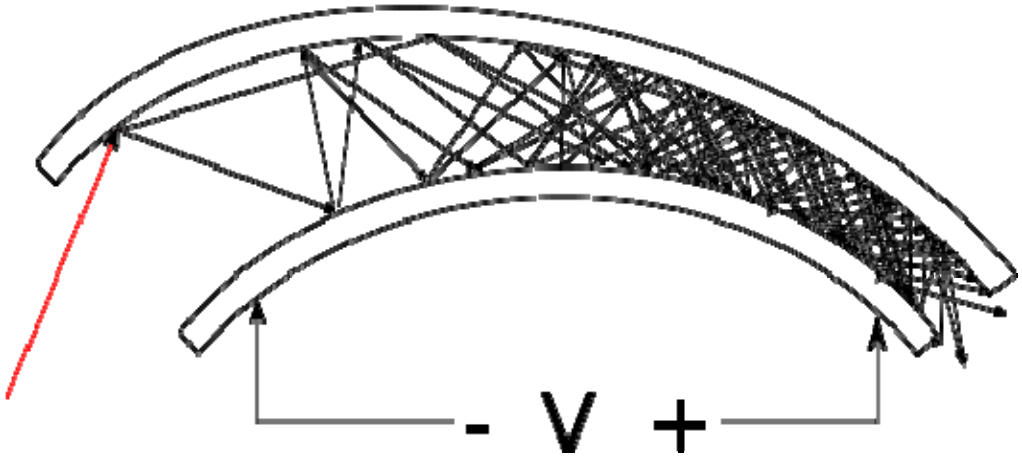
α-氰基-4-羟基肉桂酸
(alpha-cyano-4-hydroxycinnamic acid)

检测器

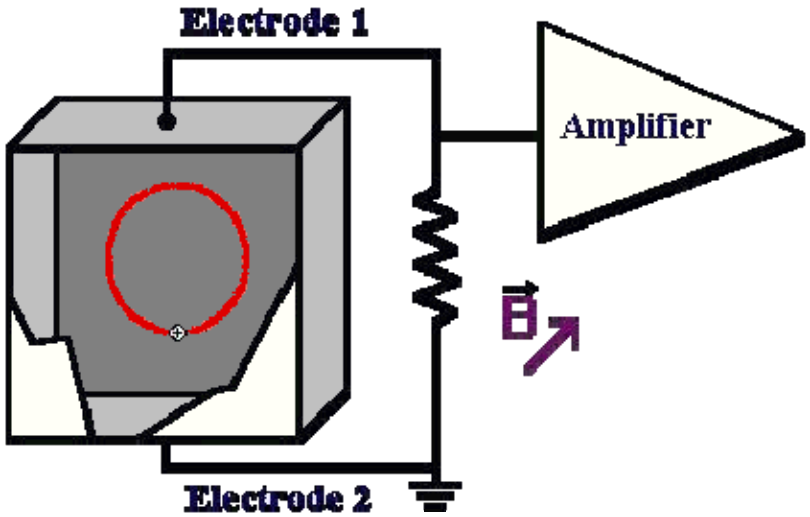

Ion Source
Mass Analyzer
Detector
电子倍增器
像电流
intensity
mass/charge

检测器

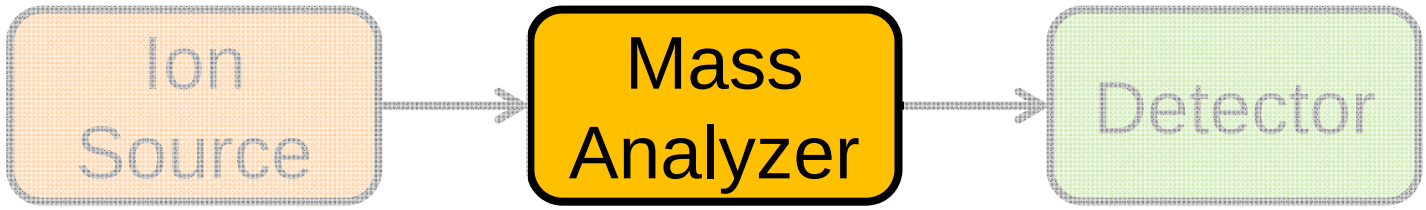
电子倍增检测器



像电流检测器



质量分析器



- 飞行时间 (Time-of-flight, TOF)
- 扇形磁场 (Magnetic Sector)
- 四级杆 (quadrupoles, Q)
- 离子阱 (Ion trap, IT)
- 离子回旋共振 (ion cyclotron resonance, ICR)
- 轨道阱 (Orbitrap)

质谱原理

$$\bar{F} = m\bar{a} = m\frac{d\bar{v}}{dt} = z(\bar{E} + \bar{v} \times \bar{B})$$

$$\frac{m}{z}\frac{d\bar{v}}{dt} = \bar{E} + \bar{v} \times \bar{B}$$

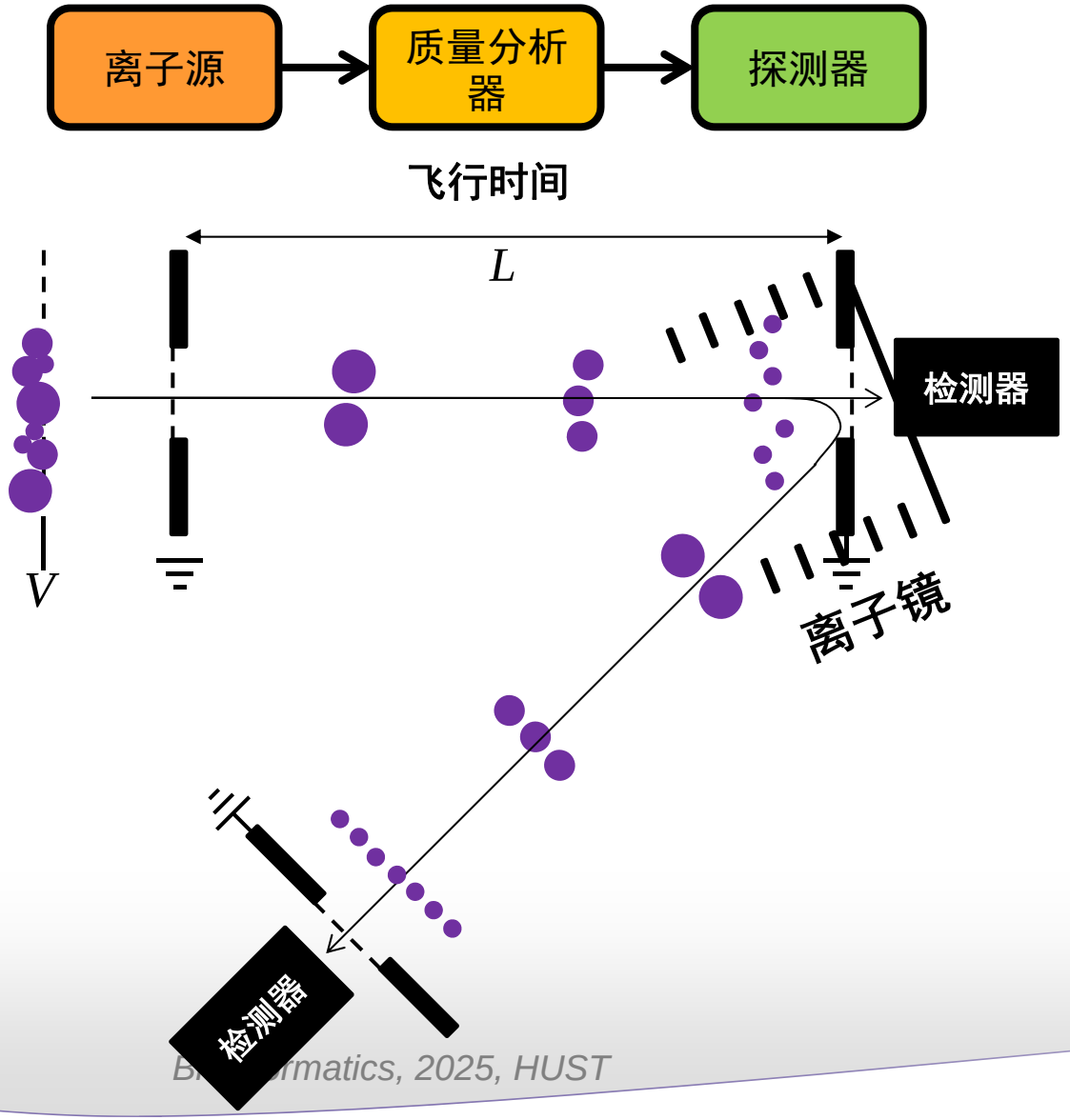
- ***z*** 电荷数
- ***E*** 电场强度
- ***v*** 速度
- ***B*** 磁场强度

Time-of-Flight, TOF

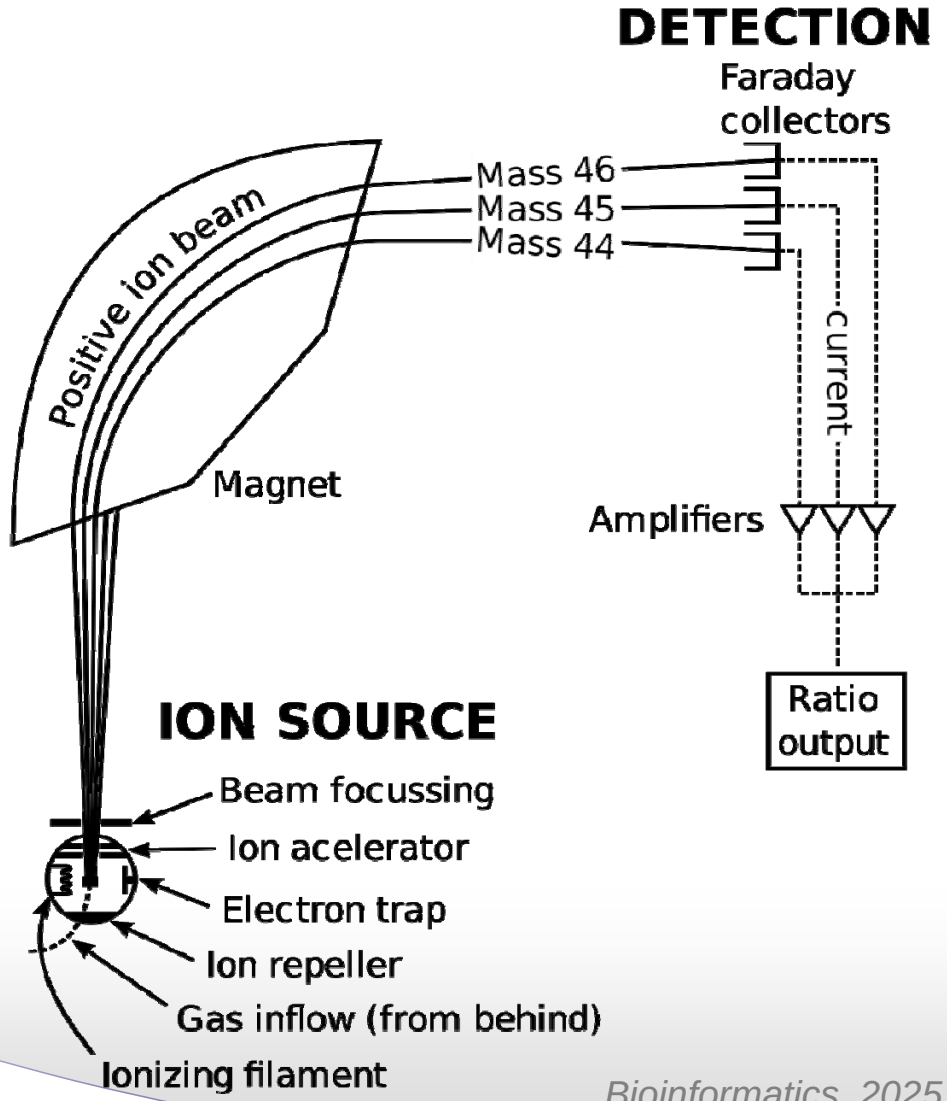
$$zV = \frac{mv^2}{2}$$

$$L = vt$$

$$t = \frac{L}{\sqrt{2V}}\sqrt{\frac{m}{z}}$$



扇形磁场

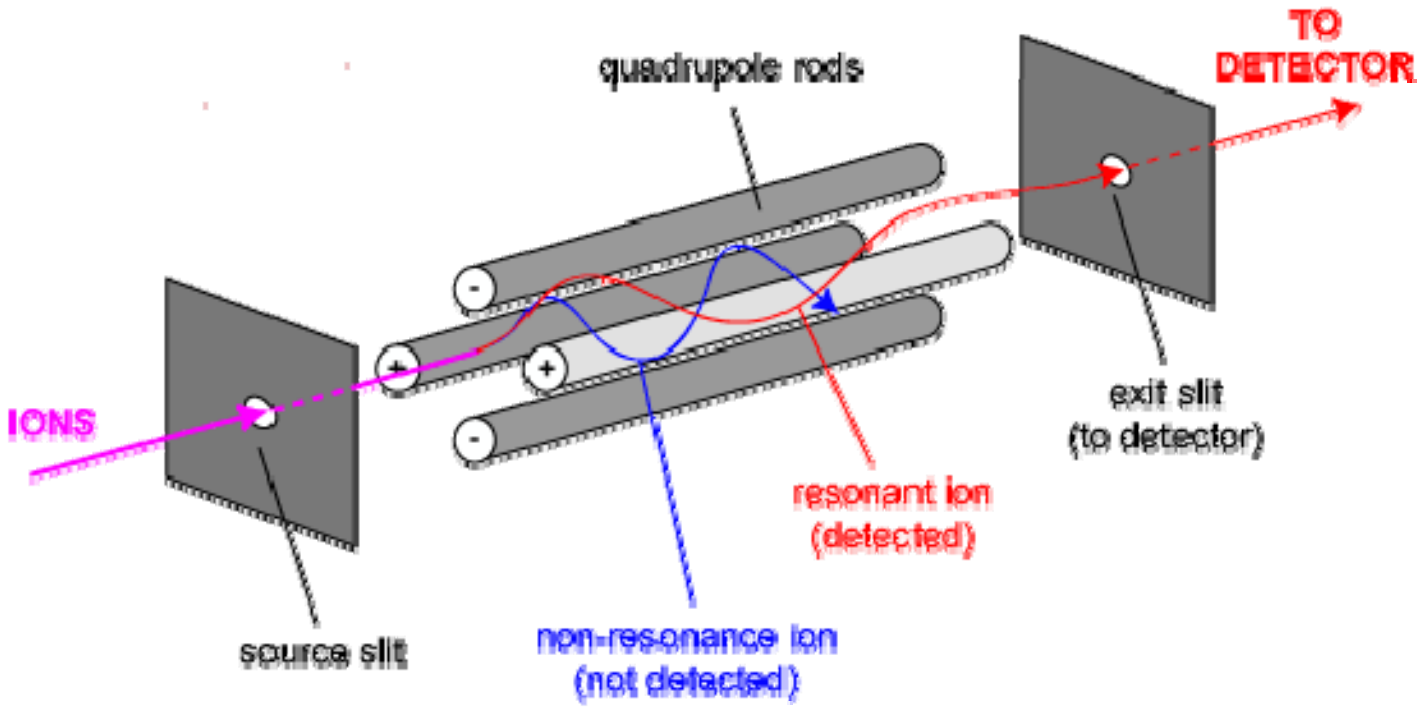


$$\overline{F} = \frac{mv^2}{R} = zvB$$

$$zV = \frac{mv^2}{2}$$

$$R = \sqrt{\frac{m}{z}}\sqrt{\frac{2V}{B}}$$

四级杆



$$-U - V\cos(\omega t)$$

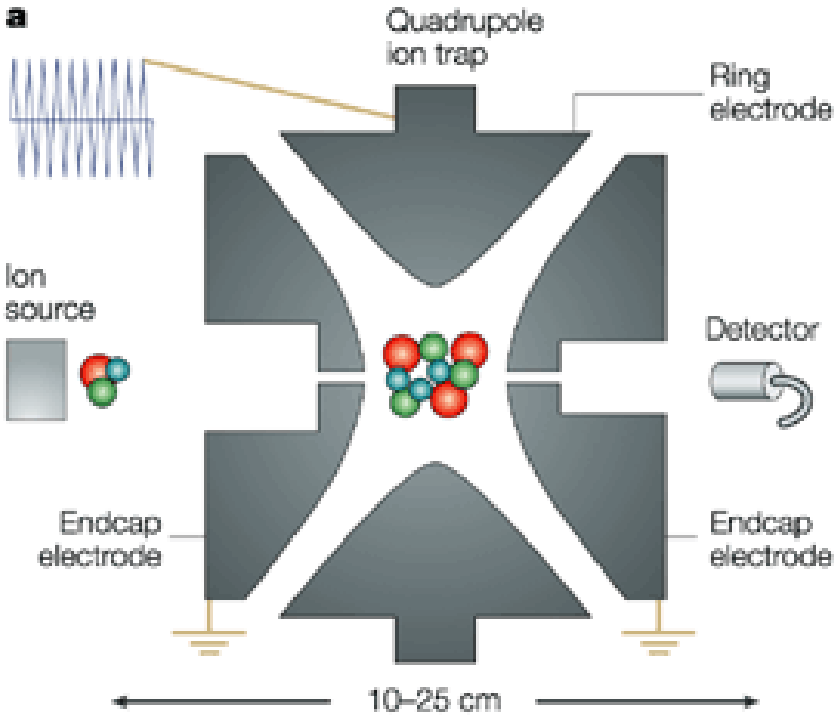
$$U + V\cos(\omega t)$$

y

x

离子阱

❒ 四级杆离子阱


a
Quadrupole
ion trap
Ring
electrode
Ion
source
Detector
Endcap
electrode
Endcap
electrode
10–25 cm

傅里叶变换的离子回旋共振

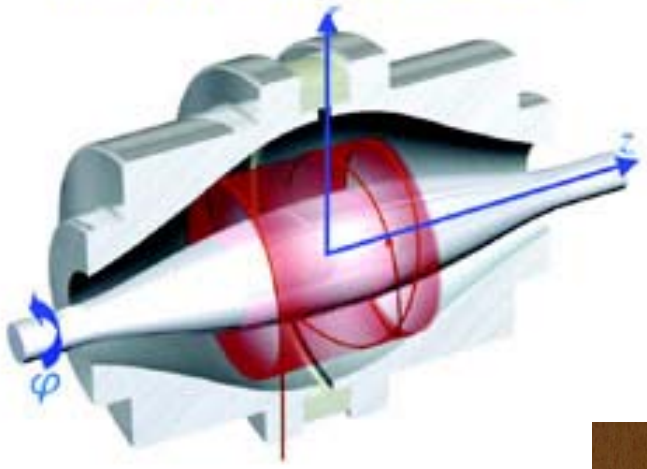
$$\overline{F} = \frac{mv^2}{R} = zvB$$

$$\boxed{\frac{m}{z}} = \frac{R}{v}B = \frac{B}{\omega}$$

轨道阱 (Orbitrap)

LTQ Orbitrap XL

- Characteristic frequencies:
  - *Frequency of rotation* ω_φ
  - *Frequency of radial oscillations* ω_r
  - *Frequency of axial oscillations* ω_z



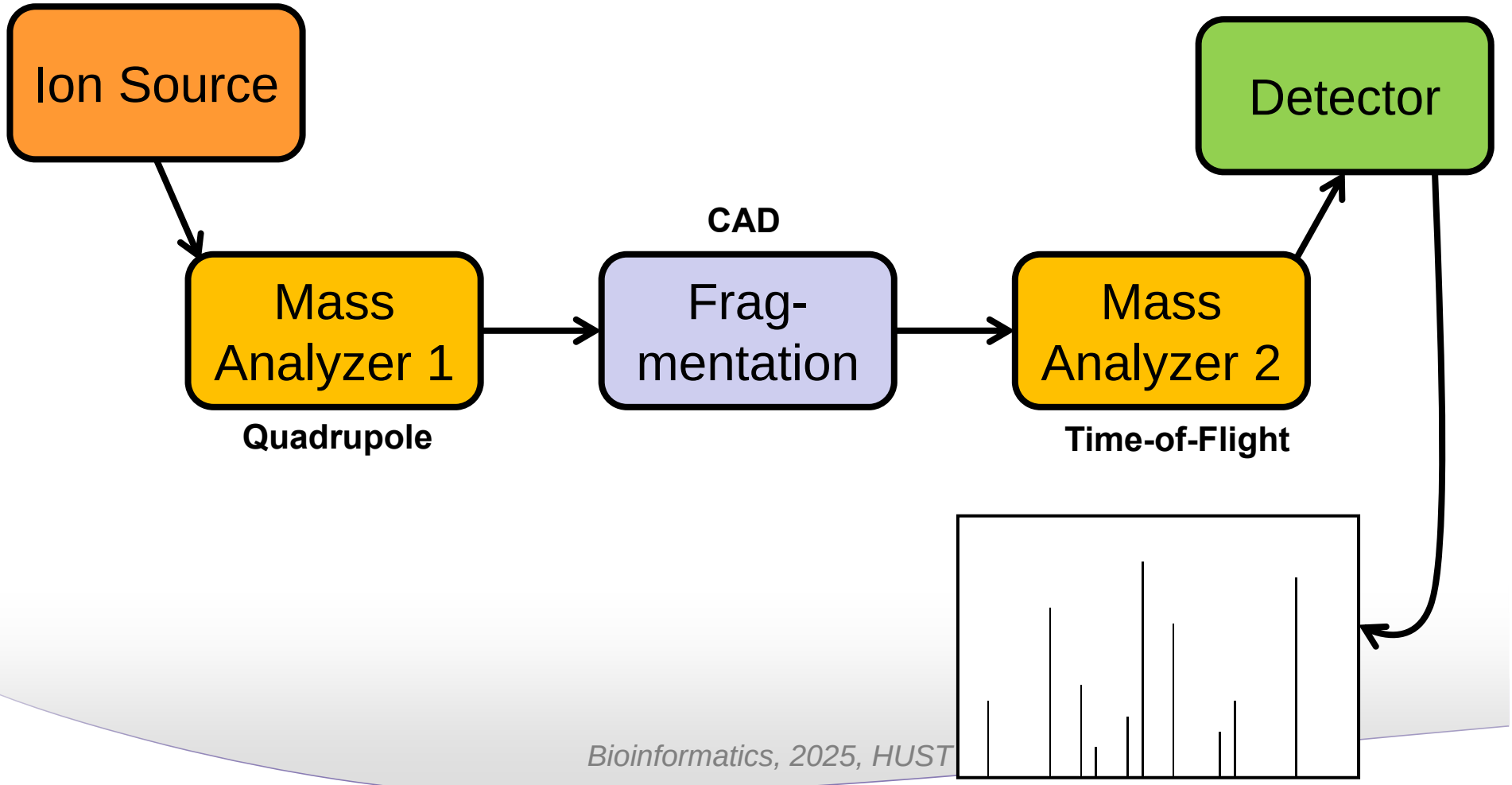
$$\omega_\varphi = \frac{\omega_z}{\sqrt{2}}\sqrt{\left(\frac{R_m}{R}\right)^2 - 1}$$

$$\omega_r = \omega_z\sqrt{\left(\frac{R_m}{R}\right)^2 - 2}$$

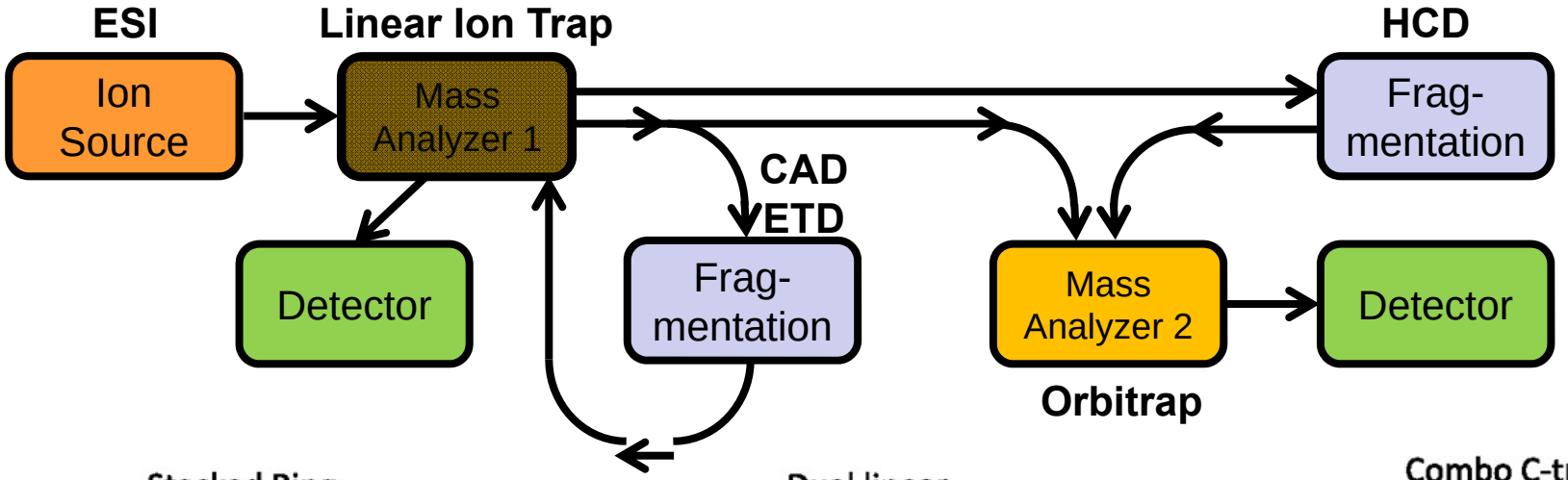
$$\omega_z = \sqrt{\frac{k}{m/z}}$$

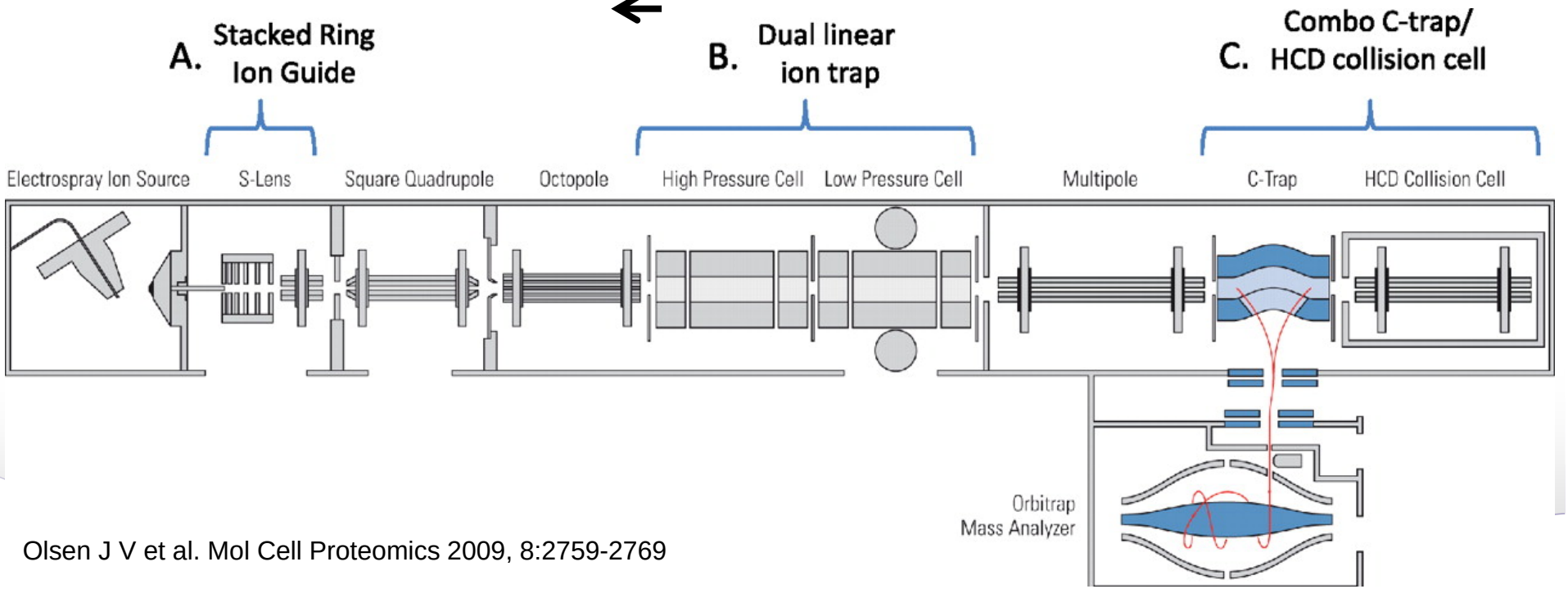
四级杆-时间飞行串联质谱

- 碰撞活化离解 (Collision Activated Dissociation)



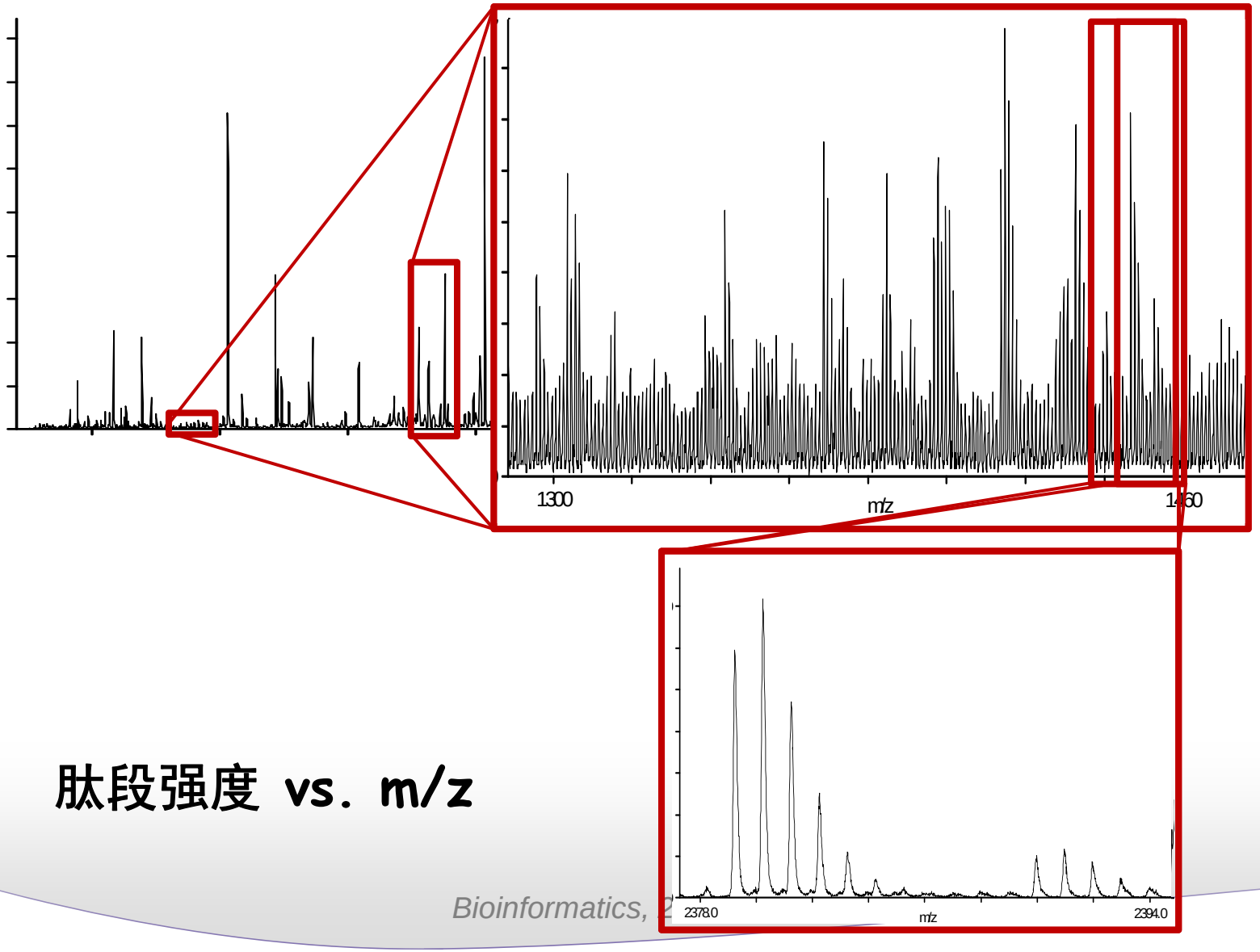
Linear Ion Trap / Orbitrap





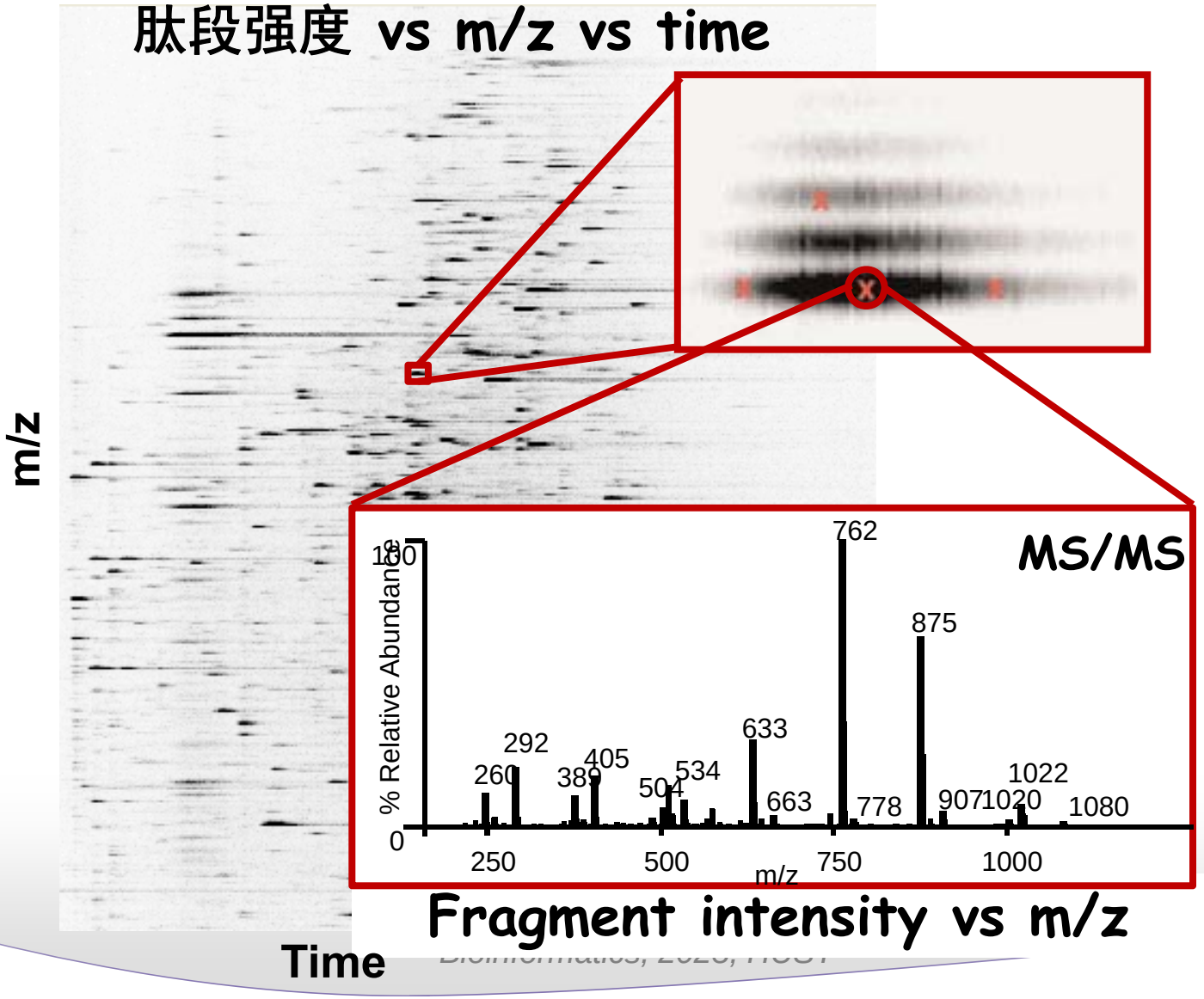
Olsen J V et al. Mol Cell Proteomics 2009, 8:2759-2769

质谱图：MALDI-TOF

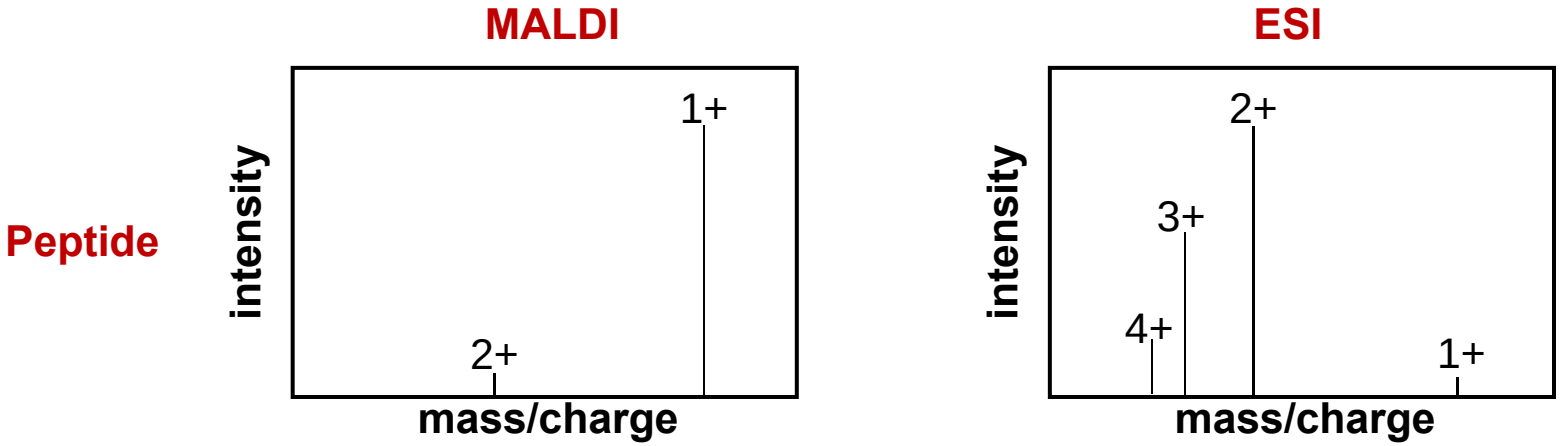


肽段强度 vs. m/z

质谱图：ESI-LC-MS/MS



电荷态分布

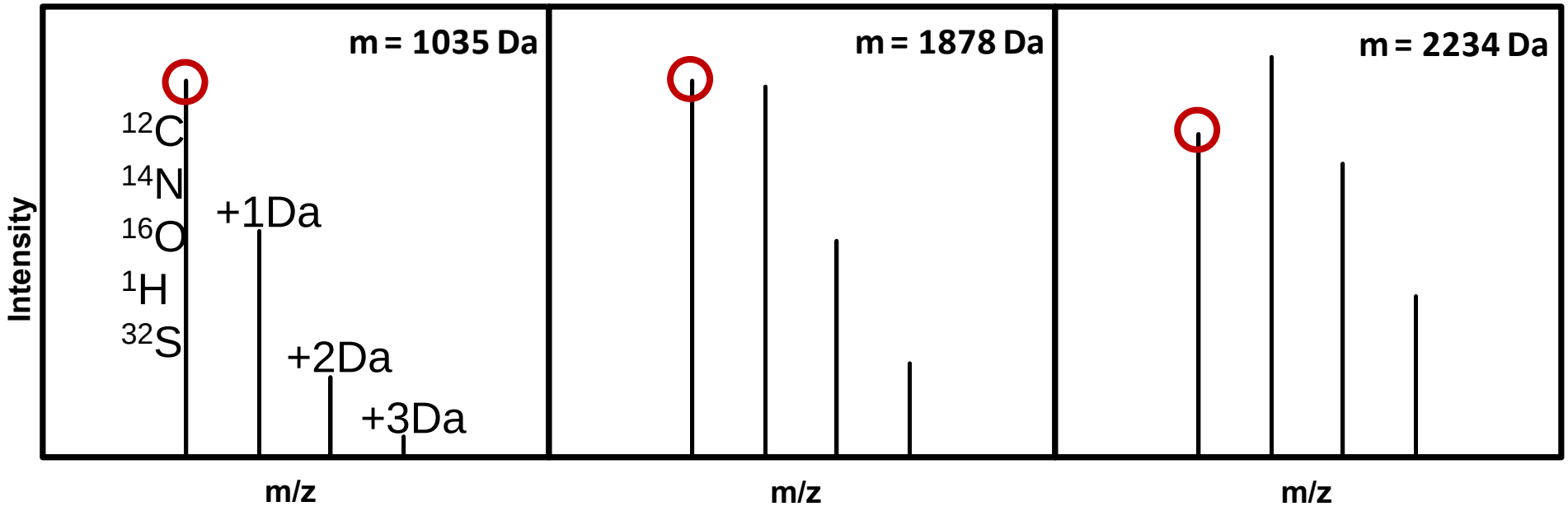


$$\frac{m}{z} = \frac{M + nH}{n}$$

M – 分子量
n – 电荷数
H – 质子的分子量

- 分子量898 Da的肽段
  - 携带1个H^+：(898 + 1) / 1 = 899 m/z
  - 携带2个H^+：(898 + 2) / 2 = 450 m/z
  - 携带3个H^+：(898 + 3) / 3 = 300.3 m/z

同位素分布



0.015% ^{2}H
1.11% ^{13}C
0.366% ^{15}N
0.038% ^{17}O, 0.200% ^{18}O,
0.75% ^{33}S, 4.21% ^{34}S, 0.02% ^{36}S

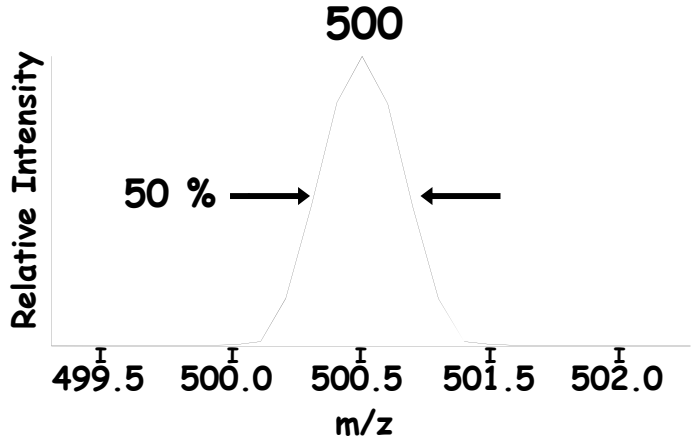
仅有^{12}C and ^{13}C:
p=0.0111
n：肽段中C的个数
m：肽段中^{13}C的个数
T_m：肽段m中^{13}C的相对强度

$$T_m = \binom{n}{m} p^m (1-p)^{n-m}$$

分辨率 (Resolution)

ΔM = full width at half maximum (FWHM)

ΔM = 最小质量峰的分离能力，允许区分两种不同类型的离子



分辨率 (R) = M/ΔM = 500/0.5 = 1000

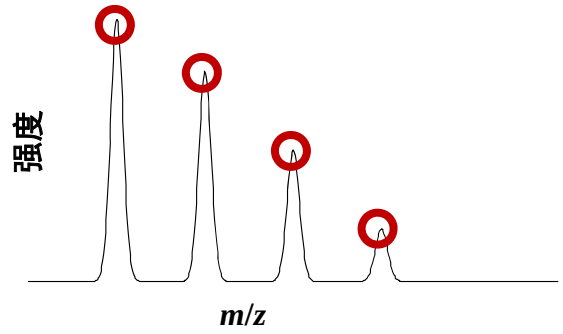
- 多高的分辨率能够区分乙酰化 (acetylation, +42.0100)和三甲基化 (Trimethylation, 42.0464)?
  - R = 1600/0.0364 = 43,956
  - 1/R = 22.75 ppm

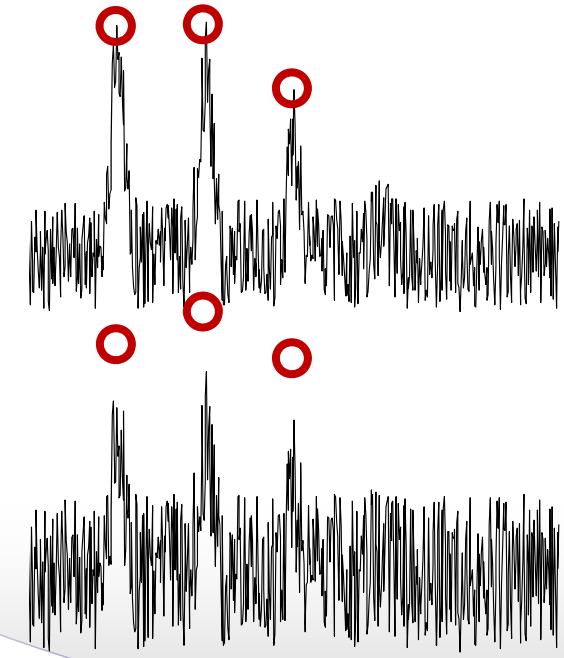
质谱仪-质量准确性

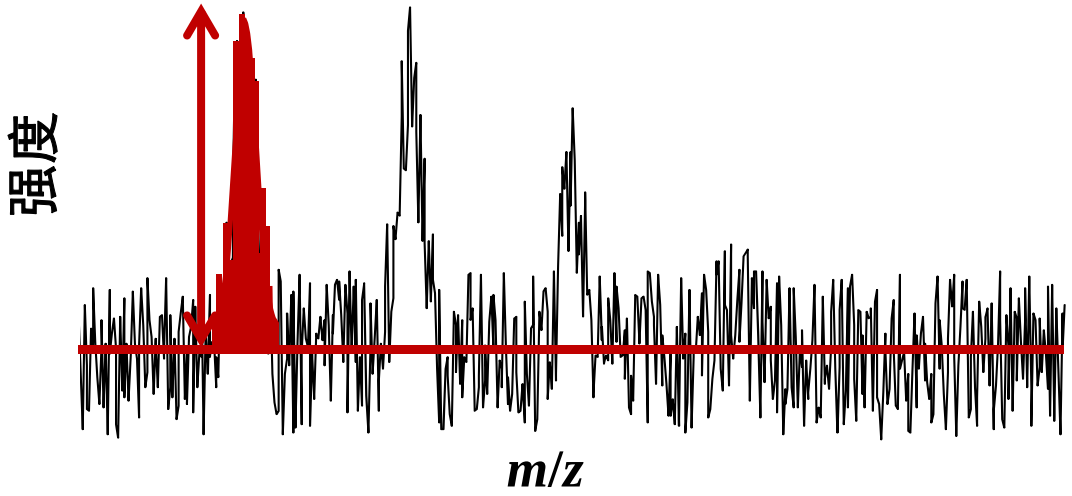
- ppm = $\Delta M/M = 10^{-6}$

Instrument	Applications	Resolution	Mass accuracy	Sensitivity	Dynamic range	Scan rate
LIT (LTQ)	Bottom-up protein identification in high-complexity, high-throughput analysis, LC-MS^n capabilities	2000	100 ppm	Femtomole	1e4	Fast
TQ (TSQ)	Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)	2000	100 ppm	Attomole	1e6	Moderate
LTQ-Orbitrap	Protein identification, quantification, PTM identification	100,000	2 ppm	Femtomole	1e4	Moderate
LTQ-FTICR, Q-FTICR	Protein identification, quantification, PTM identification, top-down protein identification	500,000	<2 ppm	Femtomole	1e4	Slow, slow
Q-TOF, IT-TOF	Bottom-up, top-down protein identification, PTM identification	10,000	2–5 ppm	Attomole	1e6	Moderate, fast
Q-LIT	Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)	2,000	100 ppm	Attomole	1e6	Moderate, fast

质量峰发现







峰高 (Peak height)：

抗干扰，统计规律性较差

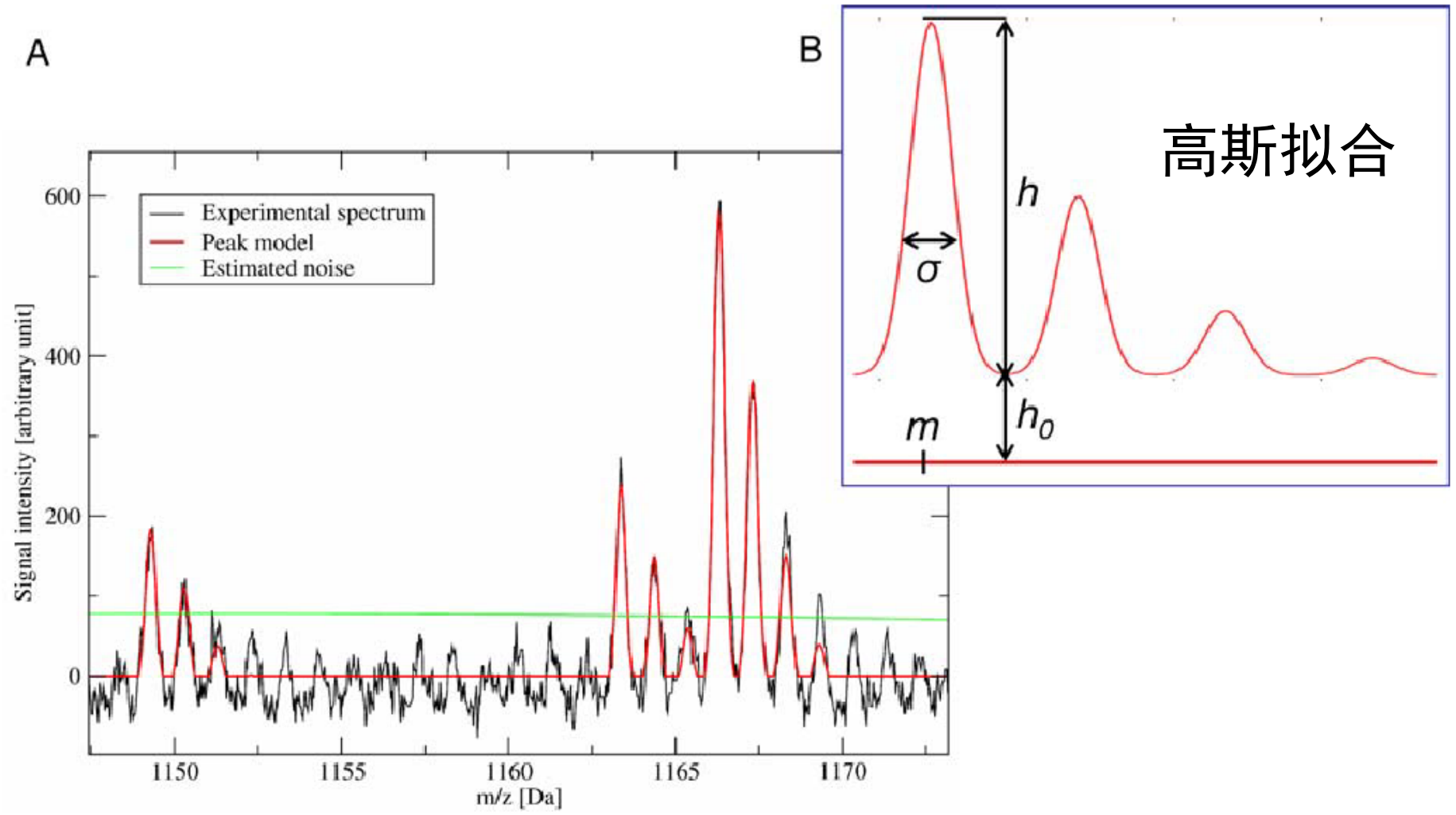
峰面积 (Peak area)：

统计规律性较好，抗干扰差

曲线拟合 (Curve fitting)：

统计规律性较好，需要知道质量峰的形状，运算速度慢

质量峰发现：高斯拟合

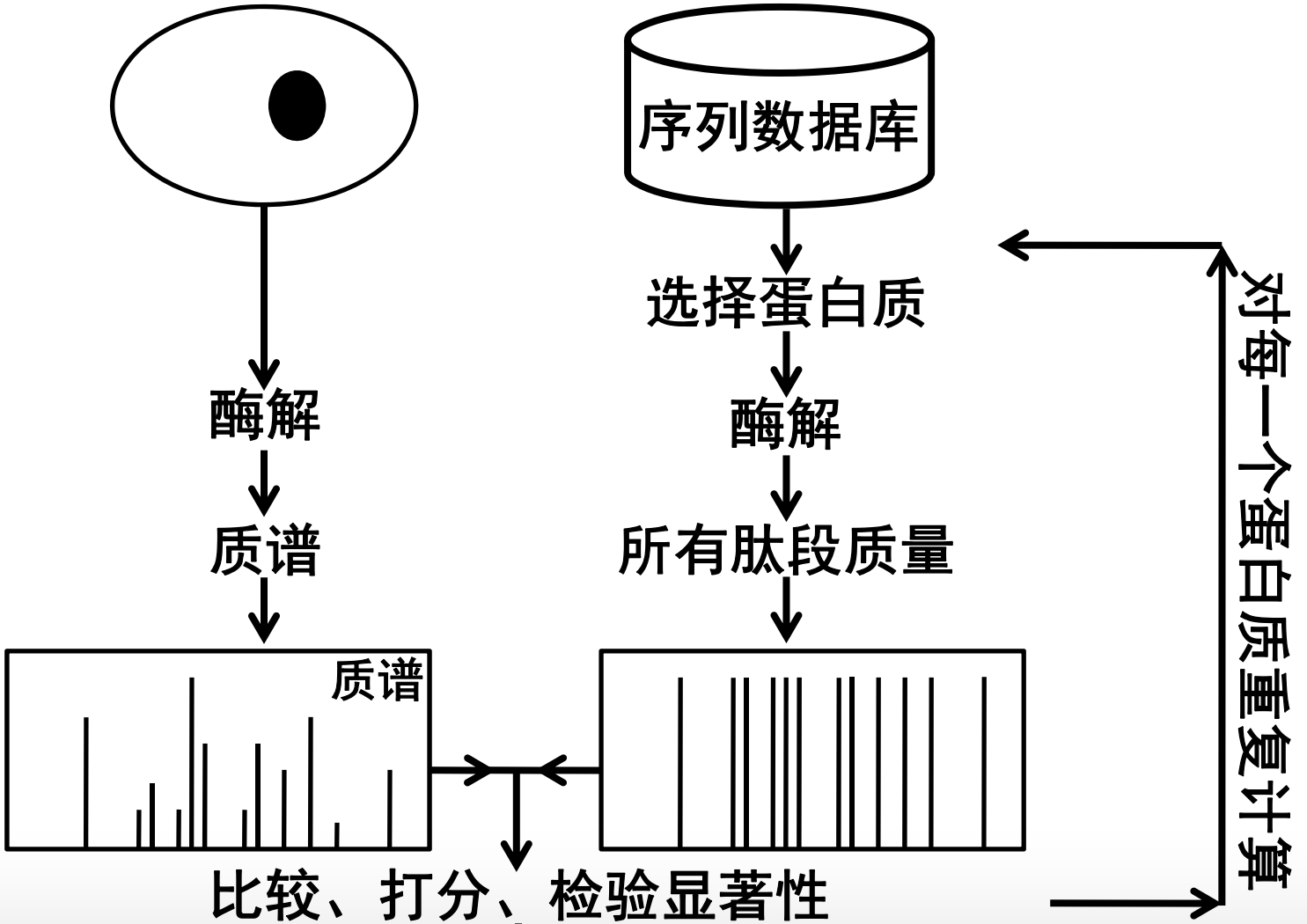


doi:10.1371/journal.pcbi.0030114.g003

蛋白质鉴定：肽段指纹谱

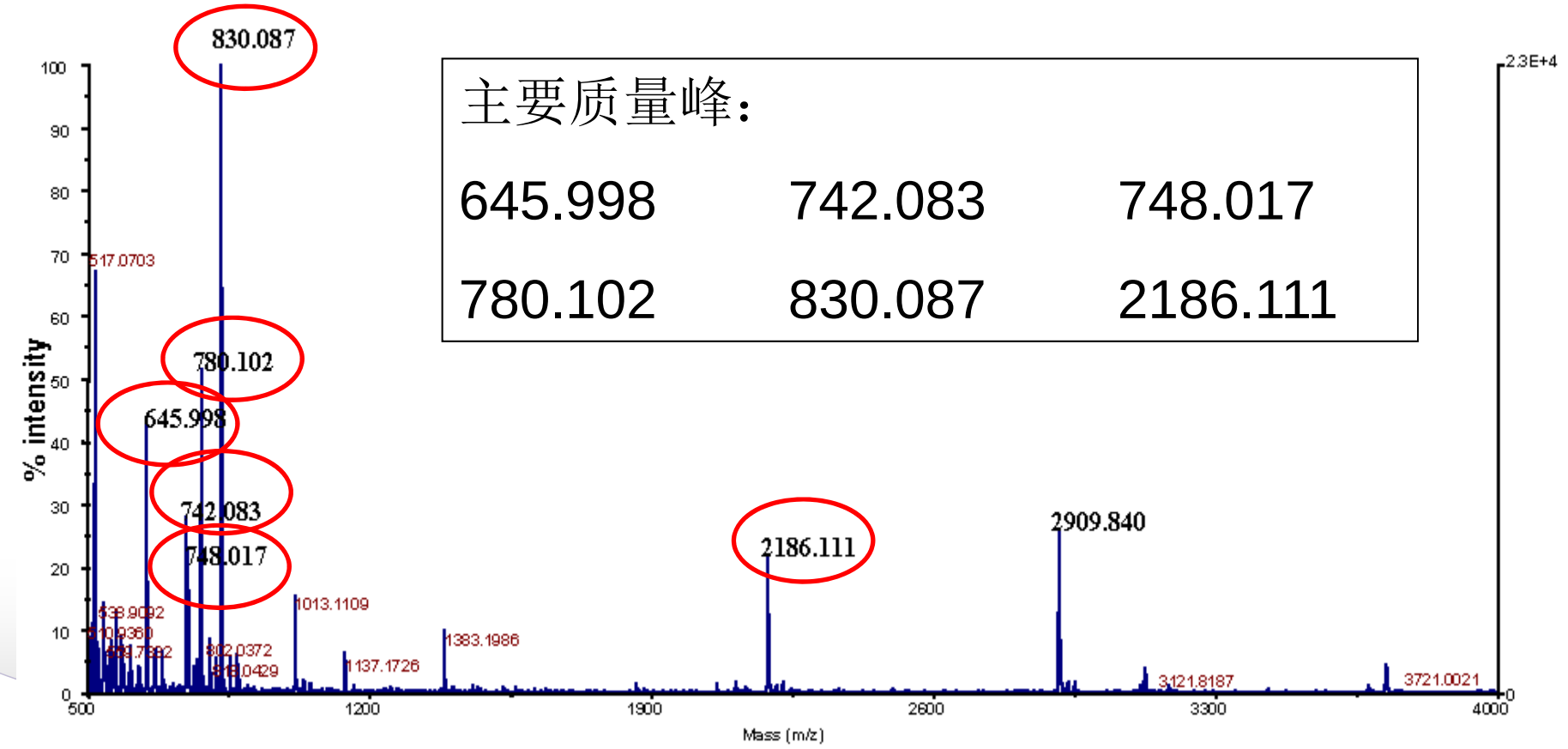
- Peptide Mass Fingerprinting (PMF)
- 一级谱MS数据
- 从基因组测序中可获得预测的蛋白质序列
- 之前蛋白质不需要被测序
- 仅需要知道基因的开放阅读框 (open reading frame) 即可

肽段指纹谱



肽段指纹谱鉴定

- 牛β-casein的胰蛋白酶切



主要质量峰：

645.998	742.083	748.017
780.102	830.087	2186.111

理论酶切肽段生成

```
>sp|P02666|CASB_BOVIN Beta-casein OS=Bos taurus GN=CSN2 PE=1 SV=2
MKVLILACLVALALARELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDEL
QDKIHPFAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVPPFLQPEVMGVSKVKEAMAPK
HKEMPFPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQPHQPLPPTVMFPPQSVLSL
SQSKVLPVPQKAVPYPQRDMPIQAFLLYQEPVLGPVRGPFPIIV
```

PeptideMass

PeptideMass [references] cleaves a protein sequence from the UniProt Knowledgebase (Swiss isoelectric point and mass values for the protein of interest. If desired, PeptideMass can return th variants.

Instructions are available.

Enter a UniProtKB protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P0

```
MADSRDPASDQMQHWKEQRAAQKADVLTTGAGNPVGDKLNVITVGPRGPL
LVQDVVFTDEMAHFDRERIP
ERVVHAKGAGAFGYFEVTHDITKYSKAKVFEHIGKKTPIAVRFSTVAGES
GSADTVRDPRGFAVKFYTED
GNWDLVGNNTPIFFIRDPILFPSFIHSQKRNPQTHLKDPDMVWDFWSLRP
ESLHQVSFLFSDRGIPDGHR
HMNGYGSHTFKLVNANGEAVYCKFHYKTDQGIKNLSVEDAARLSQEDPDY
GIRDLFNAIATGKYPSWTFY
IQVMTFNQAETFPFNPFDLTKVWPHKDYPLIPVGKLVLNRNPVNYFAEVE
QIAFDPSNMPPGIEASPDKM
LQGRLFAYPDTHRHRLGPNYLHIPVNCPYRARVANYQRDGPMCMQDNQGG
APNYYPNSFGAPEQQPSALE
```

Reset the fields. Perform the cleavage of the protein.

The peptide masses are
with cysteines treated with: nothing (in reduced form)
- with acrylamide adducts
- with methionines oxidized
- $[M+H]^+$ or $[M]$ or $[M-H]^-$ or $[M+2H]^{2+}$ or $[M+3H]^{3+}$.
- average or monoisotopic.

Select an enzyme : Trypsin

The peptide masses from your sequence are:

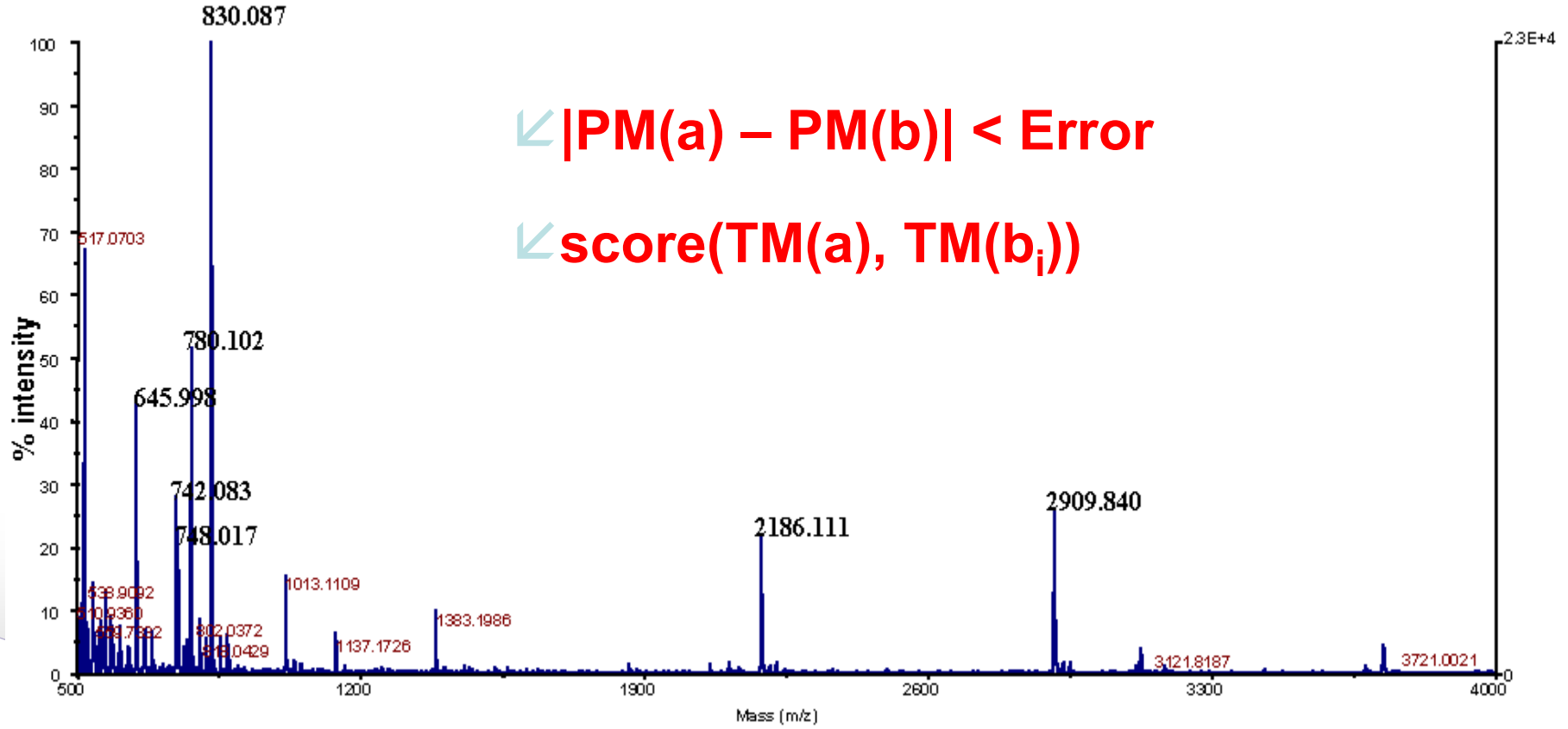
[Theoretical pI: 6.90 / Mw (average mass): 59756.17 / Mw (monoisotopic mass): 59718.75]

mass	position	#MC	modifications	peptide sequence
4569.9768	389-430	0		DGPMCMQDNQGGAPNYYPNS FGAPEQQPSALEHSIQYSGE VR
3432.6285	274-301	0		YPSWTFYIQVMTFNQAETFP FNPFDLTK
3209.5149	178-203	0		DPDMVWDFWSLRPESLHQVS FLFSDR
3176.4880	321-349	0		NPVNYFAEVEQIAFDPSNMP PGIEASPDK
2518.2037	136-156	0		FYTEDGNWDLVGNNTPIFFI R
2189.0695	48-66	0		GPLLVQDVVFTDEMAHFDR
2184.0720	481-499	0		NFTEVHPDYGSHIQALLDK
1755.8999	366-380	0		LGPNYLHIPVNCPYR
1712.8278	78-93	0		GAGAFGYFEVTHDITK
1708.8877	507-522	0		NAIHTFVQSGSHLAAR
1528.8158	157-169	0		DPILFPSFIHSQK
1493.6979	432-444	0		FNTANDDNVTQVR
1483.7023	113-127	0		FSTVAGESGSADTVR
1481.7383	445-456	0		AFYVNVLNEEQR
1414.7172	24-38	0		ADVLTTGAGNPVGDK
1342.5844	6-16	0		DPASDQMQHWK
1292.6117	253-263	0		LSQEDPDYGIR
1280.6303	222-233	0		LVNANGEAVYCK
1278.5684	211-221	0		HMNGYGSHTFK
1119.5581	355-363	0		LFAYPDTHR
1097.5772	459-468	0		LCENIAGHLK
1049.5625	264-273	0		DLFNAIATGK
1001.5666	307-315	0		DYPLIPVGK
974.4901	244-252	0		NLSVEDAAR
968.5887	39-47	0		LNVITVGPR
962.5305	469-476	0		DAQIFIQK
849.4465	500-506	0		YNAEKPK
837.4577	171-177	0		NPQTHLK

http://web.expasy.org/peptide_mass/

质量峰选择

- 实际质量与理论质量之差小于允许误差
- 将实际峰与理论峰比较打分



- $|PM(a) - PM(b)| < Error$
- $score(TM(a), TM(b_i))$

MOWSE Score

Mowse

The Mowse scoring algorithm is described in [Pappin, 1993].

The first stage of a Mowse search is to compare the calculated peptide masses for each entry in the sequence database with the set of experimental data. Each calculated value which falls within a given mass tolerance of an experimental value counts as a match. A molecular weight range for the intact protein can be used as a pre-filter.

Rather than just counting the number of matching peptides, Mowse uses empirically determined factors to assign a statistical weight to each individual peptide match. The matrix of weighting factors is calculated during the database build stage, as follows:

A frequency factor matrix, **F**, is created, in which each row represents an interval of 100 Da in peptide mass, and each column an interval of 10 kDa in intact protein mass. As each sequence entry is processed, the appropriate matrix elements $f_{i,j}$ are incremented so as to accumulate statistics on the size distribution of peptide masses as a function of protein mass. The elements of **F** are then normalised by dividing the elements of each 10 kDa column by the largest value in that column to give the Mowse factor matrix **M**:

$$m_{i,j} = \frac{f_{i,j}}{\left| f_{i,j} \right|_{\text{max in column } j}}$$

After searching the experimental mass values against a calculated peptide mass database, the score for each entry is calculated according to:

$$\text{Score} = \frac{50{,}000}{M_{\text{Prot}} \times \prod_{n} m_{i,j}}$$

Where M_{Prot} is the molecular weight of the entry and the product term is calculated from the Mowse factor elements for each match between the experimental data and peptide masses calculated from the entry.

肽段出现的频率表

- 打分原则：基于肽段出现的频率
- 给定参考序列数据库：建立频率表

蛋白质 / 肽段	0 – 10 k Da	10 – 20 k Da	20 – 30 k Da	…..
0-100 Da	54	23	7	…..
100-200 Da	34	12	23	
…	…..	….	…..	….

MASCOT PMF

- 设置参数
  - SwissProt
  - Trypsin
  - Mammals
  - Input
    - 645.998
    - 742.083
    - 748.017
    - 780.102
    - 830.087
    - 2186.111

http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF

MASCOT PMF

MATRIX SCIENCE Mascot Search Results

User : Yu Xue
Email : xueyu@hust.edu.cn
Search title : PP1
Database : SwissProt 2017_02 (553655 sequences; 198177566 residues)
Taxonomy : Mammalia (mammals) (66570 sequences)
Timestamp : 4 Mar 2017 at 08:51:57 GMT
Top Score : 82 for CASB_BOVIN, Beta-casein OS=Bos taurus GN=CSN2 PE=1 SV=2

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 61 are significant (p<0.05).

1. CASB_BOVIN **Mass:** 25091 **Score:** 82 **Expect:** 0.00047 **Matches:** 6
Beta-casein OS=Bos taurus GN=CSN2 PE=1 SV=2
CASB_BUBBU **Mass:** 25090 **Score:** 82 **Expect:** 0.00047 **Matches:** 6
Beta-casein OS=Bubalus bubalis GN=CSN2 PE=2 SV=1
B3A3_RAT **Mass:** 135322 **Score:** 38 **Expect:** 9.4 **Matches:** 6
Anion exchange protein 3 OS=Rattus norvegicus GN=Slc4a3 PE=1 SV=1
CENPJ_MOUSE **Mass:** 152985 **Score:** 30 **Expect:** 68 **Matches:** 6
Centromere protein J OS=Mus musculus GN=Cenpj PE=2 SV=2
CASB_CAPHI **Mass:** 24849 **Score:** 44 **Expect:** 2.4 **Matches:** 4
Beta-casein OS=Capra hircus GN=CSN2 PE=2 SV=1
CASB_SHEEP **Mass:** 24859 **Score:** 44 **Expect:** 2.4 **Matches:** 4
Beta-casein OS=Ovis aries GN=CSN2 PE=1 SV=3
NFAC4_MOUSE **Mass:** 95723 **Score:** 34 **Expect:** 27 **Matches:** 5
Nuclear factor of activated T-cells, cytoplasmic 4 OS=Mus musculus GN=Nfatc4 PE=1 SV=2
LANC2_MOUSE **Mass:** 50745 **Score:** 33 **Expect:** 30 **Matches:** 4
LanC-like protein 2 OS=Mus musculus GN=Lancl2 PE=1 SV=1

CASB_BOVIN

MATRIX SCIENCE MASCOT Search Results

Protein View: CASB_BOVIN

Beta-casein OS=Bos taurus GN=CSN2 PE=1 SV=2

Database:	SwissProt
Score:	82
Expect:	0.00047
Monoisotopic mass (M_r):	25091
Calculated pI:	5.26
Taxonomy:	Bos taurus

Sequence similarity is available as an NCBI BLAST search of CASB_BOVIN against nr.

Search parameters

Enzyme:	Trypsin: cuts C-term side of KR unless next residue is P.
Mass values searched:	6
Mass values matched:	6

Protein sequence coverage: 23%

Matched peptides shown in ***bold red***.

```
  1 MKVLILACLV ALALARELEE LNVPGEIVES LSSSEESITR INKKIEKFQS
 51 EEQQQTEDEL QDKIHPFAQT QSLVYPFPGP IPNSLPQNIP PLTQTPVVVP
101 PFLQPEVMGV SKVKEAMAPK HKEMPFPKYP VEPFTESQSL TLTDVENLHL
151 PLPLLQSWMH QPHQPLPPTV MFPPQSVLSL SQSKVLPVPQ KAVPYPQRDM
201 PIQAFLLYQE PVLGPVRGPF PIIV
```

Unformatted sequence string: 224 residues (for pasting into other applications).

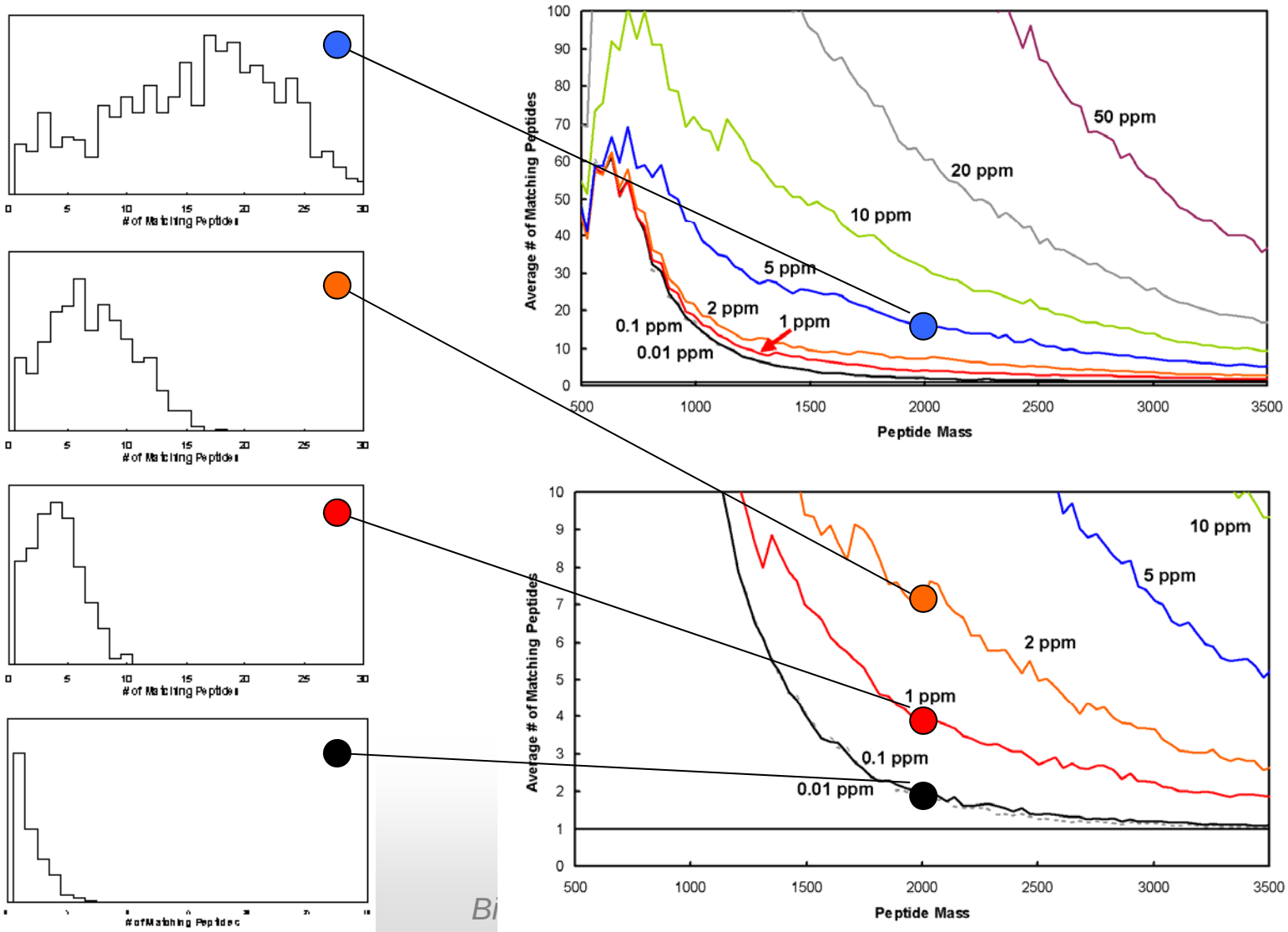
Sort by ◉ residue number ○ increasing mass ○ decreasing mass
Show ◉ matched peptides only ○ predicted peptides also

Start – End	Observed	Mr(expt)	Mr(calc)	Delta	M	Peptide
115 – 120	645.9980	644.9907	645.3156	-0.3248	0	K.EAMAPK.H
123 – 128	748.0170	747.0097	747.3625	-0.3528	0	K.EMPFPK.Y
185 – 191	780.1020	779.0947	779.4905	-0.3958	0	K.VLPVPQK.A
192 – 198	830.0870	829.0797	829.4446	-0.3649	0	K.AVPYPQR.D
199 – 217	2186.1110	2185.1037	2185.1605	-0.0568	0	R.DMPIQAFLLYQEPVLGPVR.G
218 – 224	742.0830	741.0757	741.4425	-0.3668	0	R.GPFPIIV.-

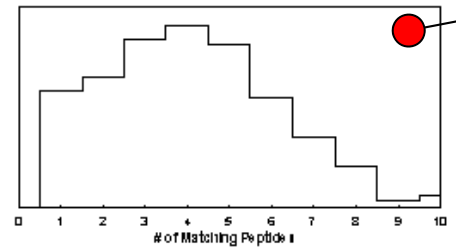
肽段指纹谱的优缺点

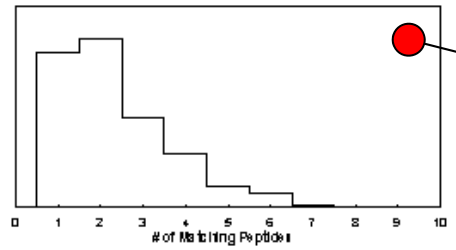
- 优点：
  - 快速、简便、灵敏度高
  - 不需要二级质谱MS/MS数据
  - 简单的质谱仪即可获得数据
- 缺点：
  - 不适合蛋白质混合物的分析
  - 可信度较低
  - 依赖于质谱的分辨率和质量准确性
  - 依赖于序列数据库的大小

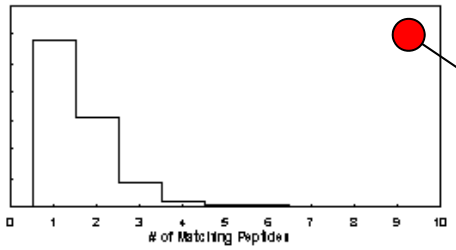
肽段鉴定 *vs.* 质量准确性

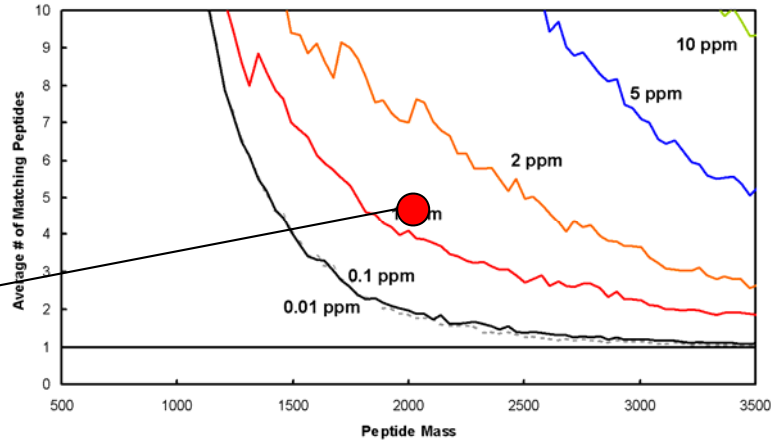


数据库大小

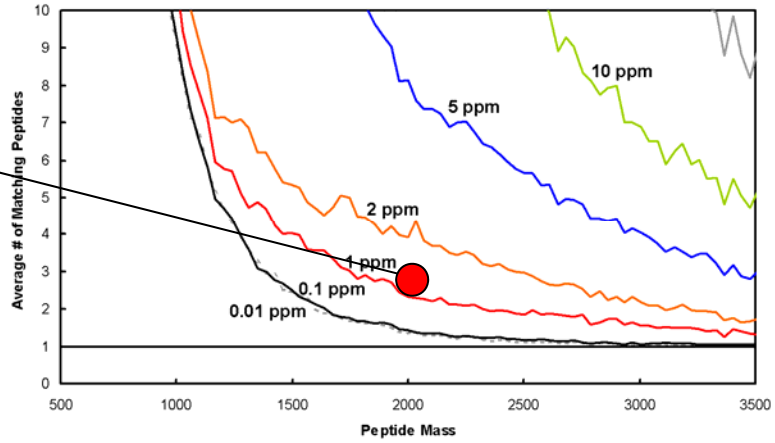




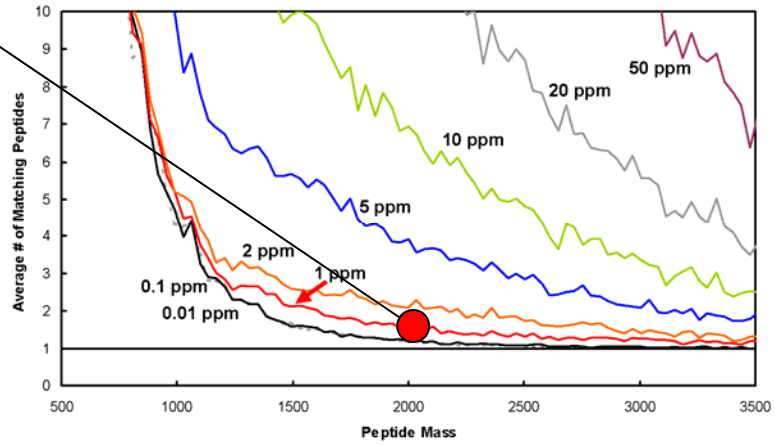




Human



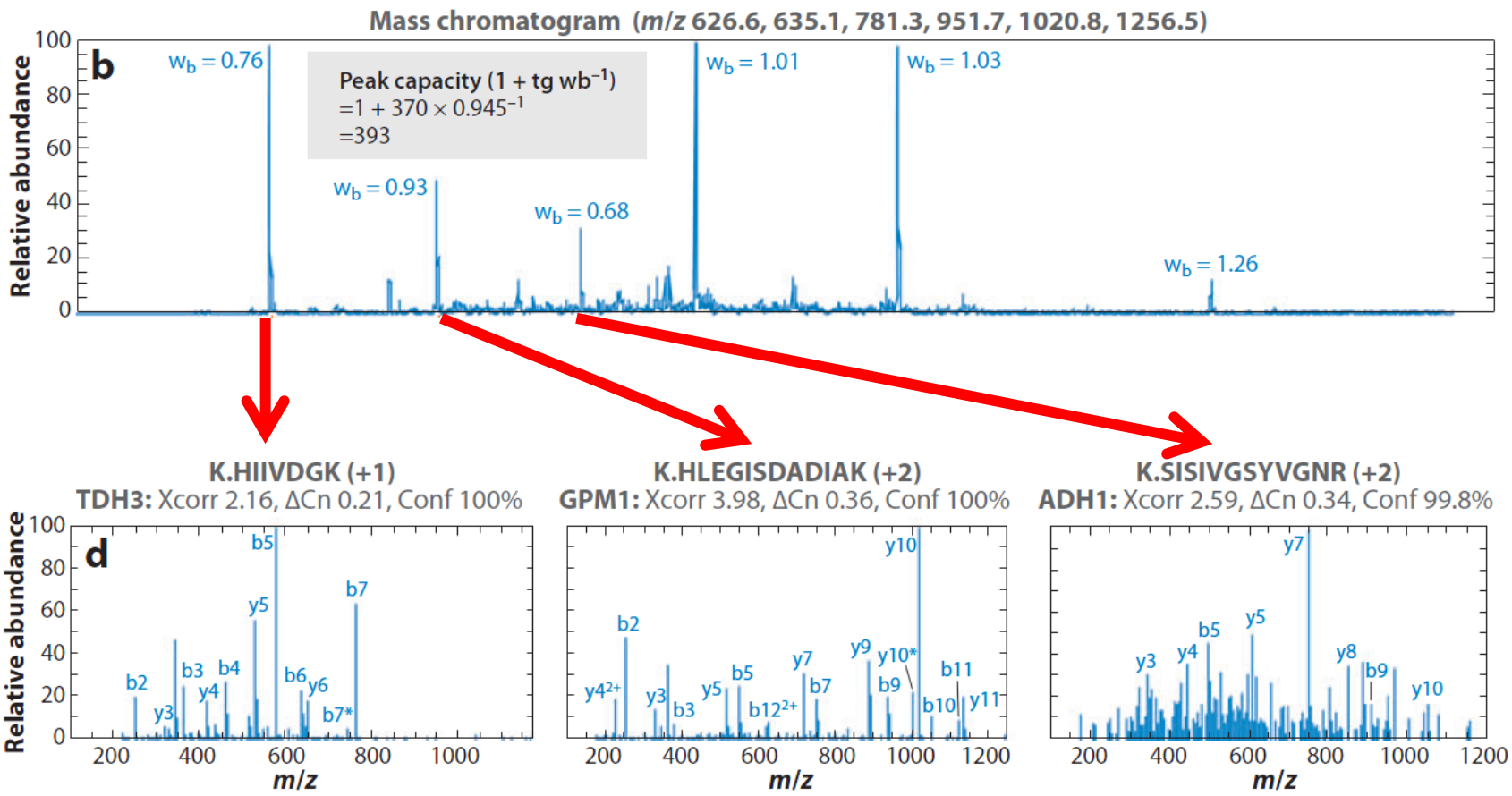
C. elegans



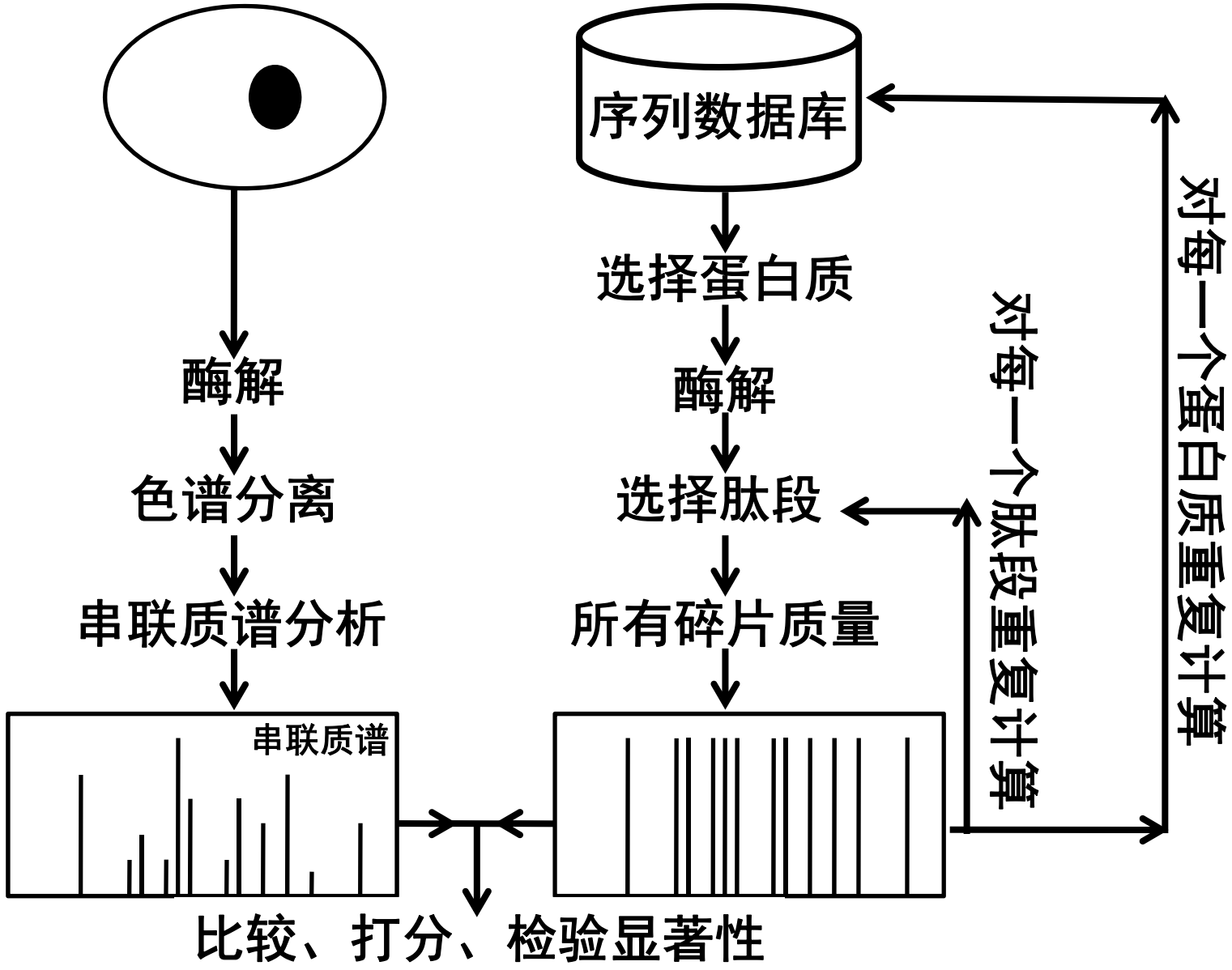
S. cerevisiae

串联质谱MS/MS

- 蛋白质酶解成肽段之后继续碎裂
- 二级谱

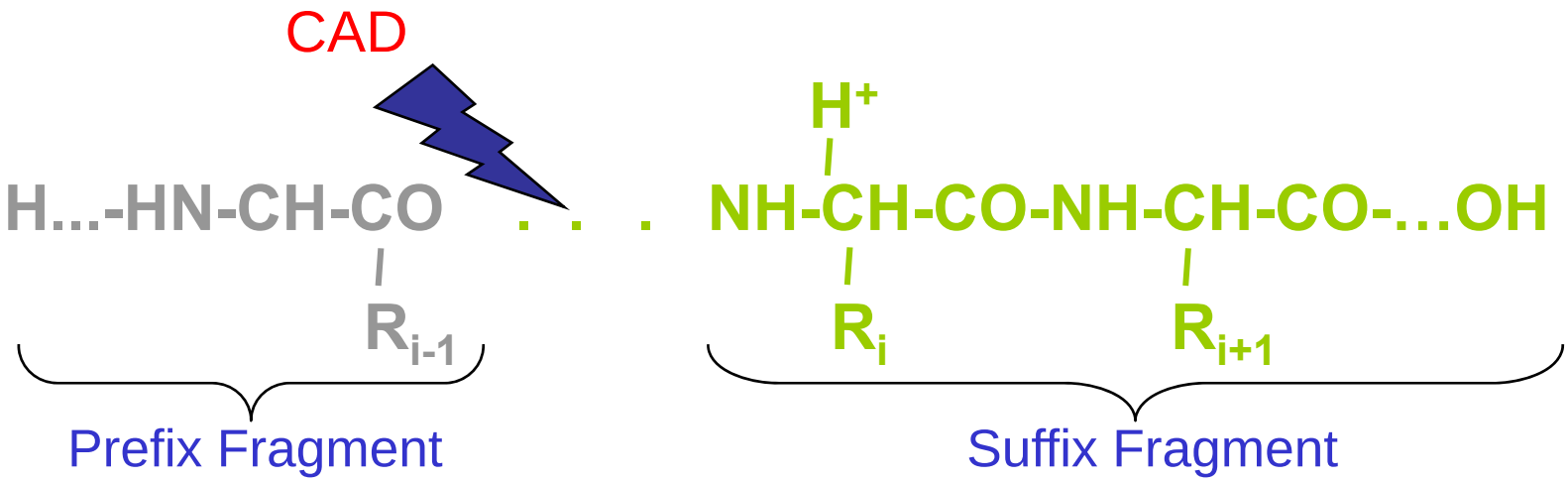


串联质谱：数据库搜索


序列数据库
选择蛋白质
酶解
选择肽段
所有碎片质量
对每一个蛋白质重复计算
对每一个肽段重复计算
酶解
色谱分离
串联质谱分析
串联质谱
比较、打分、检验显著性

肽段碎裂

- 碰撞活化离解，CAD

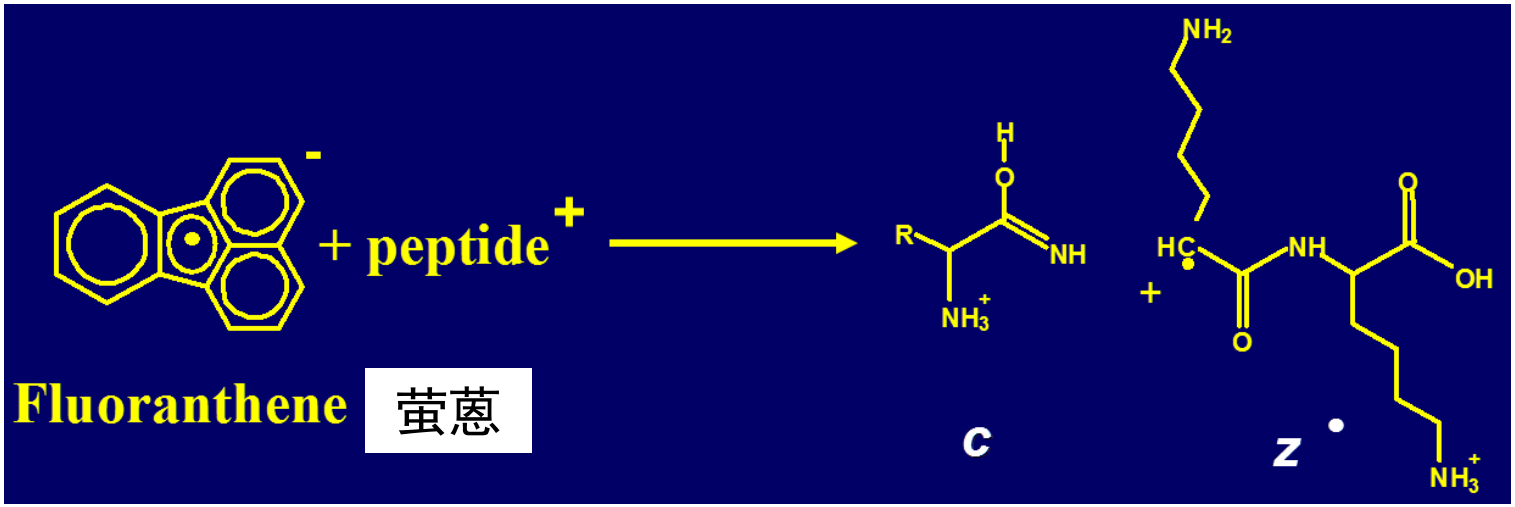


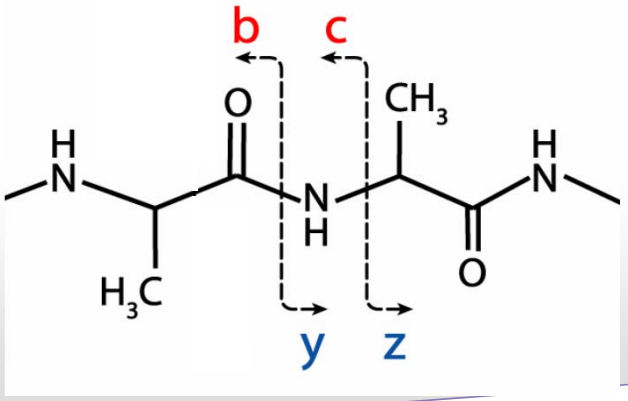
- 肽段倾向于在主链发生碎裂
- 中性丢失：丢失中性的基团如NH_3和H_2O

电子转移解离

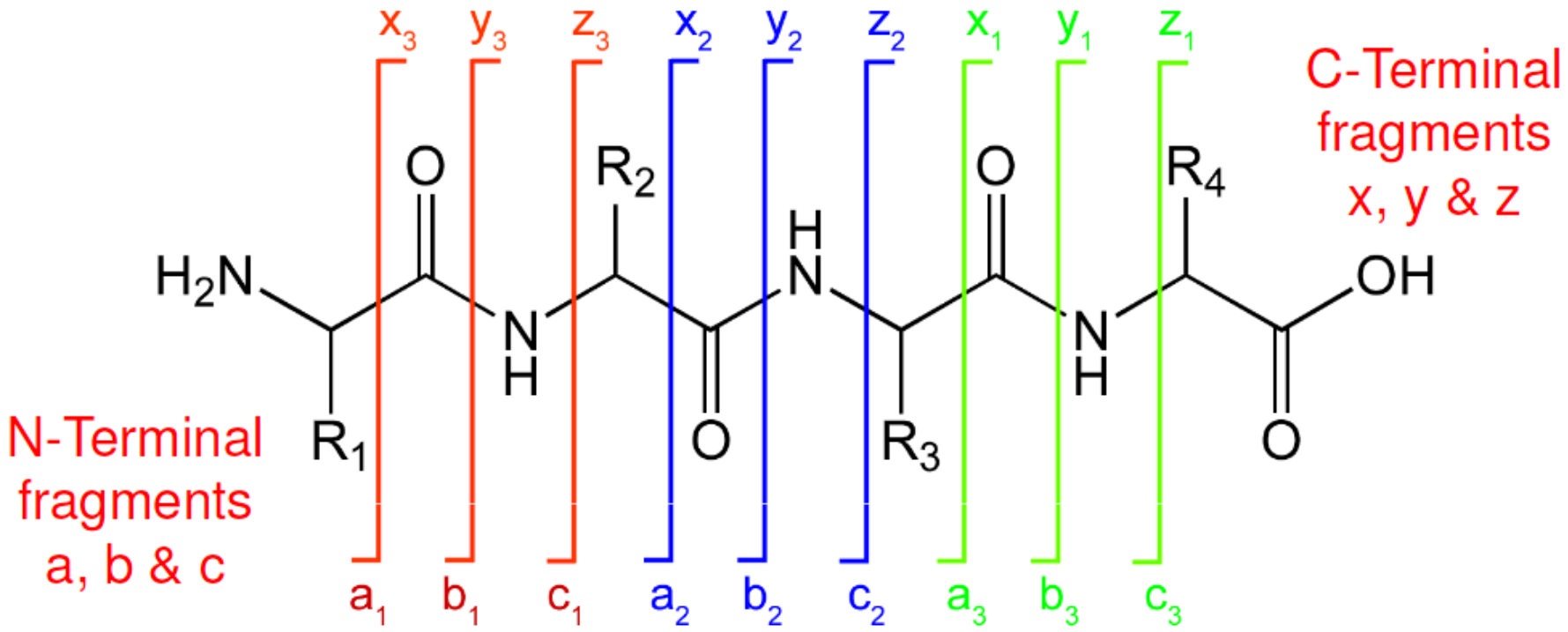
Electron Transfer Dissociation, ETD

- *c*, *z*离子



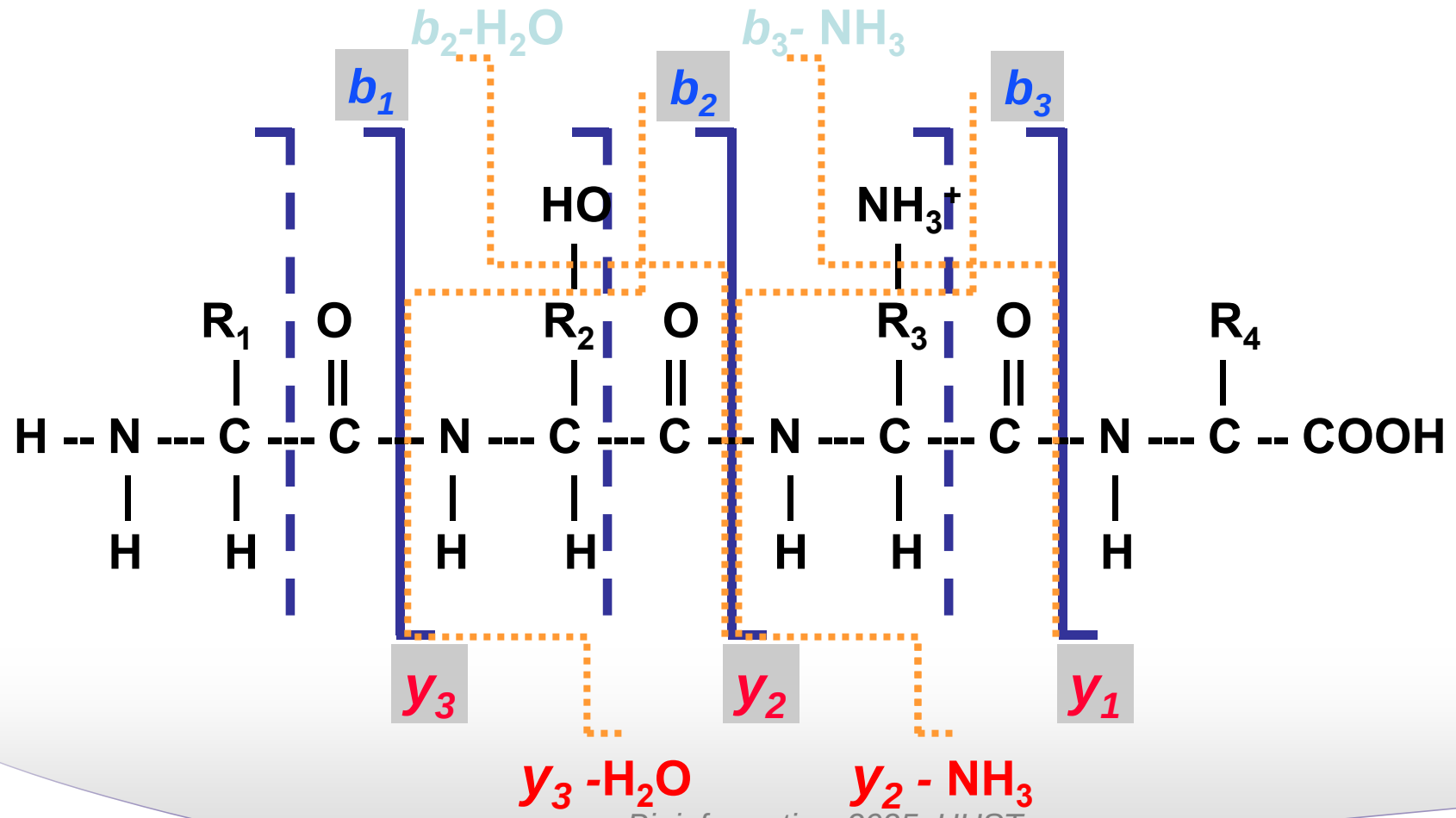


肽段碎裂的离子类型


x3 y3 z3 x2 y2 z2 x1 y1 z1
C-Terminal fragments x, y & z
N-Terminal fragments a, b & c
H2N
O
R2
H
N
O
R4
OH
R1
O
R3
N
H
O
a1 b1 c1 a2 b2 c2 a3 b3 c3

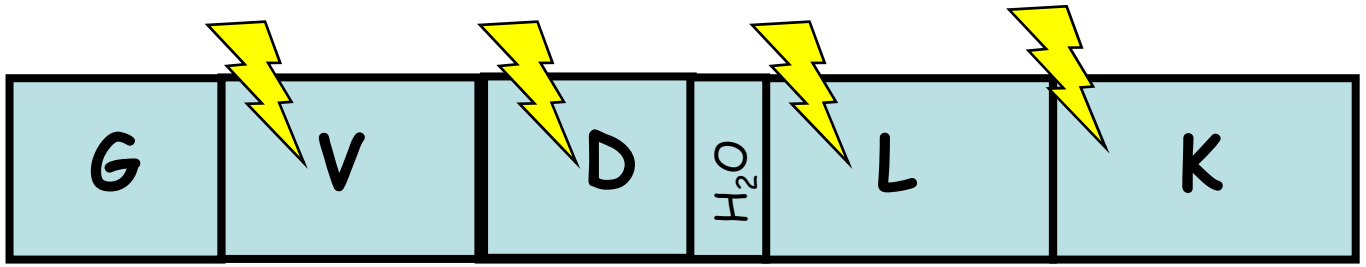
碎裂的主要离子类型

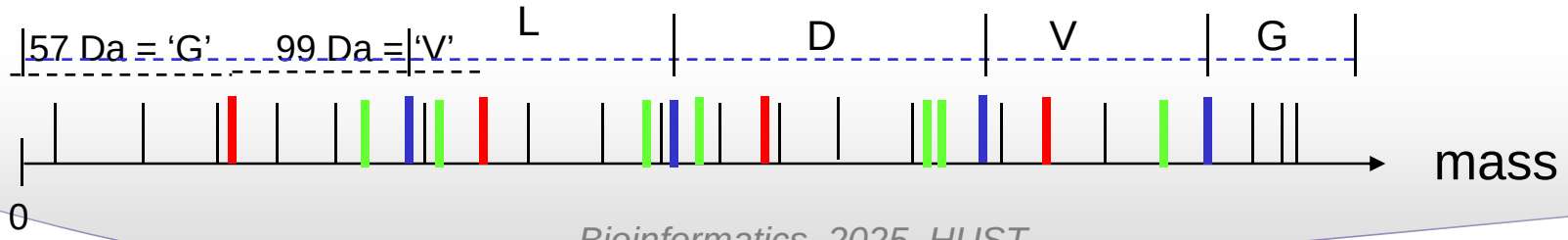
- CAD：主要是*b*、*y*型



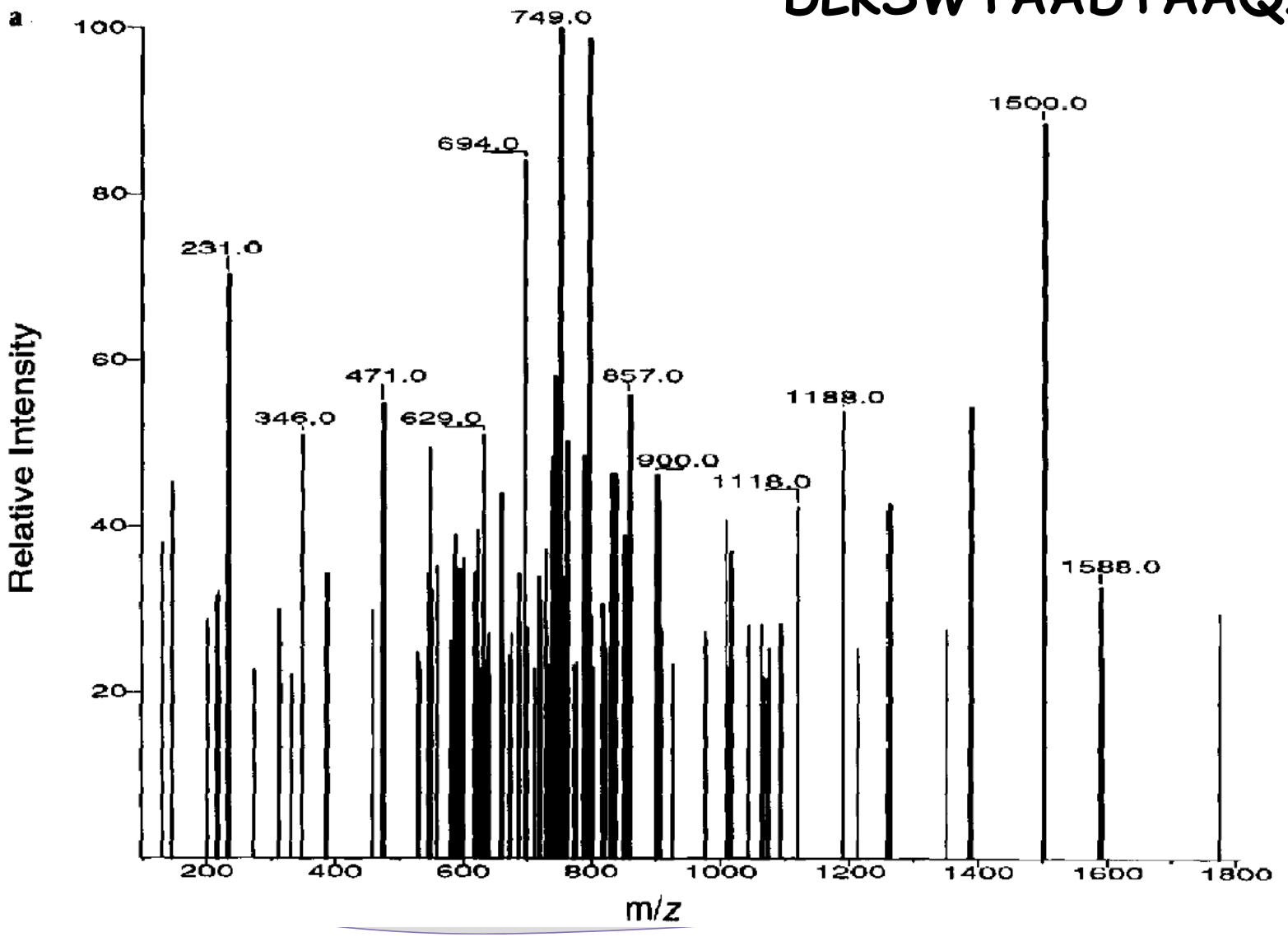
二级谱图

- Mass Spectra
  - 前缀和后缀片段
  - 中性丢失的片段 ($-H_2O$, $-NH_3$)
  - 噪音、缺失的质谱峰





真实的二级谱图


DLKSWTAADTAAQISQ
a
Relative Intensity
100
80
60
40
20
749.0
694.0
1500.0
231.0
471.0
857.0
1188.0
346.0
629.0
900.0
1118.0
1588.0
200
400
600
800
1000
1200
1400
1600
1800
m/z

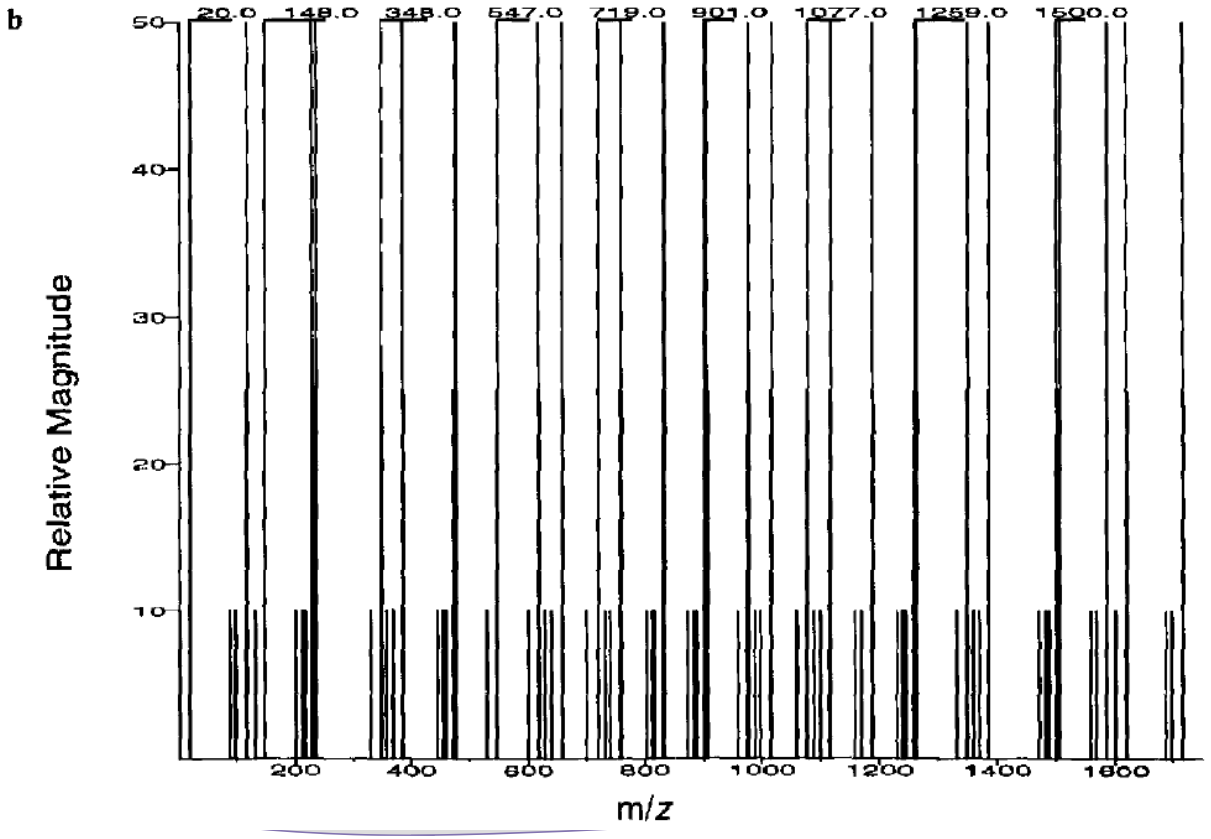
SEQUEST算法

- 理论谱图生成规则
  - ***b*, *y*: 50；*b*, *y* ± 1: 25**
  - **中性丢失 & *a*: 10**

$$b_n = \sum a_n + 1$$

$$y_n = \mathrm{MW} - \sum a_n$$

***a*: 氨基酸分子量**



谱图比较

$$S_p = \left(\sum i_m\right) \times n_m \times (1+\beta) \times (1+\rho)/n_t$$

匹配片段离子强度之和

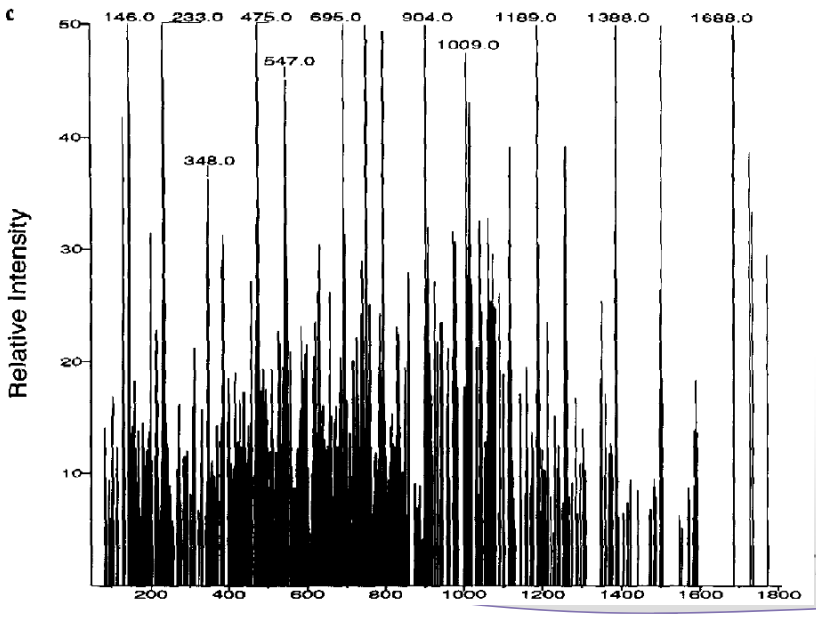
匹配的片段离子数量

β=0.075
ρ=0.15

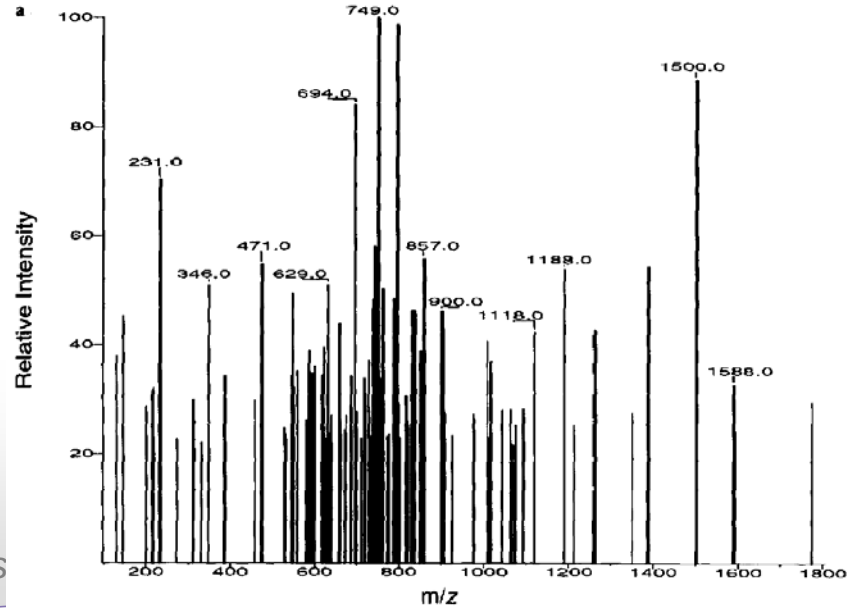
所有片段离子

- 首次打分：前200个得分最高的肽段

理论谱图



实验谱图



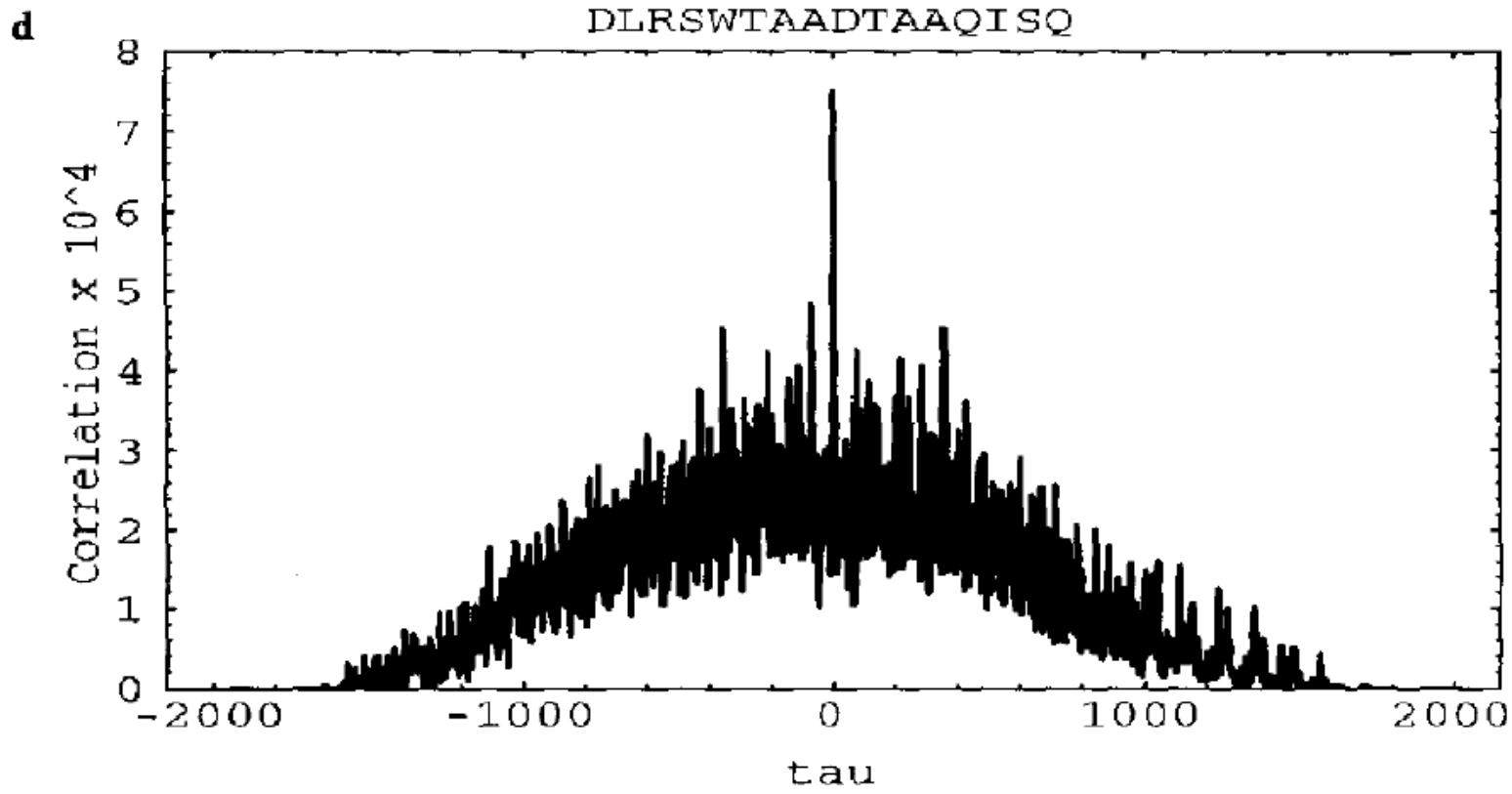
相关性分析

- 第二轮打分：针对首轮200个得分最高的肽段
- 交叉相关系数
  - 理论谱*a*左移或右移*t*之后与实验谱*e*比较

$$(a*e)(t)=\frac{\sum_{i=0}^{N}(a_i-E(a))(e_{i+t}-E(e))}{\sqrt{Var(a)Var(e)}}$$

$$X_{corr}=\frac{(a*e)(0)-E\{(a*e)(t);t\in[-75;75]\}}{(a*a)(0)-E\{(a*a)(t);t\in[-75;75]\}}.$$

理论谱与实验谱的相关性



MASCOT

- Matrix Science, http://www.matrixscience.com
- 基本假设：肽段匹配是独立的伯努利过程

$$L = \log\left(\prod_{i=1}^{n}\left[\prod_{\theta\in M(s,i)}\frac{p_\theta}{r_\theta}\prod_{\theta\in S(s,i)-M(s,i)}\frac{1-p_\theta}{1-r_\theta}\right]\right)$$

p_θ 实验谱的片段与理论谱的匹配概率

r_θ 实验谱的片段与随机谱的匹配概率

从头测序

- ***de novo* sequencing**

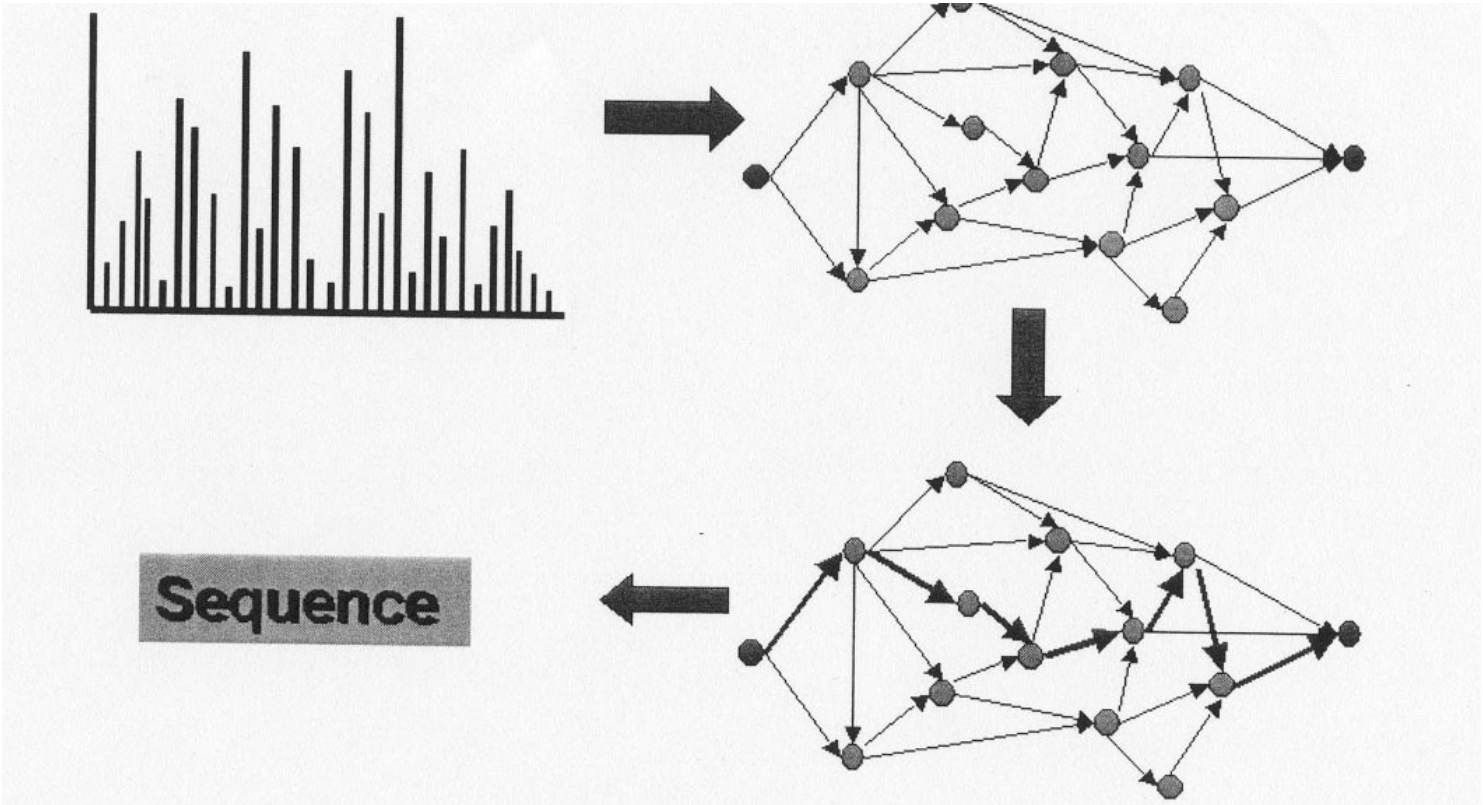
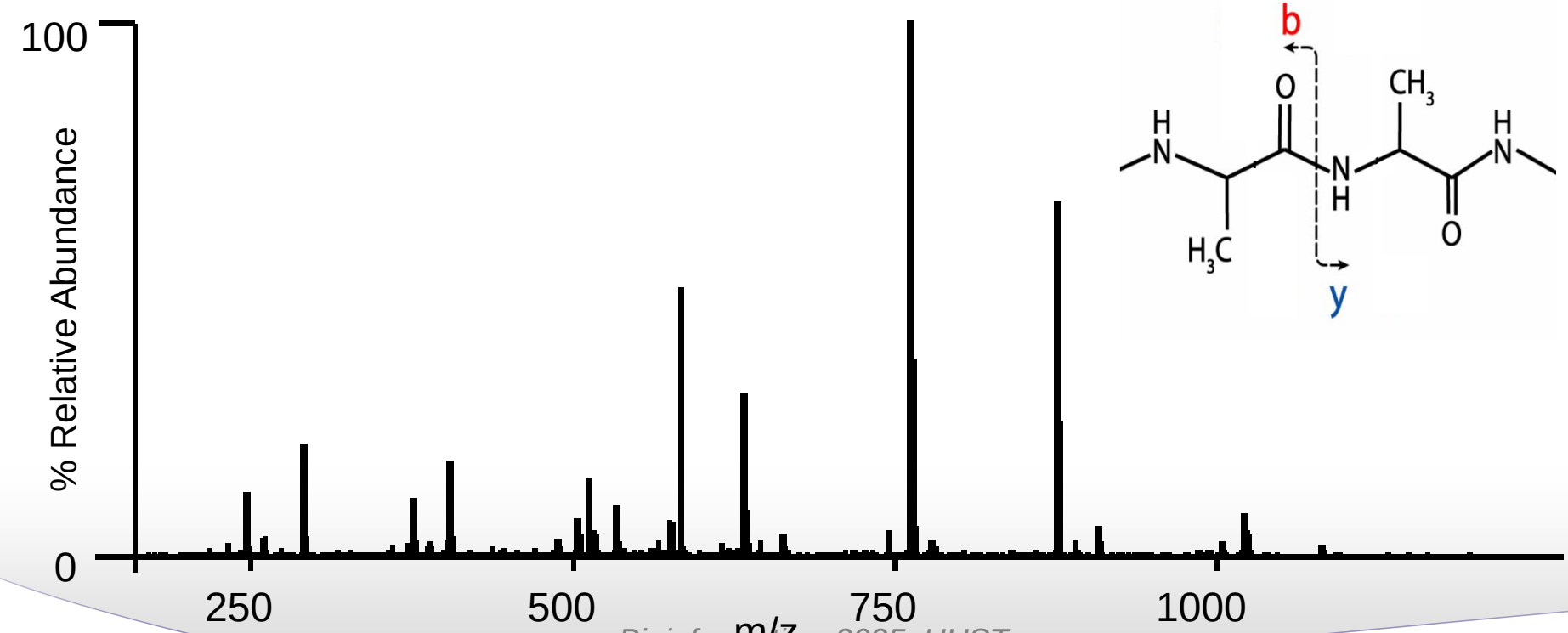


Figure 11.5: Noise in a spectrum generates many "false" edges and vertices in the spectrum graph and disguises edges corresponding to the real peptide sequence. Sequence reconstruction corresponds to finding an optimal path in the spectrum graph.

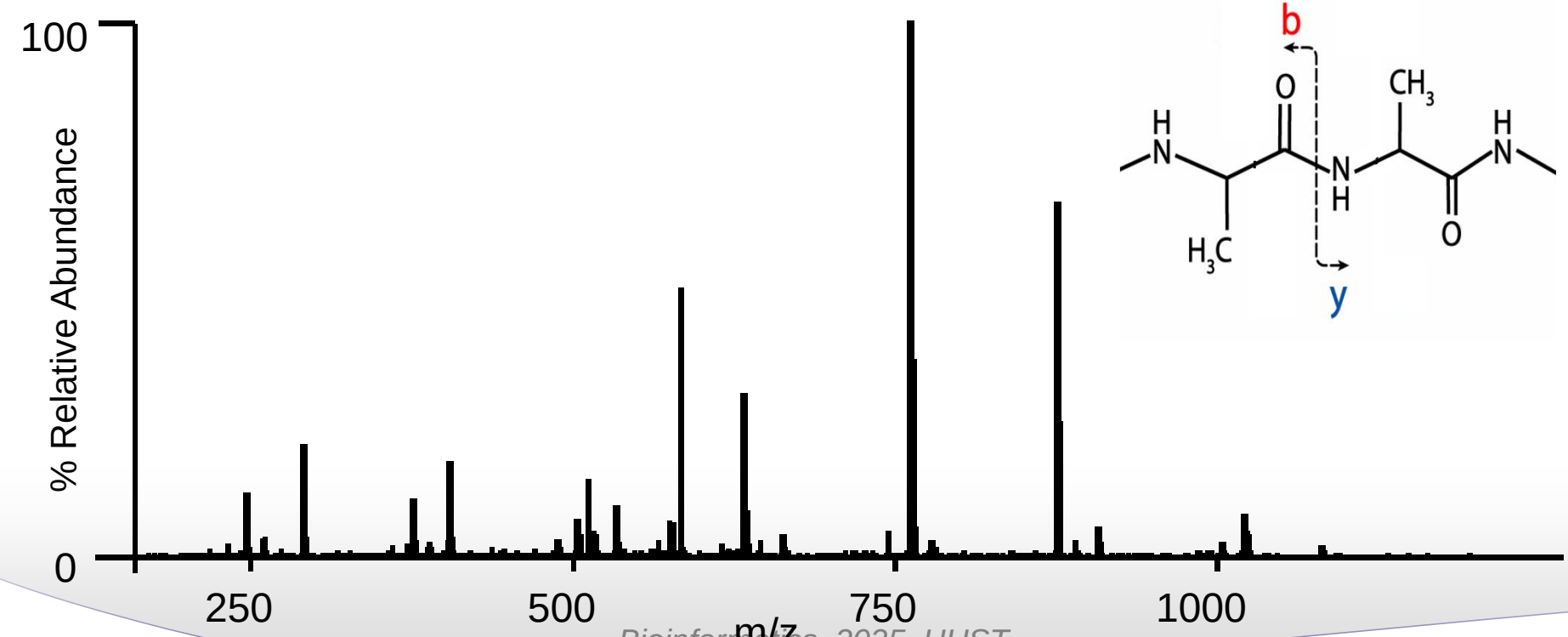
手工验证

S G F L E E D E L K



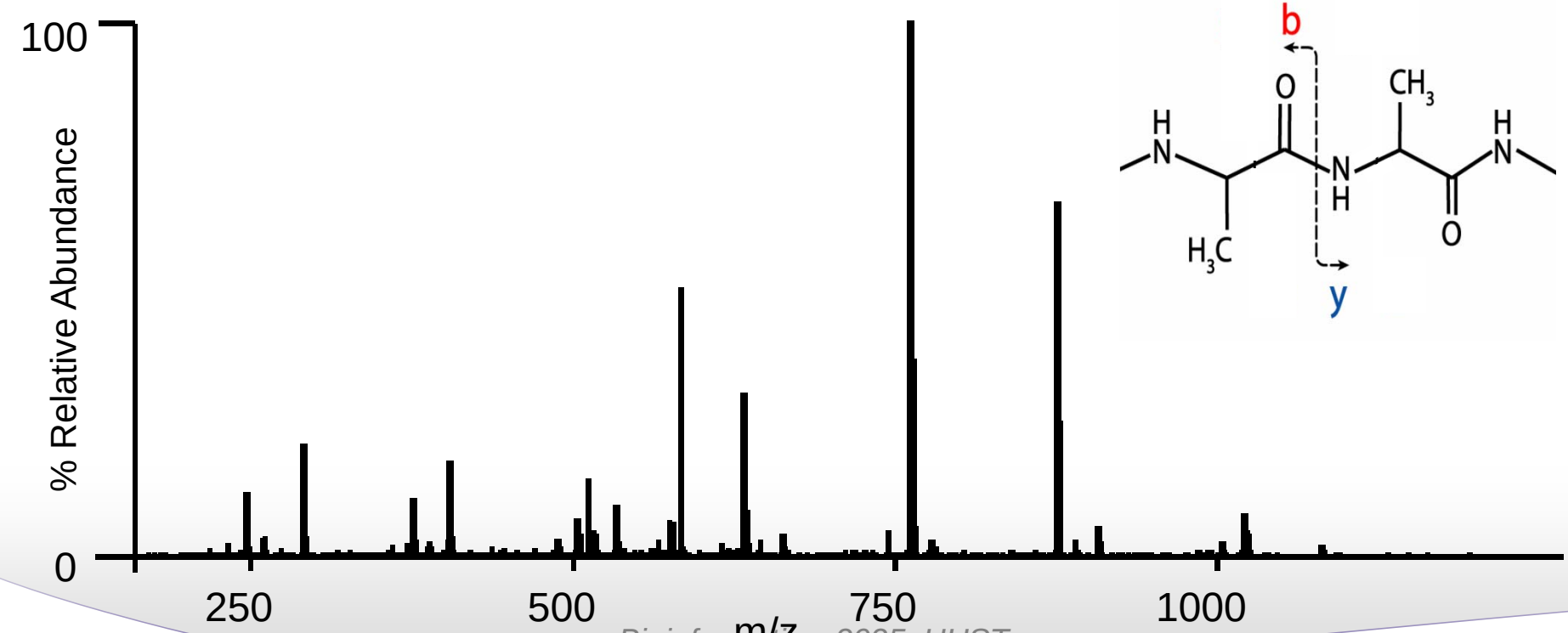
手工验证

S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions



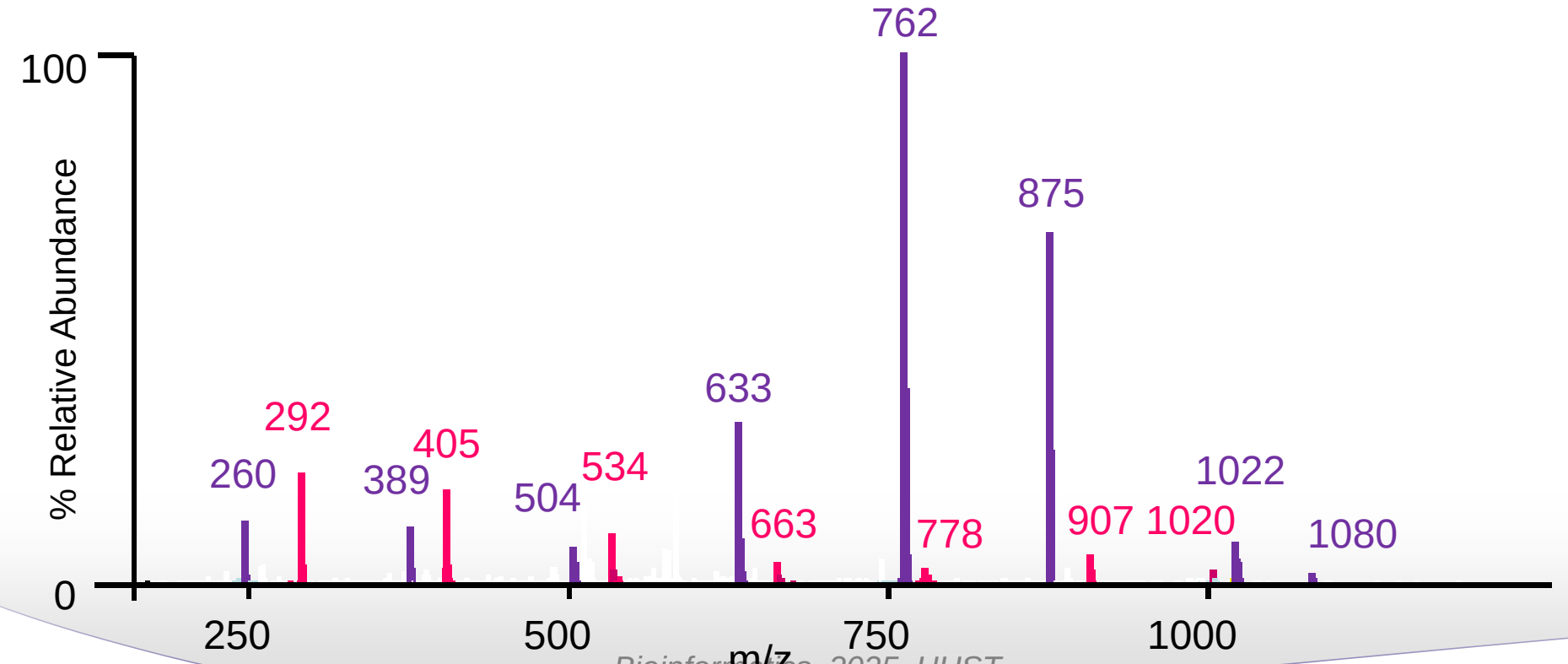
手工验证

S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions



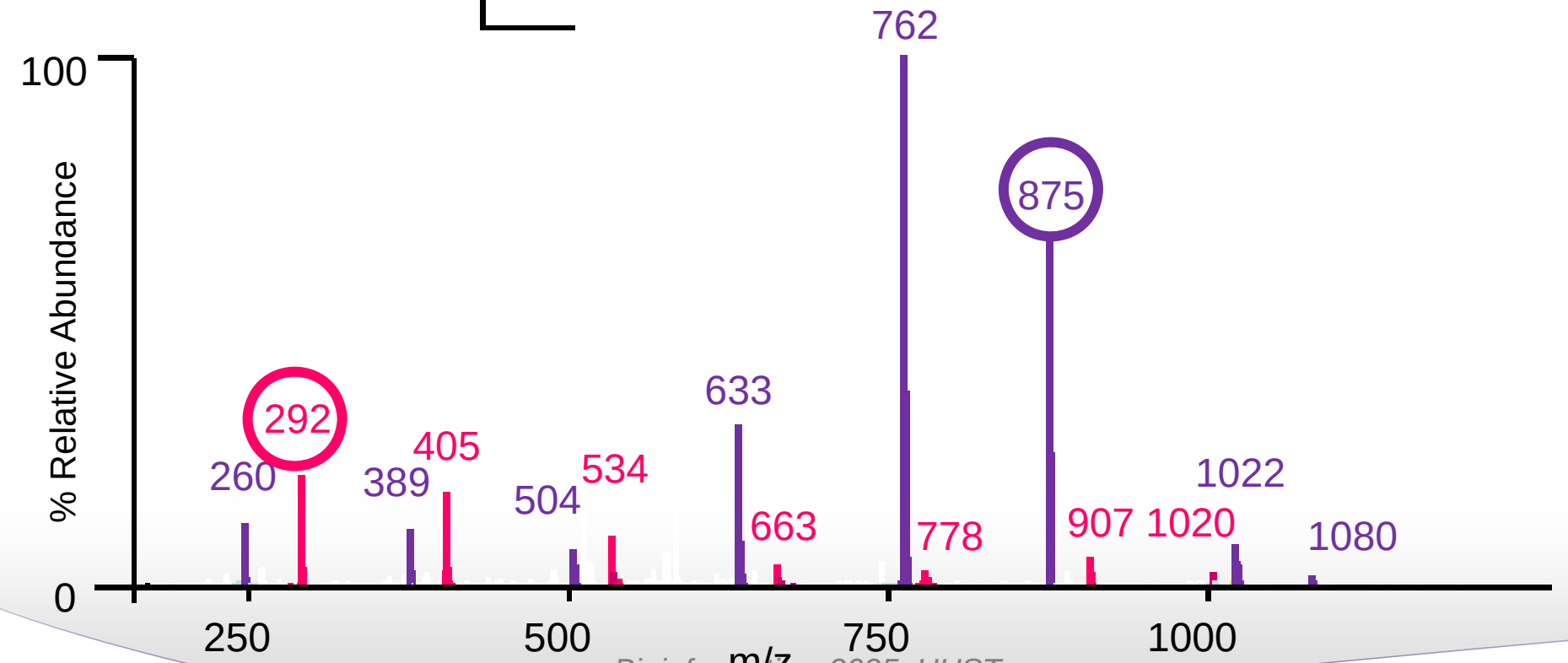
手工验证

S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions

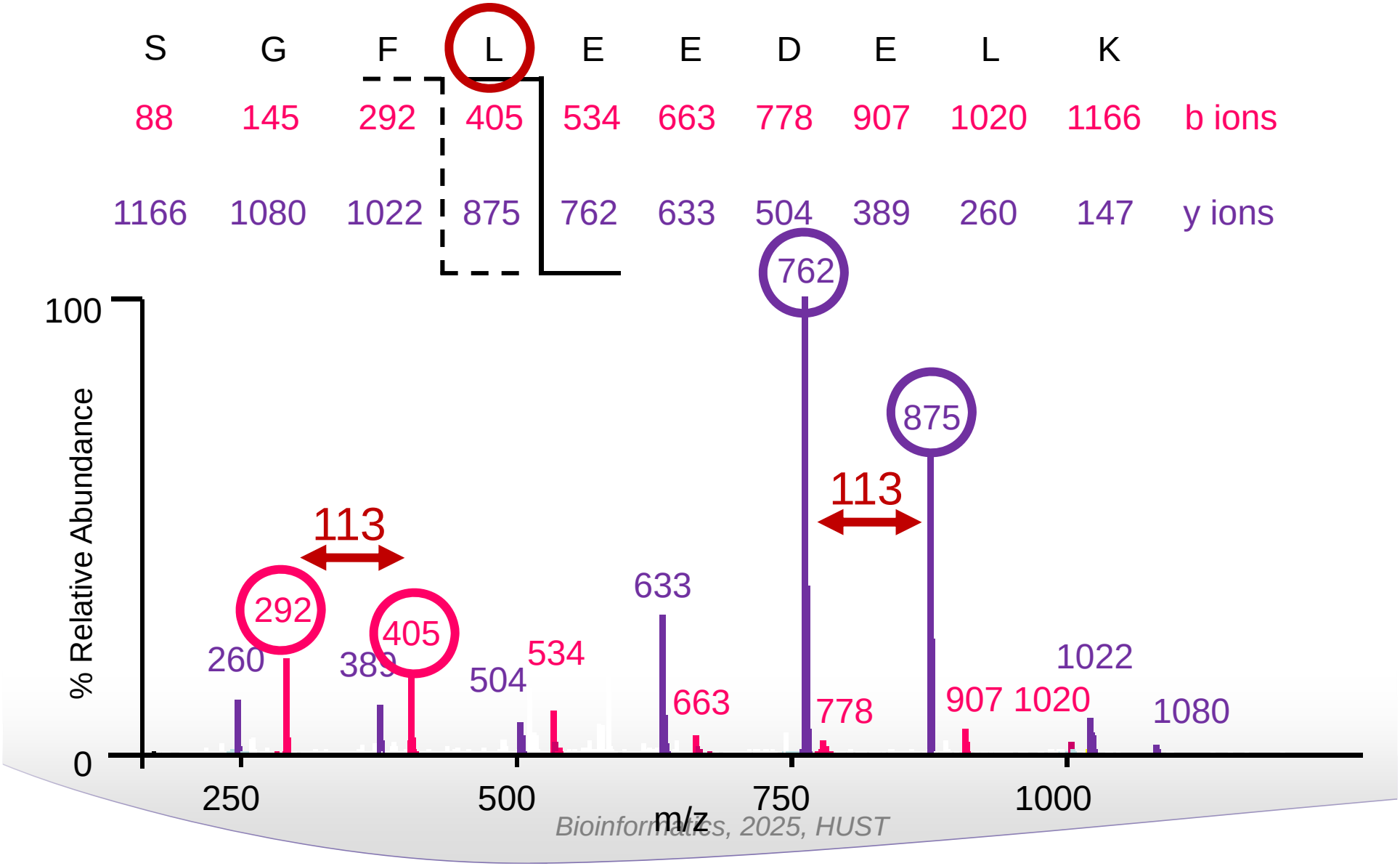


手工验证

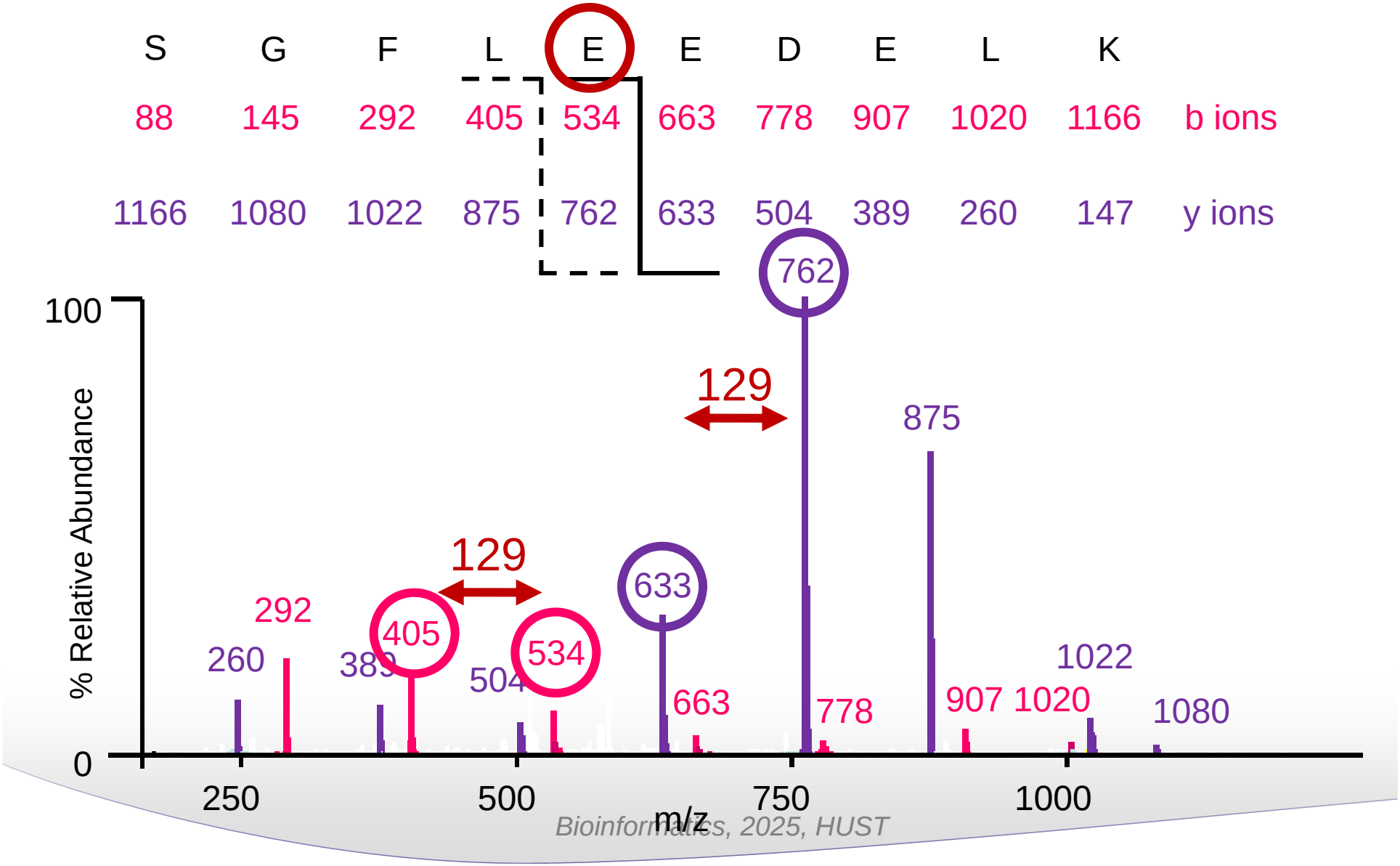
S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions



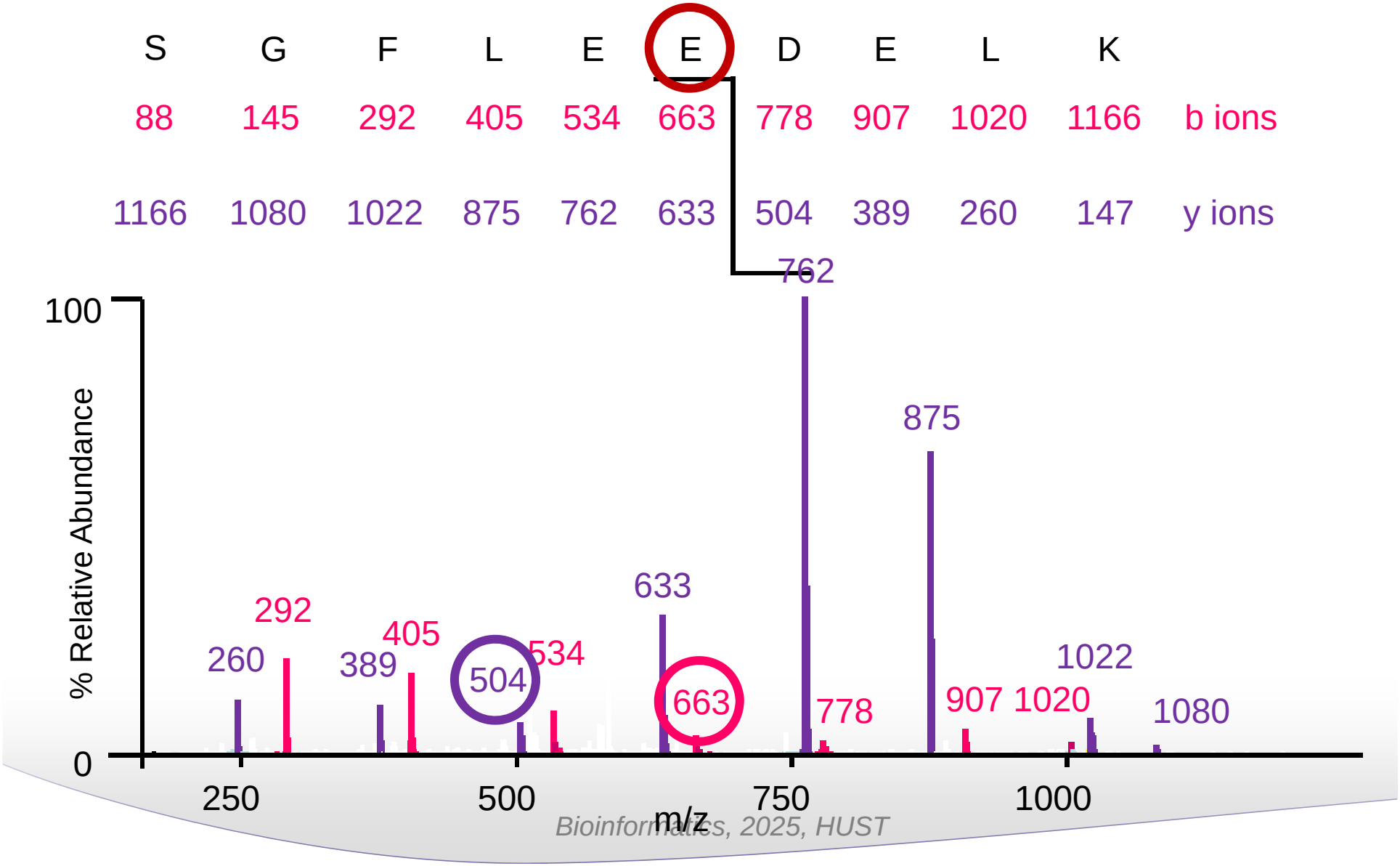
手工验证


S G F L E E D E L K
88 145 292 405 534 663 778 907 1020 1166 b ions
1166 1080 1022 875 762 633 504 389 260 147 y ions
762
875
113
113
292
405
260
389
504
534
633
663
778
907
1020
1022
1080
100
0
% Relative Abundance
250
500
750
1000
m/z

手工验证

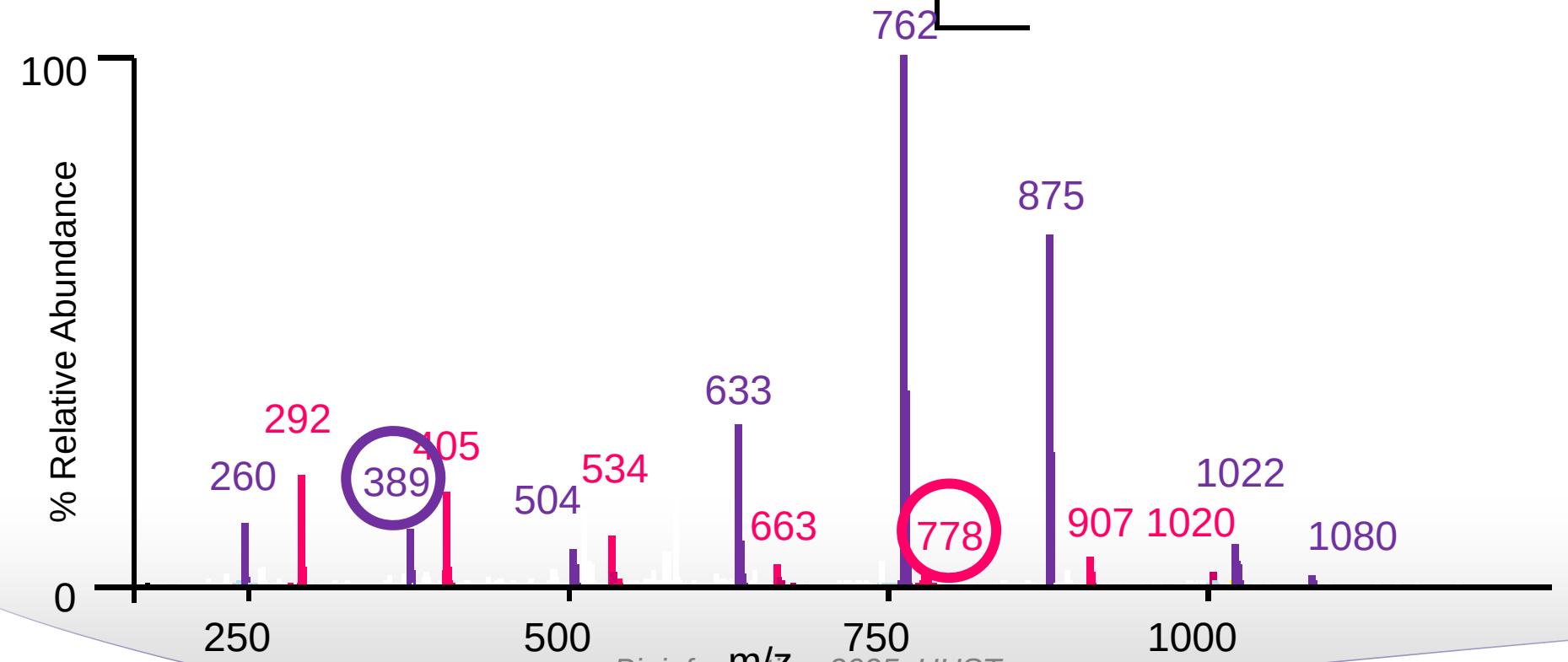

S G F L E E D E L K
88 145 292 405 534 663 778 907 1020 1166 b ions
1166 1080 1022 875 762 633 504 389 260 147 y ions
762
129
875
129
633
292
405
534
260
389
504
663
778
907
1020
1022
1080
100
% Relative Abundance
0
250
500
750
1000
m/z

手工验证


S G F L E E D E L K
88 145 292 405 534 663 778 907 1020 1166 b ions
1166 1080 1022 875 762 633 504 389 260 147 y ions
762
875
633
292
405
260
389
504
534
663
778
907
1020
1022
1080
100
0
% Relative Abundance
250
500
750
1000
m/z

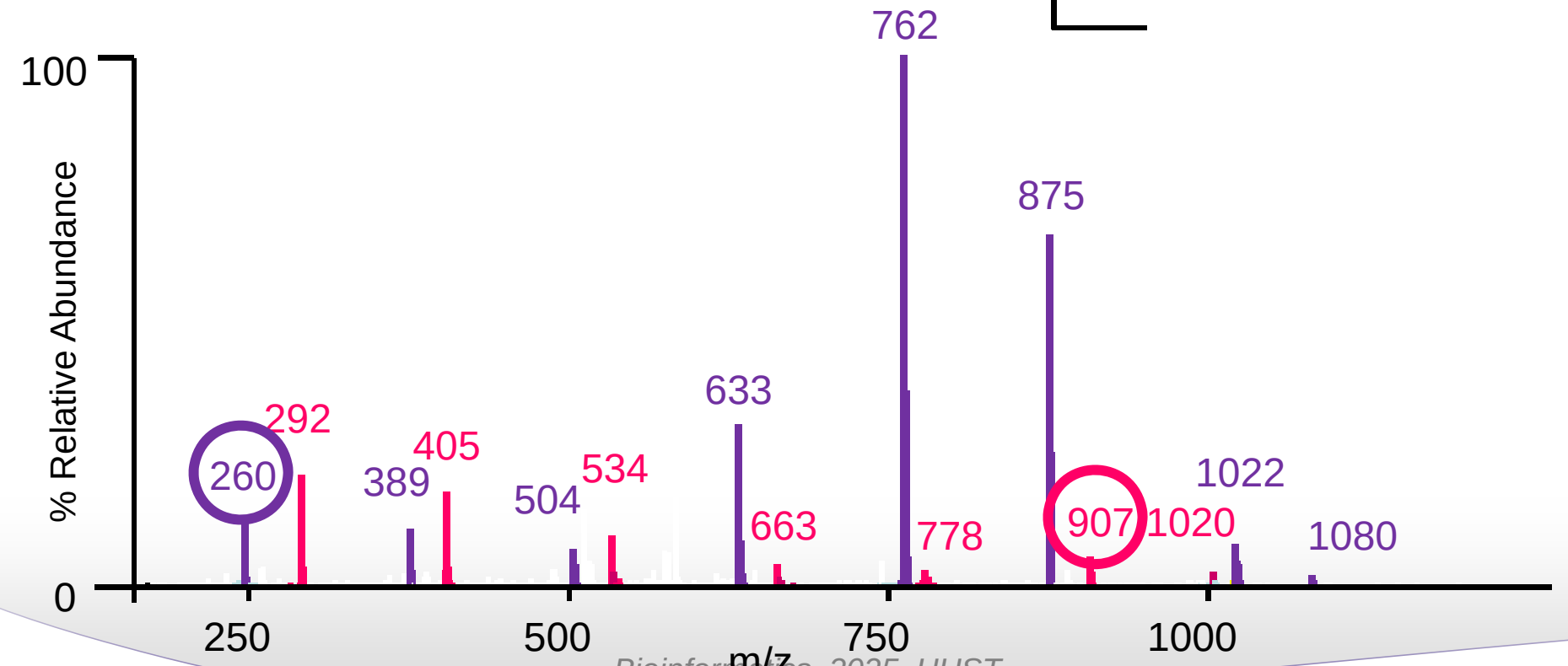
手工验证

S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions

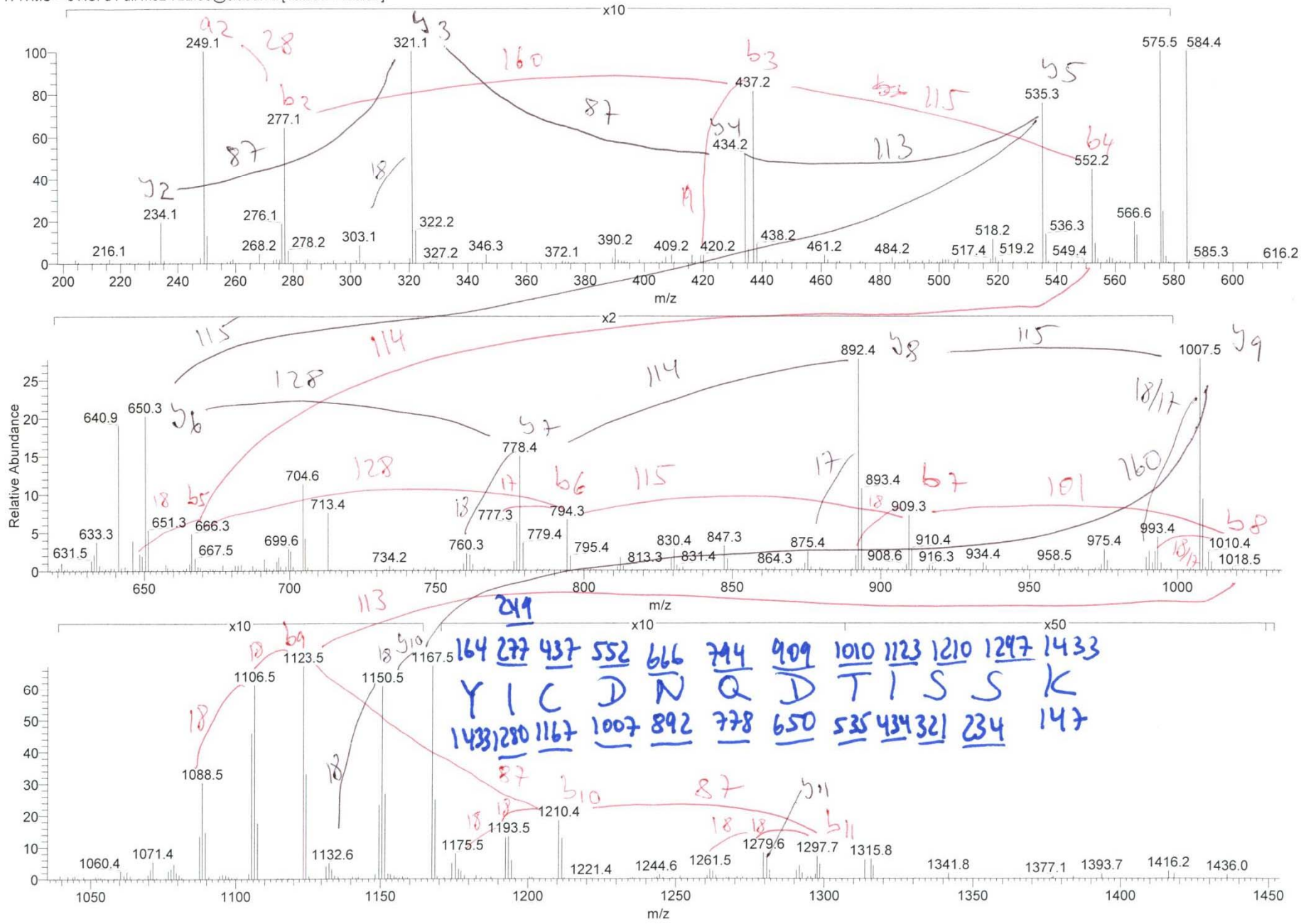


手工验证

S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions



E:\for lecture\030713_jrc_std4
30 fmole AV, BSA, AIIP
3/8/2013 11:09:33 AM
030713_jrc_std4 #1725 RT: 18.01 AV: 1 NL: 7.45E4
T: ITMS + c NSI d Full ms2 722.33@cid35.00 [185.00-1455.00]



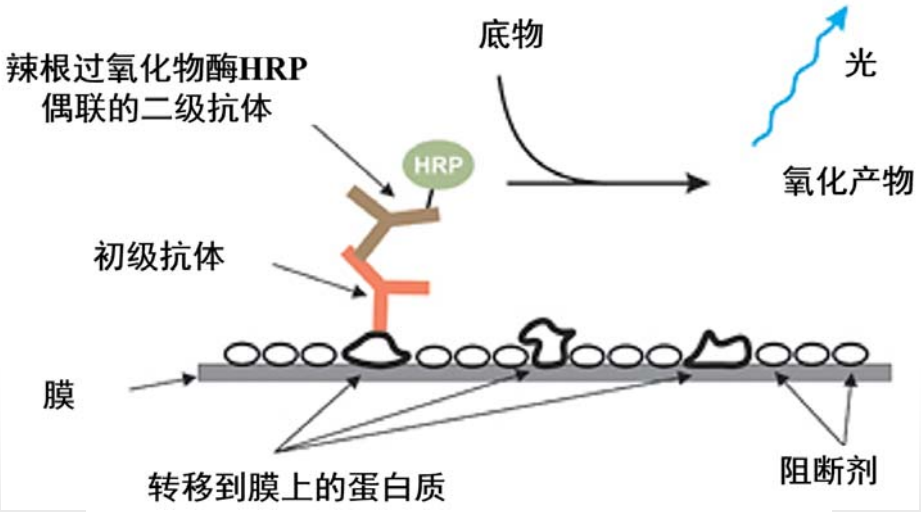
表达蛋白质组学

- 定量蛋白质组学
  - 检测生物样本中蛋白质的表达水平或丰度
- 基于图像的定量技术
  - 二维凝胶电泳、蛋白质芯片、影像质谱流式术IMC
  - 将表达信息转化为图像信号，再利用图像处理的方法进行定量分析
- 基于标记的定量技术
  - 应用于ESI-LTQ-Orbitrap等质谱仪的蛋白质组定量中
  - 无标记定量：根据MS/MS质量峰面积确定蛋白质丰度
  - 标记定量：对蛋白质或肽段进行同位素化学标记或代谢标记

蛋白质芯片

❒ 蛋白质微阵列

- 将已知蛋白质、抗体或抗原点印并固定在普通玻璃载体、多孔凝胶覆盖及微孔芯片介质上
- 由高密度蛋白质或多肽分子组成的微阵列
- 利用荧光物质、酶或化学发光物质标记的生物分子与芯片上探针反应

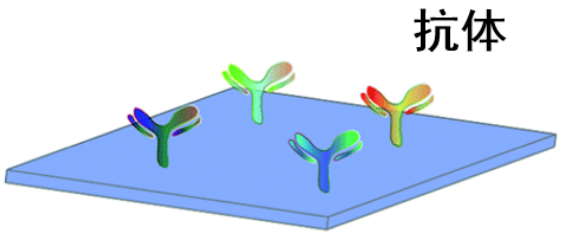


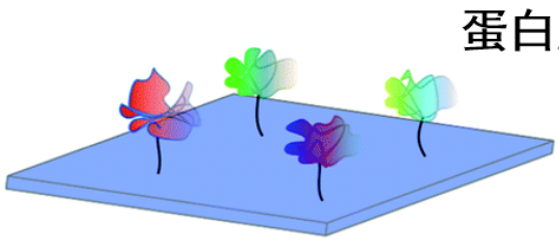
基于抗体的蛋白质芯片

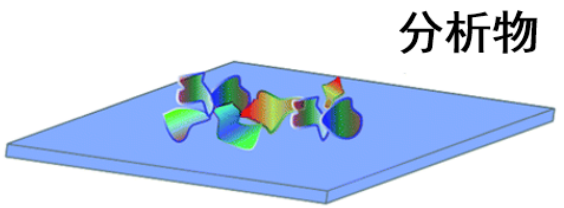
蛋白质芯片的类型

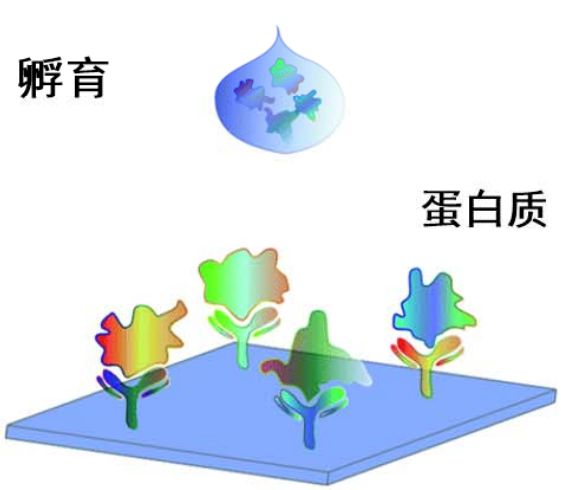
- 分析型蛋白质芯片
  - 密度相对较低，用于生物分子的大量、快速检测
  - 抗体芯片、抗原芯片
- 功能型蛋白质芯片
  - 高密度芯片，载体固定天然或融合蛋白质
  - 鉴定相互作用的蛋白质及生物分子，并确定丰度
- 反向蛋白质芯片
  - 将生物样品如组织或细胞裂解物点样在芯片载体上
  - 用蛋白质抗体间接地检测蛋白质的存在及丰度

蛋白质芯片的类型

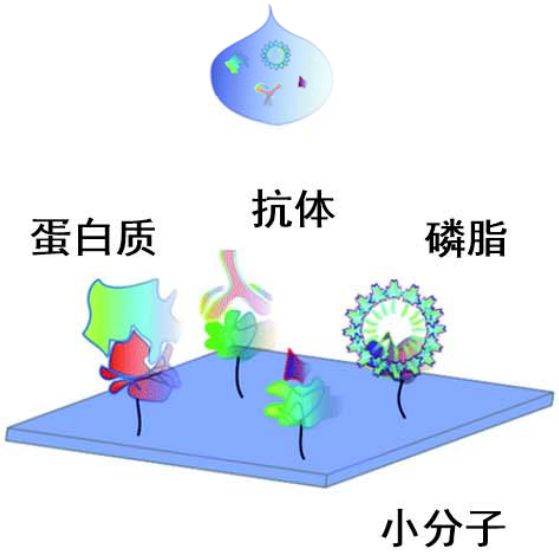




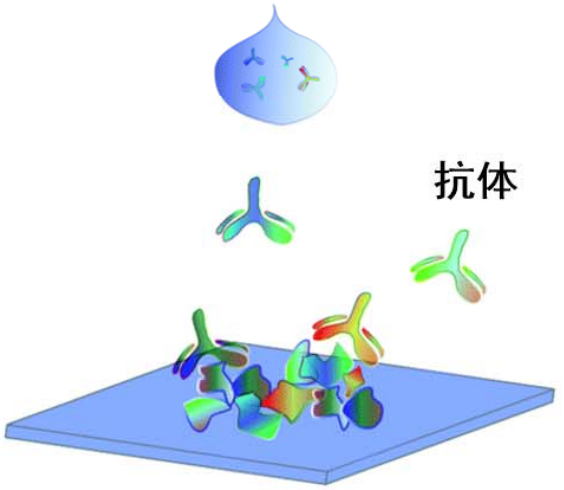




分析型（抗体芯片）



功能型



反向

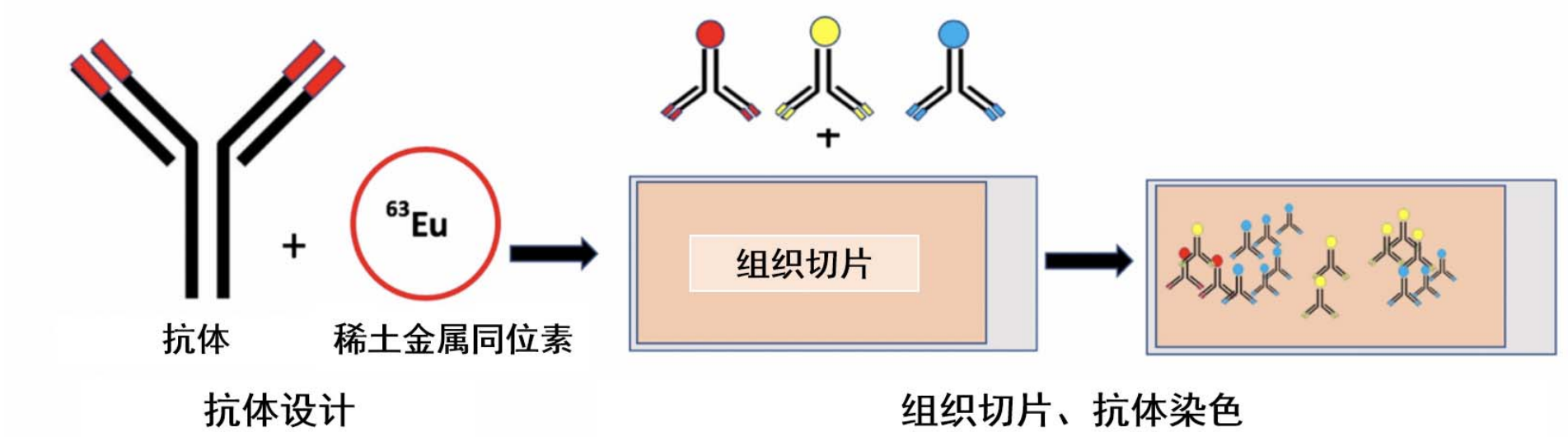
蛋白质芯片的优点

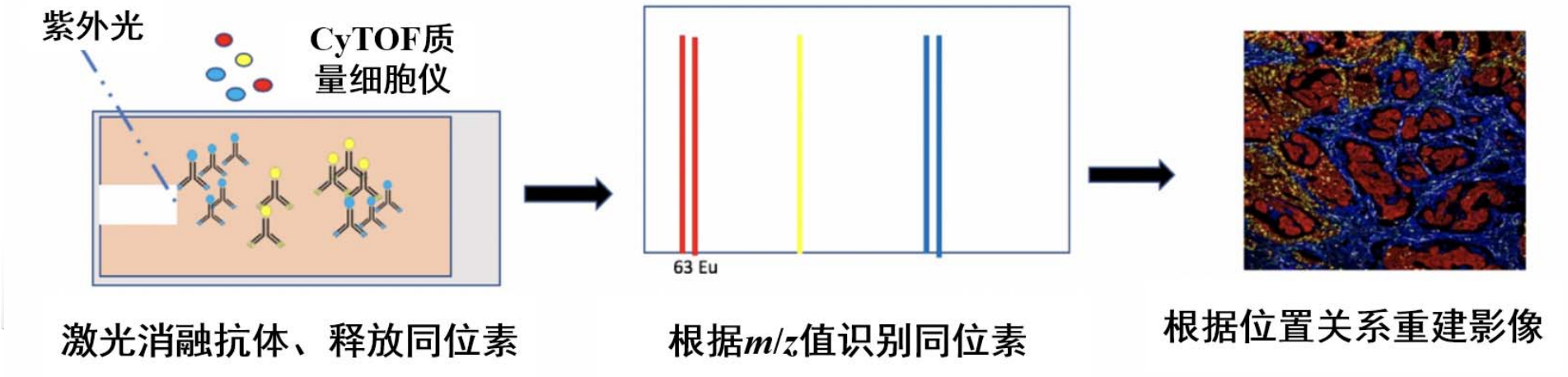
- 特异性较强，芯片检测的准确性由抗原与抗体之间、蛋白质与配体之间特异性结合决定
- 灵敏度较高，可检测出样品中微量蛋白质（纳克级）的存在
- 通量较高，单次实验可以同时检测上千种目标蛋白质
- 重复性较好，不同批次实验间的差异较小
- 应用性较强，样品的前处理比较简单，只需对少量实际样本进行沉降分离和标记后，即可加于芯片上进行分析和检测
- 适用范围较广，适用于包括组织、细胞及体液等多种生物样品

影像质谱流式术（IMC）

- 流式细胞术
  - 使用荧光标记的抗体或探针标记单个细胞
  - 主要用于细胞悬浮液的分析
- 影像质谱流式术
  - 质量细胞术（mass cytometry）的延伸和拓展
  - 将抗体与金属螯合的高分子聚合物偶联
  - 标记整张组织切片
  - 可对石蜡包埋的组织切片进行分析
  - 可用于患者队列的回顾性研究

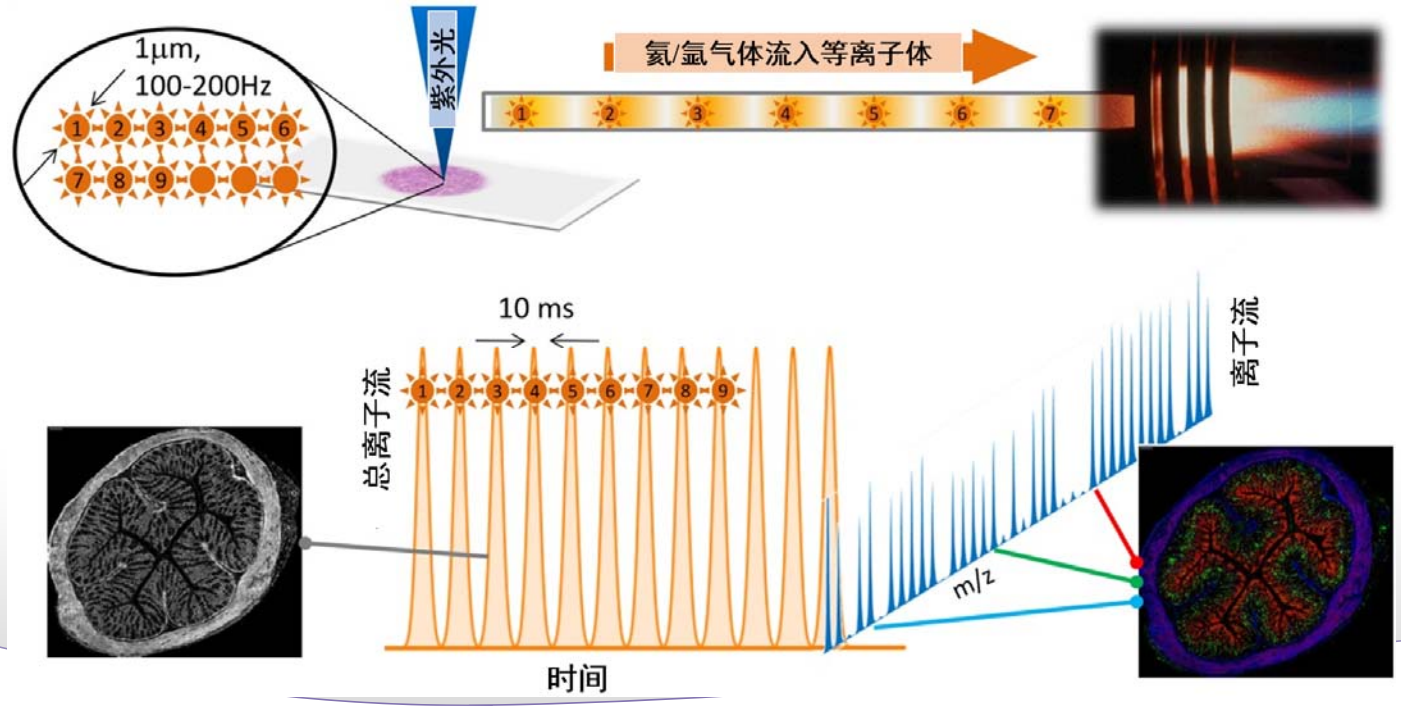
影像质谱流式术的实验流程





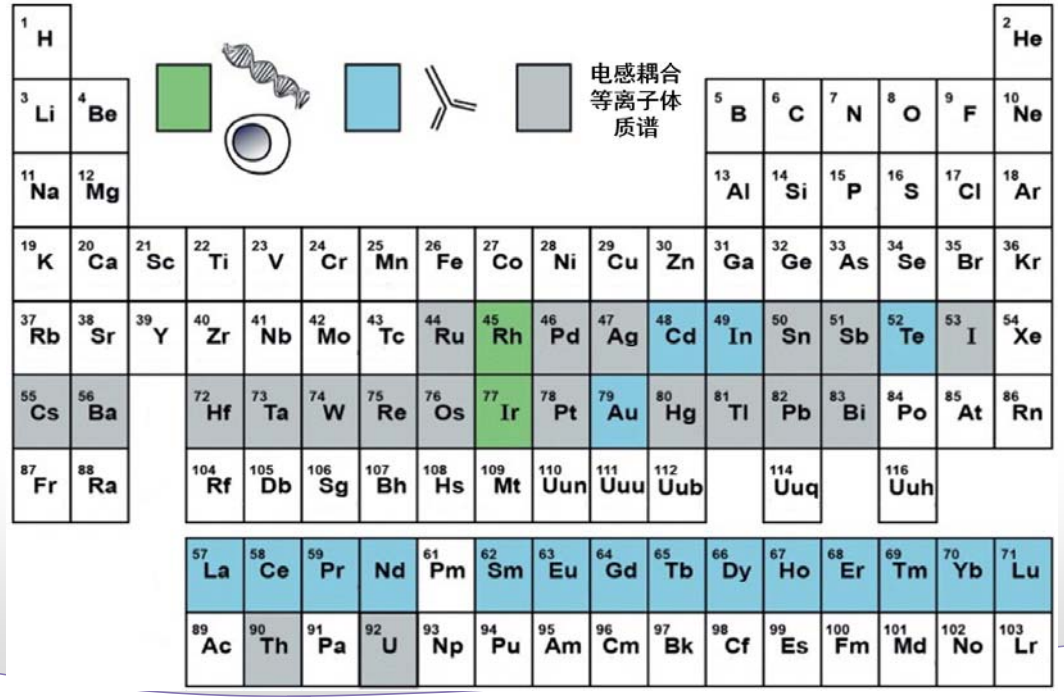
影像质谱流式术的原理

- 电感耦合等离子体
- 随惰性气体的流动进入飞行时间（TOF）质谱进行检测
- 获得每个小区域的金属同位素的丰度
- 得到切片上目标蛋白质丰度的分布
- 根据多离子束成像技术重建整张切片的图像

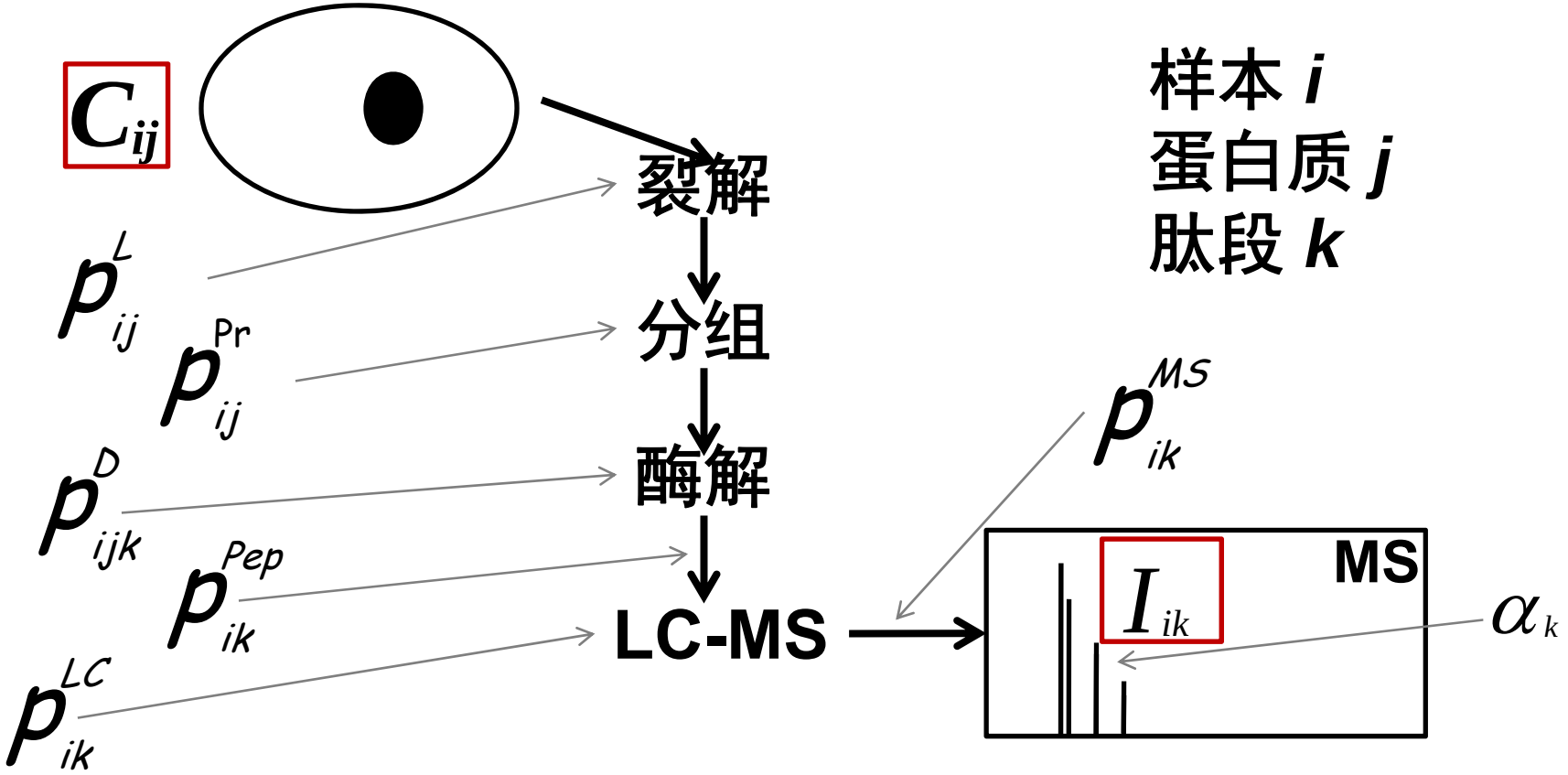


影像质谱流式术的优缺点

- 40种金属同位素
  - 37种镧系元素，3种非镧系元素铋、金和铂
  - 单次实验可以标记抗体数量的上限是40个
  - 优点：可以开展单细胞蛋白质组分析
  - 缺点：单次实验中可分析的蛋白质数量较少，通量较低



基于标记的蛋白质定量



$$I_{ik} = \alpha_k \sum_j \left[C_{ij}\, p_{ij}^{L}\, p_{ij}^{Pr}\, p_{ijk}^{D} \right] p_{ik}^{Pep}\; p_{ik}^{LC}\; p_{ik}^{MS}$$

$$C_{ij}^{k} = \frac{I_{ik}}{\alpha_k\, p_{ij}^{L}\, p_{ij}^{Pr}\, p_{ijk}^{D}\, p_{ik}^{Pep}\, p_{ik}^{LC}\, p_{ik}^{MS}}$$

非标记定量

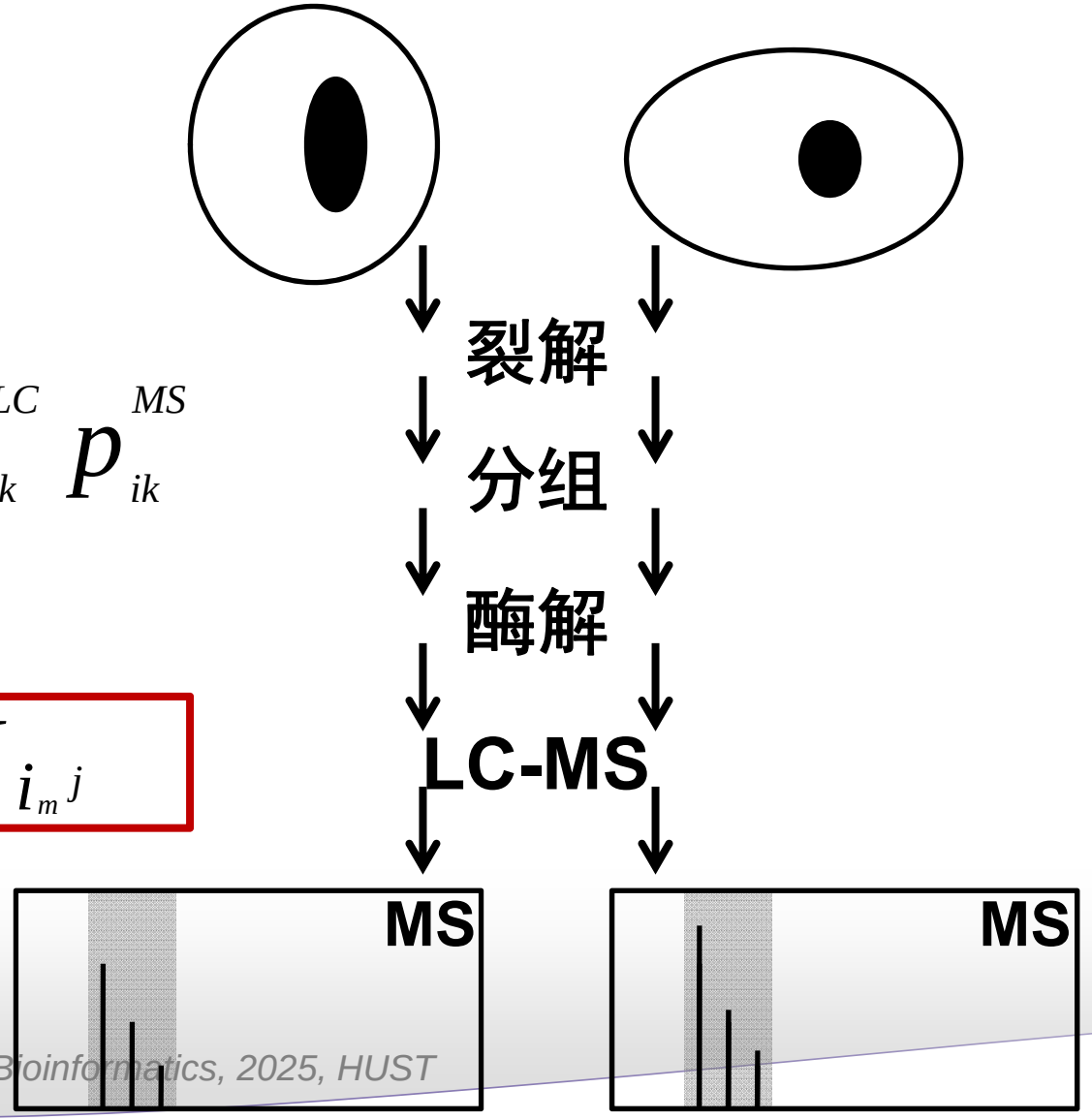
样本 *i*
蛋白质 *j*
肽段 *k*

基本假设:

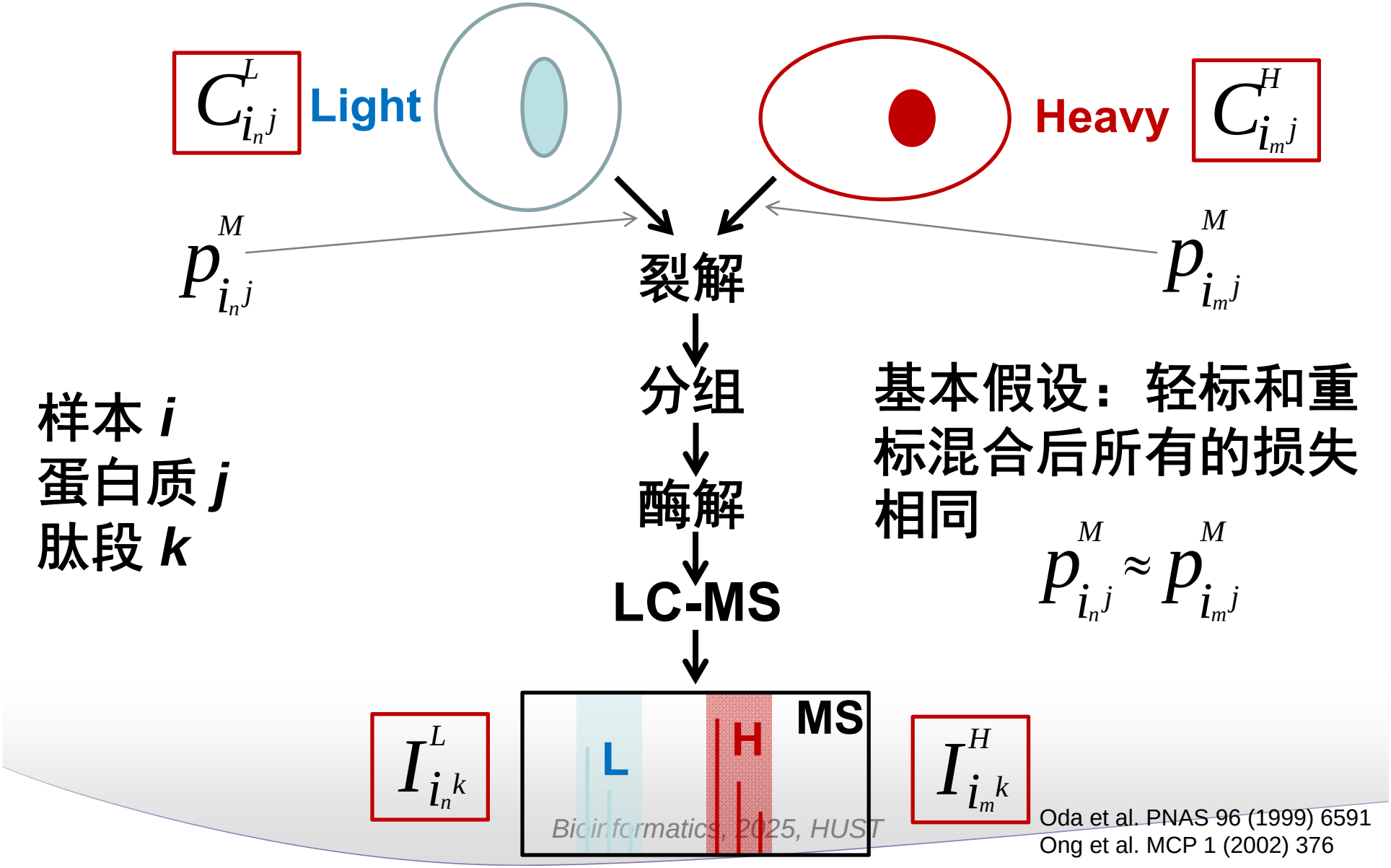
$$\alpha_k \; p_{ij}^{L} \; p_{ij}^{\Pr} \; p_{ijk}^{D} \; p_{ik}^{Pep} \; p_{ik}^{LC} \; p_{ik}^{MS}$$

在所有样本中恒定

$$C_{i_n j} / C_{i_m j} = I_{i_n j} / I_{i_m j}$$



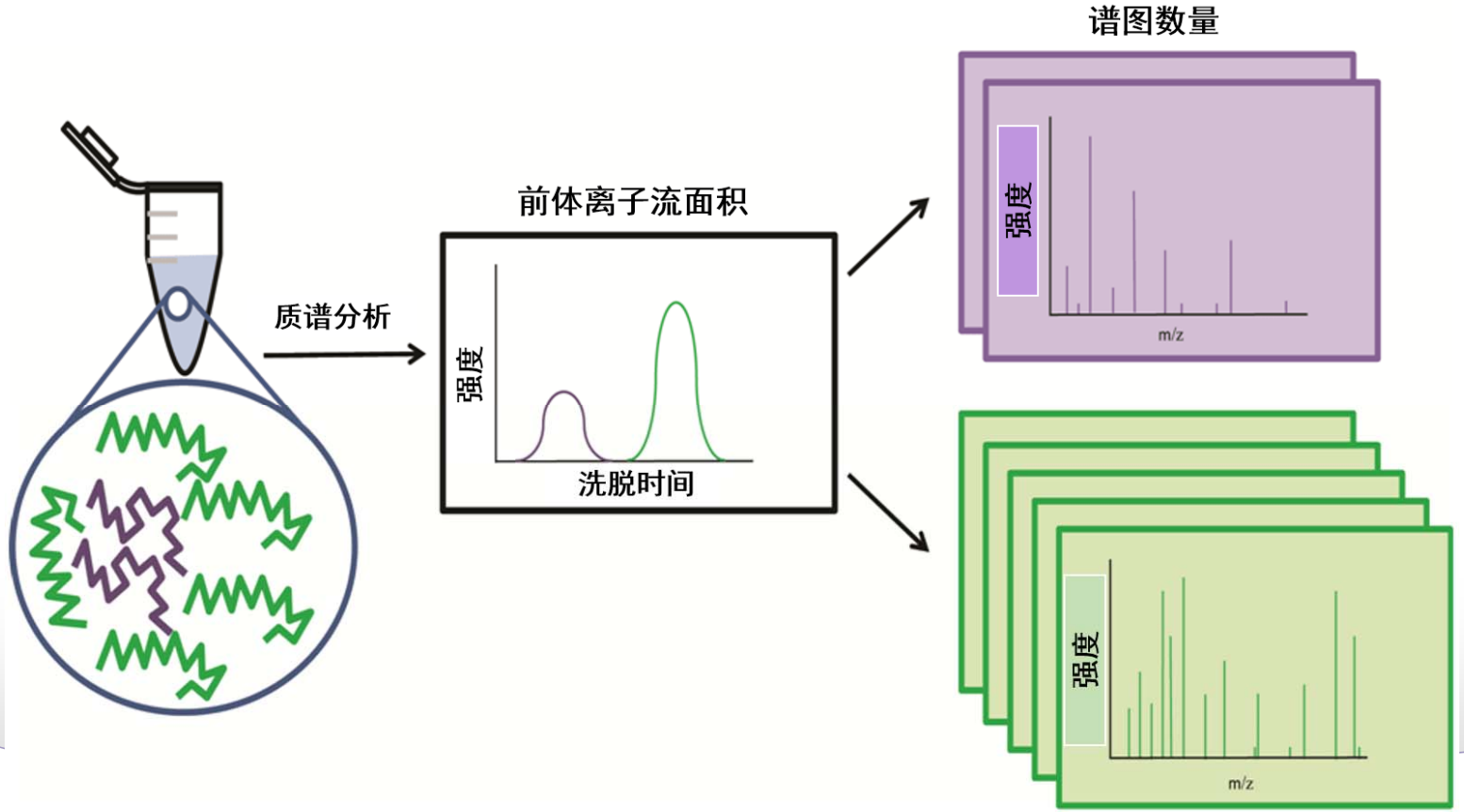
标记定量



Oda et al. PNAS 96 (1999) 6591
Ong et al. MCP 1 (2002) 376

无标记定量

- 两类定量算法
  - 根据前体离子流面积（肽段丰度）
  - 根据能够回贴到蛋白质序列的谱图数量



无标记定量算法

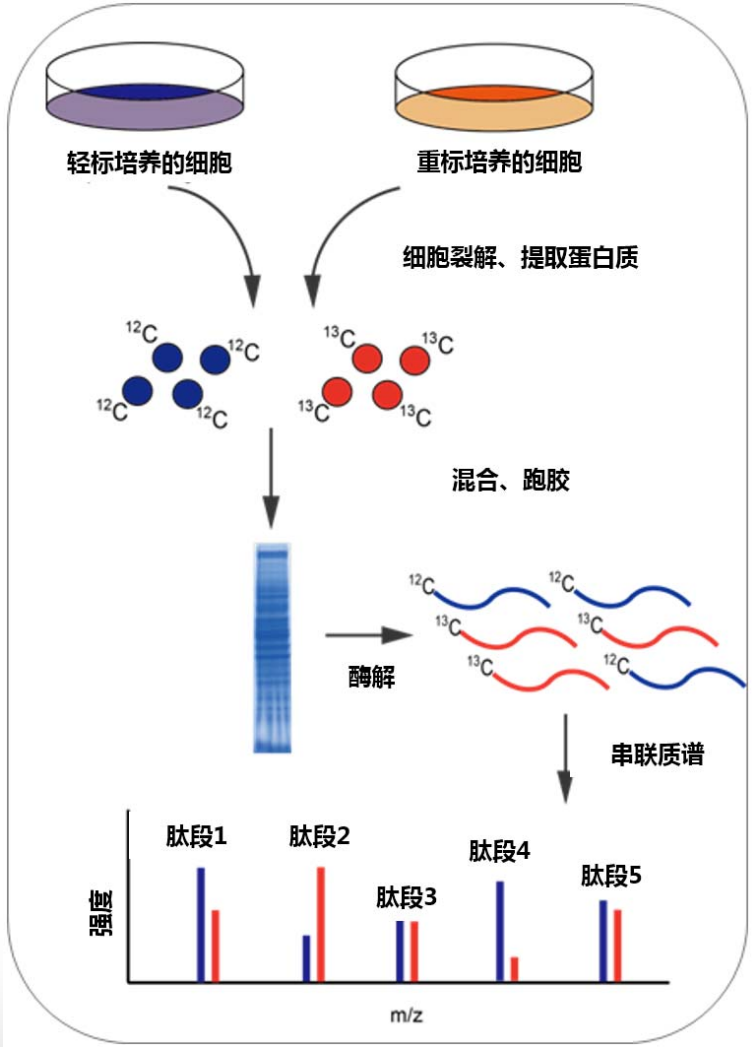
- 根据肽段丰度推测蛋白质丰度
  - “前三”（TOP3）法：将质谱检测的、质量峰面积最大的三条肽段的强度值累加再计算平均值，作为蛋白质的表达丰度值
  - “基于强度的绝对定量”（intensity-based absolute quantification，iBAQ）法：将蛋白质上所有质谱检测到的肽段的强度值累加，再除以该蛋白质理论酶解之后的肽段数量，计算结果即为蛋白质的表达丰度

无标记定量技术的优缺点

- 优点
  - 实验操作简便
  - 蛋白质标记存在效率问题，没有标记就不存在损失
  - 无额外的试剂或材料，成本更低
  - 蛋白质定量较为准确
- 缺点
  - 质谱鉴定的肽段对于蛋白质整体的覆盖度较低
  - 不同人操作质谱仪的习惯不同，参数设置也不完全相同，因此不同样品的检测中存在较强的批次效应（batch effect）

代谢标记定量

- 细胞培养稳定同位素标记（SILAC）
  - 在细胞培养液中加入同位素标记的特定氨基酸
  - 溶液中同位素标记的氨基酸含量，比细胞内蛋白质中不含标记的同种氨基酸浓度高
  - 在细胞正常的新陈代谢活动中，蛋白质中不含标记的氨基酸会被替换成同位素标记的同种氨基酸
  - 将不同标记的样品等量混合、裂解、酶解



SILAC实验流程

常用的同位素标记的氨基酸

- 赖氨酸（K）：分子式为$C_6H_{12}ON_2$，残基分子量为128.1
- 精氨酸（R）：分子式为$C_6H_{12}ON_4$，残基分子量为156.1
- Trypsin酶解之后的肽段，C末端是赖氨酸或精氨酸，很容易被质谱检测

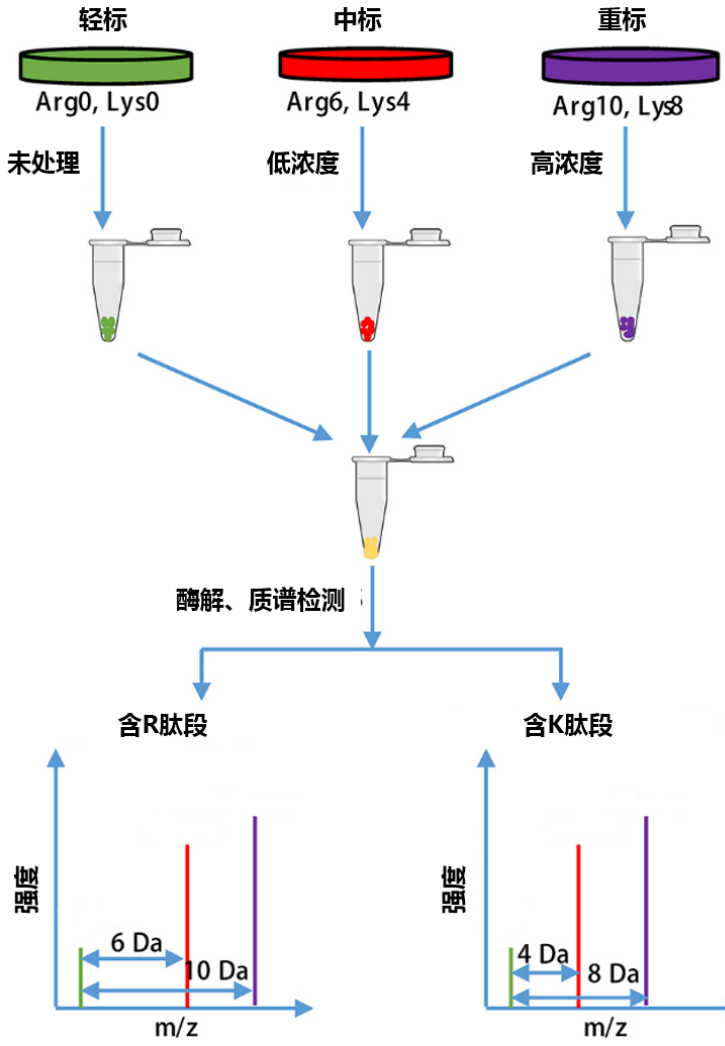
K	Lys	$C_6H_{12}ON_2$	128.095	128.174
R	Arg	$C_6H_{12}ON_4$	156.101	156.188

常用的SILAC标记组合

- 单次实验中，能够同时标记样品的上限是3个
  - 未经同位素标记的$^{12}C_6{}^{14}N_2$-赖氨酸和$^{12}C_6{}^{14}N_4$-精氨酸组合，记为“K0、R0”（“轻标”）
  - 部分碳同位素标记的$^{13}C_4{}^{14}N_2$-赖氨酸和所有碳同位素标记的$^{13}C_6{}^{14}N_4$-精氨酸，其分子量要比轻标的同类氨基酸分别高4和6个道尔顿，记为“K4、R6”（“中标”）
  - 全部碳、氮同位素标记的$^{13}C_6{}^{15}N_2$-赖氨酸和$^{13}C_6{}^{15}N_4$-精氨酸，记为“K8、R10”（“重标”）

SILAC的应用场景

- 经典的“处理”和“对照”研究
  - “重标”：高浓度药物处理
  - “中标”：低浓度药物处理
  - “轻标”：未处理



利用SILAC研究药物的分子效应

SILAC技术的优缺点

- 优点
  - 实验可重复性高
  - 同位素标记的效率可达到95-98%
  - 标记在样本处理之前，消除了样本处理过程中可能造成的偏差
  - 相对定量更精准、方差小
  - 处理细胞样品比较方便
- 缺点
  - 单次实验能够标记的样品较少
  - 同位素标记的赖氨酸和精氨酸，实验成本更高
  - 组织样本难以被标记

同位素化学标记技术

- 等重同位素标签相对和绝对定量（iTRAQ）技术
  - AB SCIEX公司研发
- 串联质量标记（Tandem mass tag, TMT）
  - Thermo Fisher公司研发

iTRAQ

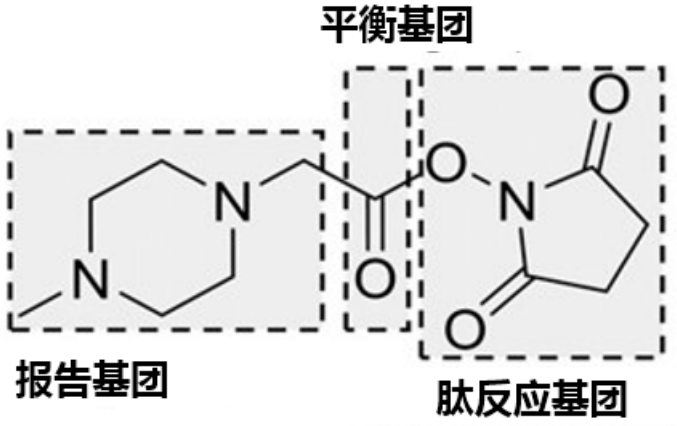
TMT

同位素化学标记技术的原理

- iTRAQ和TMT标签：包括报告基团、平衡基团和肽反应基团
- 将酶解后肽段的N端连接上不同分子量的标签
- 在报告基团和平衡基团之间预留了HCD的断裂点
- 在MS/MS检测中，离子通过第二重质量分析器的时候发生HCD碎裂
- 第三重质量分析器根据碎裂之后的*b*离子和*y*离子来推断肽段的组成
- 再根据报告基团的质量峰来推断肽段的丰度

4-plex的iTRAQ标签

- 4-plex的iTRAQ标签
  - 4种报告基团：分子量分别为114、115、116和117 Da
  - 4种平衡基团：分子量分别为31、30、29和28 Da，使得报告基团和平衡基团的总分子量为145 Da
  - 肽反应基团：与肽段N端的氨基反应形成共价键连接
- 8-plex的报告基团分子量为114~121 Da



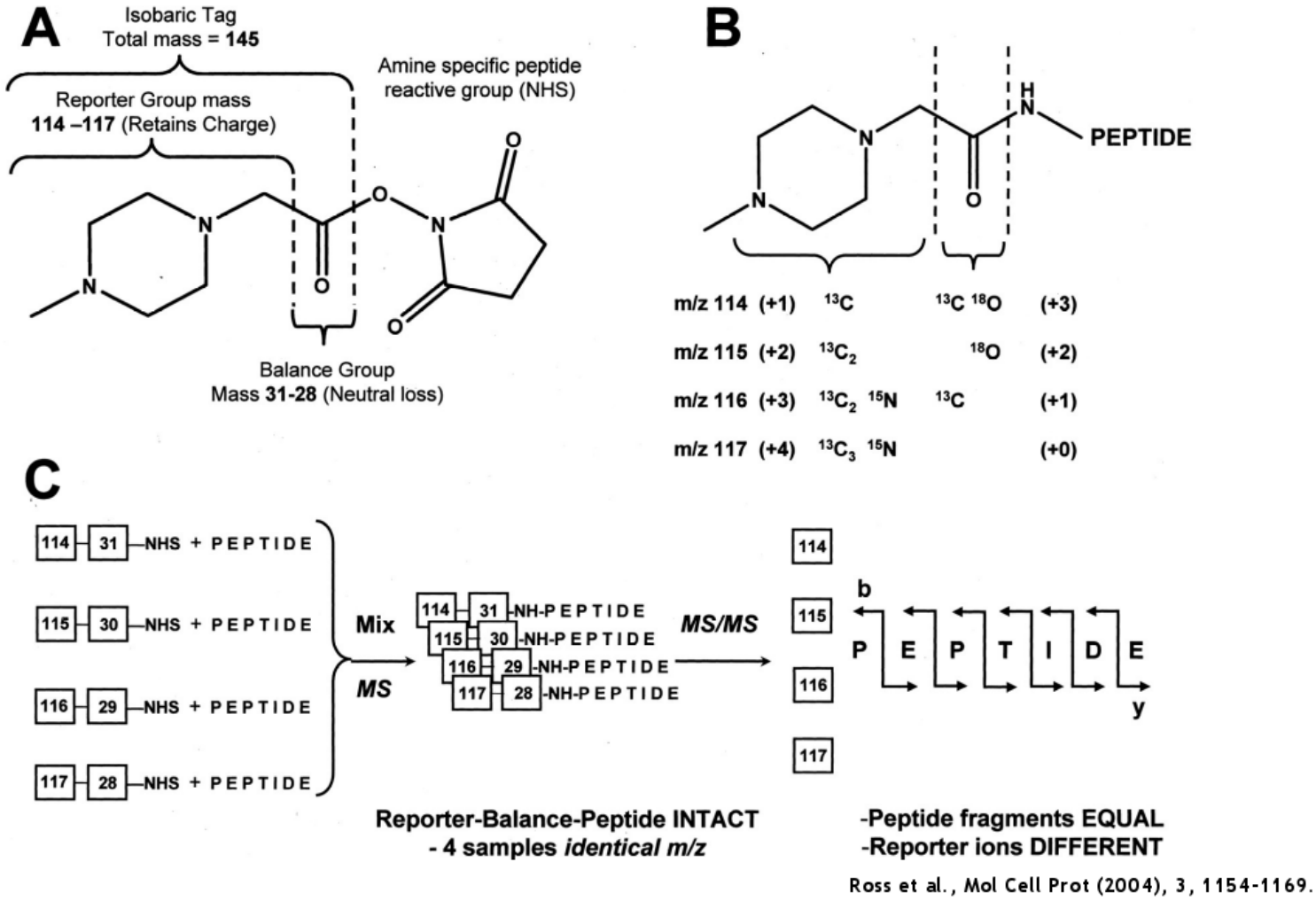
iTRAQ标签的组成

iTRAQ 4-plex标签

	报告基团	平衡基团
iTRAQ 114	$^{13}C_1$	$^{13}C_1{}^{18}O_1$
iTRAQ 115	$^{13}C_2$	$^{18}O_1$
iTRAQ 116	$^{13}C_2{}^{15}N_1$	$^{13}C_1$
iTRAQ 117	$^{13}C_2{}^{15}N_2$	

iTRAQ标签与肽段连接

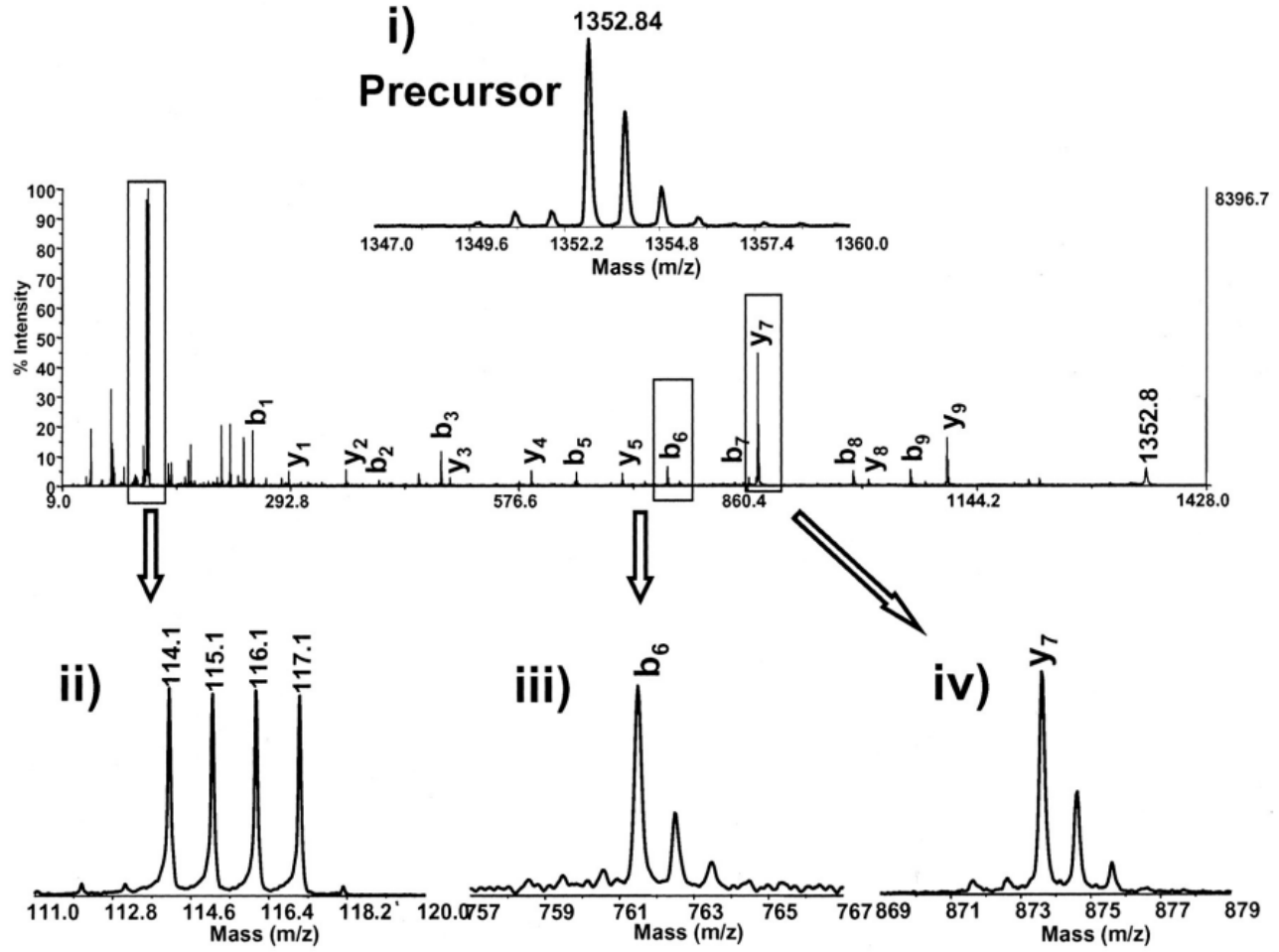
iTRAQ



Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

iTRAQ标记的谱图解析

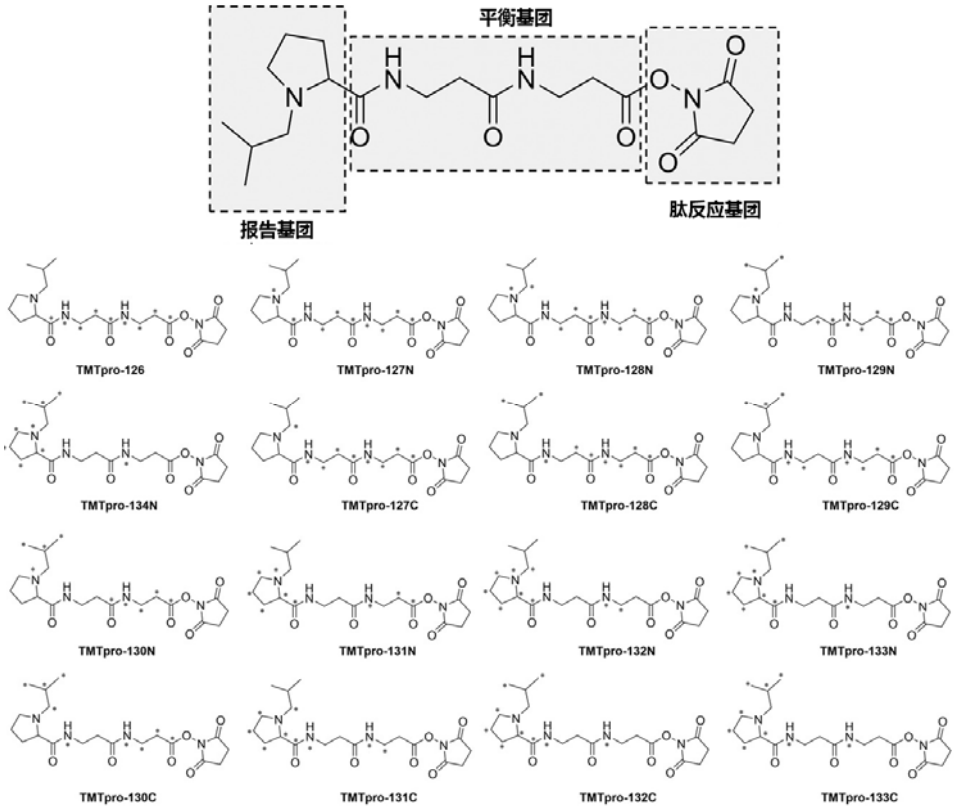
iTRAQ



Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

TMT的4种试剂

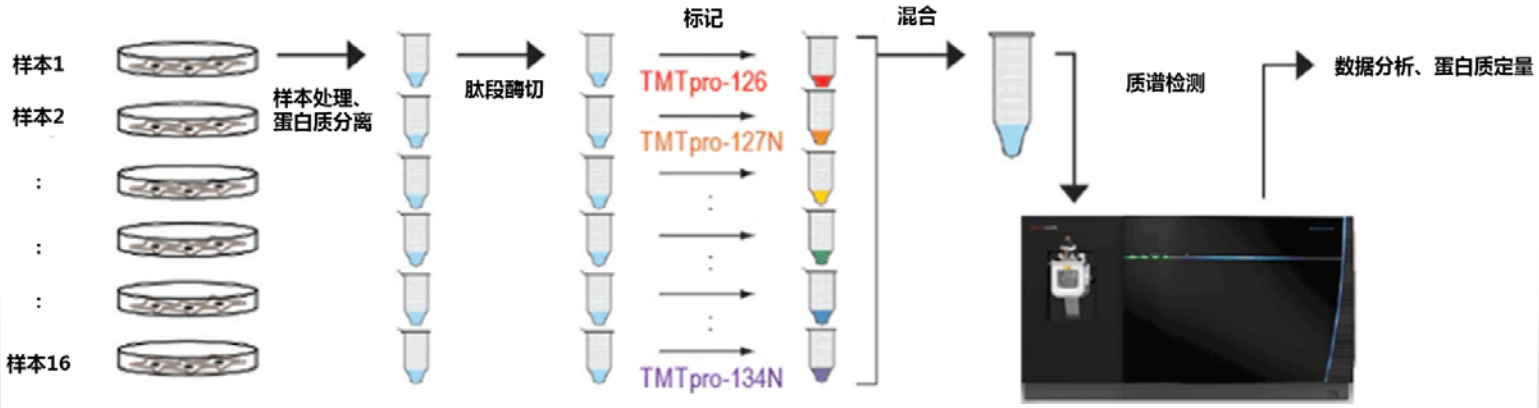
- TMT 6、10、11以及TMTpro 16-plex
  - 可分别同时标记6、10、11和16组样品
- 16-plex的TMT标签
  - 分子量为126-134 Da
  - 每隔0.5 Da一个标签
  - 质谱检测时，不同标签对应的质量峰靠得更近



TMTpro 16-plex标签

TMTpro 16-plex的应用场景

- 同时标记的样品更多，应用更广泛
- 主流标记蛋白质组学技术，大量临床样本分析
- 第一个标签：“混样”，每份样本中取等量少部分混合作为对照样品
- 剩余15个标签：标记不同的临床样本
- 数据分析时，用“混样”中蛋白质丰度对其他样本做归一化处理，即可消除实验的批次效应



利用TMTpro 16-plex标签开展定量蛋白质组分析

蛋白质组的绝对定量

- 无标记、代谢标记和同位素化学标记技术　蛋白质组的相对定量
- 绝对定量　在样品制备的时候加入已知质量的标准品蛋白质作为对照
- 蛋白质的绝对定量算法
  - “蛋白质组之尺”（proteomic ruler）　估算单个细胞中每种蛋白质的拷贝数量
  - 2014年，德国Matthias Mann提出

蛋白质组之尺的算法思想

- 单个细胞内所有蛋白质的质量为T_m，蛋白质P的质量P_m，单个细胞内P的数量P_c可由下述公式计算：

$$\frac{P_m}{T_m} = \frac{I_P}{I_T} \Rightarrow P_m = T_m \times \frac{I_P}{I_T} \Rightarrow P_c = \frac{N_A}{M} \times T_m \times \frac{I_P}{I_T}$$

- 其中，I_p是蛋白质P的质谱丰度，I_T是所有蛋白质丰度的加和，N_A是阿伏伽德罗常数（约为6.02×10^{23}），M是蛋白质的摩尔质量也就是分子量

蛋白质组之尺的算法思想

- 单个细胞内所有组蛋白（histone）的质量总和H_m，可根据下述公式计算

$$\frac{H_m}{T_m} = \frac{I_H}{I_T}$$

- I_H是所有组蛋白丰度的加和。在哺乳动物的细胞里，DNA缠绕在组蛋白8聚体复合物上形成核小体，且染色体DNA的质量D_m与组蛋白的质量H_m几乎相等，因此上述公式可转化为：

$$\frac{H_m}{T_m} = \frac{I_H}{I_T} => \frac{D_m}{T_m} = \frac{I_H}{I_T} => T_m = D_m \times \frac{I_T}{I_H}$$

$$=> P_c = \frac{N_A}{M} \times D_m \times \frac{I_T}{I_H} \times \frac{I_P}{I_T}$$

$$=> P_c = \frac{N_A}{M} \times D_m \times \frac{I_P}{I_H}$$

蛋白质组之尺的算法思想

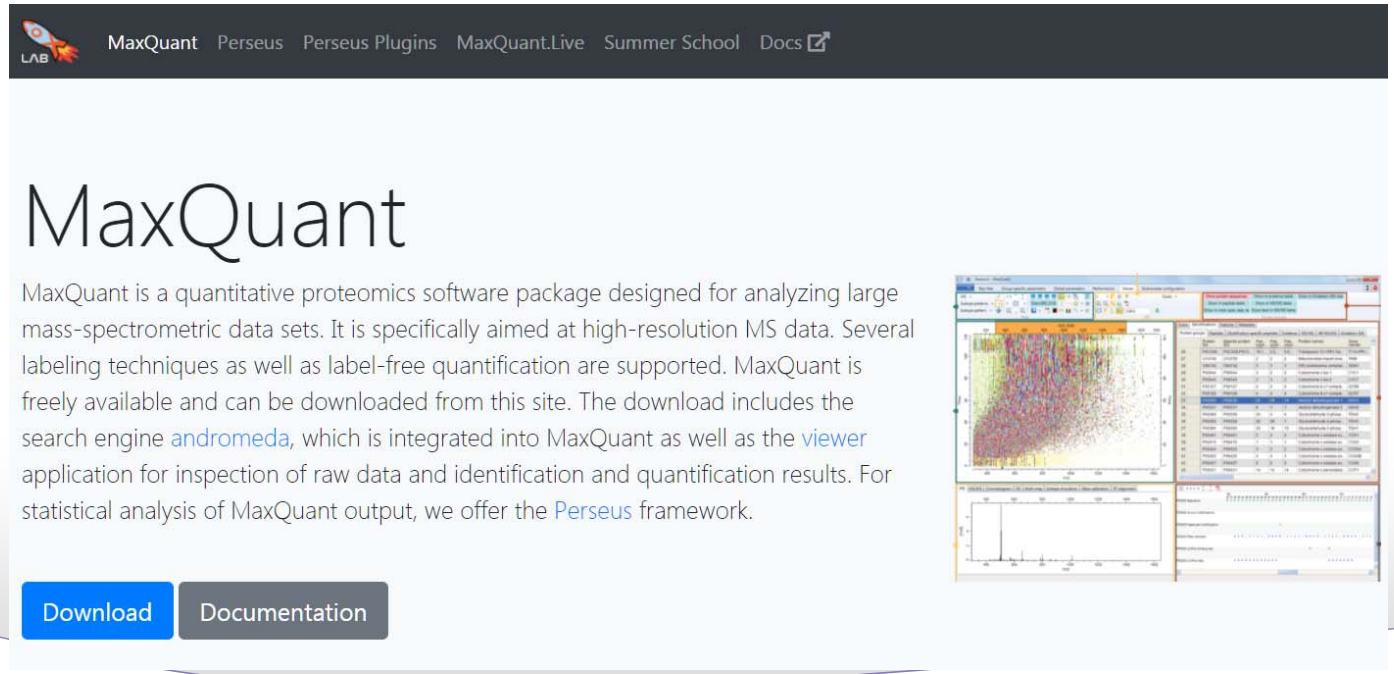
- 根据实验测量的结果，双倍体人类细胞中DNA质量约为6.5 pg = 6.5×10^{-12}克，代入上述公式：

$$P_c = \frac{6.02 \times 10^{23}}{M} \times 6.5 \times 10^{-12} \times \frac{I_P}{I_H} = 3.91 \times 10^{12} \times \frac{I_P}{M \times I_H}$$

- “蛋白质组之尺”也可以用来估算细胞内RNA的分子数量R_c，这是因为细胞内核糖体RNA分子约占总RNA分子的80%，而核糖体RNA分子与核糖体蛋白质分子数量接近1：1，因此可以用上述公式估算核糖体蛋白质分子数量，从而推导出细胞内RNA分子的总拷贝数

MaxQuant

- 主流软件：可处理无标记定量和标记定量的蛋白质组学数据
  - 搜索引擎Andromeda：用于肽段鉴定的数据库搜索
  - Viewer插件：可视化MS/MS数据，便于手工验证
  - Perseus软件：对数据做更深入的统计和计算分析



靶向蛋白质组

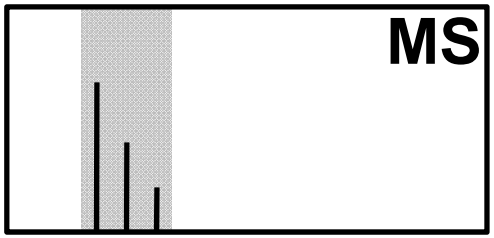
Shotgun proteomics ← LC-MS → Targeted MS

Digestion

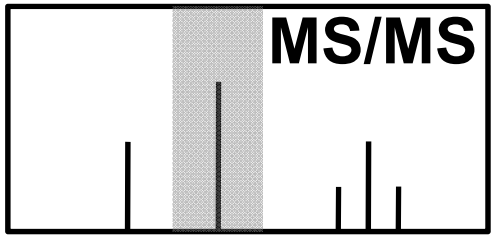
Fractionation

Lysis

1. 记录m/z



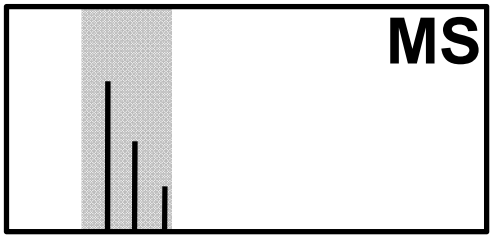
2. 基于丰度和片段选择肽段



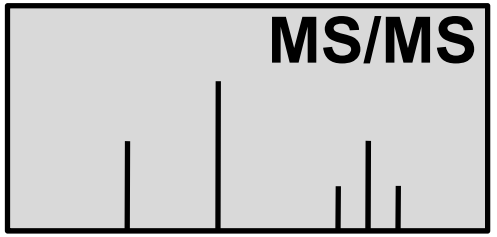
3. 数据库搜索：蛋白质鉴定

Data Dependent Acquisition (DDA)

1. 选择前体离子



2. 前体碎片化

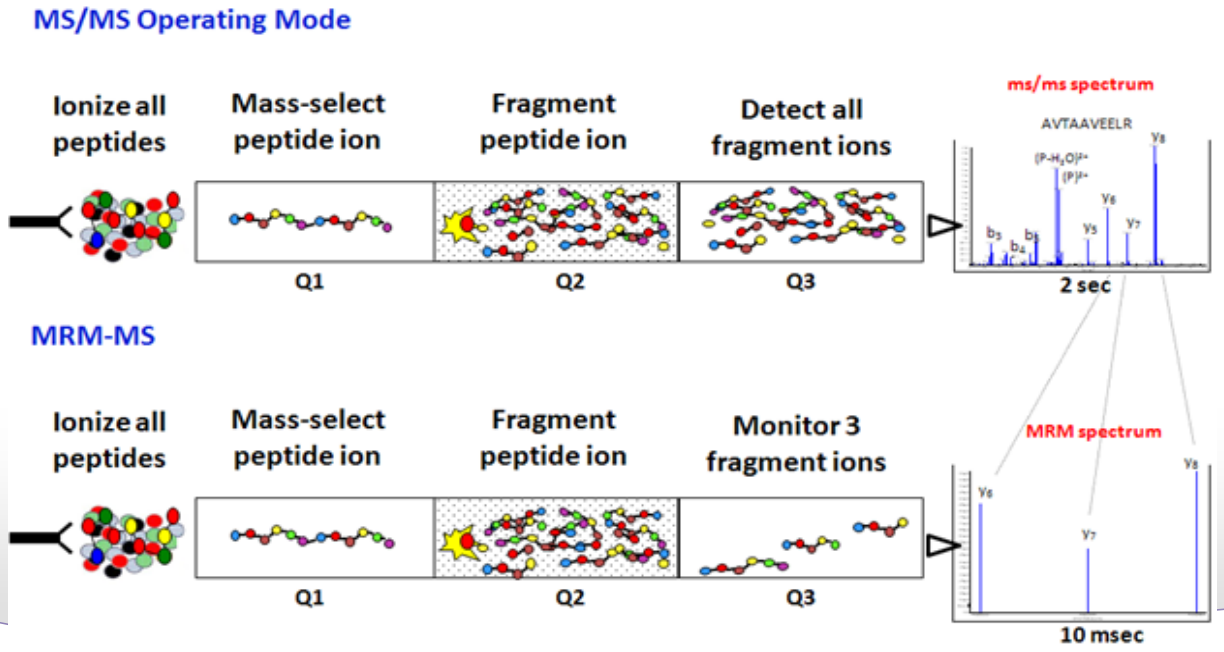


3. 利用前体-碎片对鉴定

肽段集需要事先确定

多重反应监测

- Multiple Reaction Monitoring
- 离子过滤器：三重四级杆
  - Q1：选择前体
  - Q2: CAD碎片化
  - Q3：特异性检测3个碎片离子



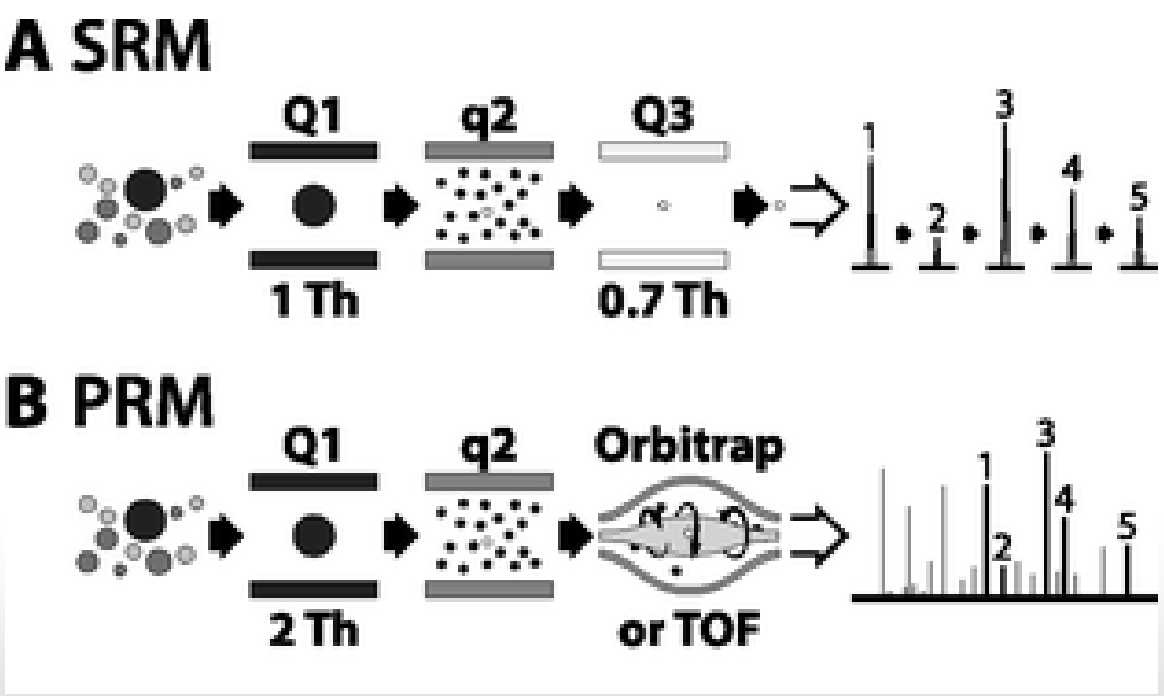
MRM的优缺点

- 优点
  - 快速检测多个离子态：10 ms/离子态
  - 单次试验可分析许多肽段 (100s)，多次检测每个肽段的离子态
  - 高灵敏度、高可重复性
  - 可检测低丰度蛋白质
  - 定量的黄金标准 (Golden standard)
- 缺点
  - 只能检测事先确定的肽段集，需要知道电荷状态、保留时间、相对的离子强度
  - 每次检测的反应数量有限

并行反应监测

Parallel Reaction Monitoring, PRM

- Q3: 高分辨率质量解析器，检测所有靶离子
- 产生高分辨率、全覆盖的二级质谱数据
- 不需要事先选择离子



MRM的应用

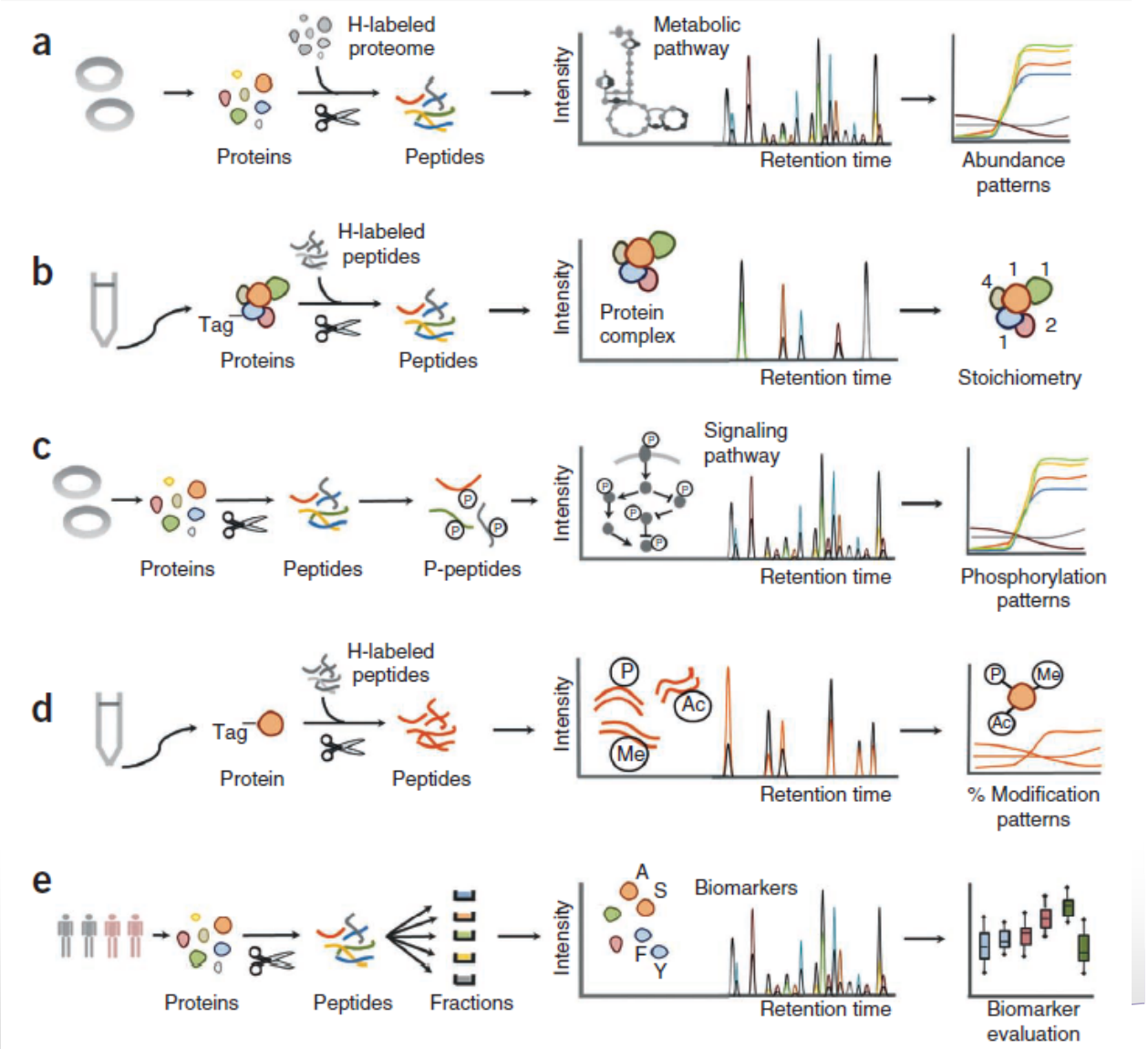
代谢通路分析

蛋白质复合物分析

磷酸化

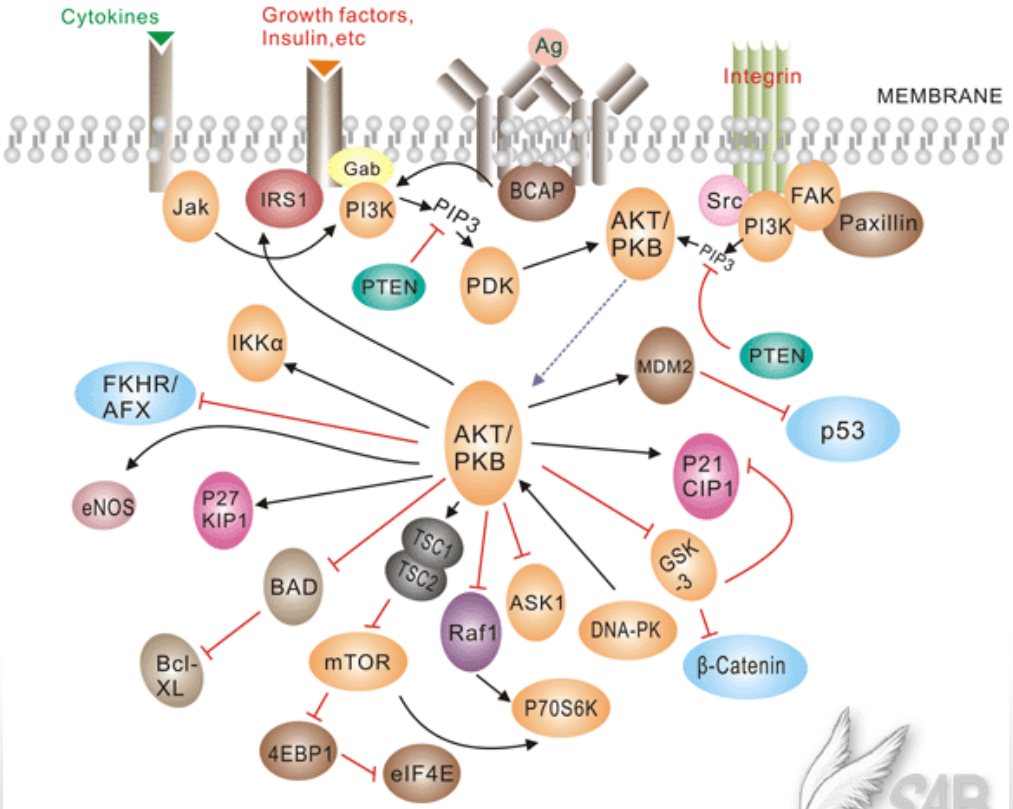
蛋白质修饰

分子标志物：疾病相关蛋白质标志物的验证



靶标选择

- 根据临床或生物学问题选择特定的蛋白质
  - AKT1和乳腺癌



- AKT
- PDK
- BAD
- MDM2
- GSK3
- mTOR
- RAF1

肽段选择

- 每个蛋白质选择3条代表性肽段
- 有独特的序列
- 能够被LC-MS稳定的检测到
- 8-25 aa
- 离子化效率高
- *m/z*在仪器可检测的范围内
- 没有丢失的酶解
- 不能过于亲水 (难以存留)，或疏水 (粘柱)

特征肽段选择

蛋白质集合

Trypsin

肽段

RefSeq
Ensembl
Uniprot

特征肽段

Step 1: 蛋白质全长序列

```
MPKKKPTPIQLNPAPDGSAVNGTSSAETNLEALQKKLEELELDEQQRKRLEAFLTQKQKVGELKDDDFEK
ISELGAGNGGVVFKVSHKPSGLVMARKLIHLEIKPAIRNQIIRELQVLHECNSPYIVGFYGAFYSDGEIS
ICMEHMDGGSLDQVLKKAGRIPEQILGKVSIAVIKGLTYLREKHKIMHRDVKPSNILVNSRGEIKLCDFG
VSGQLIDSMANSFVGTRSYMSPERLQGTHYSVQSDIWSMGLSLVEMAVGRYPIPPPDAKELELMFGCQVE
GDAAETPPRPRTPGRPLSSYGMDSRPPMAIFELLDYIVNEPPPKLPSGVFSLEFQDFVNKCLIKNPAERA
DLKQLMVHAFIKRSDAEEVDFAGWLCSTIGLNQPSTPTHAAGV
```

Step 2: 模拟酶切

PTPIQLNPAPDGSAVNGTSSAETNLEALQK
LEAFLTQK
PSNIVLVNSR
LEELELDEQQR
DDDFEK…..

Step 3: 与参考序列数据库相比

-仅能匹配到单个基因的肽段

PTPIQLNPAPDGSAVNGTSSAETNLEALQK
LEAFLTQK
PSNIVLVNSR
LEELELDEQQR
DDDFEK…..

肽段匹配

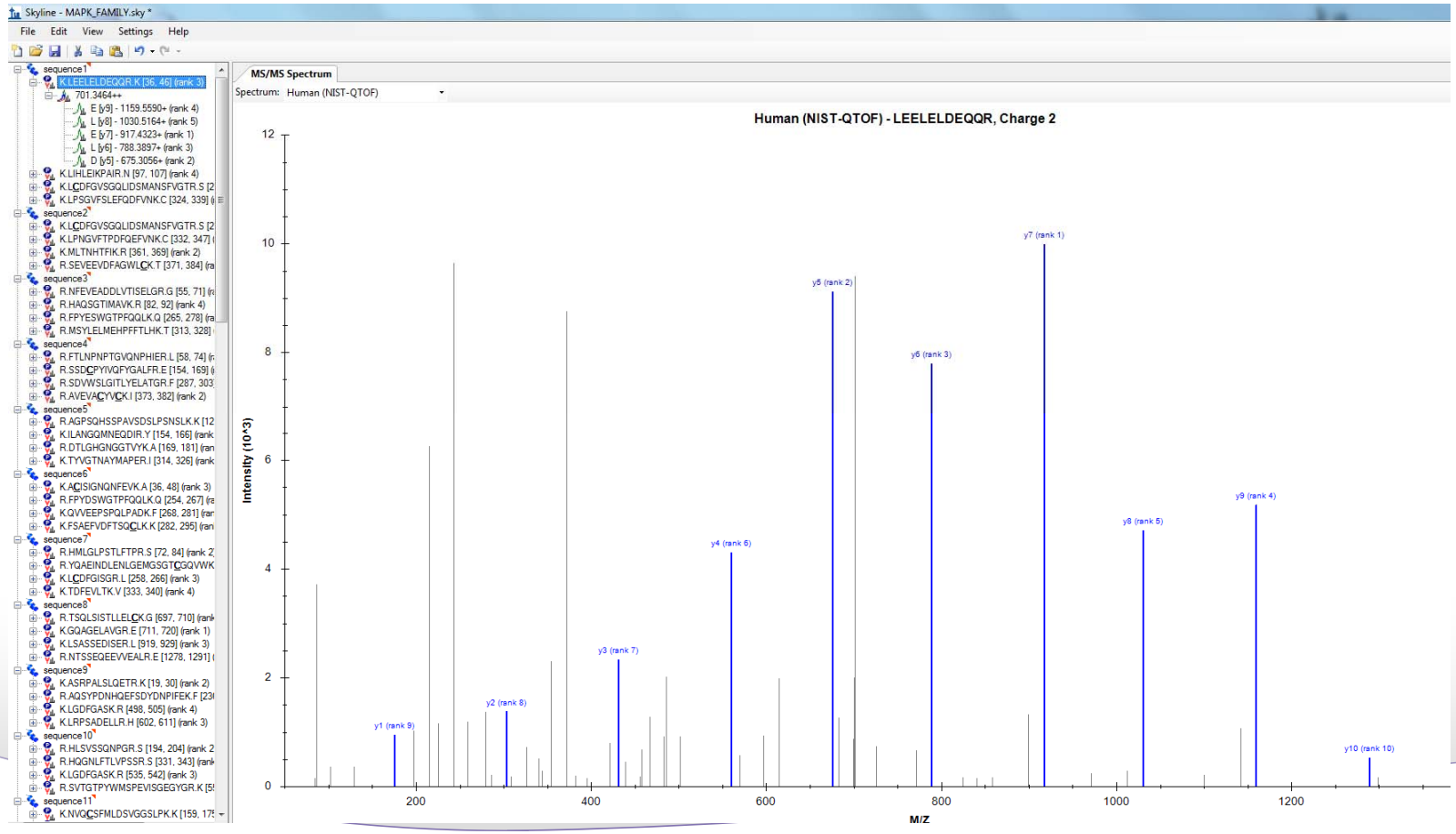
(Reference Protein DB)

蛋白质到基因的匹配

(Using protein names and genomic DB)

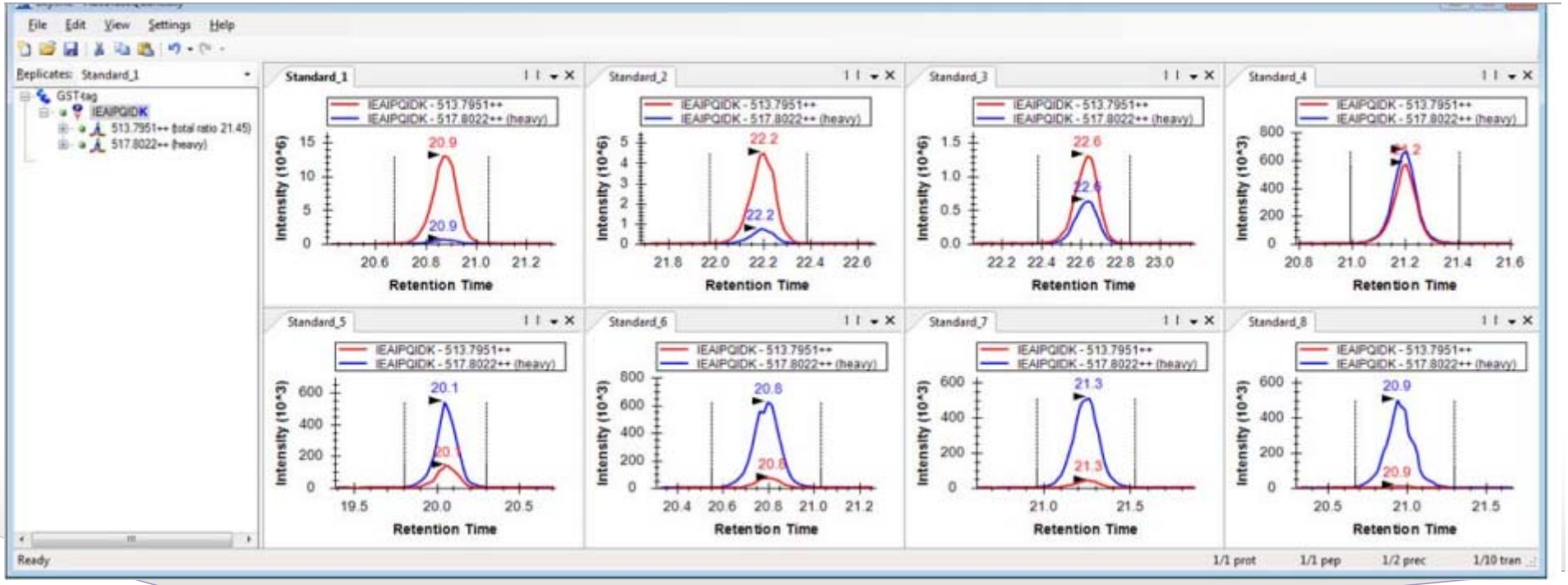
LC/MS性质：Skyline

- 与二级谱图库比较
- 发现最常出现的离子态



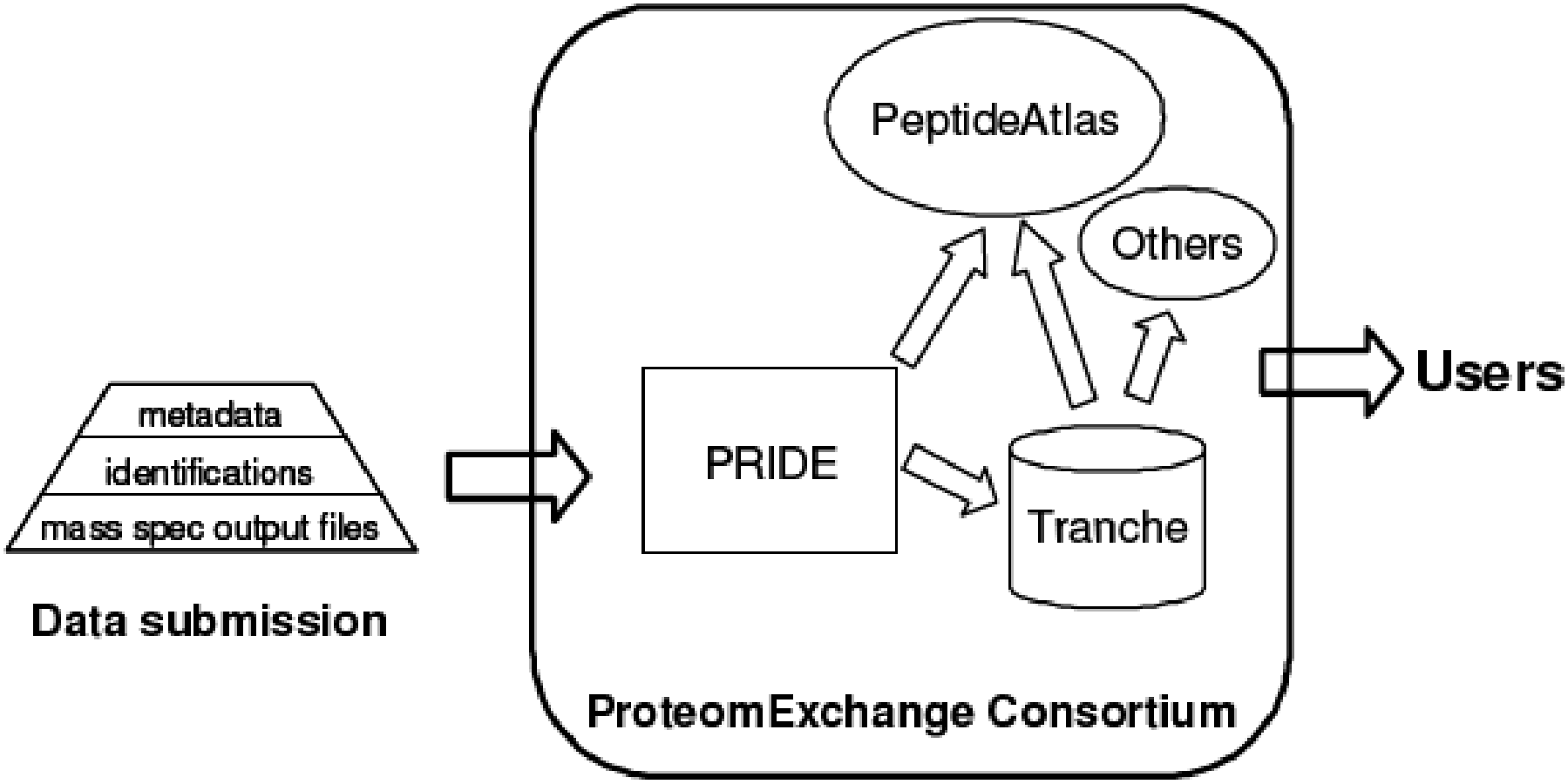
Skyline定量

- 导入质谱鉴定的原始文件
- 选择感兴趣的肽段
- 根据曲线确定样品量



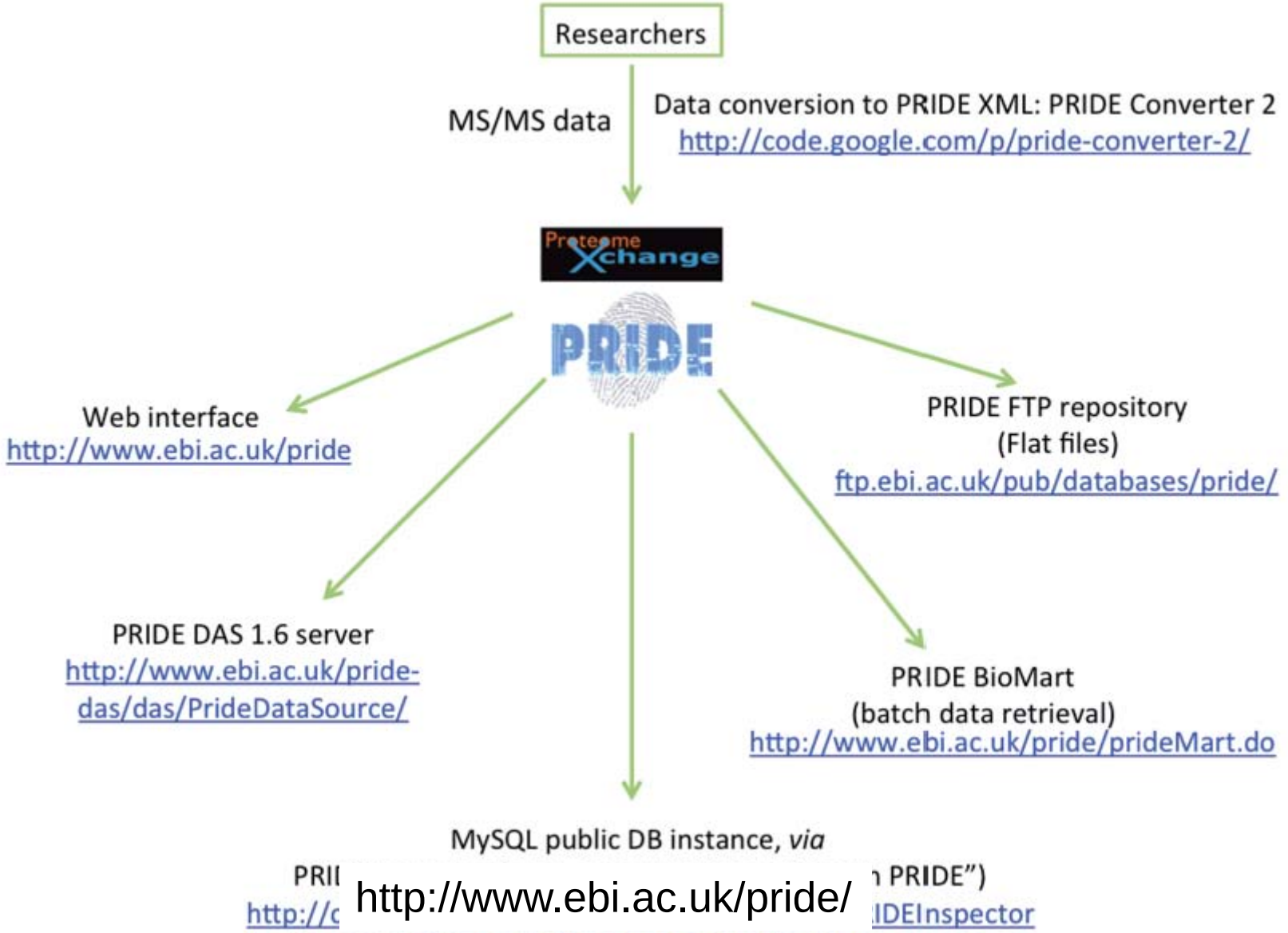
质谱数据库

ProteomeExchange

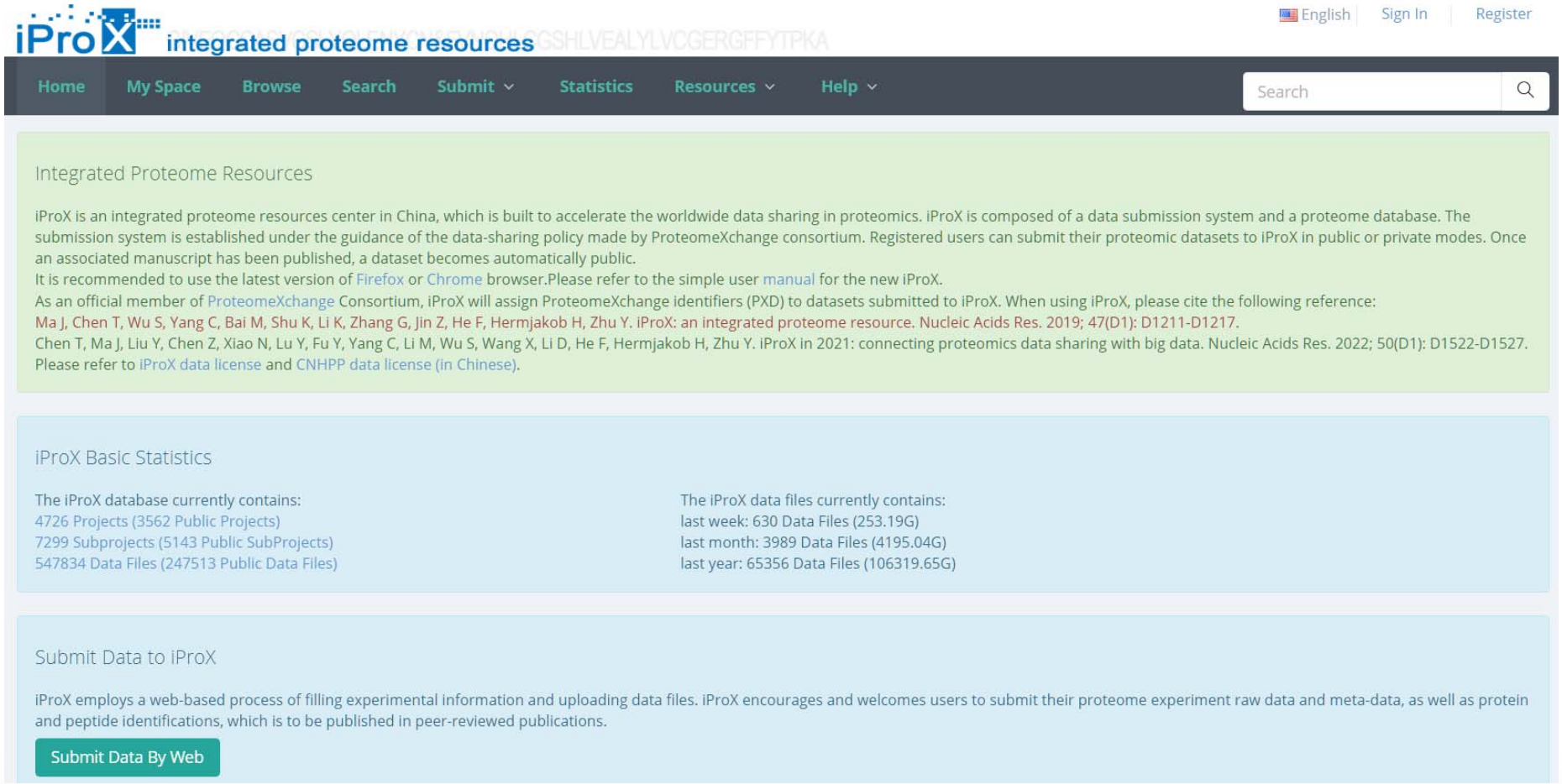


http://www.proteomeexchange.org/

PRIDE


Researchers
MS/MS data
Data conversion to PRIDE XML: PRIDE Converter 2
http://code.google.com/p/pride-converter-2/
ProteomeXchange
PRIDE
Web interface
http://www.ebi.ac.uk/pride
PRIDE FTP repository
(Flat files)
ftp.ebi.ac.uk/pub/databases/pride/
PRIDE DAS 1.6 server
http://www.ebi.ac.uk/pride-das/das/PrideDataSource/
PRIDE BioMart
(batch data retrieval)
http://www.ebi.ac.uk/pride/prideMart.do
MySQL public DB instance, via
http://www.ebi.ac.uk/pride/

iProX



https://www.iprox.cn/

PeptideAtlas

http://www.peptideatlas.org/

GPMDB

gpmdb

accession | gpm # | sequence | keyword | GO
BTO | Chr # | SNAP | pSYT | lists
home | statistics | species | thegpm | about

the gpm

Information
about the GPM
about gpmdb
send us email

Search sites

Eukaryote proteomes
1 2 3 4 5 6 7

Boutique proteomes
human mouse
cow bacteria
plant rat

Algorithms
X! P3 X! Hunter

Information
gpmDB wiki
review lists

Some species


gpmDB statistics for Sun Mar 3 11:49:49 2013 UTC (#3315)

models = 217,125
proteins = 84,408,917
distinct proteins = 1,724,816
protein redundancy = 48.9 ×
peptides = 687,211,623
distinct peptides = 4,286,043
peptide redundancy = 160.3 ×
residues = 9,620,962,722
statistics archive: GPMDB
pages viewed: global map
US visits map
European visits map
Asian visits map
Oceania visits map
South American visits map
African visits map

http://gpmdb.thegpm.org

GPM sponsors
- Proteome Software
- Beavis Informatics
- MCPSB, UM
- LMSGIC, RU

data
- Tranche
- PeptideAtlas
- PRIDE

projects
- iMOP
- HPP
- C-HPP
- HPFP
- The HPA

general info
- ENSEMBL
- STRING DB
- Unimod
- NCTA

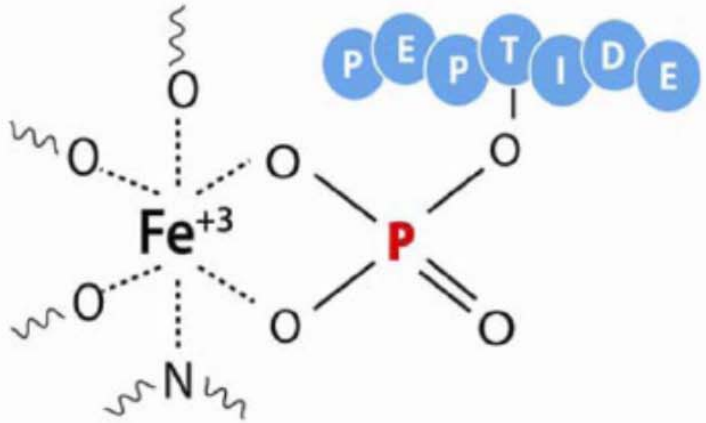
pathways
- KEGG
- Reactome

磷酸化蛋白质组学

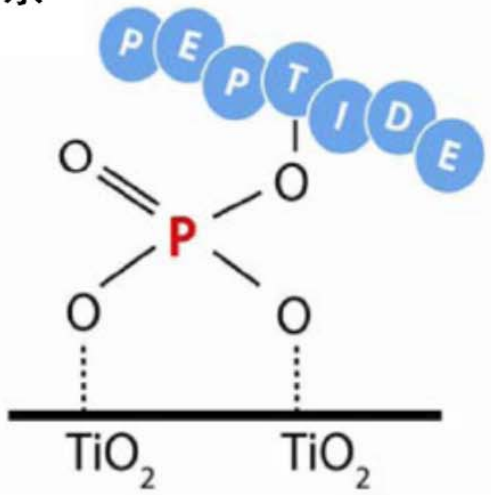
- 鉴定磷酸底物和位点的高通量技术
  - 结合磷酸肽富集和质谱鉴定技术
  - 溶液中的磷酸根为HPO_3^{2-}离子，带两个负电荷
- 磷酸肽富集方法
  - 金属固载的亲和色谱（IMAC），主要包括Fe^{3+}-IMAC、Ti^{4+}-IMAC和Zr4+-IMAC，其中Ti^{4+}-IMAC中的钛为四价阳离子
  - 金属氧化物的亲和色谱（MOAC），如TiO_2-MOAC，在溶液中电离出阳离子
  - 强阳离子交换（SCX）树脂
  - 识别磷酸化酪氨酸的抗体

磷酸肽富集的4种方法

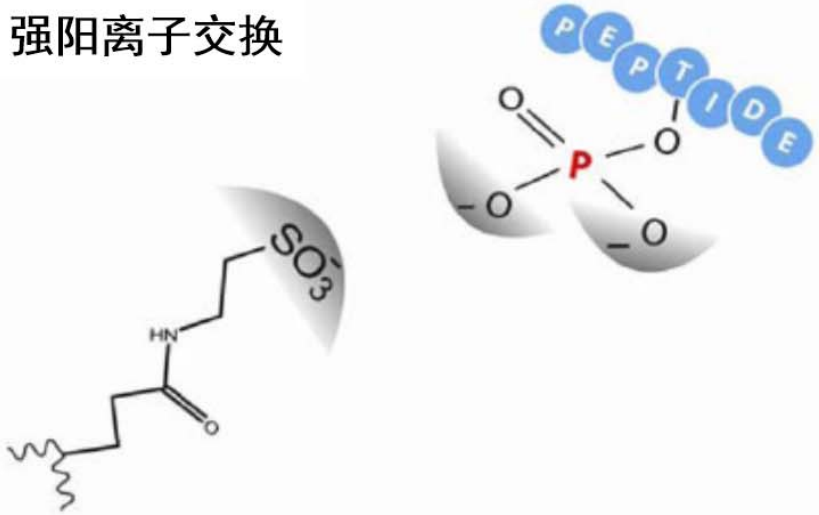
金属固载的亲和色谱



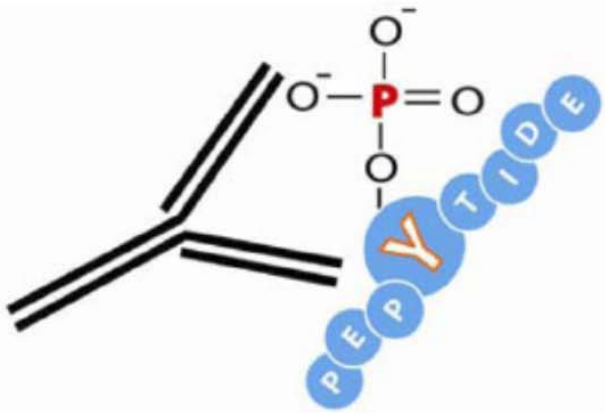
金属氧化物的亲和色谱



强阳离子交换



识别磷酸化酪氨酸的抗体



磷酸肽富集方法的特点

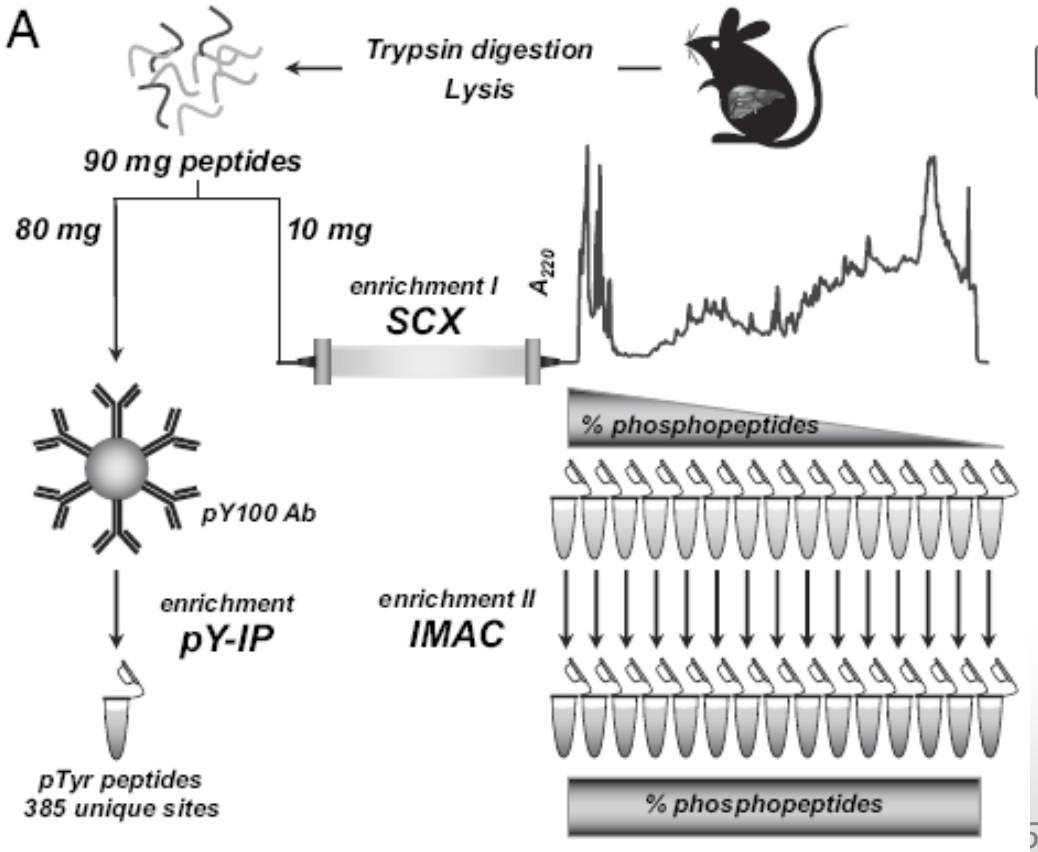
- 磷酸肽的富集方法
  - 液相色谱的分离中，IMAC和MOAC的材料与磷酸根有较强的亲和力，因此磷酸肽在后期被分离
  - SCX带负电，与磷酸根有较强的排斥力，因此磷酸根在前期被分离
  - 酪氨酸与磷酸根的亲和力较弱，酸性环境中共价键常常会发生断裂，因此需要用特异性抗体进行磷酸肽的富集
- 磷酸肽的质谱鉴定
  - 需要根据二级质谱的信息
  - 推断磷酸肽的组成及磷酸化位点的位置
  - 通常磷酸肽的丰度即为所包含位点的磷酸化水平

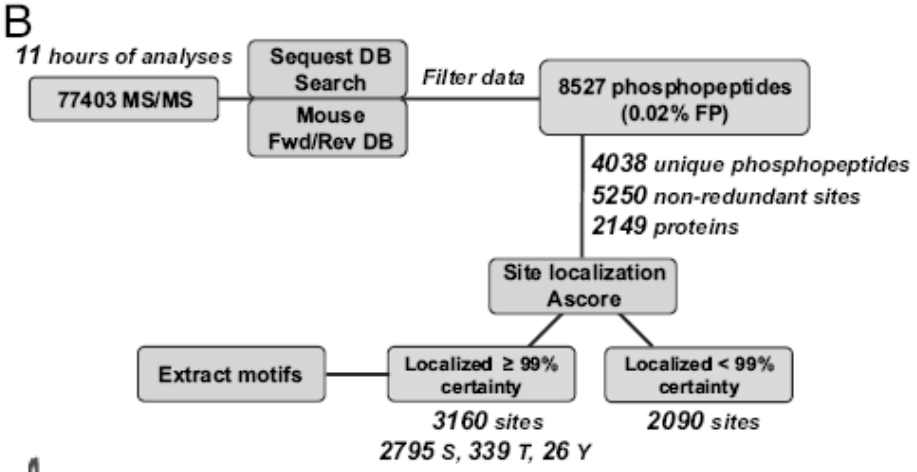
小鼠肝脏的磷酸化组分析

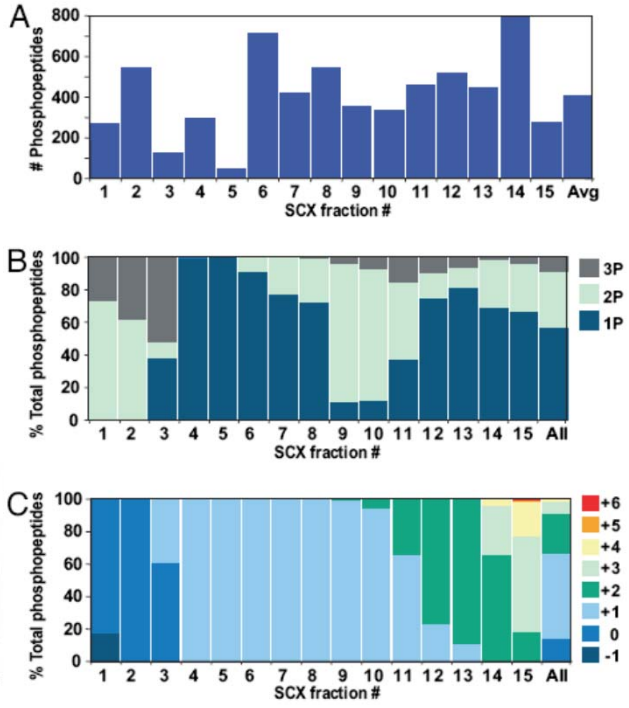
- 21天小鼠的肝脏
  - 90 mg肝脏蛋白质 trypsin酶切
  - 10mg: SCX & IMAC
  - 80mg: pY-IP
- 5,635 个磷酸化位点，2,328个蛋白质
- 磷酸化模体的发现
  - RRxs (PKA), LxRxxs (CaMK), and RxRxxs (AKT)
  - Casein kinase II:sxDxExE, sxxEE, sDxE, and sDxD
  - MAPK: PxsP and PxtP
  - “Dipolar”: Rxxsxx[DE]; R碱性，D/E酸性

磷酸肽富集

- SCX + IMAC
  - 15个SCX组分





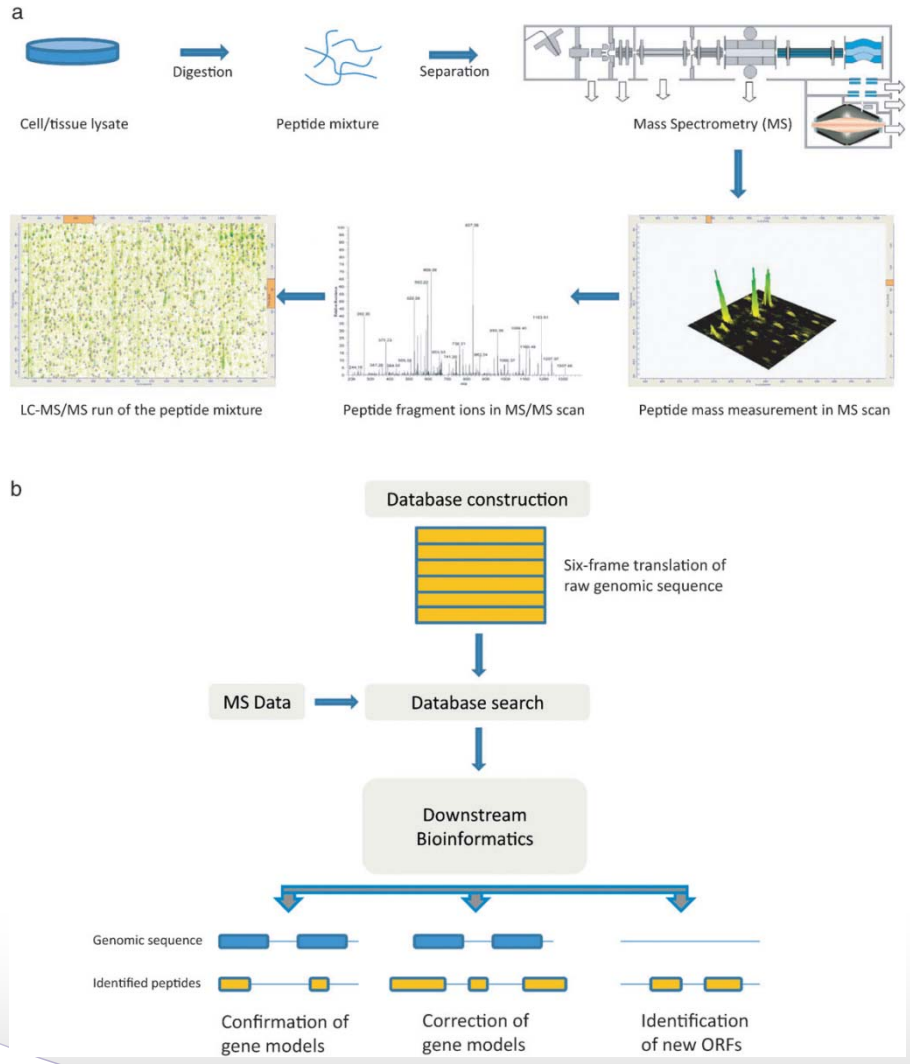


蛋白基因组 (Proteogenomics)

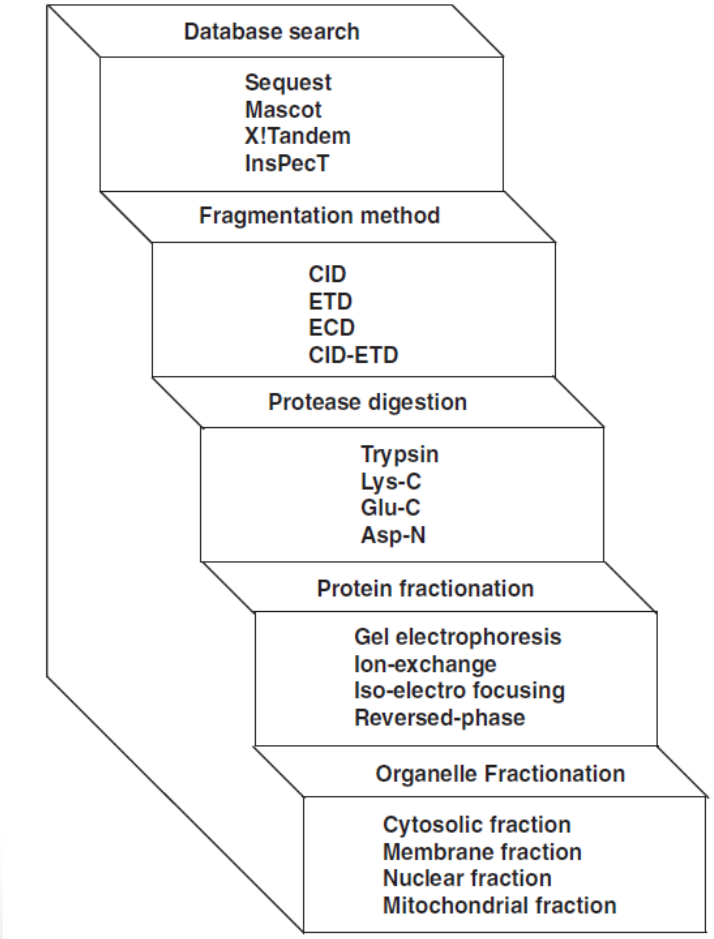
- 基因组注释
  - 验证已有的基因模型，纠正错误的基因模型
  - 发现新基因和新的可变剪接异构体
- 遗传变异对蛋白质组的影响
- 分子标志物鉴定

- **Confirmation of existing gene models**
  - Peptides mapping to predicted genes confirm the existence of annotated gene models
  - N-terminally acetylated peptides can be used to confirm the N-termini of proteins
  - Exon–exon spanning peptides confirm splice isoforms
- **Correction of existing gene models**
  - Peptides mapping to annotated intronic regions signify the presence of a novel exon or extension of an existing exon
  - Peptides mapping to exon–intron boundaries extend the existing exon
  - Peptides mapping to different reading frames indicates identification of additional reading frames
  - Peptides overlapping a novel exon–exon junction represent novel splice isoforms
- **Identification of novel genes**
  - Mapping of peptides in the non-genic region indicates a novel gene
  - Peptides mapping to an annotated pseudogene indicates a wrongly annotated pseudogene
- **Others**
  - Fusion protein identification from peptide mapping to two different protein-coding genes
  - Peptides with sequences that diverge from known genomic sequences reflect sequence polymorphisms in the genome

蛋白基因组分析流程

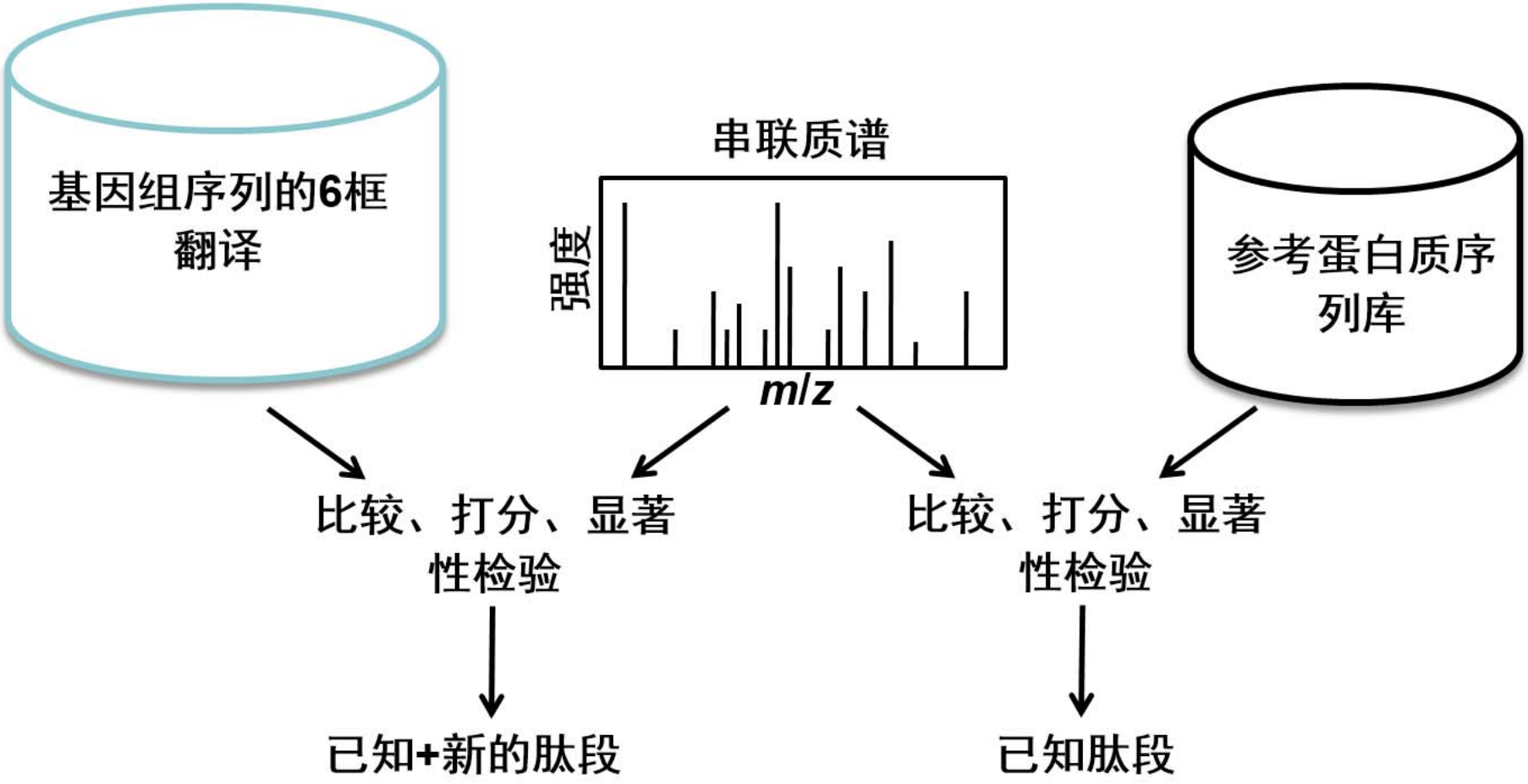


Krug K., Nahnsen S, Macek B, Molecular Biosystems 2010



Renuse S, Chaerkady R and A Pandey, Proteomics. 2011

基于基因组的序列数据库

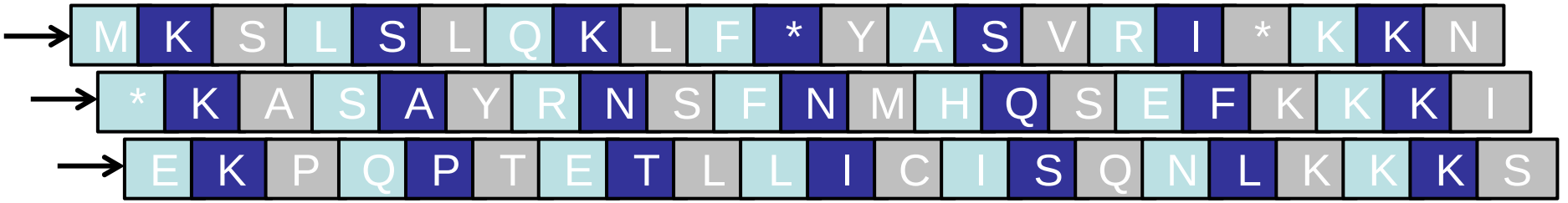


结合基因组数据的新蛋白发现

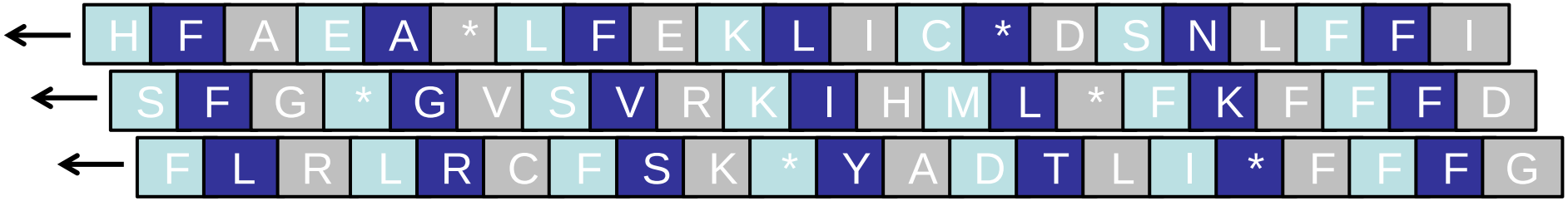
6框翻译数据库

ATGAAAAGCCTCAGCCTACAGAAACTCTTTTAATATGCATCAGTCAGAATTTAAAAAAAAAATC

正链



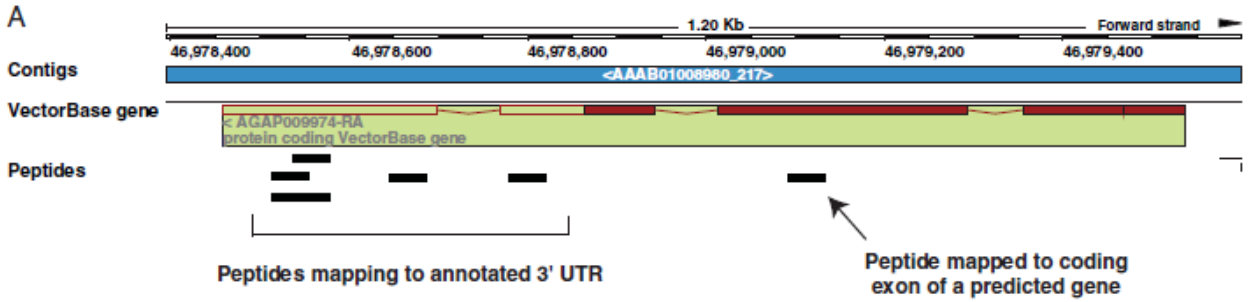
负链



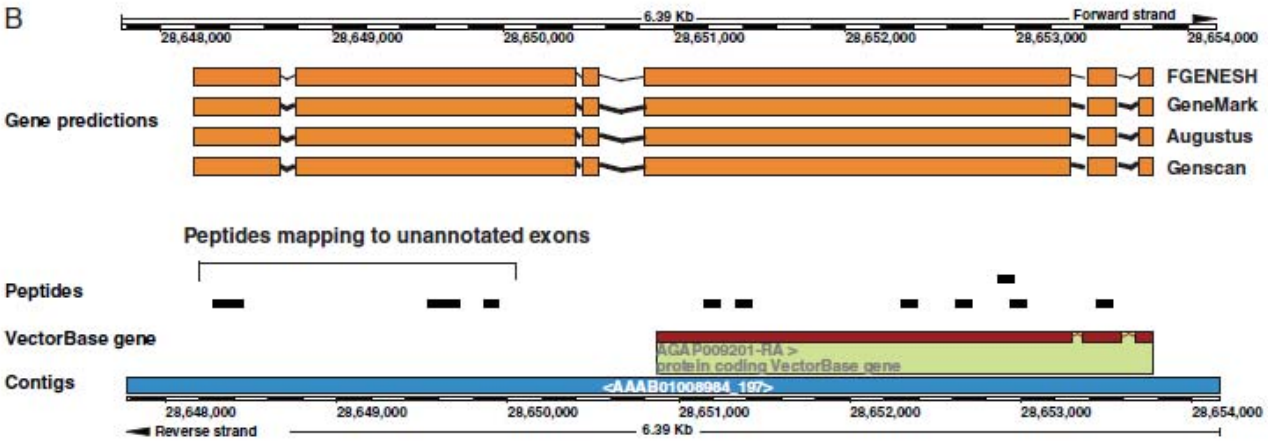
- 软件
  - Peppy: 建库+搜索
  - BCM Search Launcher
  - InsPecT: Perl脚本

基因组注释：*A. gambiae*

Peptides mapping to annotated 3’ UTR

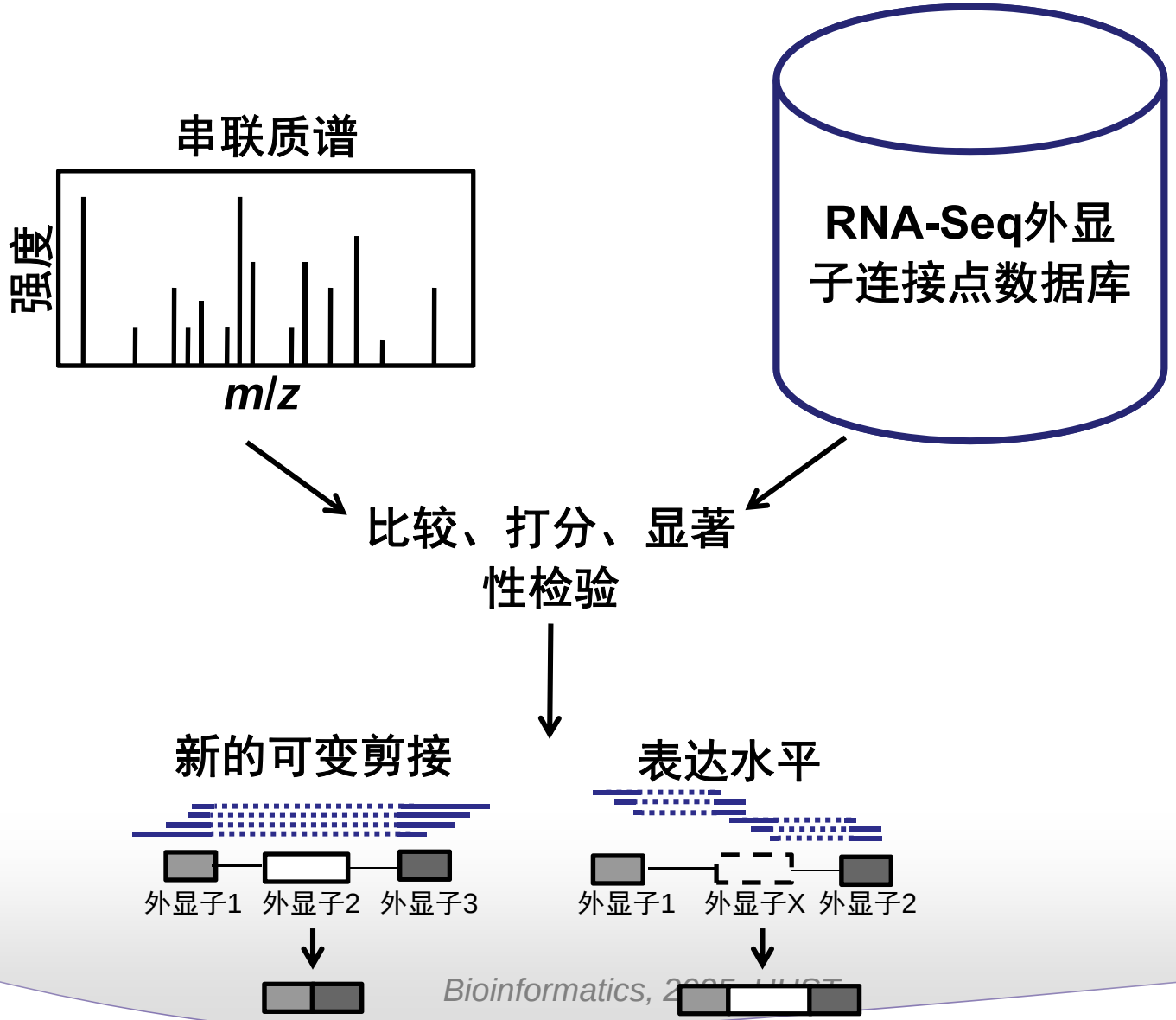


Peptides mapping to novel exon within an existing gene

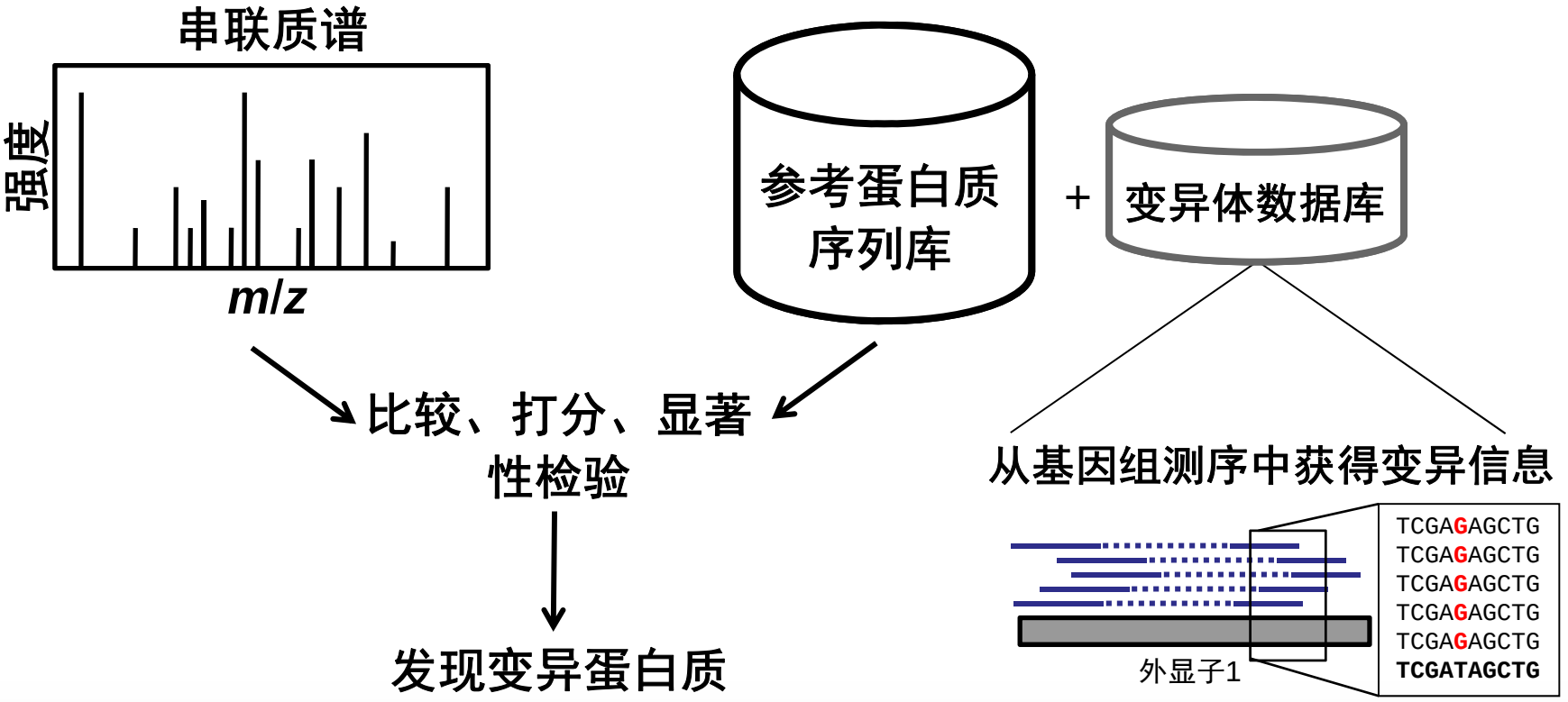


Renuse S, Chaerkady R and A Pandey, Proteomics. 11(4) 2011

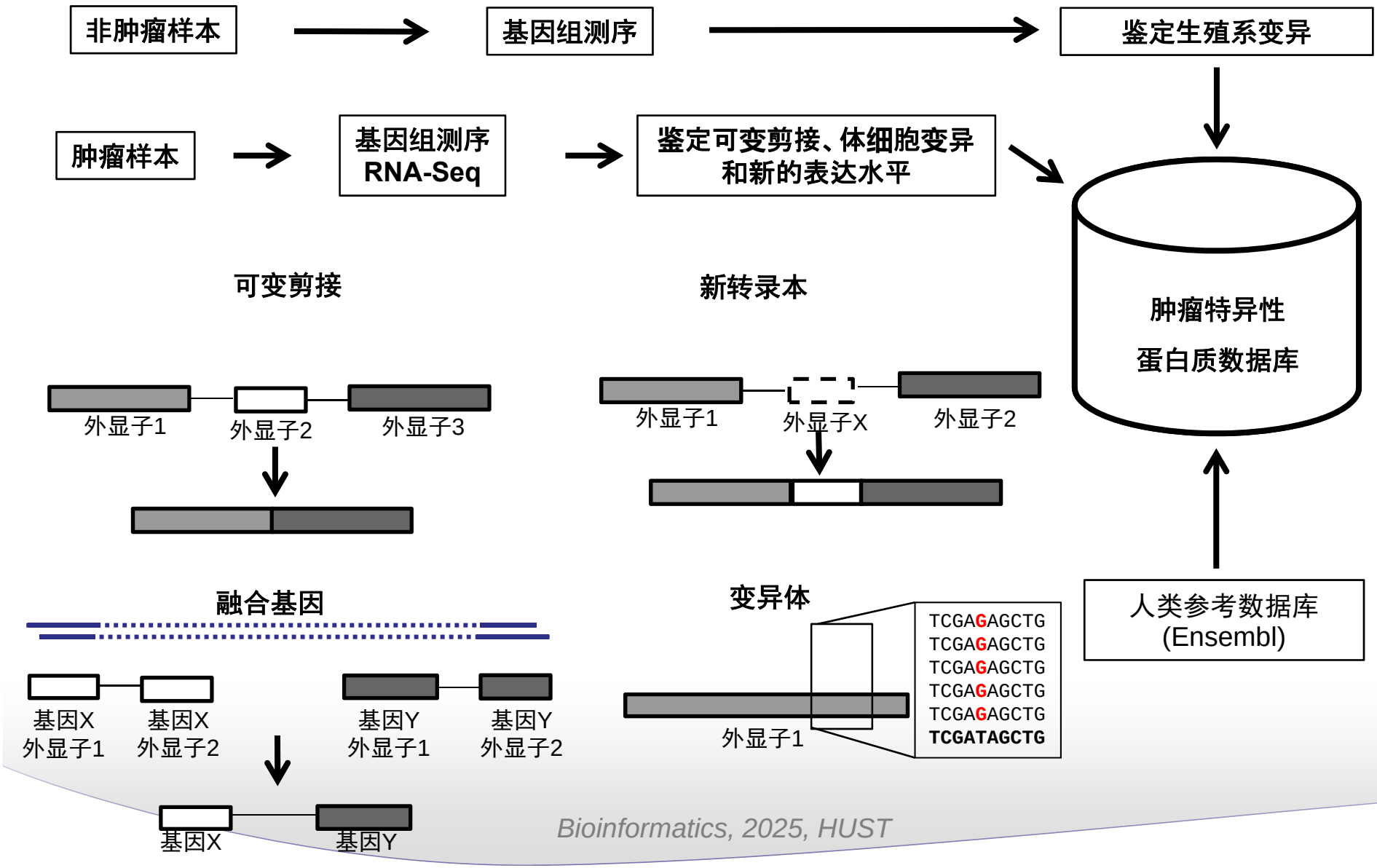
基于RNA的序列数据库



单核苷酸变异数据库

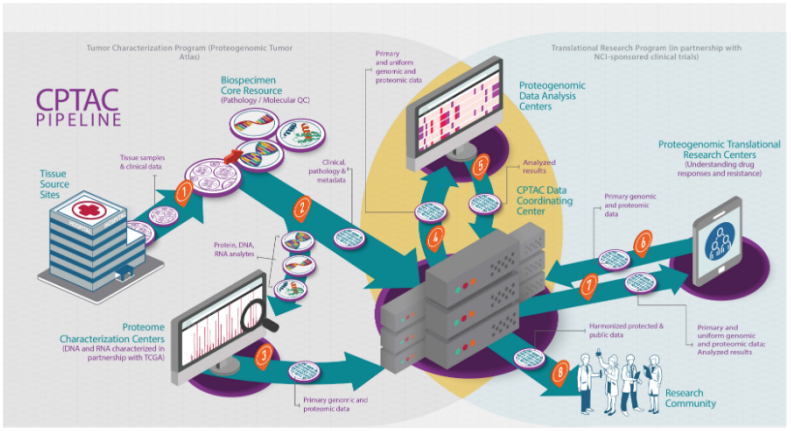

串联质谱
强度
m/z
参考蛋白质序列库
+
变异体数据库
比较、打分、显著性检验
发现变异蛋白质
从基因组测序中获得变异信息
TCGAGAGCTG
TCGAGAGCTG
TCGAGAGCTG
TCGAGAGCTG
TCGAGAGCTG
TCGATAGCTG
外显子1

肿瘤特异性蛋白质数据库



CPTAC：临床肿瘤蛋白质组

CPTAC



The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics.

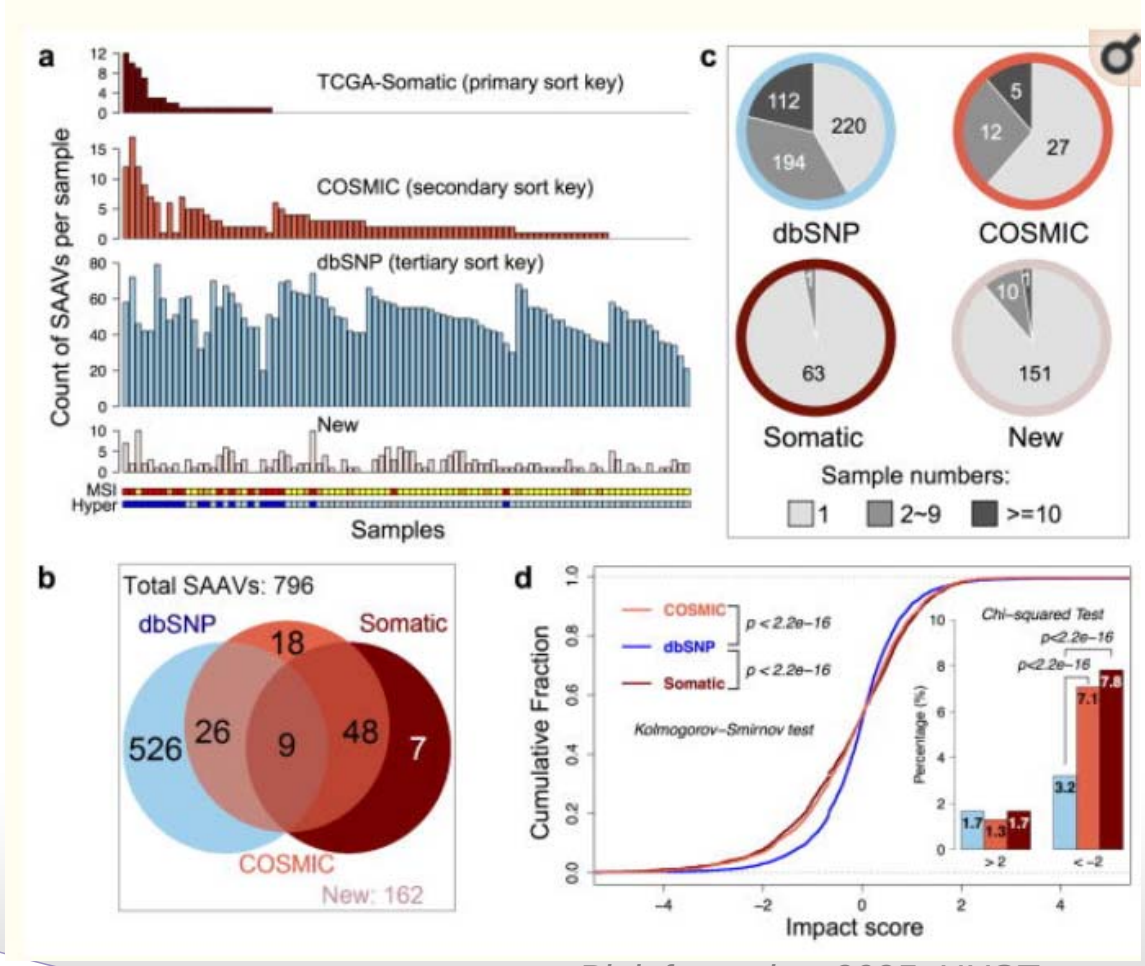
Study Name	Description	Journal	Data
Colon Cancer Therapeutic Opportunities	Proteogenomic study on a prospectively collected **colon cancer** cohort with **110 paired** tumor and normal adjacent tissues.	**Nature**	**2019**
Proteogenomics of Gastric Cancer	Proteogenomic analysis was performed on a cohort of **80 patients** with early-onset **gastric cancer** recruited from the Korean population.	**Cancer Cell**	**2019**
Reproducible Proteome and Phosphoproteome Workflow	Optimized workflow for global proteome and phosphoproteome analysis of tissues or cell lines that uses isobaric tags (**TMT** (tandem mass tags)-10) for multiplexed analysis	**Nature Protocal**	**2018**
Proteogenomics of Ovarian Cancer	Comprehensive proteomic characterization of **174 ovarian high-grade serous carcinomas** (HGSC) previously characterized by The Cancer Genome Atlas (TCGA) along with evaluation of post-translational modifications, phosphosites, in 69 of the 174 tumors.	**Cell**	**2016**
Breast Cancer Proteogenomics Landscape Study	Proteogenomic analysis of **primary breast cancer** tumors was used to elucidate the functional consequences of somatic mutations. Proteome and phosphoproteome mass spectrometry raw data from **105 TCGA tumors** are available along with supplementary data on copy number alterations.	**Nature**	**2016**

结直肠癌的蛋白基因组学分析

- 样本来源：95组TCGA肿瘤样本, 30例正常样本
- 组学鉴定：
  - 基因组
  - 转录组
  - 蛋白组 （非标）
- 相关分析：
  - 氨基酸突变 – 碱基突变
  - 蛋白丰度 – 基因表达
  - 拷贝数变异 – 基因表达、蛋白丰度
  - 蛋白组分型 – 转录组分型
  - 蛋白组分析 – 蛋白标志物

氨基酸突变 *vs.* 碱基突变

- 氨基酸突变大部分来自于碱基突变

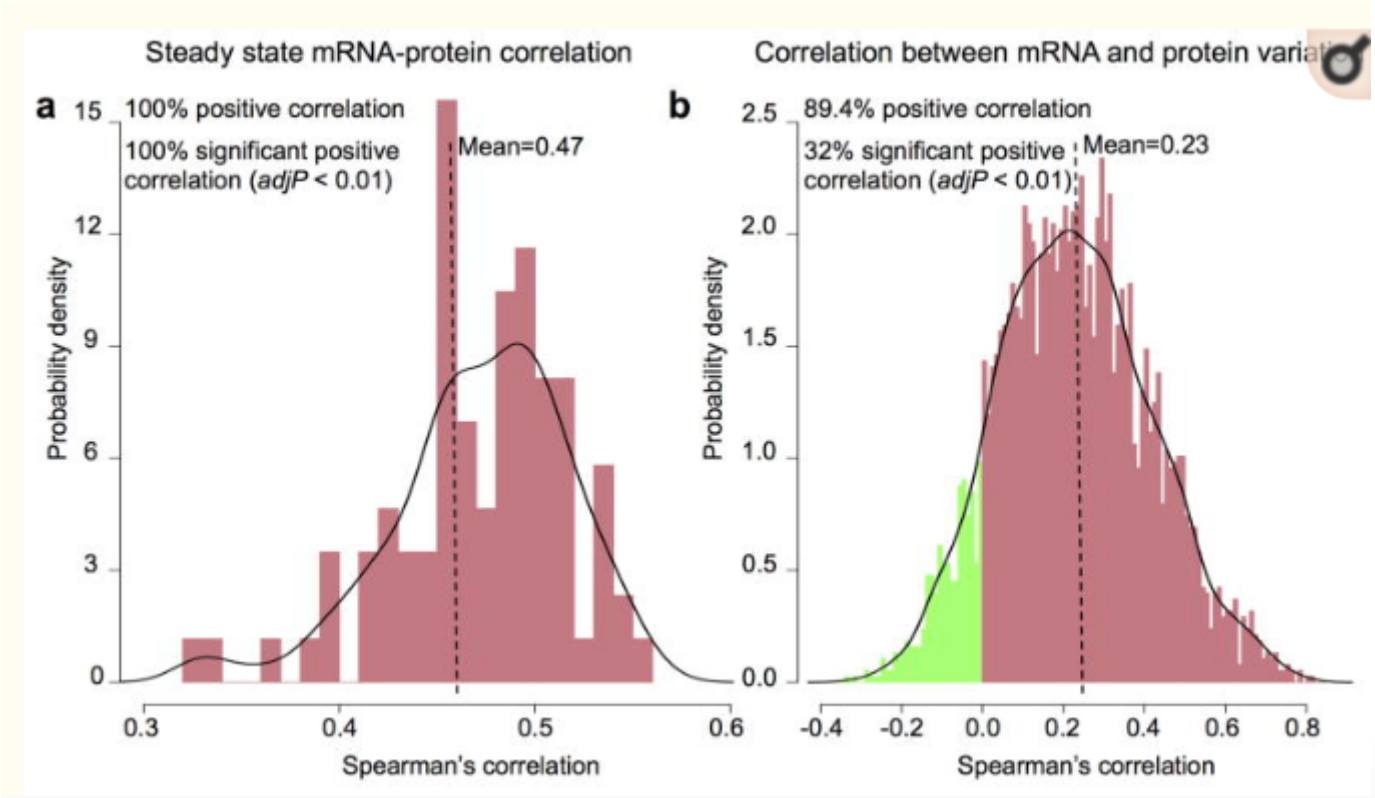


796 SAAVs

64 TCGA
101 COSMIC
526 dbSNP

mRNA丰度 *vs.* 蛋白质丰度

❒ 用mRNA丰度预测蛋白丰度并不可靠

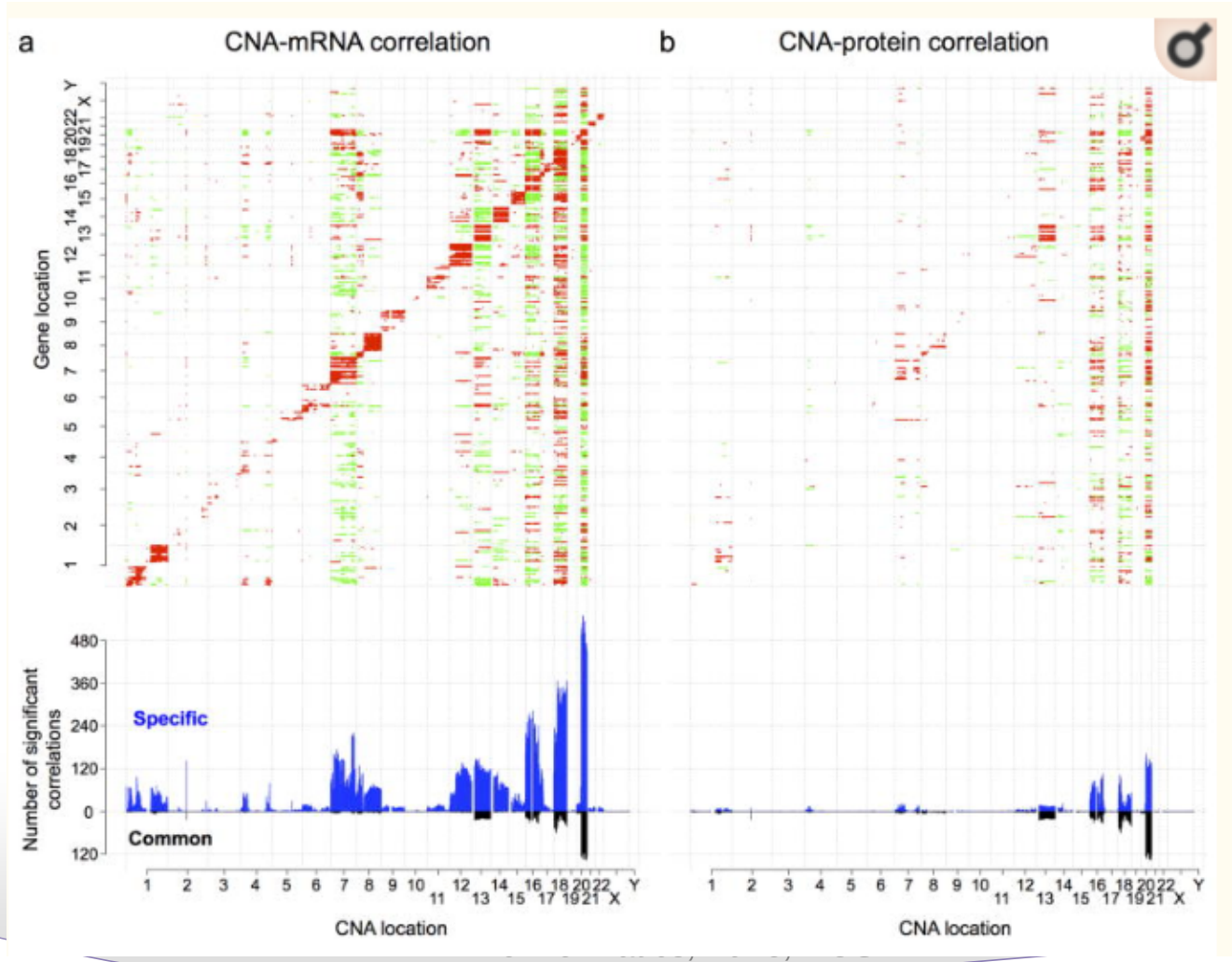


稳定状态mRNA-蛋白
显著正相关：**100%**
平均相关系数：**0.47**

变化状态mRNA-蛋白
显著正相关：**32%**
平均相关系数：**0.23**

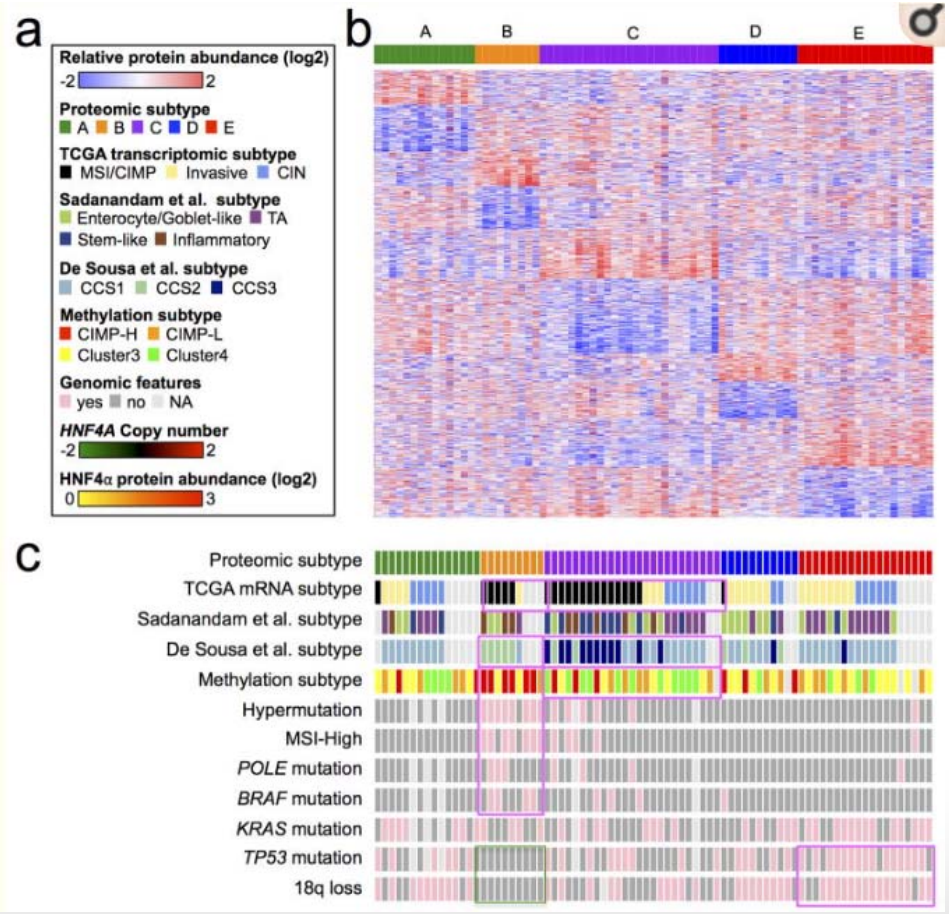
拷贝数变异分析

- **CNAs对转录影响大，对蛋白层面影响较低**



蛋白质组的分子分型

- 蛋白组分型结果与转录组分型结果部分吻合

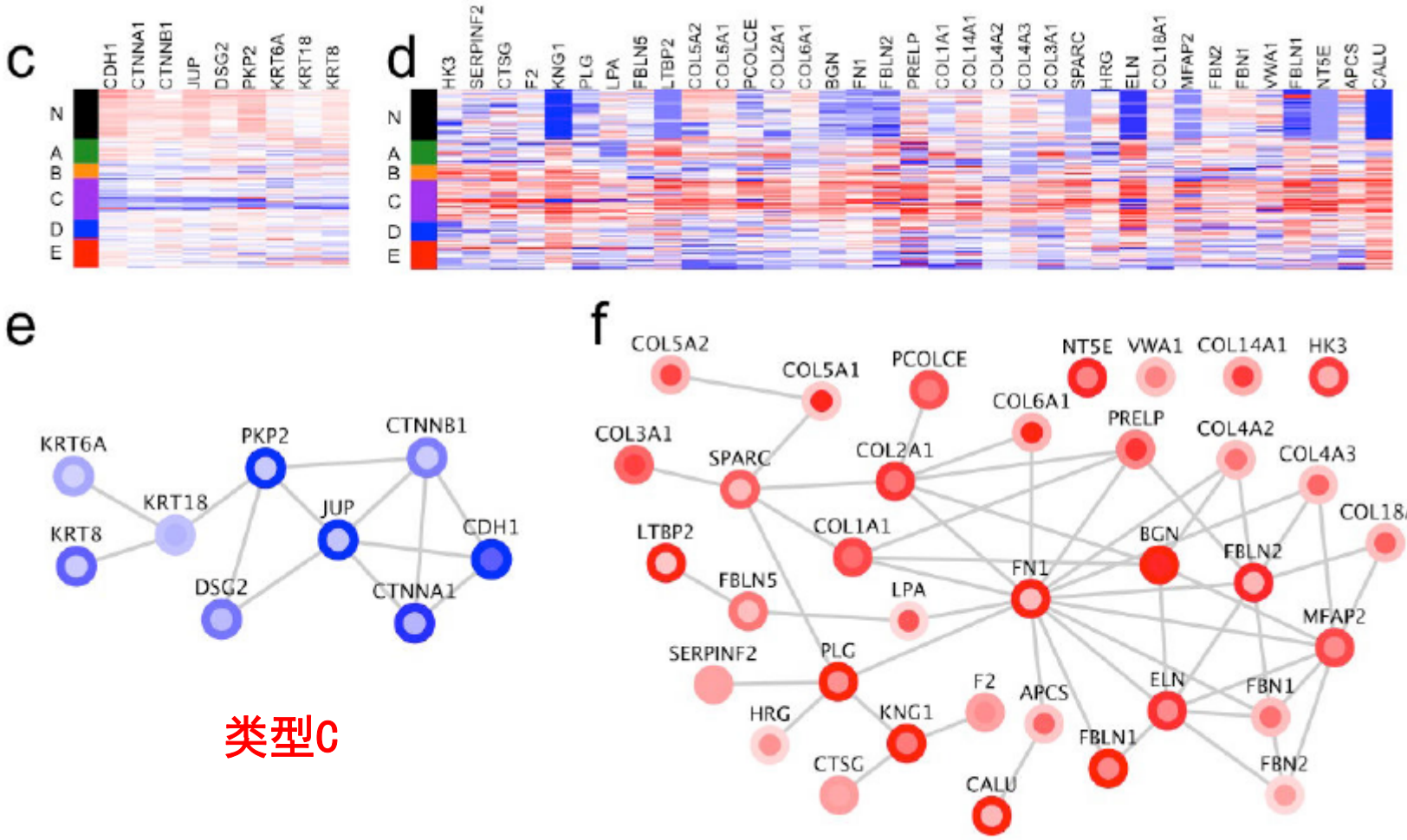


蛋白组**类型B**与转录组**类型CIMP-H**显著相关

蛋白组**类型E**与**TP53突变**及**18q缺失**显著相关

富集分析

❒ 各分型分子标记物及通路富集



类型C