# KSBi-BIML 2024

## Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

## 생명정보학 & 머신러닝 워크샵 (온라인)

# 질량분석 및 단백체 데이터 분석

김민식 _ DGIST

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

# KSBi-BIML 2024

## Bioinformatics & Machine Learning(BIML)
## Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

**한국생명정보학회장 이 인 석**

**강의개요**

# 질량분석 및 단백체 데이터 분석

## (이론) Mass Spectrometry-based Proteomics
## (실습) Proteomics Data Analysis

발달, 노화, 및 질병 발생 과정 동안, 장기 및 조직의 위치에 따라 단백질을 만들어 내고 있는 것은 매우 흥미로운 일이다. 이를 통해 다세포 생물의 하나인 인간 몸속의 수많은 세포가 같은 유전체를 보유하더라도 각기 다른 일을 유기적으로 할 수 있는 것일 것이다. 유전자에 내재된, 그러나 이해되지 못한 표현형을 해석하는 것은 생물학적으로 의학적으로 매우 중요한 일이다.

최근 질량분석법을 기반으로 하는 단백질 집합(통칭 단백체, Proteome)에 대한 연구 기술이 급격히 발달하고 있으며 가까운 시일 내에 NGS 수준의 방대한 데이터 양을 생산하는 날이 가까워지고 있다. 단백질은 세포 내에서의 위치(즉, 핵, 세포질)와 단백질 변형(즉, PTM, cleavage 등)에 따라 다른 기능을 하고 있으며 이를 이해하는 것은 생명체가 시간적, 그리고 공간적으로 어떻게 외부 환경에 반응하고, 어떻게 내부적으로 짜여진 프로그램을 영위해 나가는지 알 수 있게 할 것이다. 이는 단순히 항체를 활용하여 몇몇 단백질의 발현을 관찰하는 것과는 다른 차원의 일이라 여려진다.

본 강의에서는 질량분석에 대해 이해하고 단백체 데이터 수집 방식을 공부할 것이며 단백체 데이터 분석을 위해 사용되는 플랫폼에 대해 경험하고 데이터 처리에 대한 예를 다룰 것이다. 이를 통해 빅데이터를 빠르고 손쉽게 처리할 수 있는 핵심 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

- 질량분석 개요와 단백체 실험의 개론
- 질량분석 데이터 수집 방법 및 이해
- 글로벌 단백체 데이터 및 단일세포 단백체 데이터 형태 이해

*참고강의교재: Min-Sik Kim et al. Nature 2014

*교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

*강의난이도: 초급

# Curriculum Vitae

## Speaker Name: Min-Sik Kim, Ph.D.


▶ **Personal Info**

| | |
|---|---|
| Name | Min-Sik Kim |
| Title | Associate Professor |
| Affiliation | Department of New Biology, DGIST |

▶ **Contact Information**

| | |
|---|---|
| Address | DGIST, 333 TechnoJungang-daero, Dalseong-gun, Daegu, 42988 |
| Email | mkim@dgist.ac.kr |
| Phone Number | 053-785-1630 |


---

**Research Interest**

Mass Spectrometry, Proteomics, Systems Biology, Metabolomics, Multi-Omics

**Educational Experience**

| | |
|---|---|
| 2002 | B.S. in Chemistry, Korea University, Korea |
| 2004 | M.S. in Physical Chemistry, Korea University, Korea |
| 2013 | Ph.D. in Biological Chemistry, Johns Hopkins University School of Medicine, USA |

**Professional Experience**

| | |
|---|---|
| 2013-2016 | Postdoctoral fellow, Institute of Genetic Medicine, Johns Hopkins University School of Medicine |
| 2016-2018 | Assistant Professor, Department of Applied Chemistry, Kyung Hee University |
| 2018-present | Assistant, Associate Professor, Department of New Biology, DGIST |

**Selected Publications (5 maximum)**


1. Park, G., **Jang, E**. W., ..., **Kim, M.-S.***, Lee, Y.-S.* Dysregulation of the Wnt/β-catenin signaling pathway via Rnf146 upregulation in a VPA-induced mouse model of autism spectrum disorder. ***Experimental & Molecular Medicine***.
2. Hyeon, D. Y., Nam, D...., **Kim, M.-S.**, ... Hwang, D.*, Lee, S.-W.* (2022) Proteogenomic landscape of human pancreatic ductal adenocarcinoma in an Asian population reveals tumor cell-enriched and immune-rich subtypes. ***Nature Cancer***. 4(2):290-307.
3. **Jang, E. W., Park, J. H., ... Kim, M.-S.*** (2022) Cntnap2-dependent molecular networks in autism spectrum disorder revealed through an integrative multi-omics analysis. ***Molecular Psychiatry***. Accepted.
4. Park, J.-H., Ryu, S. J., ...,, **Lee, J. H., Park, J. H., ..., Kim, M.-S.***, Hwang, D.*, Lee, Y.-S.*, and Park, S. C.* (2021) Disruption of nucleocytoplasmic trafficking as a cellular senescence driver. ***Experimental & Molecular Medicine***. 53, 1092–1108.
5. Huh, S., Hwang, D.*, **Kim MS*** (2020) Statistical modeling for enhancing discovery power of citrullination from tandem mass spectrometry data. ***Analytical Chemistry***. 92, 19, 12975–12986.

KSBI | 한국생명정보학회

# KSBi-BIML 2024

## 질량분석 및 단백체 데이터 분석

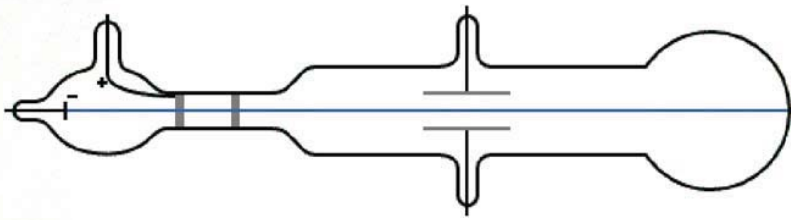
### (이론) Mass Spectrometry-based Proteomics

# Mass Spectrometry

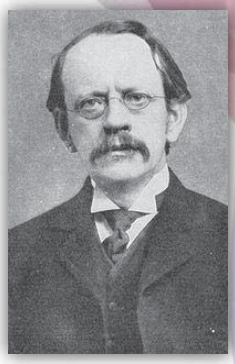
## 질량 분석기 개발의 기초 아이디어

**Nobel Prize in Physics (1906)**

**"in recognition of the great merits of his theoretical and experimental investigations on the conduction of electricity by gases."**



wikipedia



**J. J. Thomson (1856~1940)**



## 네온의 질량 스펙트럼

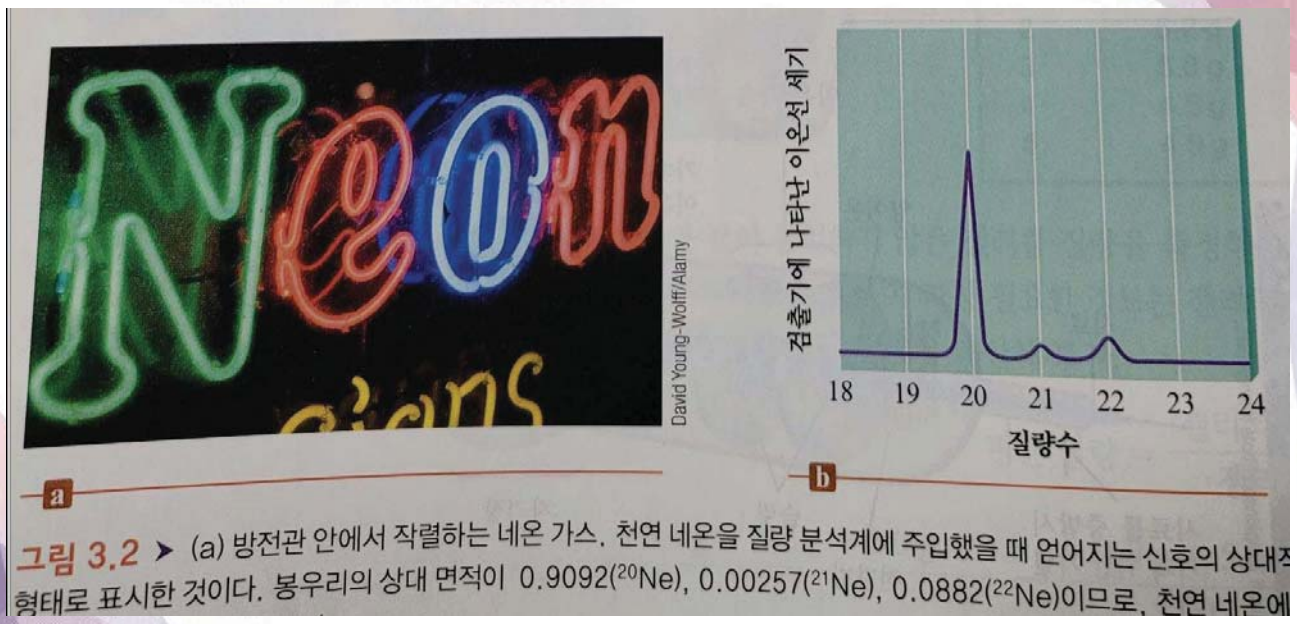
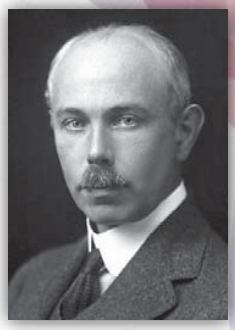



그림 3.2 ➤ (a) 방전관 안에서 작렬하는 네온 가스. 천연 네온을 질량 분석계에 주입했을 때 얻어지는 신호의 상대적 형태로 표시한 것이다. 봉우리의 상대 면적이 0.9092($^{20}$Ne), 0.00257($^{21}$Ne), 0.0882($^{22}$Ne)이므로, 천연 네온에

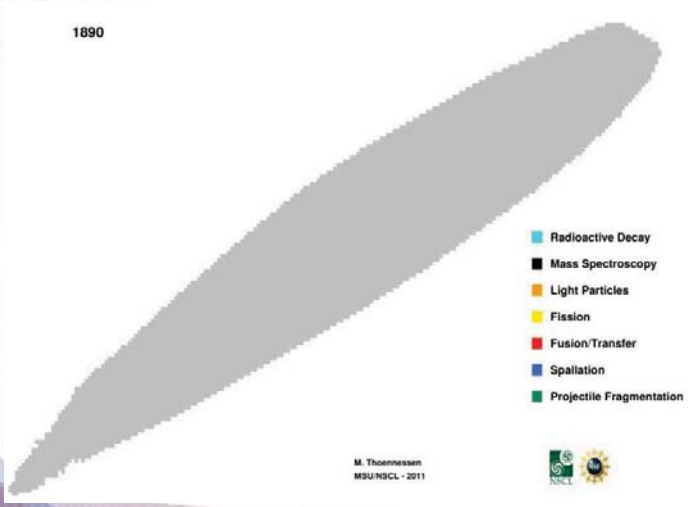
## 질량분석기의 개발과 안정한 동위원소의 발견

**Nobel Prize in Chemistry (1922)**

**"for his discovery, by means of his mass spectrograph, of isotopes, in a large number of non-radioactive elements, and for his enunciation of the whole-number rule."**



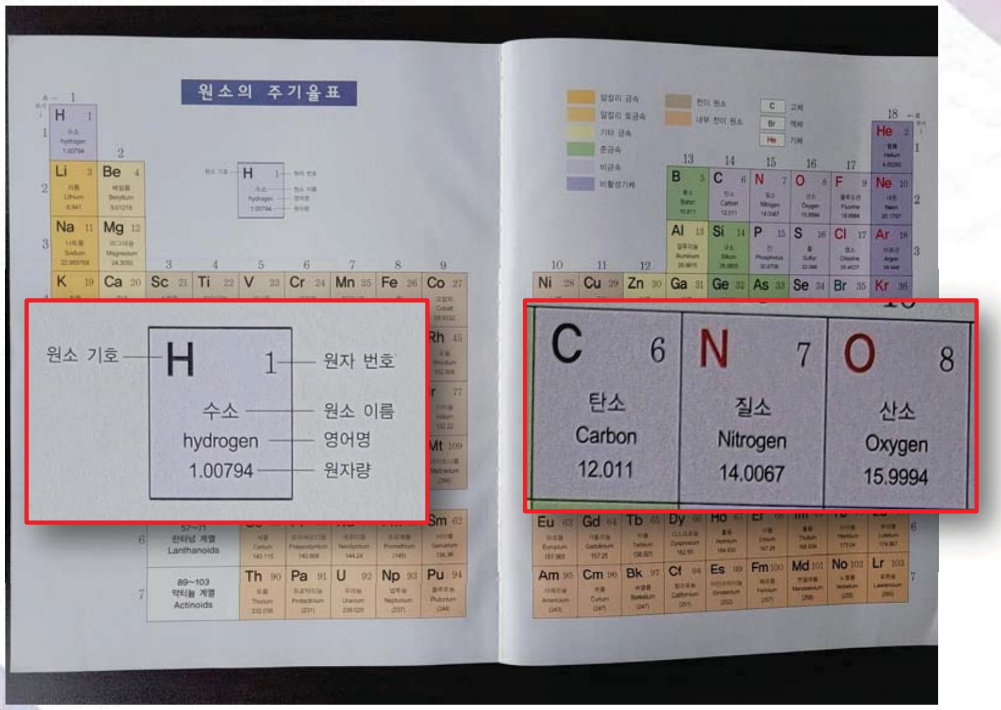
**Francis Aston (1877~1945)**




Identified
212 of the 287 naturally occurring isotopes

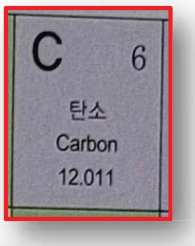


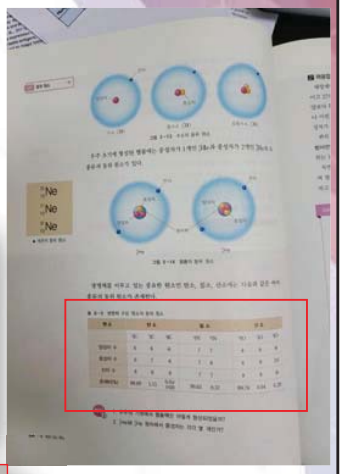
## 원자의 질량



## 탄소의 질량

- 탄소, carbon
  - (평균)원자량 12.011



교학사 화학 I

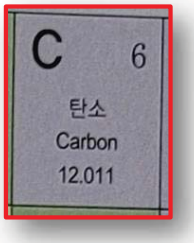
표 Ⅱ-3 생명체 구성 원소의 동위 원소

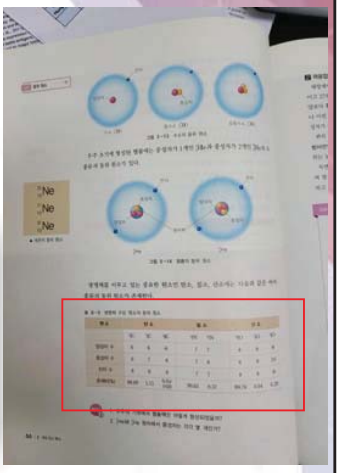
| 원소 | 탄소 | | | 질소 | | 산소 | | |
|---|---|---|---|---|---|---|---|---|
| | $^{12}_{6}C$ | $^{13}_{6}C$ | $^{14}_{6}C$ | $^{14}_{7}N$ | $^{15}_{7}N$ | $^{16}_{8}O$ | $^{17}_{8}O$ | $^{18}_{8}O$ |
| 양성자 수 | 6 | 6 | 6 | 7 | 7 | 8 | 8 | 8 |
| 중성자 수 | 6 | 7 | 8 | 7 | 8 | 8 | 9 | 10 |
| 전자 수 | 6 | 6 | 6 | 7 | 7 | 8 | 8 | 8 |
| 존재비(%) | 98.89 | 1.11 | 0.0$x$ 이하 | 99.63 | 0.37 | 99.76 | 0.04 | 0.20 |

7

## 탄소의 질량

- 탄소, carbon
  - (평균)원자량 12.011
  - $^{12}C$ - 98.89%
  - $^{13}C$ - 1.11%
  - $^{14}C$ - 거의 없음



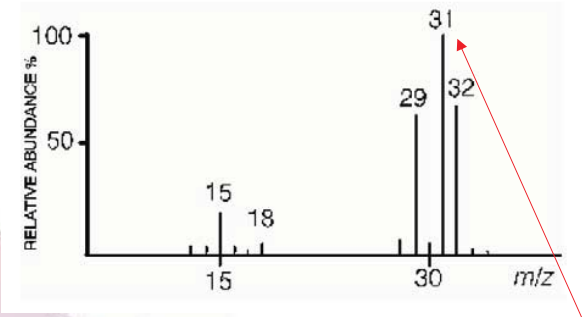
교학사 화학 I

탄소 원자의 (평균)원자량 = 12 X 0.9889 + 13 X 0.0111
= 12.0111

8

# 질량 스펙트럼

Spectrum




mass-to-charge

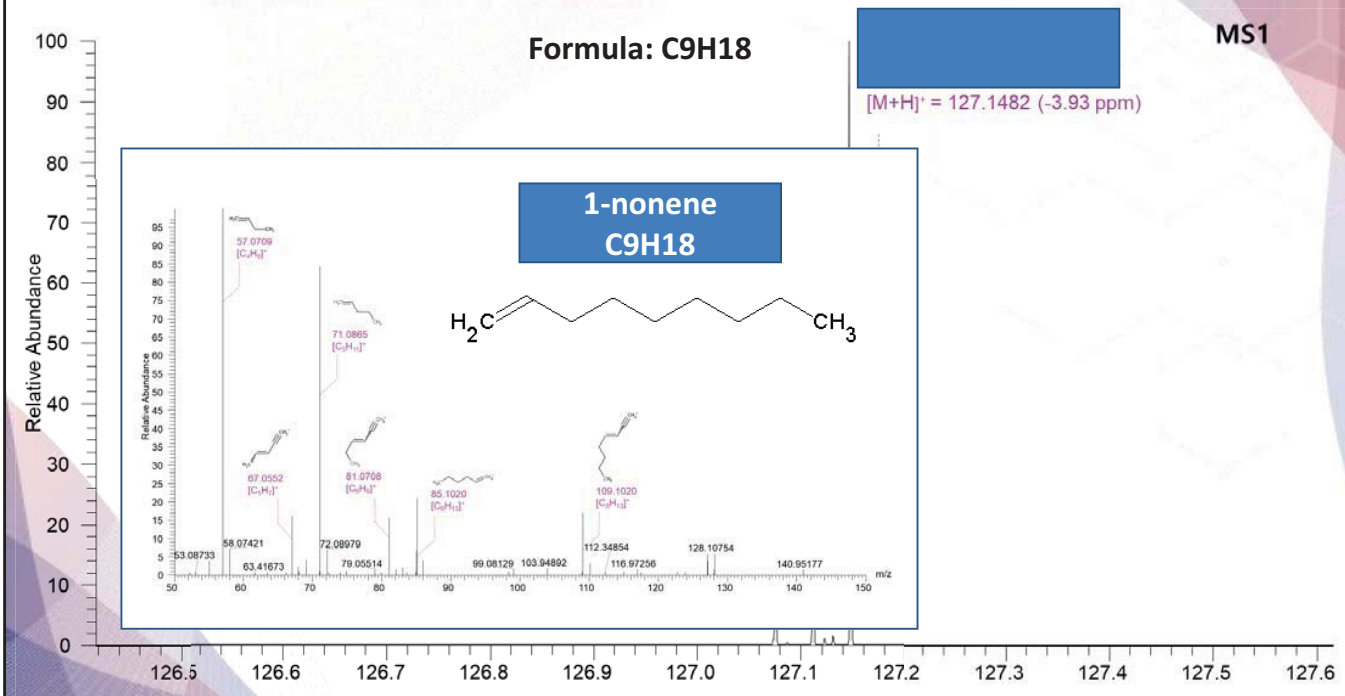
Mass table

| m/z | Relative abundance (%) | m/z | Relative abundance (%) |
|---|---|---|---|
| 12 | 0.33 | 28 | 6.3 |
| 13 | 0.72 | 29 | 64 |
| 14 | 2.4 | 30 | 3.8 |
| 15 | 13 | 31 | 100 |
| 16 | 0.21 | 32 | 66 |
| 17 | 1.0 | 33 | 0.73 |
| 18 | 0.9 | 34 | ~ 0.1 |

base peak

9

# 저분자 물질의 정성분석



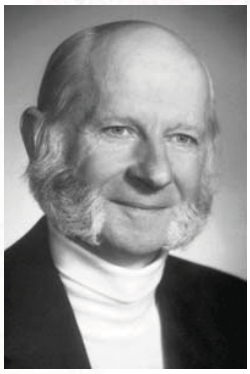

10

*Unpublished data*

## 기체 이온의 공간적 트랩핑 기술 개발

**Nobel Prize in Physics (1989)**

**"for the development of the ion trap technique."**



**Hans G. Dehmelt**
**(1913~2017)**

**Penning ion trap**

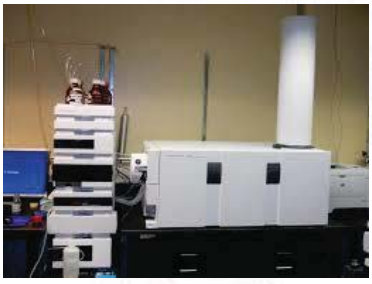
**Wolfgang Paul**
**(1913~1993)**

**Paul ion trap**

www.youtube.com

11

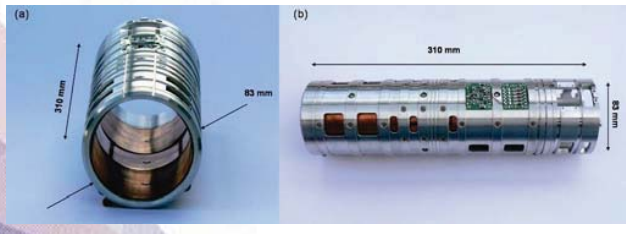
## 다양한 질량분석기



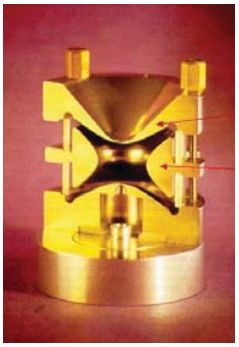
Time-of-Flight

Quadrupole

Orbitrap

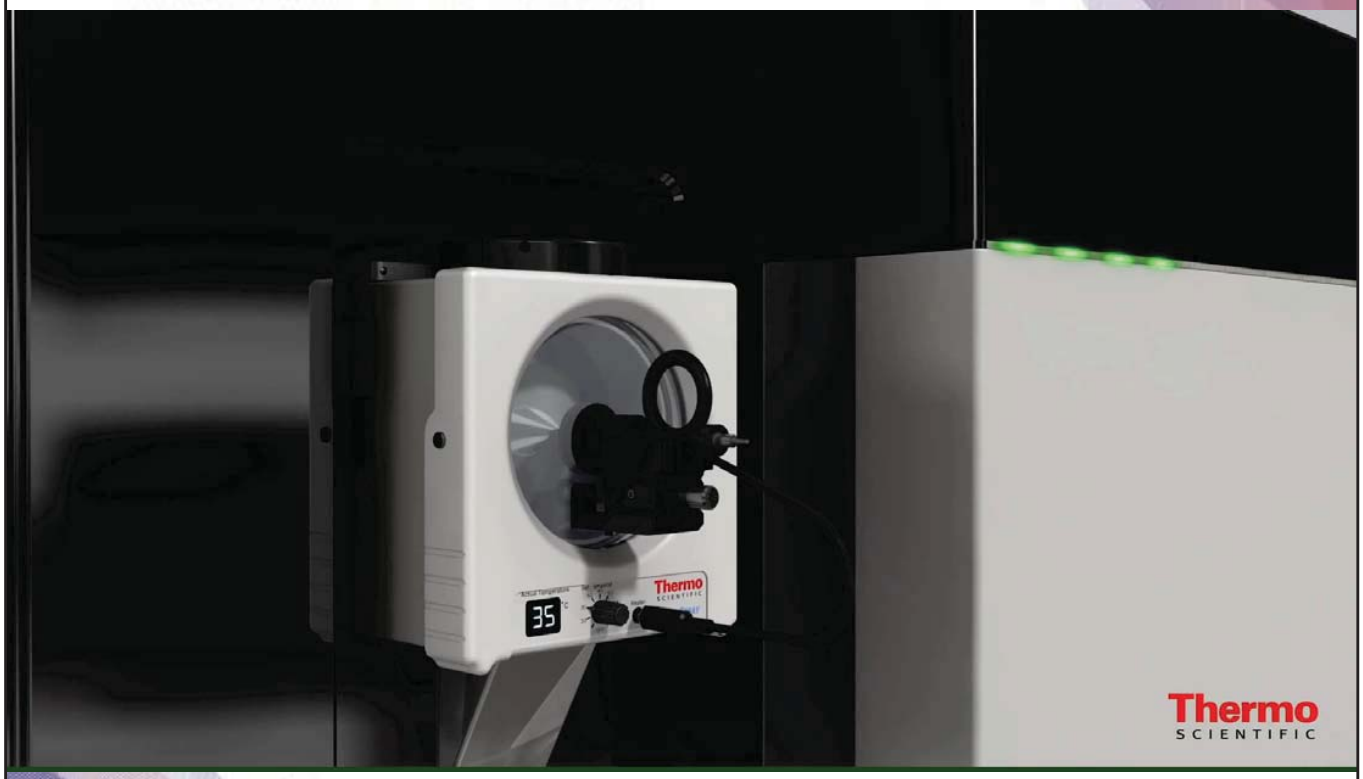



FT-ICR



Ion trap

12



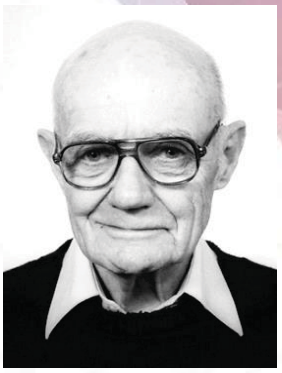

13

## 저에너지 이온화 기술 개발과 바이오 고분자 분석

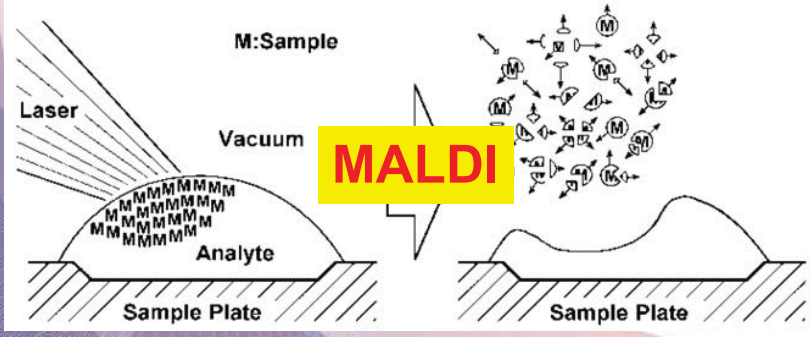
Koichi Tanaka
(1959 - )

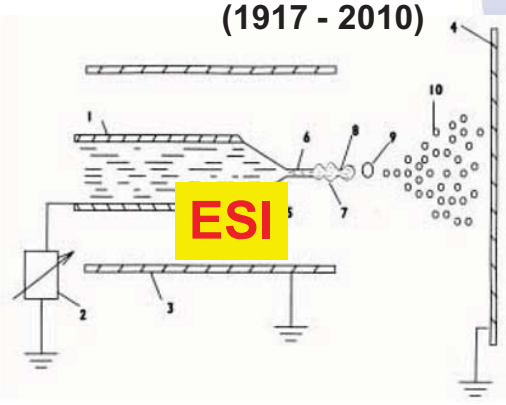
**Nobel Prize in Chemistry (2002)**

**"for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules"**

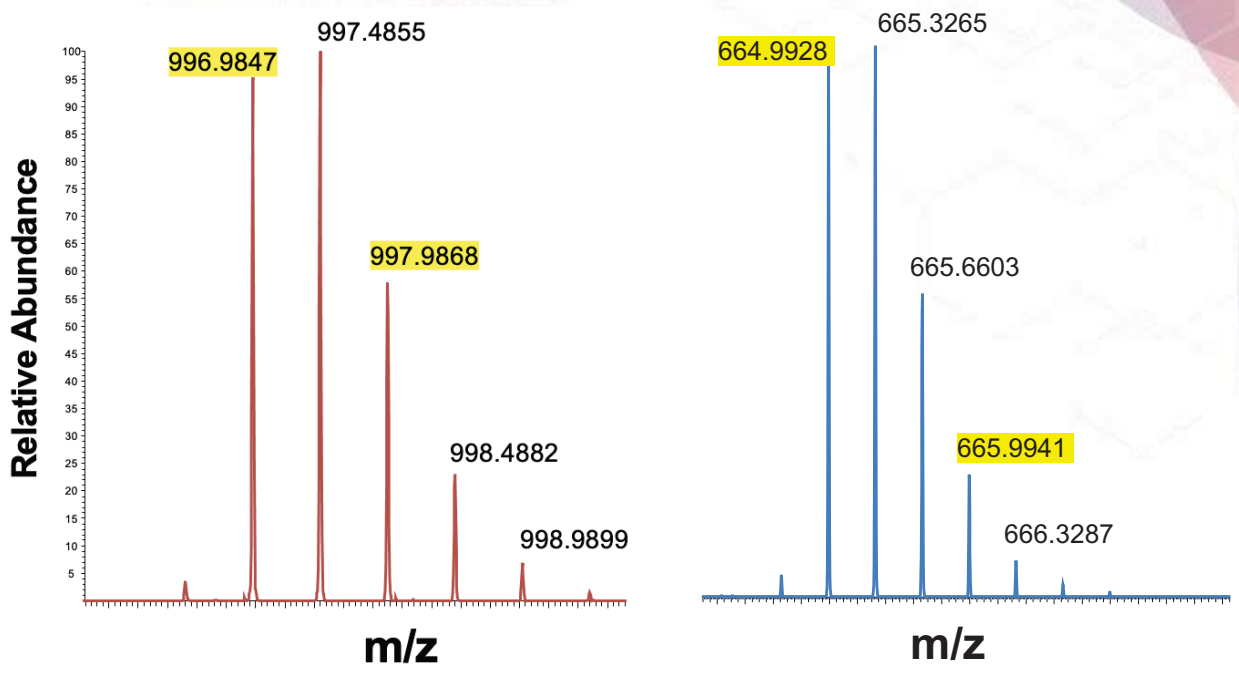


John B. Fenn
(1917 - 2010)



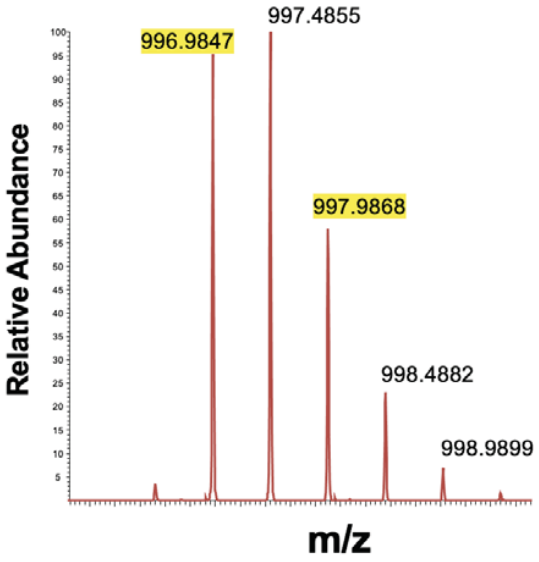



## 질량스펙트럼의 특성 – isotopic cluster (동위원소분포클러스터)




15

## 질량스펙트럼의 특성 – isotopic cluster (동위원소분포클러스터)

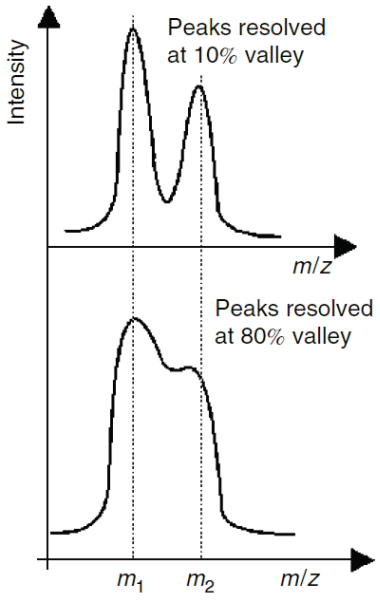



m/z = 996.9847
Charge (z) = +2

→(m) = 996.9847 x (+2) = Peptide + $2H^+$
→ precursor mass
= 996.9847 x (+2) – 2 x 1.0078
= 1991.9538

16

## 질량스펙트럼의 특성 – resolution(분해능)



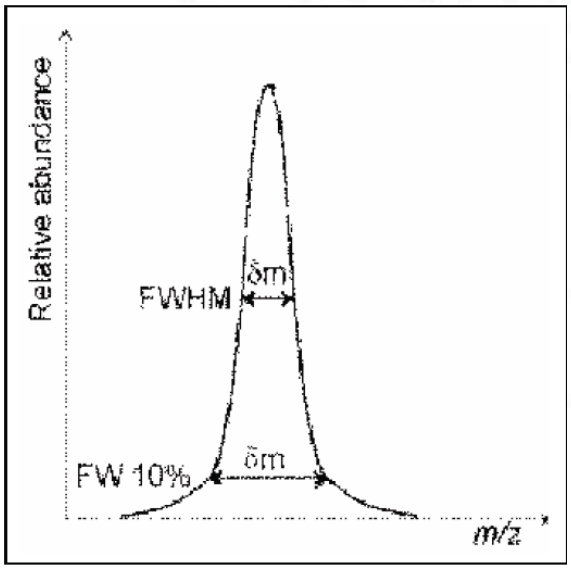
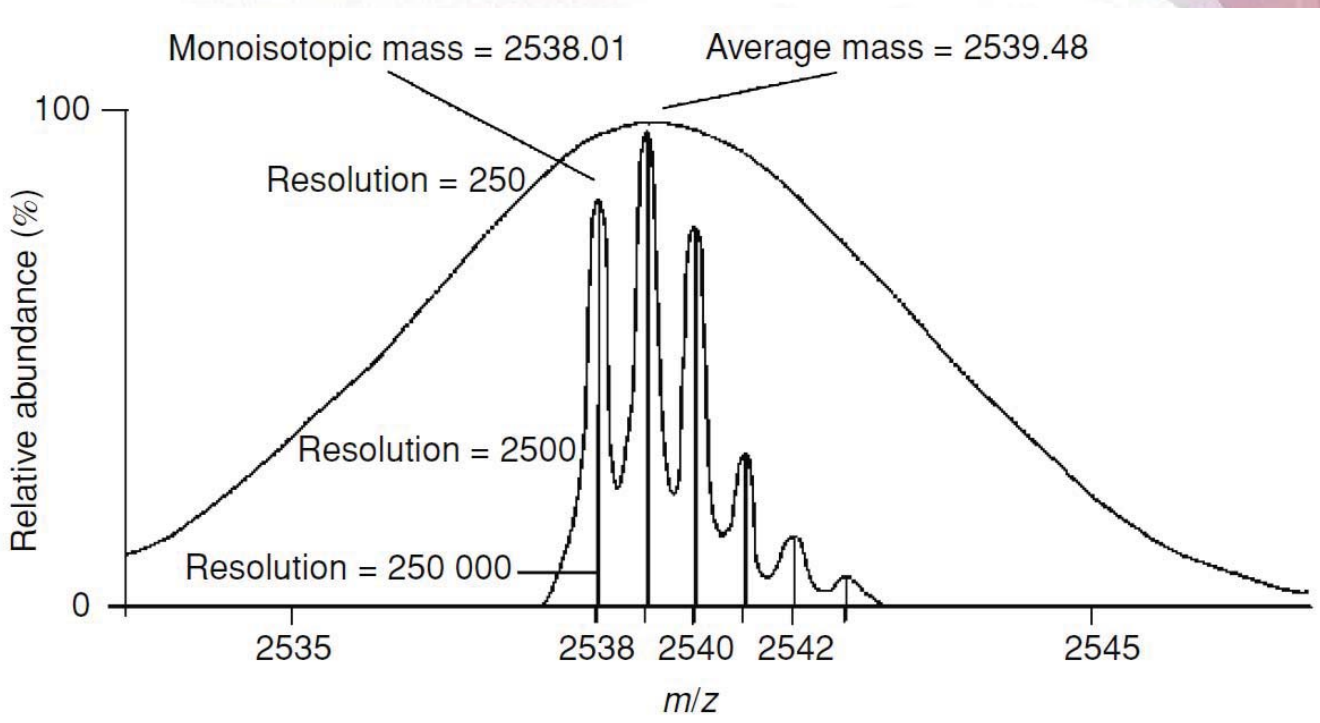




**Figure 2.1**
Diagram showing the concepts of peak resolution and valley.

17

## 질량스펙트럼의 특성 – resolution(분해능)



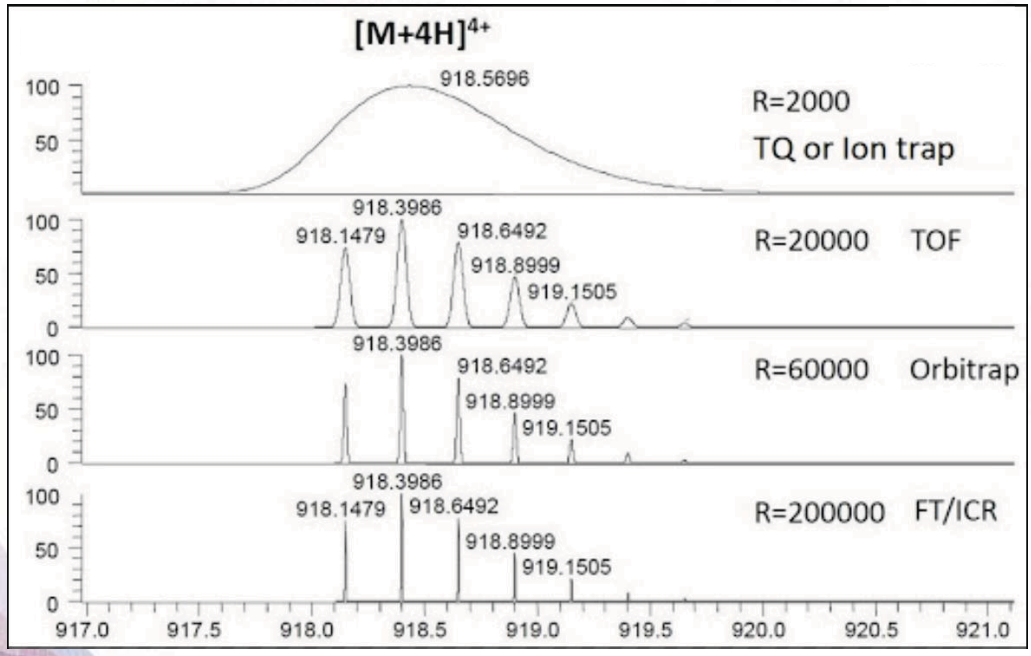

18

## 분해능의 발전

| | $m/\delta m$ | |
|---|---|---|
| 1913 | 13 | Thomson |
| 1918 | 100 | Dempster |
| 1919 | 130 | Aston |
| 1937 | 2000 | Aston |
| 1998 | 8 000 000 | Marshall and co-workers |

19

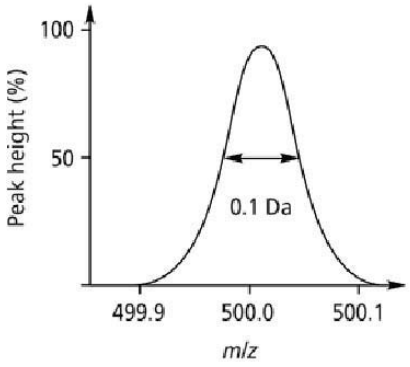
## 질량분석기별 질량분석 스펙트럼 분해능 차이




20

## 질량분석스펙트럼의 특성 - mass measurement accuracy(질량측정정확도)

True mass = 400.0000
Measured mass = 400.0020
Difference = 0.0020 or 2 mmu
Error = $\frac{0.002}{400} \times 10^6 = 5$ ppm

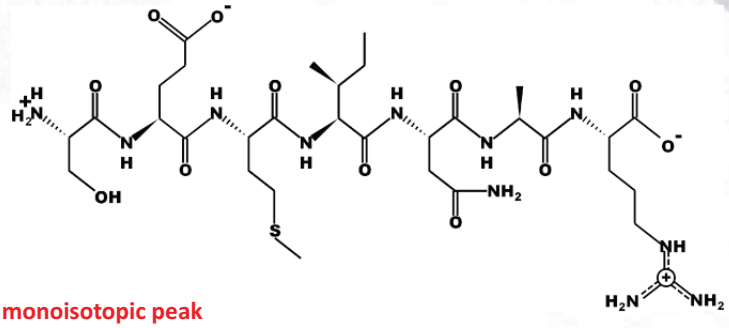



Mass = 500
Peak width (at 50%) = 0.1
Resolution (FWHM) $\frac{500}{0.1} = 5000$

21

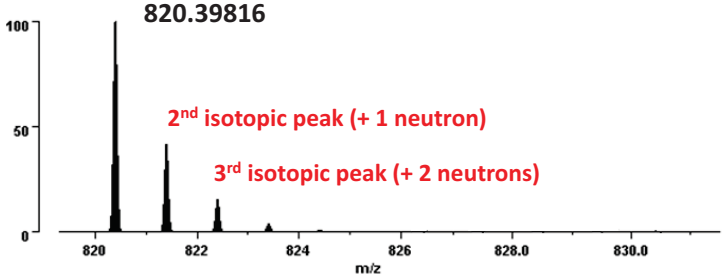
# Peptide Sequencing

## Example peptide



monoisotopic peak

820.39816



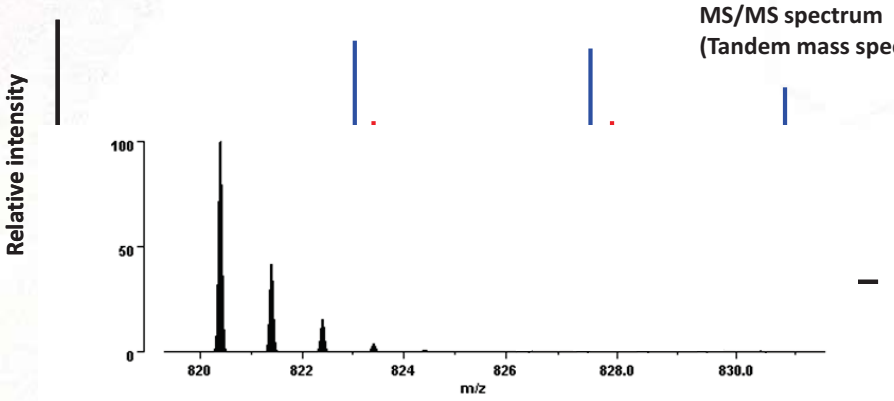

2nd isotopic peak (+ 1 neutron)

3rd isotopic peak (+ 2 neutrons)

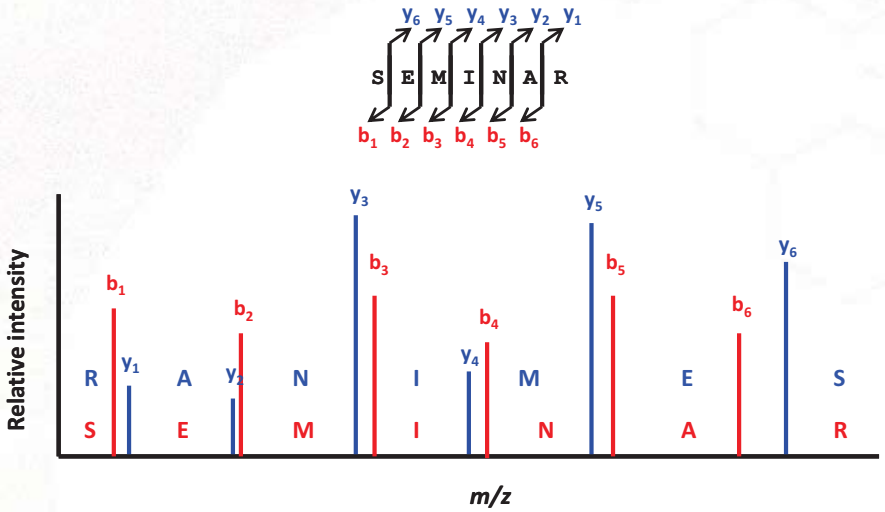


Gas phase dissociation

MS/MS spectrum
(Tandem mass spectrum)

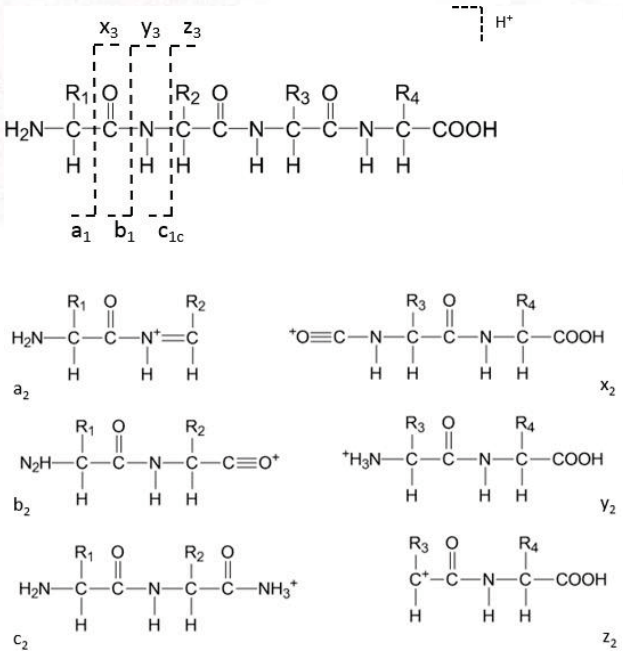


## Fragment ion assignments




25

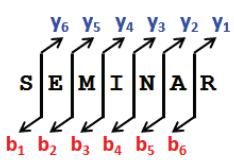
## Nomenclature of peptide fragment ions



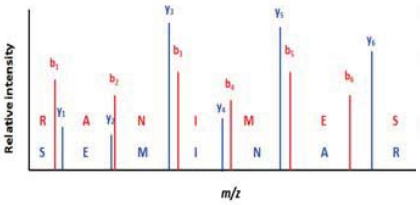

26

## PSM (Peptide-Spectrum Match)

Theoretical precursor mass




Observed precursor mass




Theoretical fragment ion mass list

217.0819
348.1224
461.2064
575.2494
646.2865
---

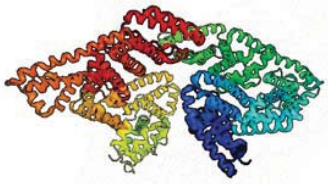
SCORE

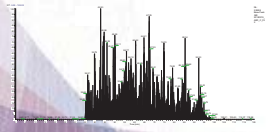
Measured fragment ion mass list

217.0823
348.1211
461.2072
575.2490
646.2875
---



## Bottom-Up Proteomics





MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLIA
FSQYLQQCPFDEHVK**LVNELTEFAK**TCVADESHAGCEKSLHTLFGDELCK
VASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPK**LKPDPNTLCDEF**
**K**ADEKKFWGKYLYEIARR**HPYFYAPELLYYANK**YNGVFQECCQAEDKGAC
LLPKIETMREKVLASSARQRLRCASIQKFGERALKAWSVARLSQKFPKAE
FVEVTK**LVTDLTK**VHK**ECCHGDLLECADDR**ADLAKYICDNQDTISSKLKE
CCDKPLLEKSHCIAEVEKDAIPENLPPLTADFAEDK**DVCKNYQEAK**DAFL
GSFLYEYSRR**HPEYAVSVLLRLAK**EYEATLEECCAKDDPHACYSTVFDKL
KHLVDEPQNLIKQNCDQFEKLGEYGFQNALIVR**YTRKVPQVSTPTLVEVS**
**RSLGK**VGTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCC
TESLVNRRPCFSALTPDETYVPKAFDEK**LFTFHADICTLPDTEK**QIKKQT
ALVELLKHKPKATEEQLKTVMENFVAFVDKCCAADDKEACFAVEGPKLVV
STQTALA

ECCHGDLLECADDR

LVNELTEFAK

DVCKNYQEAK

HPYFYAPELLYYANK

HPEYAVSVLLRLAK

LFTFHADICTLPDTEK

Data analysis

## Top-Down Proteomics

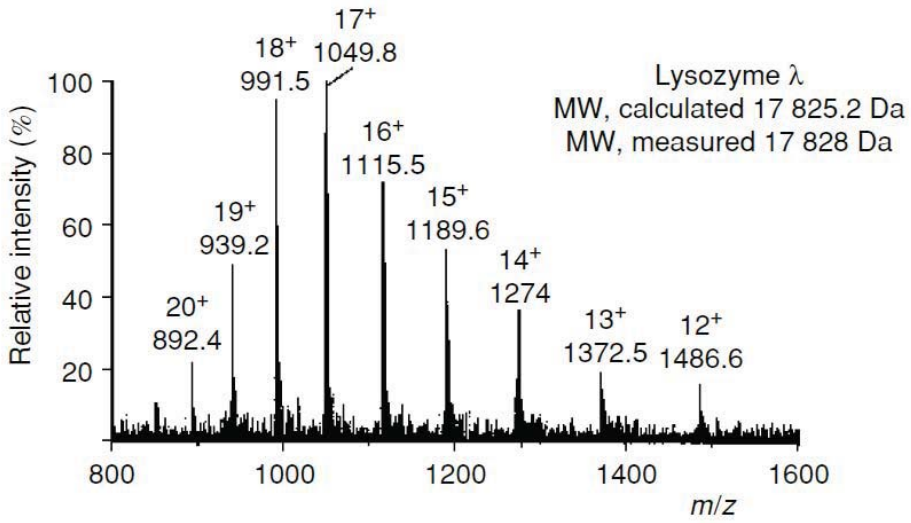
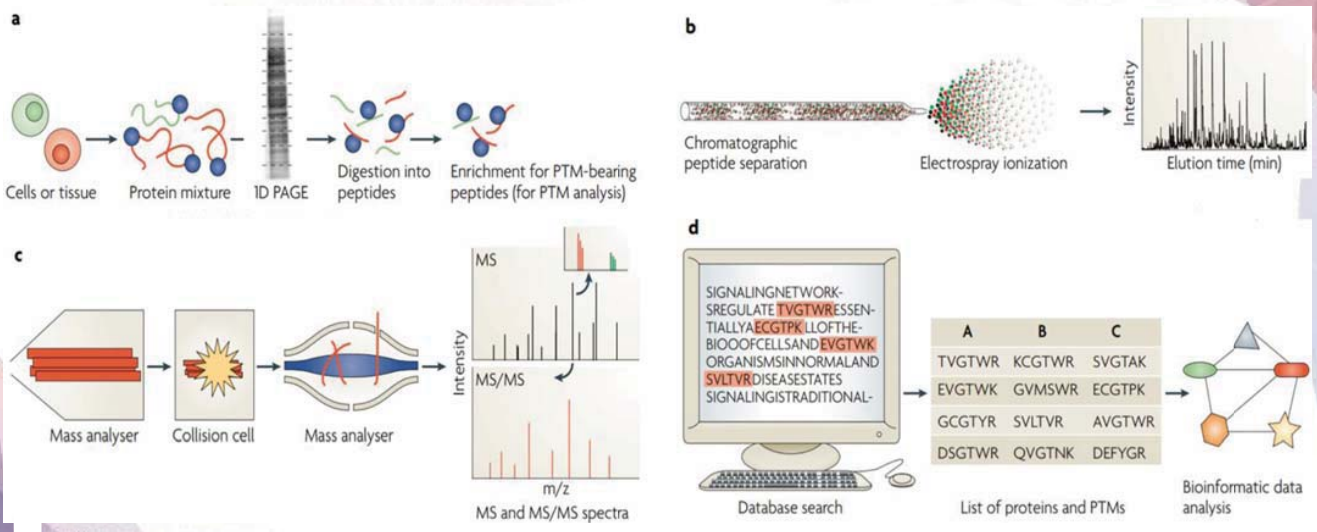



**Figure 1.23**
ESI spectrum of phage λ lysozyme; *m/z* in Th and the number of charges are indicated on each peak. The molecular mass is measured as being 17 828 ± 2.0 Da.

29

## General Proteomics Workflow




Choudhary and Mann, *Nat Rev Mol Cell Biol*, 2010

30

# Quantitation

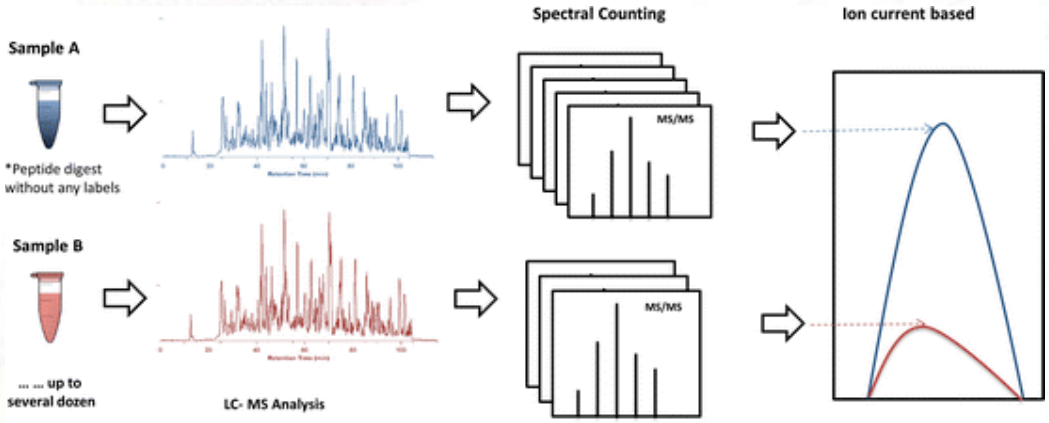
## Choice of Quantitation in Proteomics

- Label
  - Metabolic labeling
  - Chemical labeling
- Label-free
  - # spectral counting (ex. # PSM)
  - LC profile

- Model samples (ex, cell)
- Clinical samples (ex, blood)

32

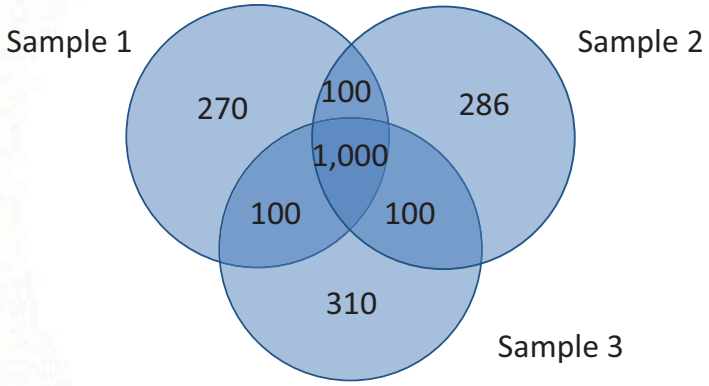
## Label-free (spectral counting or ion current area)




33

## Label-free (spectral counting or ion current area)

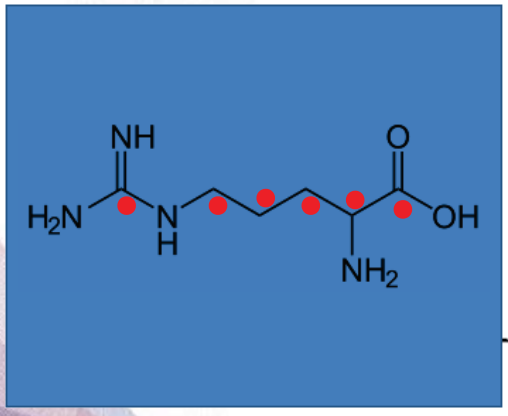
- Sample 1 : 10 ug peptides -> 1,000,000 MS/MS -> 400,000 PSMs
- Sample 2 : 10 ug peptides -> 1,100,000 MS/MS -> 380,000 PSMs
- Sample 3 : 10 ug peptides -> 950,000 MS/MS -> 390,000 PSMs

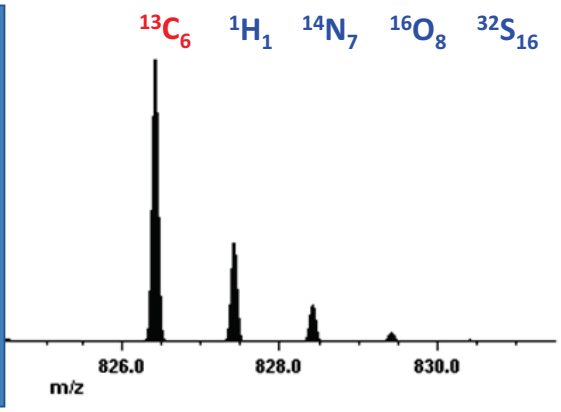



34

## Metabolic labeling

S E M I N A R*



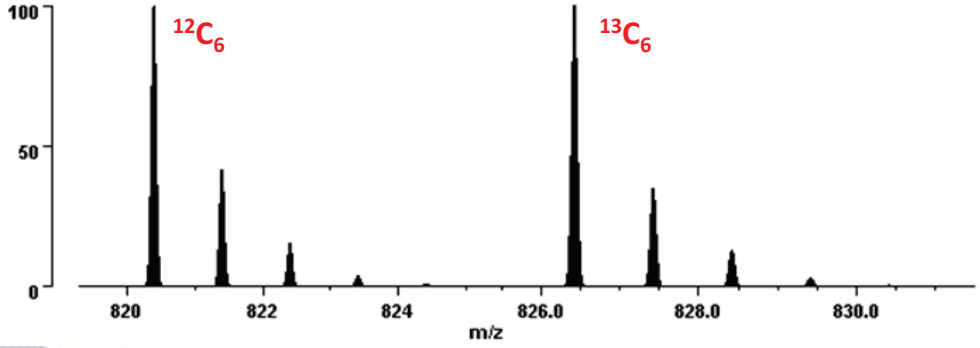

35

## SILAC-based Quantitative proteomics

SEMINAR:**SEMINAR*** = 1:1

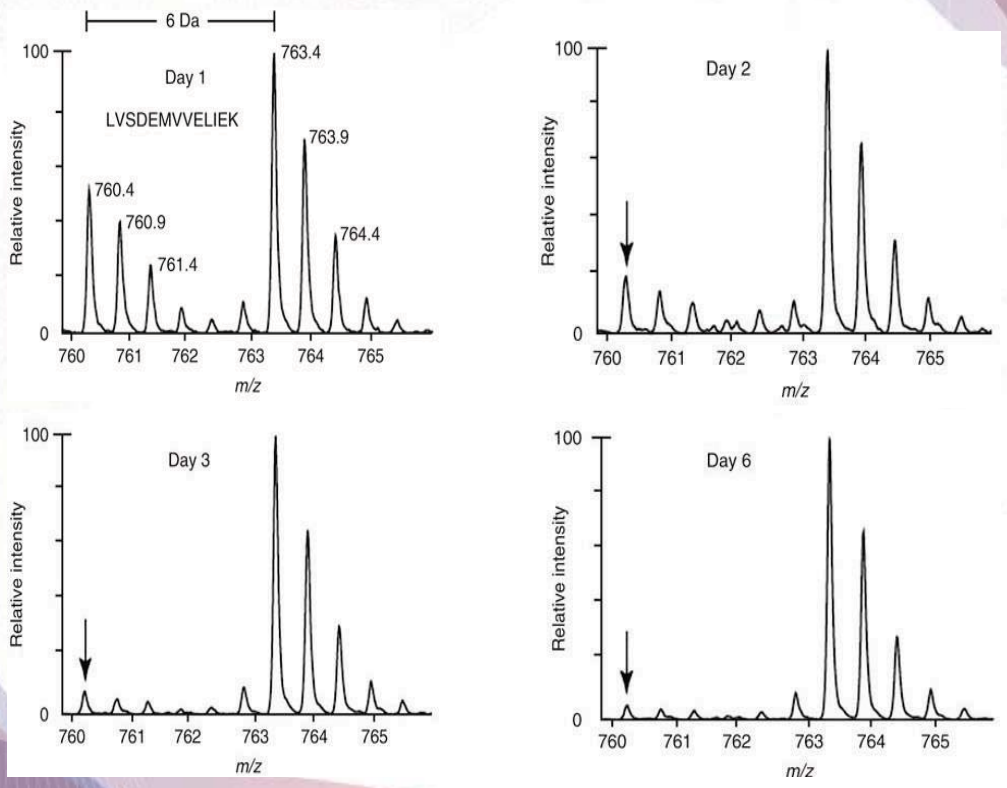
SEMINAR

SEMINAR*




36

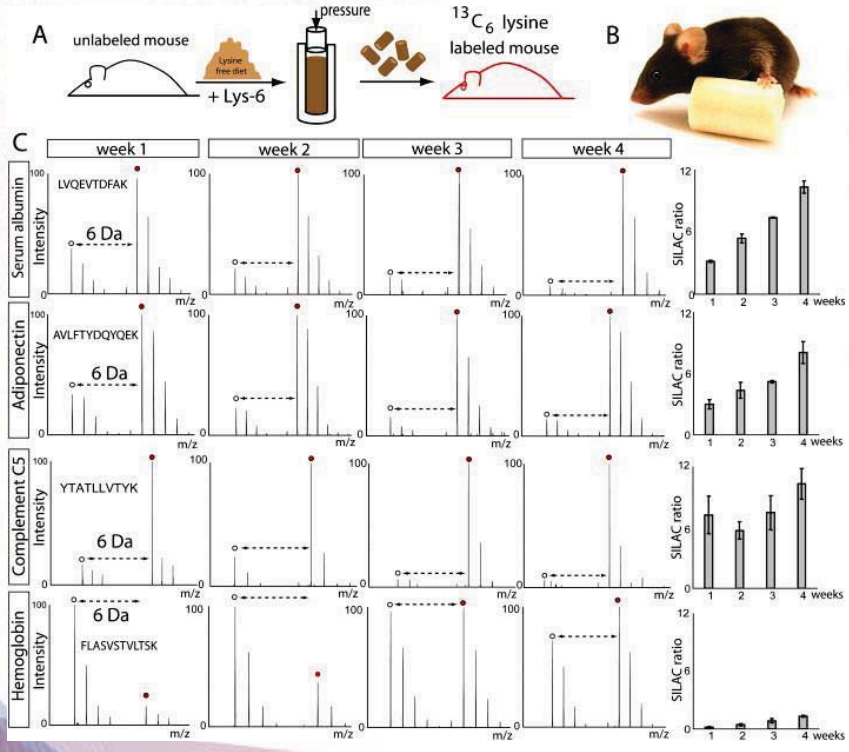
## Metabolic labeling of the whole cellular proteome




Gowda et al. *Nature Protocol*

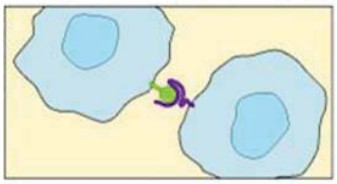
37

## In vivo SILAC labeling

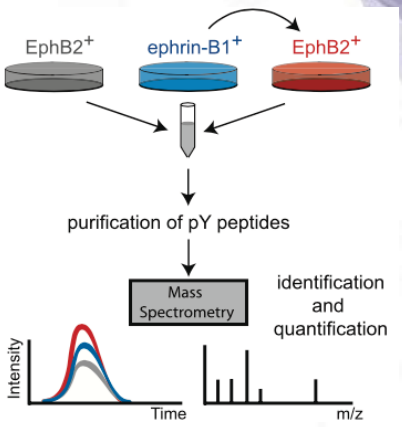


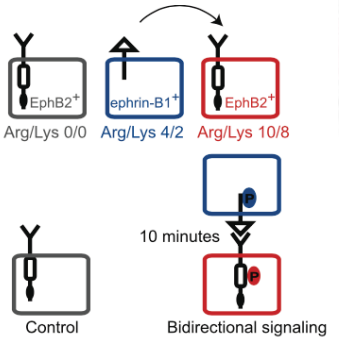

Mann et al. *Cell*

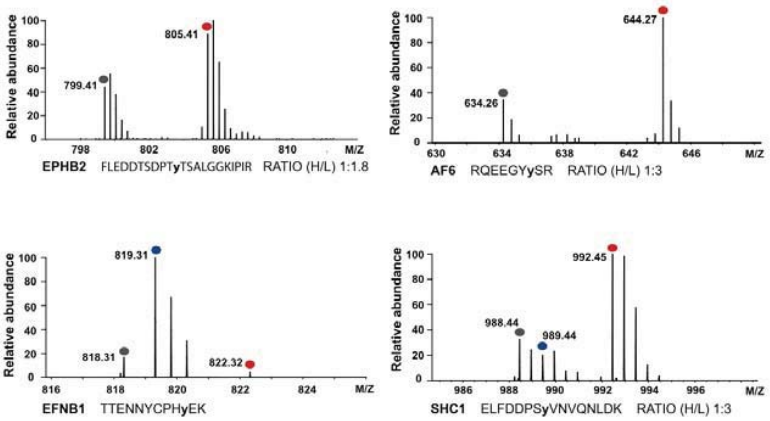
38

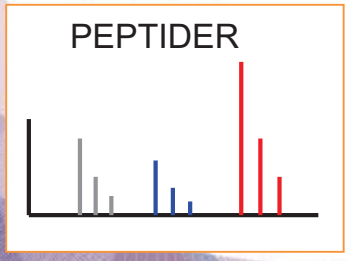


*Juxtacrine*



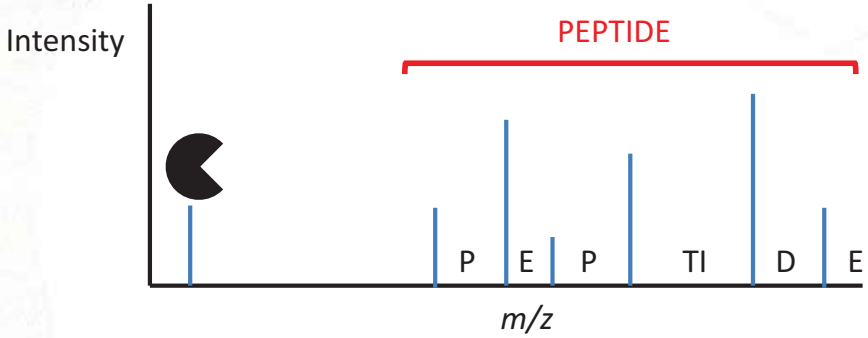



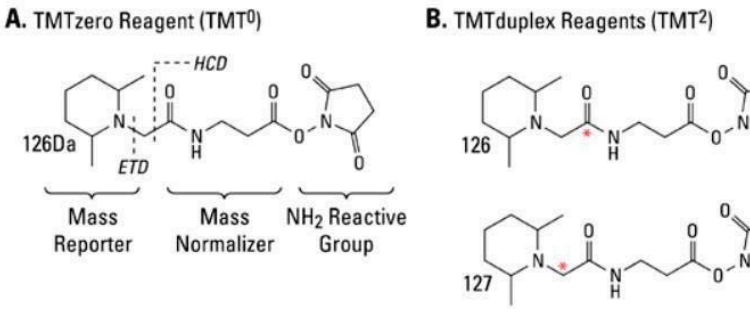

Jyrgensen et al. *Science*



## Isotope-coded chemical labeling to clinical samples

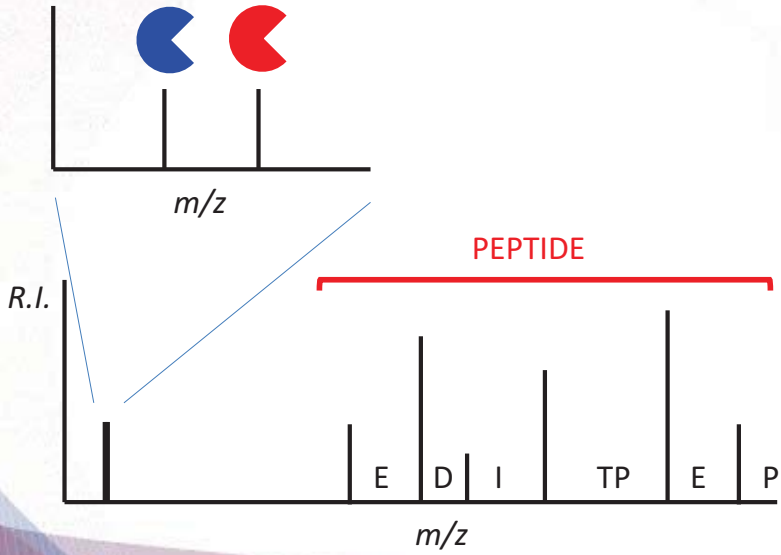
## TMT-based quantitative proteomics





41

## TMT-based quantitative proteomics


42

## Summary

- 질량분석의 기초 원리
- 질량분석 데이터의 핵심 요소
- Peptide sequencing 의 기초 원리
- Peptide 정량 기술

43

- Principle of Mass Spectrometry and Basics of Proteomics
- Applications to different research fields


44