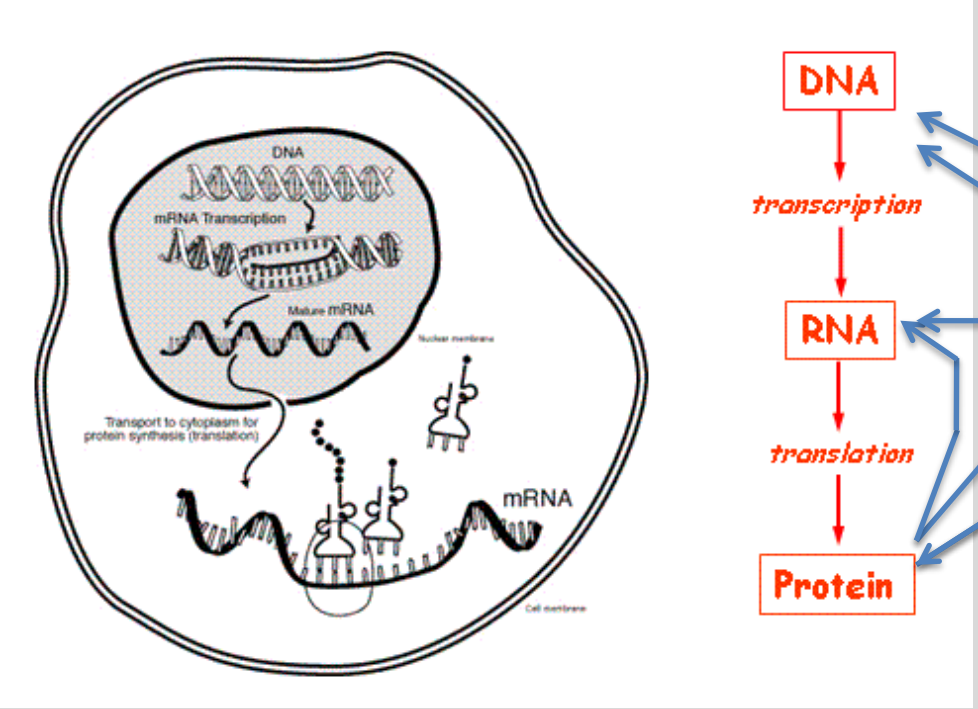
# Introduction to mass spectrometry (MS) – based proteomics and metabolomics

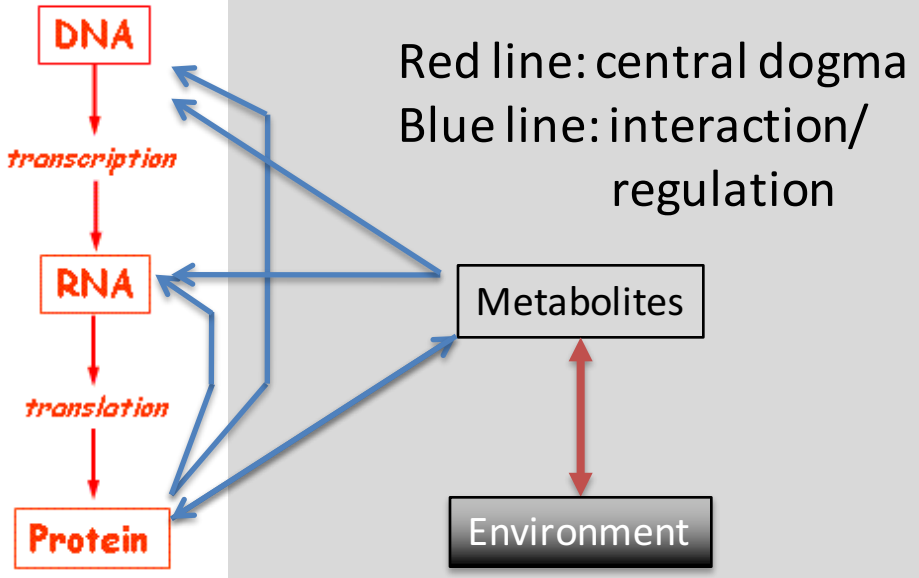
Tianwei Yu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

September 10, 2015

# Background

## High-throughput profiling of biological samples




Red line: central dogma
Blue line: interaction/ regulation

(Picture edited from http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/)

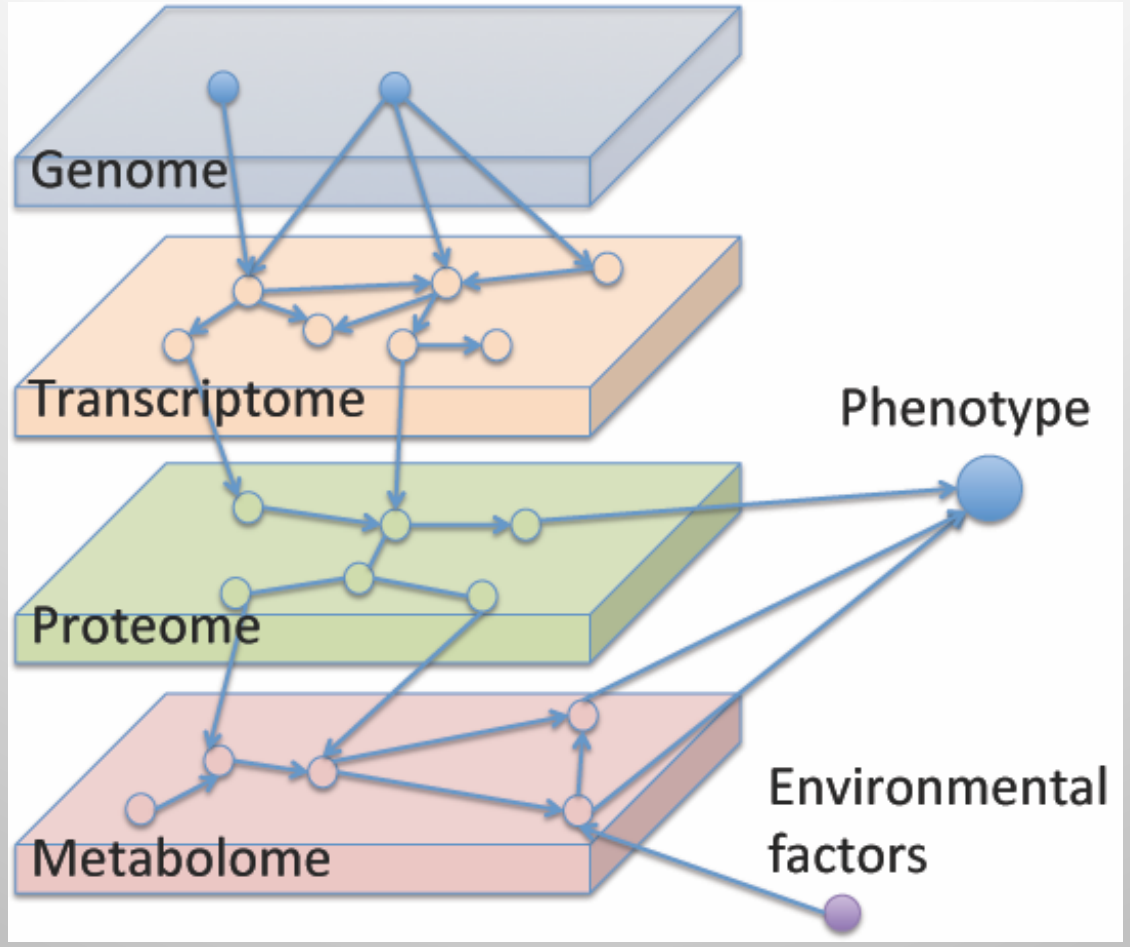
**Genome:** genotype, copy number, epigenetics …

**Transcriptome:** mRNA expression levels, alternative splicing, microRNA, lincRNA …

**Proteome:** protein concentration, modification, interaction …

**Metabolome:** metabolite concentration, dynamics, environmental chemicals …

# Background

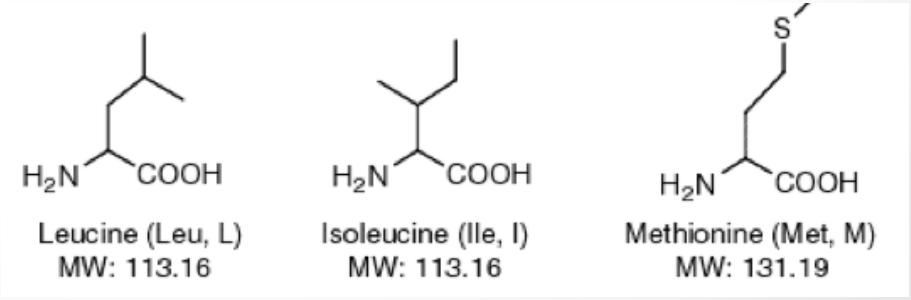

Genome
Transcriptome
Proteome
Metabolome
Phenotype
Environmental factors

# Background

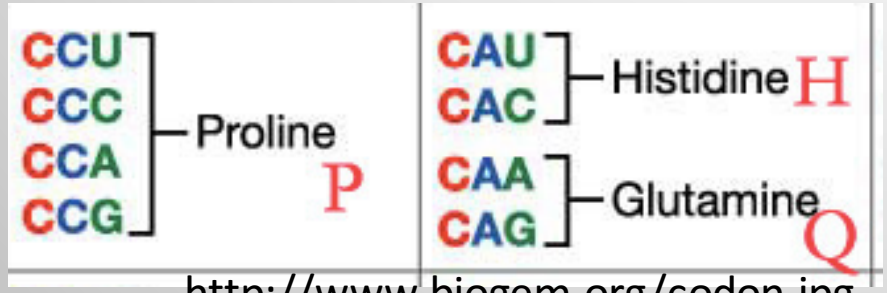
Proteins:

Made of 20 amino acids

http://matznerd.com/amino-acids/



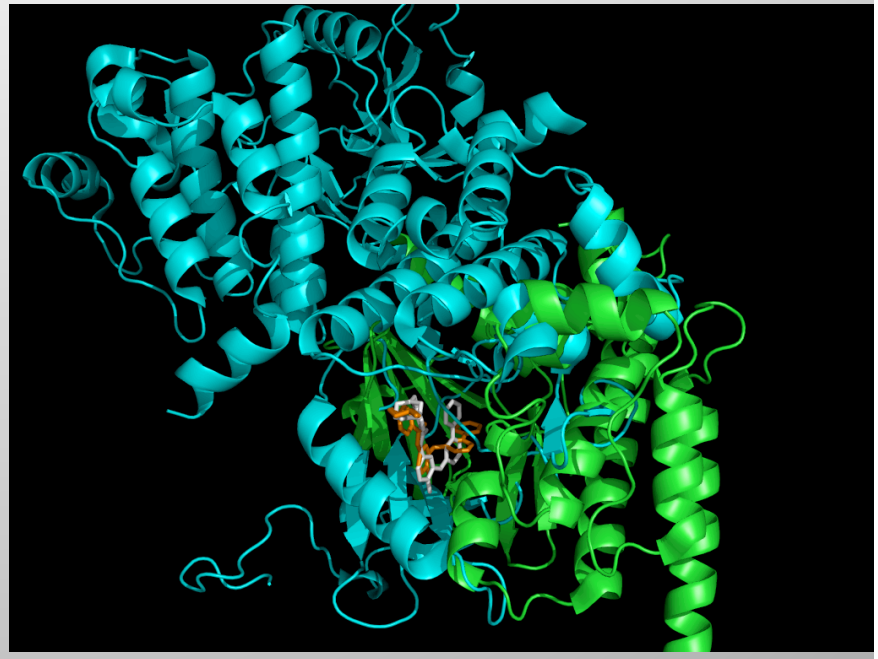

Sequence coded in DNA (Codon)




http://www.biogem.org/codon.jpg

Folded into 3 dimensional structures

Carry out biological functions like machines.



http://compbio.cs.toronto.edu/ligalign/desc.html

# Why Mass Spectrometry

Proteins/metabolites are harder to measure than DNA/RNA

- proteins are structured
- metabolites are small
- neither can be identified/quantified using sequence-based methods

mRNA measurements by RNAseq and microarrays are indeed proxy measurements of protein levels

# Why Mass Spectrometry

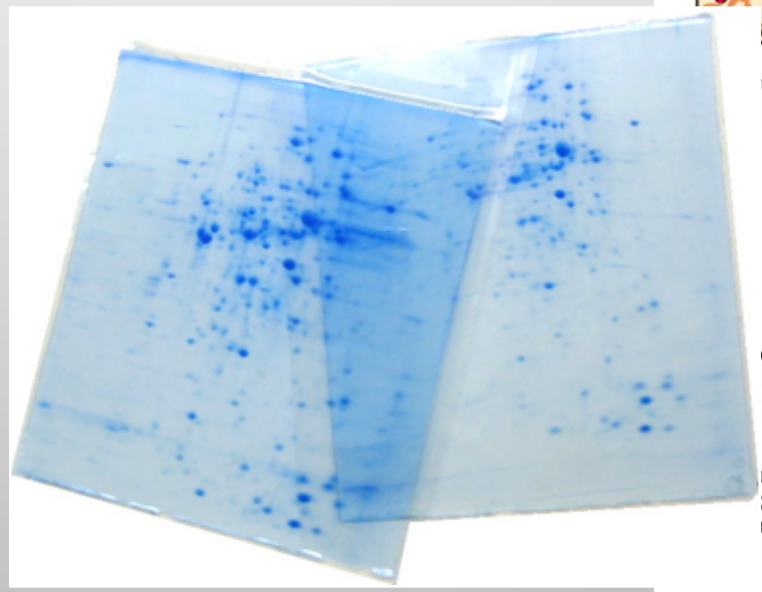
Proteins/metabolites could be separated according to their properties:

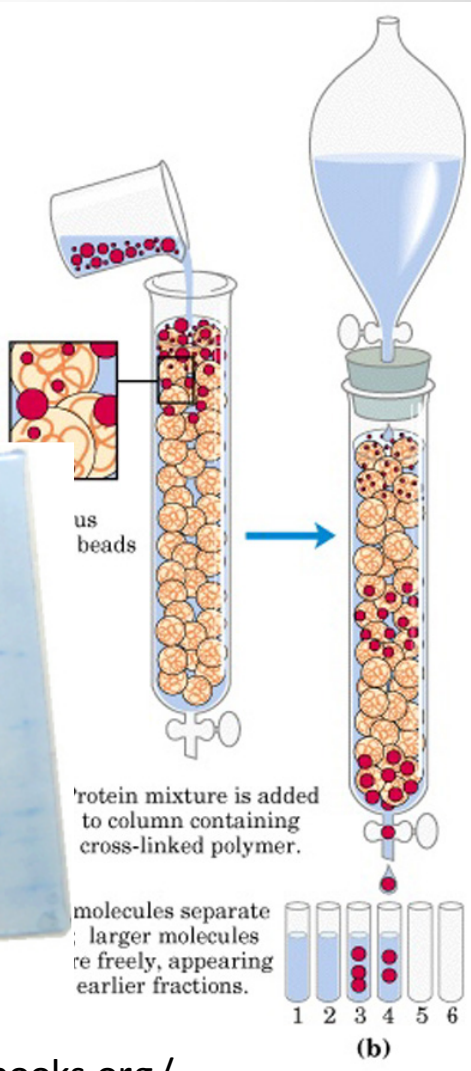
- mass/size
- hydrophilicity/hydrophobicity
- binding to specific ligands
- charge
- … … … …

Using

- Chromatography
- Electrophoresis
- … … … …

But these techniques cannot identify what is in the mixture






http://en.wikibooks.org/

# Why Mass Spectrometry

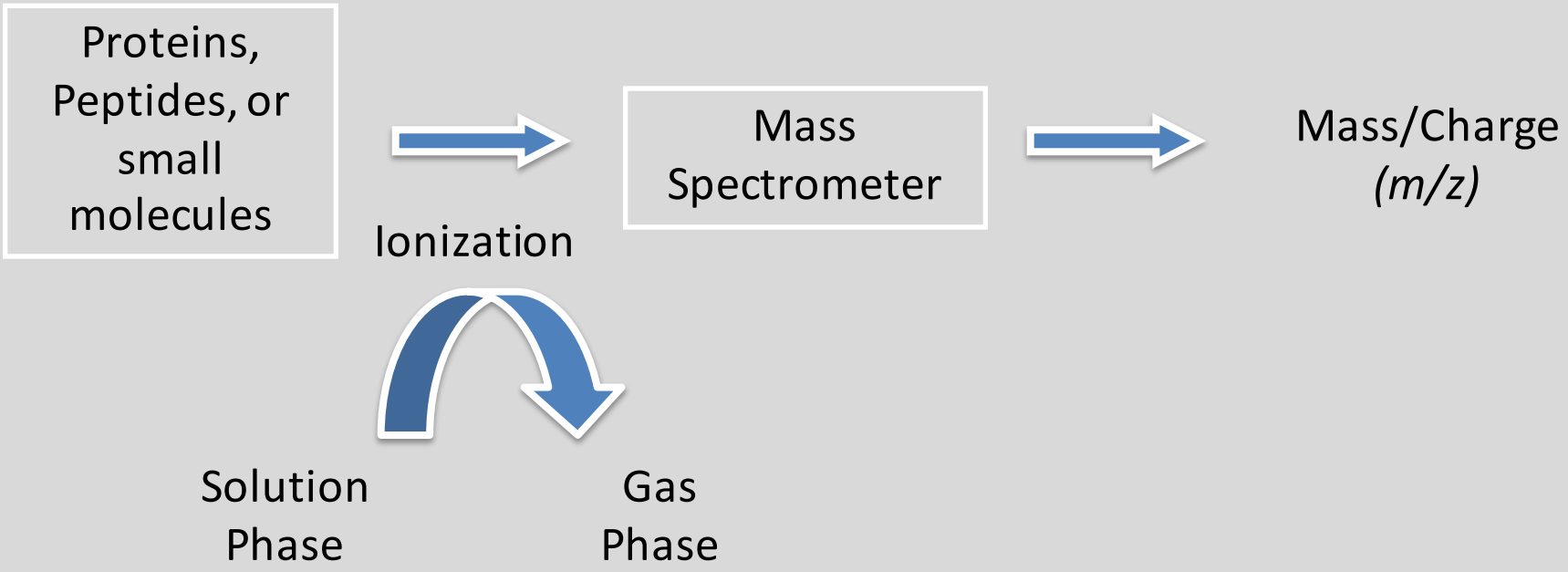
Problems with separation techniques:

- Reproducibility
- Identification / Quantification
- Inability to separate tens of thousands of iion species

Mass Spectrometry:

- Measurements based on ***mass/charge ratio (m/z)***
- Highly accurate, highly reproducible measurements
- Theoretical values easy to obtain → identification
- Can study protein modifications (small ligands attached)

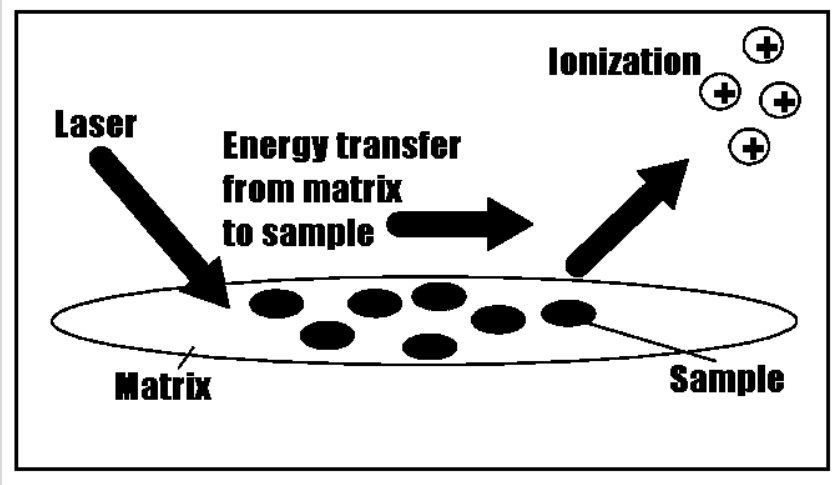
# Principle of Mass Spectrometry (MS)




Kindly provided by Dr. Junmin Peng

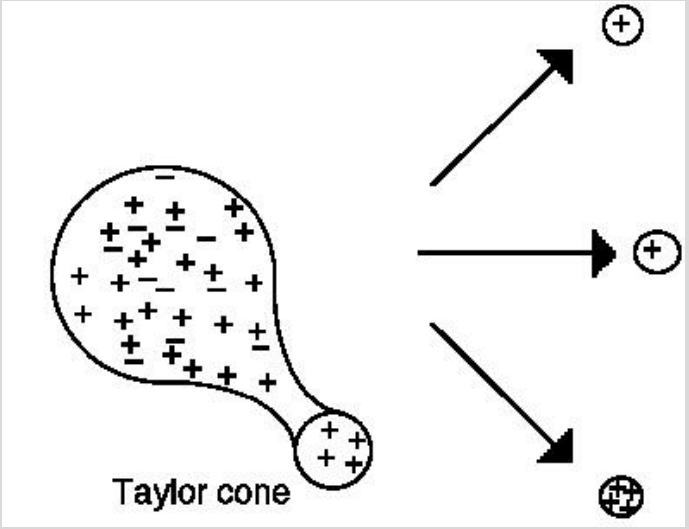
# Mass Spectrometry --- getting ion from solution to gas phase

Picture provided by Prof. Junmin Peng (Emory)




Matrix assisted laser desorption ionization (MALDI)

Electrospray ionization (ESI)

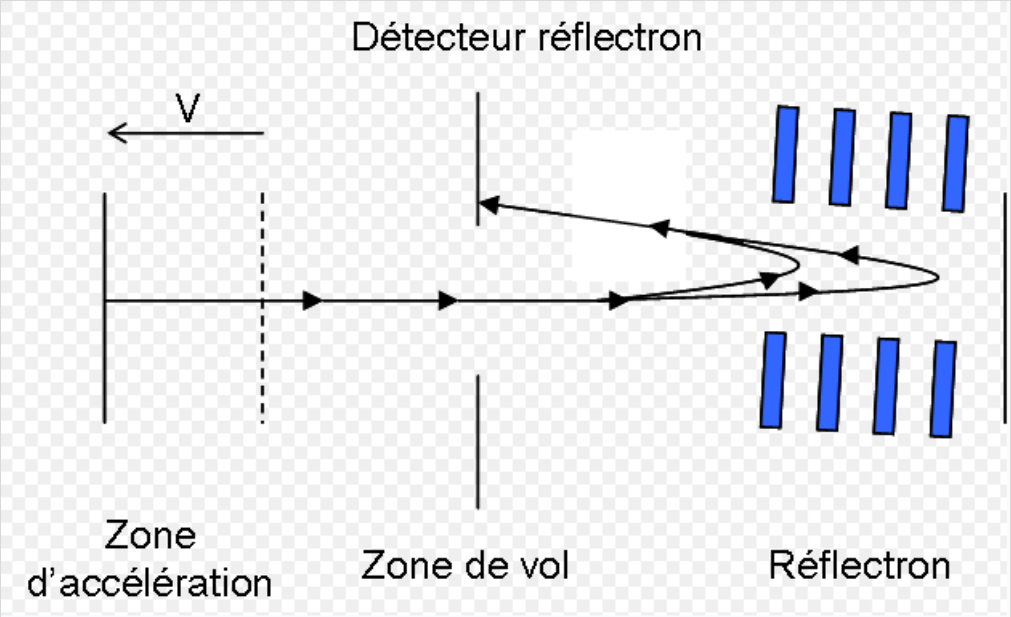


# Mass Spectrometry --- finding m/z

Time-of-flight:
Putting a charged particle in an electric field, the time of flight is

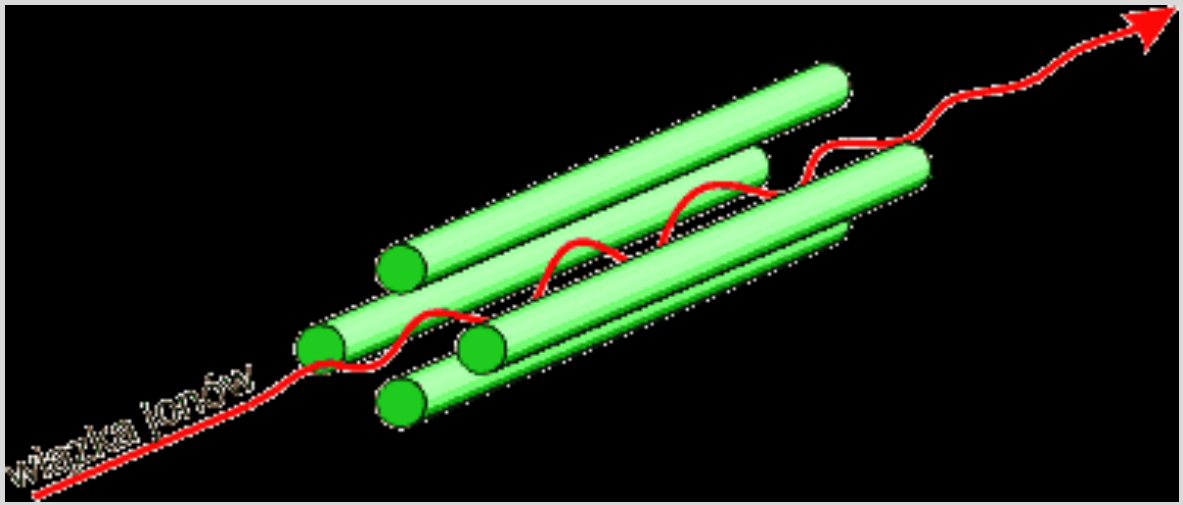
$$t = k\sqrt{\frac{m}{z}}$$

k: a constant related to instrument characteristics



# Mass Spectrometry --- finding m/z

Quadrupole:
Radio-frequency voltage applied to opposing pair of poles.
Only ions with a specific m/z can pass to the detector at each frequency.



# Mass Spectrometry --- finding m/z

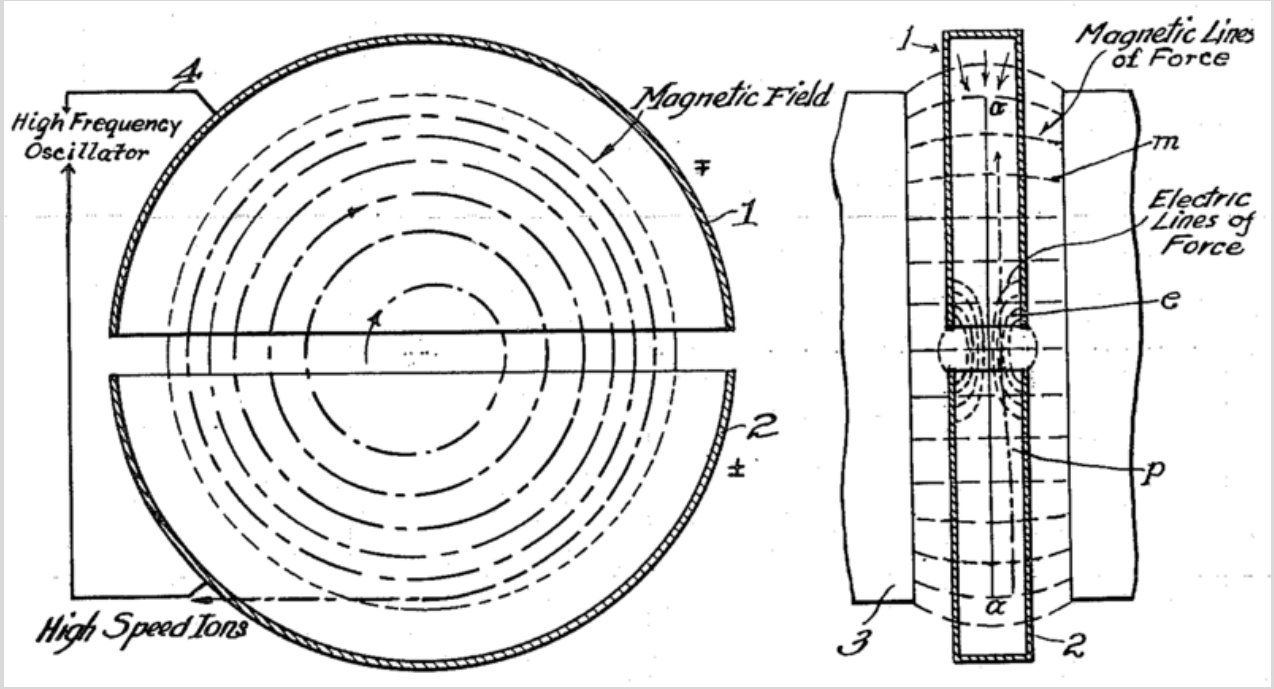
Fourier transform MS.

Ions detected not by hitting a detector, but by passing by a detecting plate. Ions detected simultaneously.

Very high resolution.

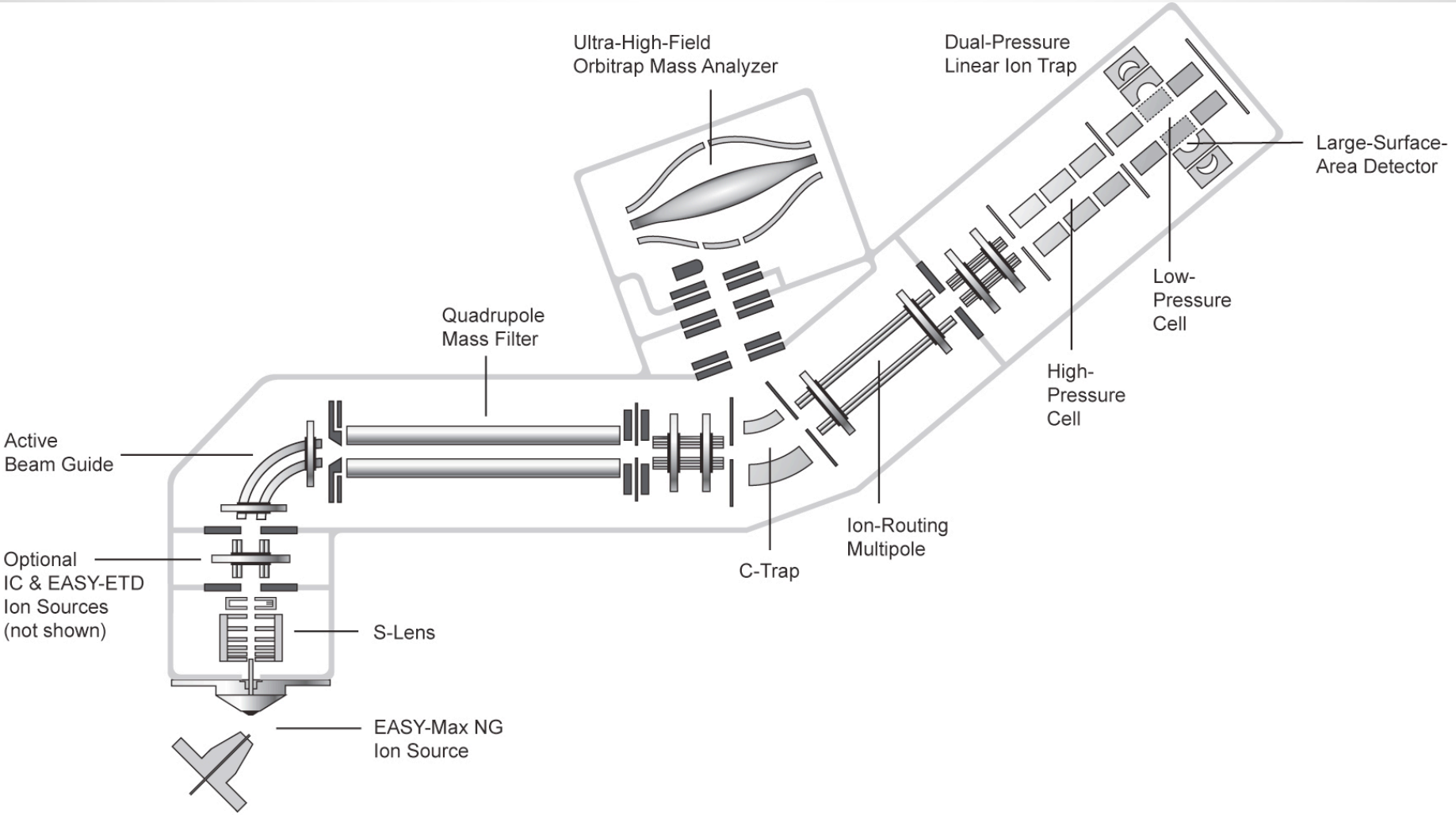
m/z detected based on the frequency of the ion in the cyclotron.

$$f \propto \frac{z}{m}$$



# Mass Spectrometry --- finding m/z

## High resolution Orbitrap Fusion



# Metabolomics - Background

Why profile metabolites?

⊙Enzymatic diseases directly influence metabolite concentrations

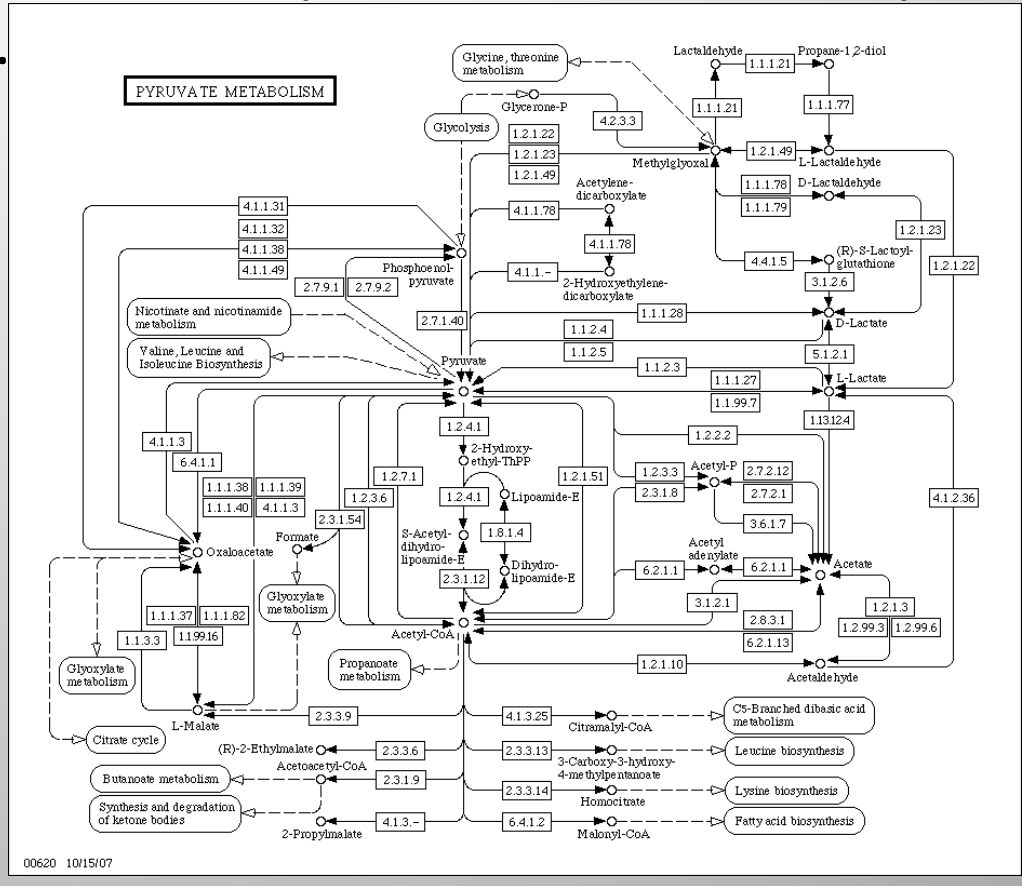
⊙Detect disease-causing chemicals (e.g. pesticides in Parkinson's) and their interactions with metabolic networks

⊙Some medicines influence the metabolic pattern, or even directly interact with some metabolites.

⊙Metabolic markers may exist for non-enzymatic diseases

⊙Elucidate the regulations in the enzyme/metabolite network

(Picture from KEGG PATHWAY)



# Metabolomics - Background

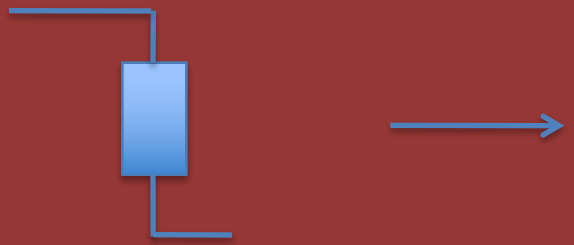
Thousands of metabolites exist in the biological system. Combined with other small molecules, high-throughput analysis of the system is a not an easy task.

| | **# entries** |
|---|---|
| Human Metabolome Database | 41,993 metabolites |
| DrugBank | 7,759 drugs (1602 FDA approved small molecules) |
| FooDB | (projected) 25,000 food components and food additives |
| | |
| Metlin | 240,964 entries |

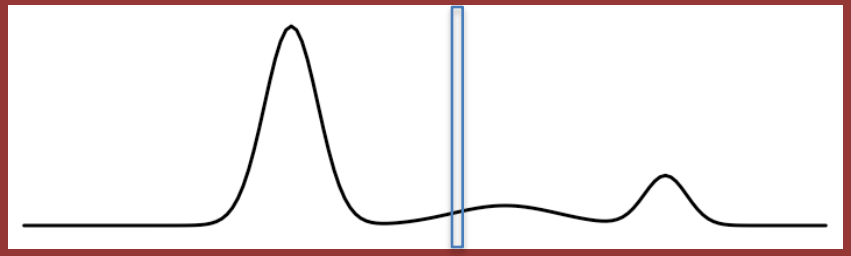
Methods used to profile the metabolome:

- Nuclear magnetic resonance (NMR)
- Gas Chromatography - Mass Spectrometry (GC/MS)
- Liquid Chromatography - Mass Spectrometry (LC/MS)
- LC/MS/MS
- …

# Metabolomics – LC/MS

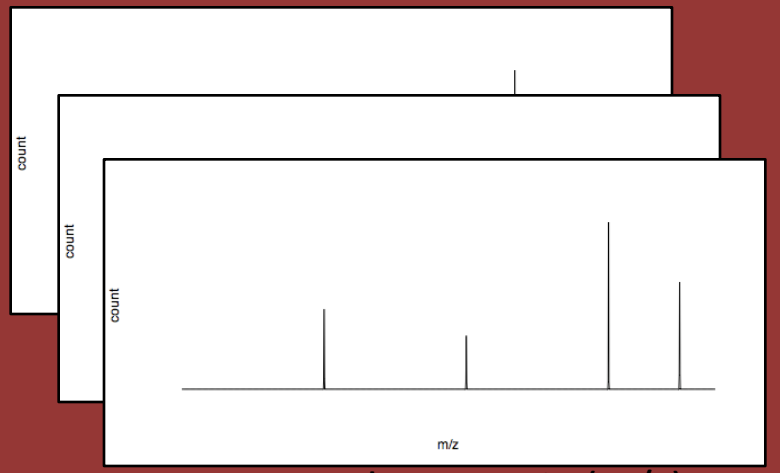


Liquid chromotography

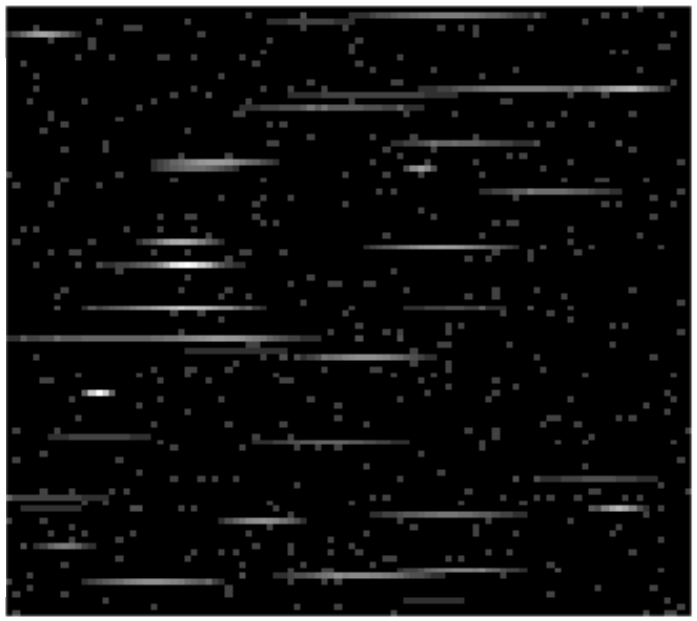


retention time

Take “slices” in retention time, send to MS

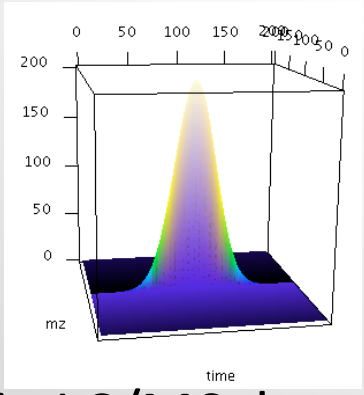


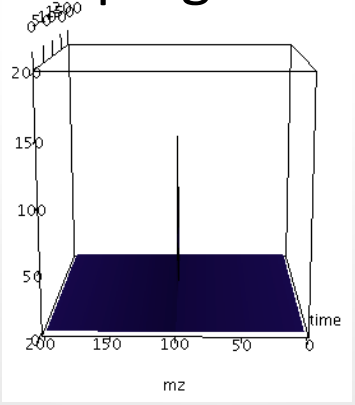

Mass-to-charge ratio (m/z)

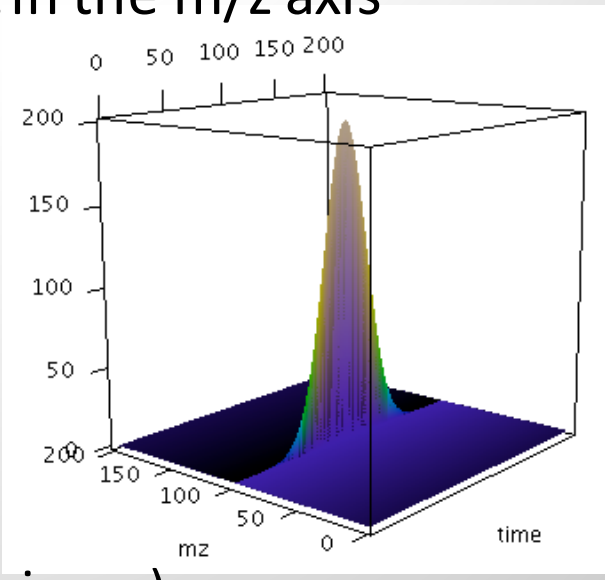


# Metabolomics - LC/MS

An ideal peak should show a Gaussian curve in intensity along the retention time axis, while keeping constant in the m/z axis



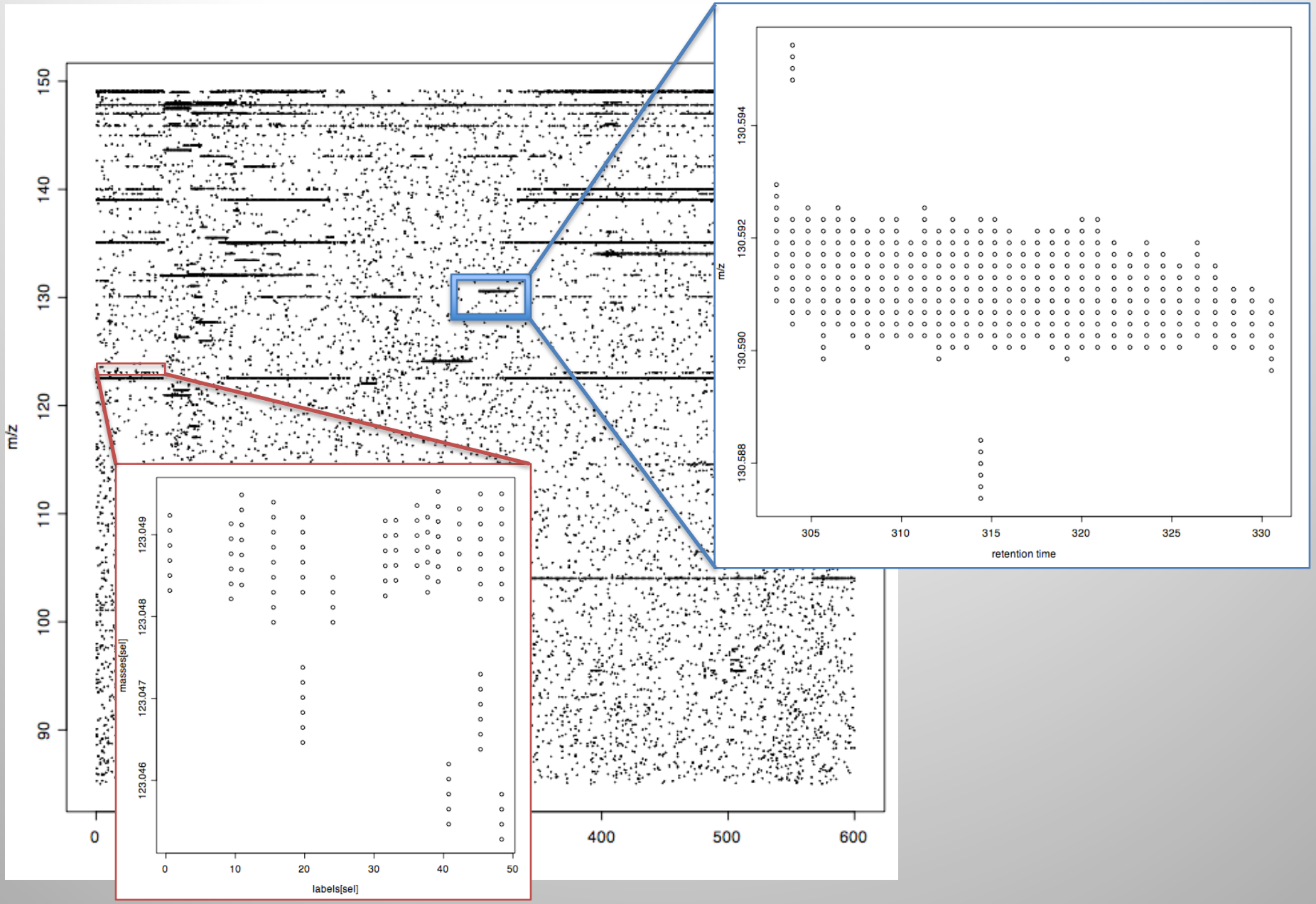




Issues in LC/MS data processing:

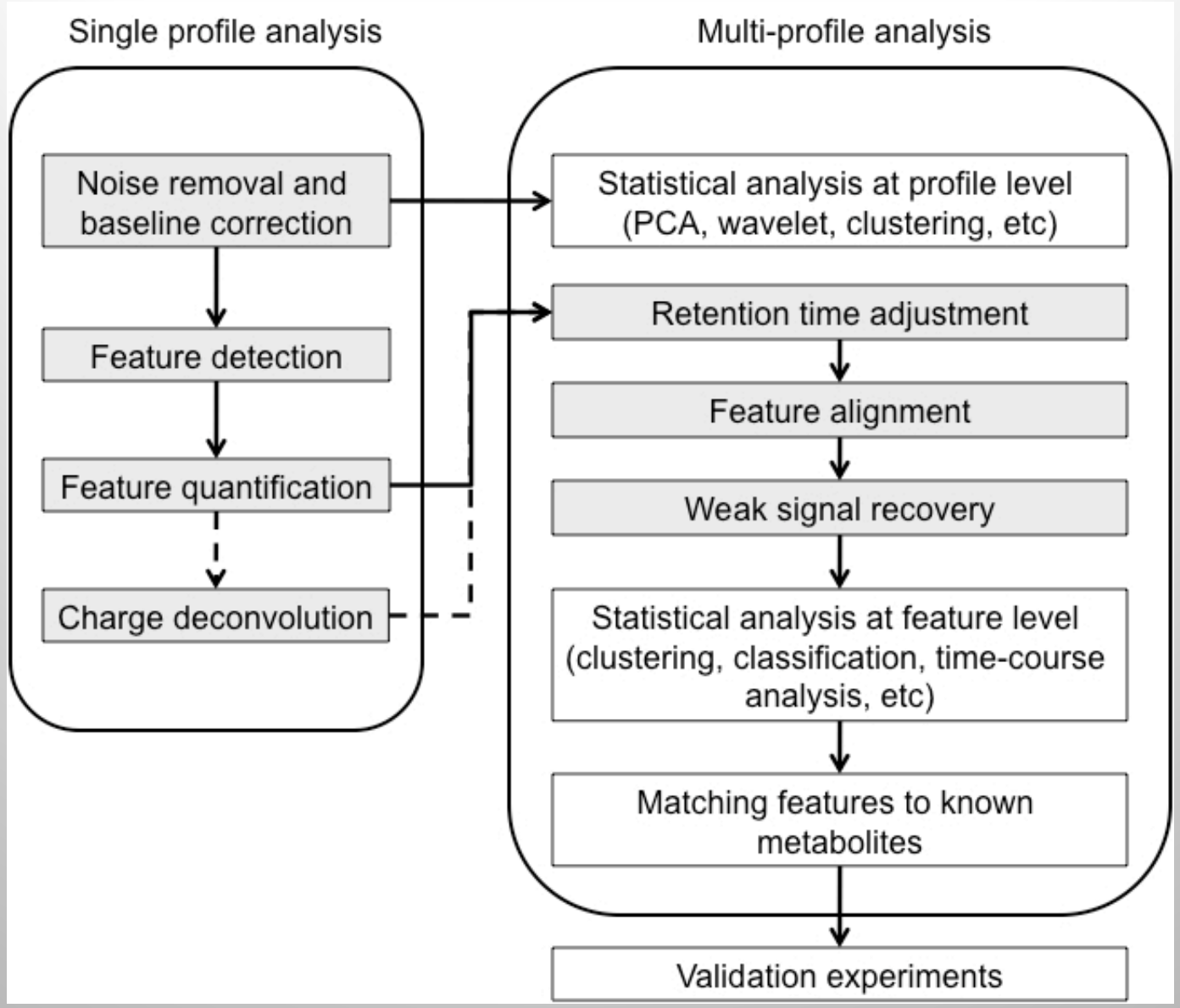
- Background noise
- Chemical noise (ridges in spectra, ion suppression,…)
- Peak intensity (along retention time axis) deviate from Gaussian curve substantially (fronting, breaks, …)
- Slight m/z shift across MS spectra (not severe with newer machines)
- Retention time shift across LC/MS spectra
- Multiple charge status of a single metabolite
- Isotopes

……

# Metabolomics – LC/FTMS data



# LC/MS data processing



# LC/MS data processing

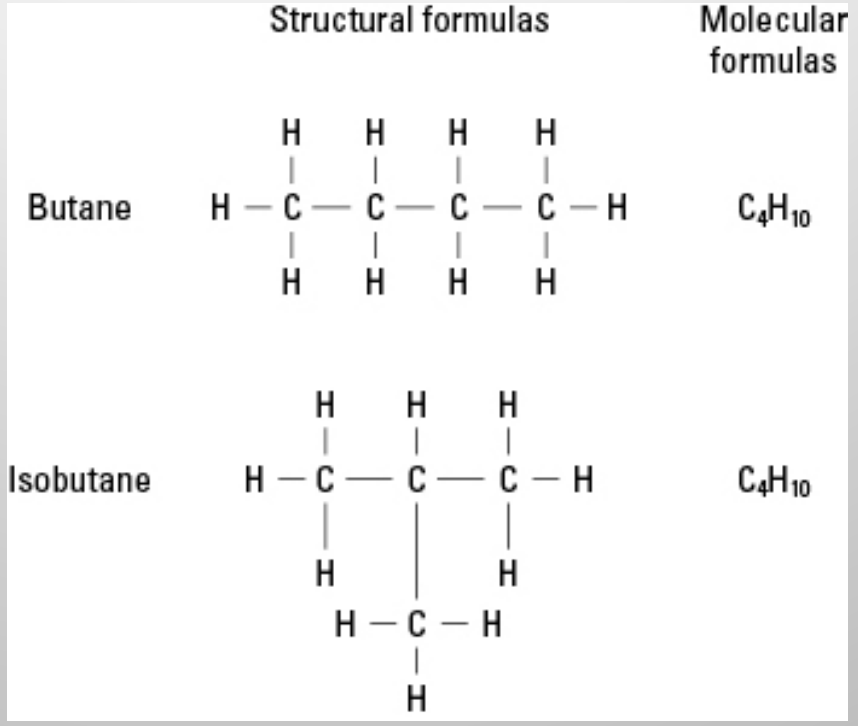
How do we know which feature is which chemical?

We have three measurement on each feature:

***m/z value,  retention time,  peak intensity***

In high-resolution data, m/z value could almost pin-point the chemical composition. However some chemicals share same chemical composition.

A targeted MS/MS experiment may be necessary to differentiate between them.




http://www.cliffsnotes.com/sciences/chemistry/chemistry/organic-compounds/structural-formulas

# Proteomics - How to identify a single protein by MS?

Digest into many peptides

Mass of many peptides

peptide mass fingerprinting (PMF)

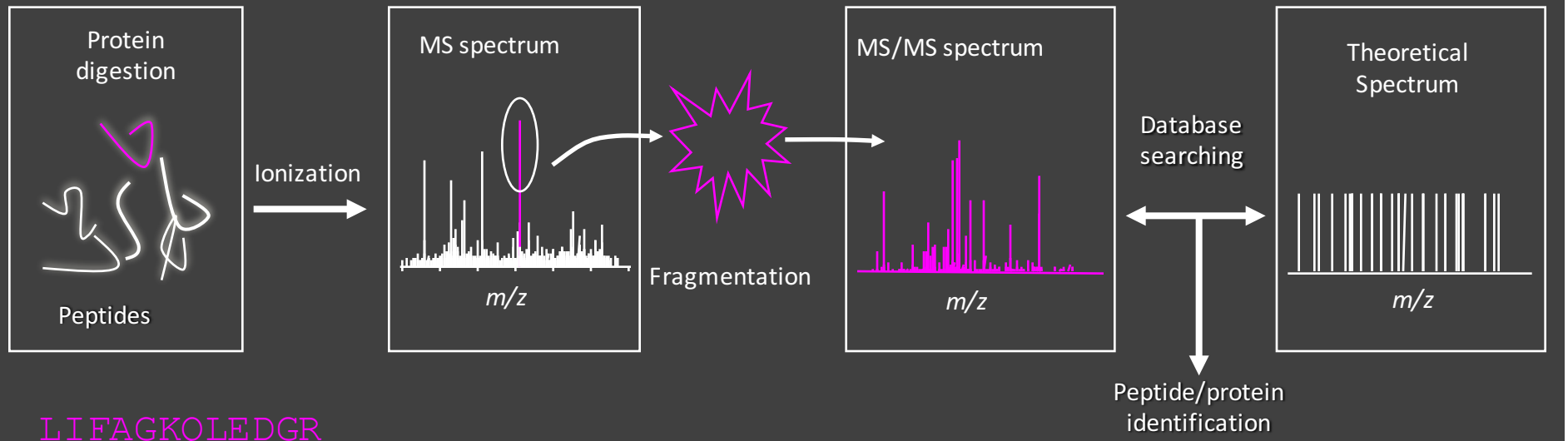
Mass/Charge *(m/z)*

Mass

Mass of many peptide fragments By Tandem Mass Spectrometry (MS/MS)

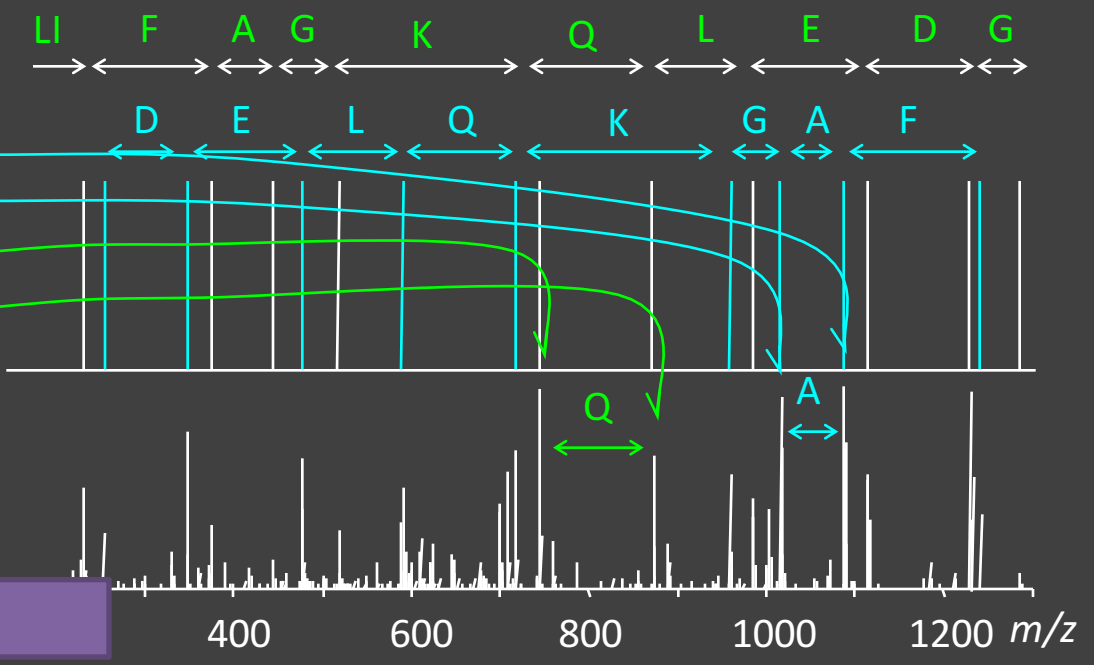
Kindly provided by Dr. Junmin Peng

# Proteomics - How to identify a single protein by MS/MS?



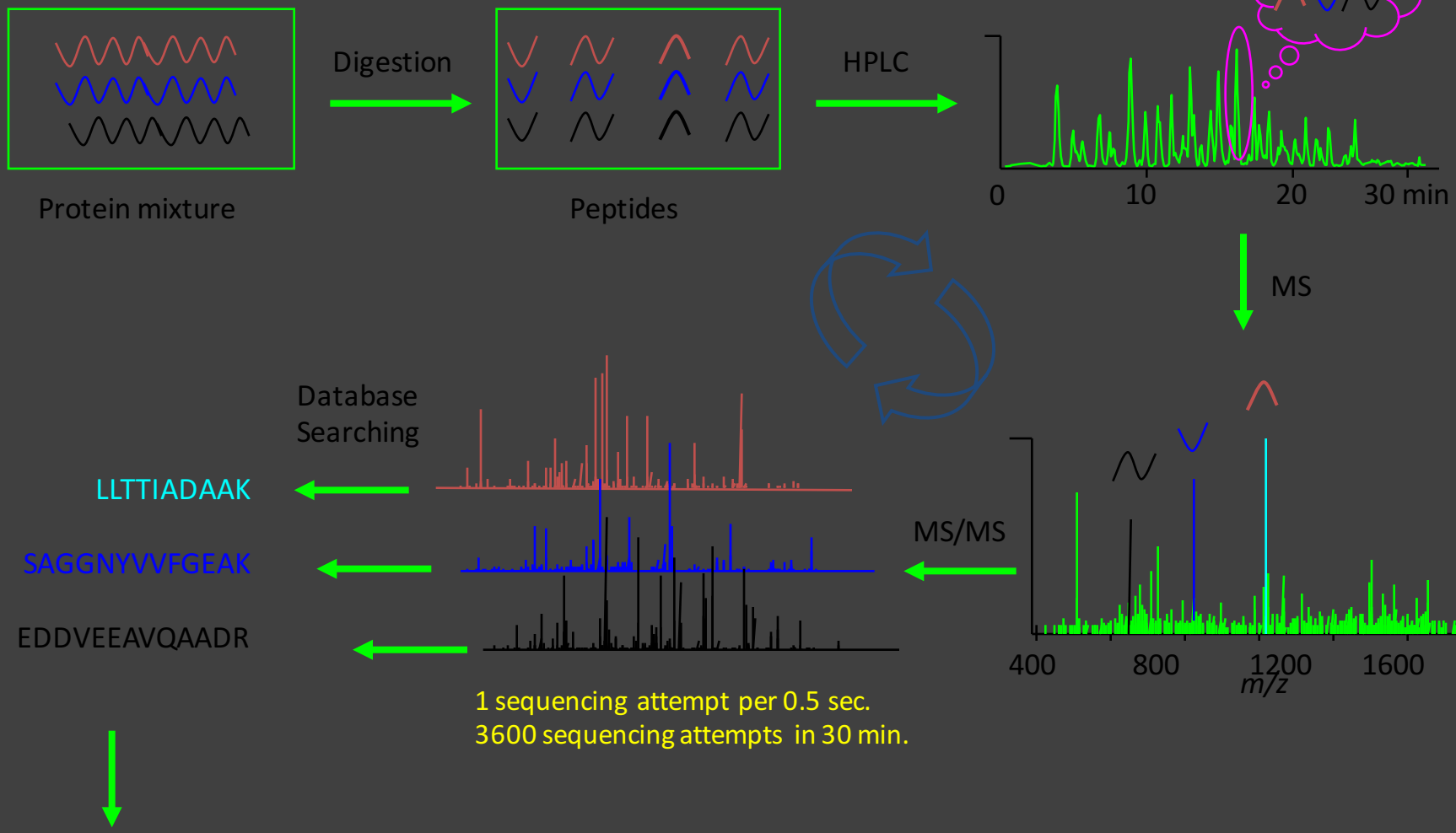

LIFAGKQLEDGR

```
b ions        y ions
1: L IFAGKQLEDGR:11
2: LI FAGKQLEDGR:10
3: LIF AGKQLEDGR :9
4: LIFA GKQLEDGR :8
5: LIFAG KQLEDGR :7
6: LIFAGK QLEDGR :6
7: LIFAGKQ LEDGR :5
8: LIFAGKQL EDGR :4
9: LIFAGKQLE DGR :3
10:LIFAGKQLED GR :2
11:LIFAGKQLEDG R :1
```




Kindly provided by Dr. Junmin Peng

# Proteomics - Analysis of protein mixture by tandem MS




All peptide sequences → Identification of many proteins

Kindly provided by Dr. Junmin Peng

# Proteomics - Analysis of complex protein mixtures

Difficulty:

In a complex biological sample (cell, tissue, serum, … ), there are several thousand protein species – tens of thousands of peptides after digestion; signal from less-abundant species may be suppressed.

Solution:

Must reduce complexity to identify and quantify proteins.

Incorporate biochemical separation techniques:

- LC-MS/MS
- LC/LC-MS/MS
- ……

Separate proteins in multiple dimensions.
Sacrifice speed.

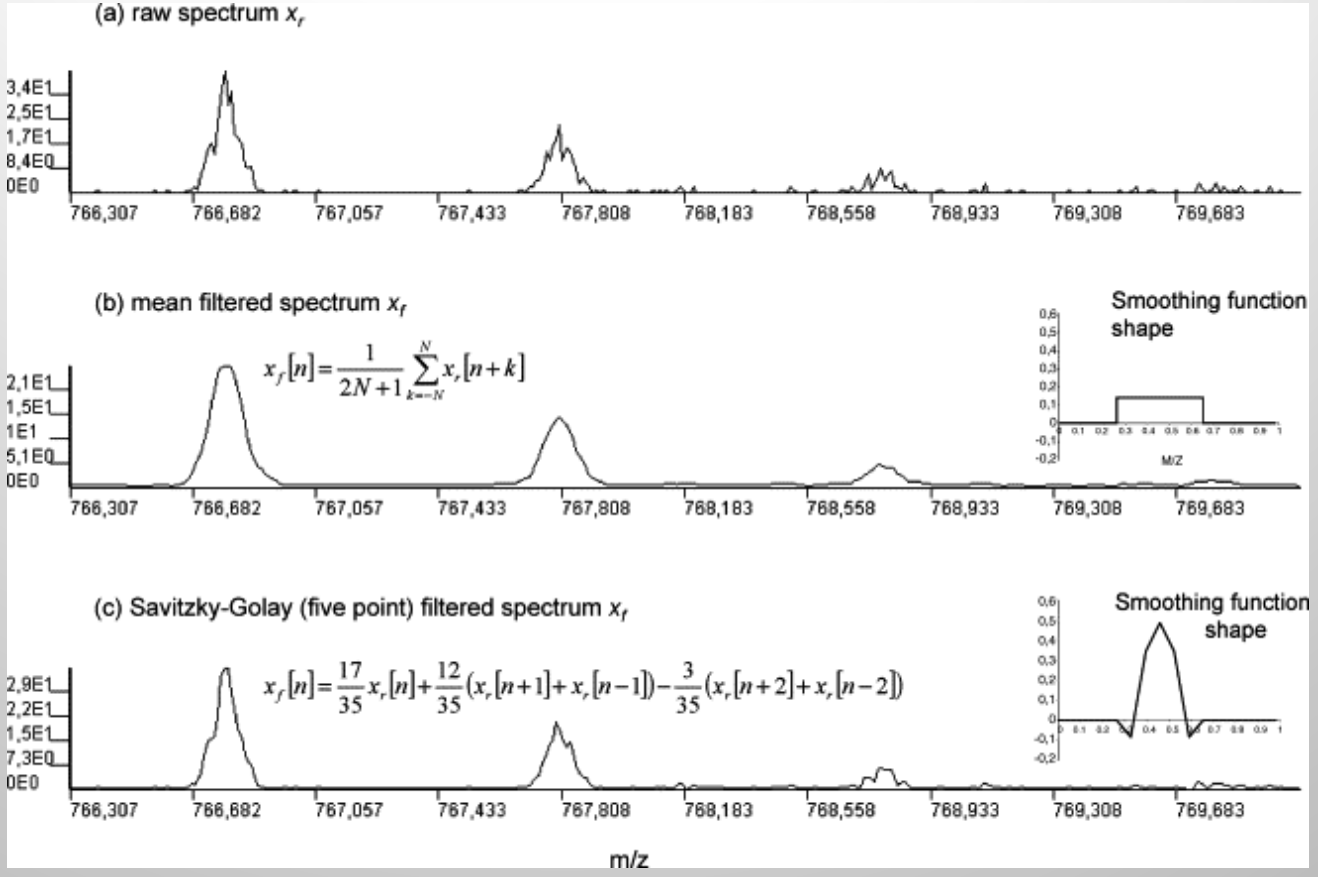
- 2D gel-MS/MS
- 2D gel/LC-MS/MS
- Affinity column separation – LC-MS/MS

Analyze a subset of proteins.
Sacrifice coverage.

# LC/MS – some example methods

**Reviewed by Katajamaa&Oresic (2007) J Chr. A 1158:318**

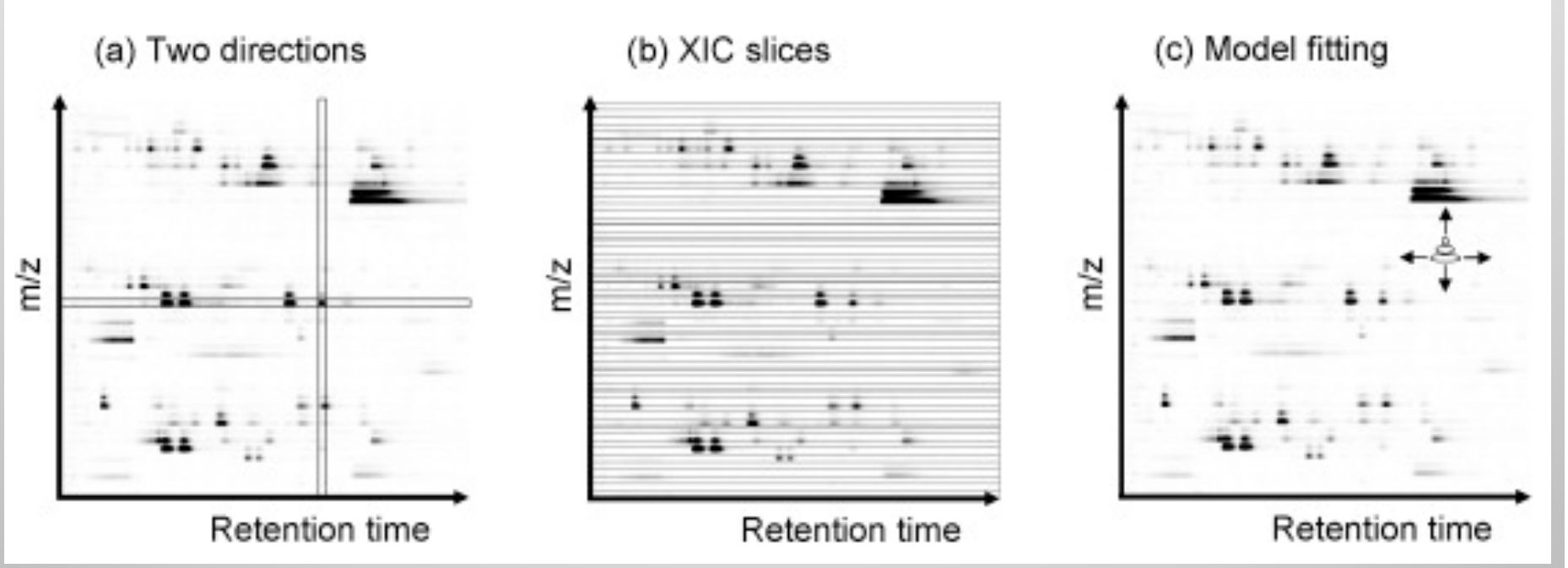
Noise reduction disregarding the time axis:



# LC/MS – some example methods

**Reviewed by Katajamaa&Oresic (2007) J Chr. A 1158:318**

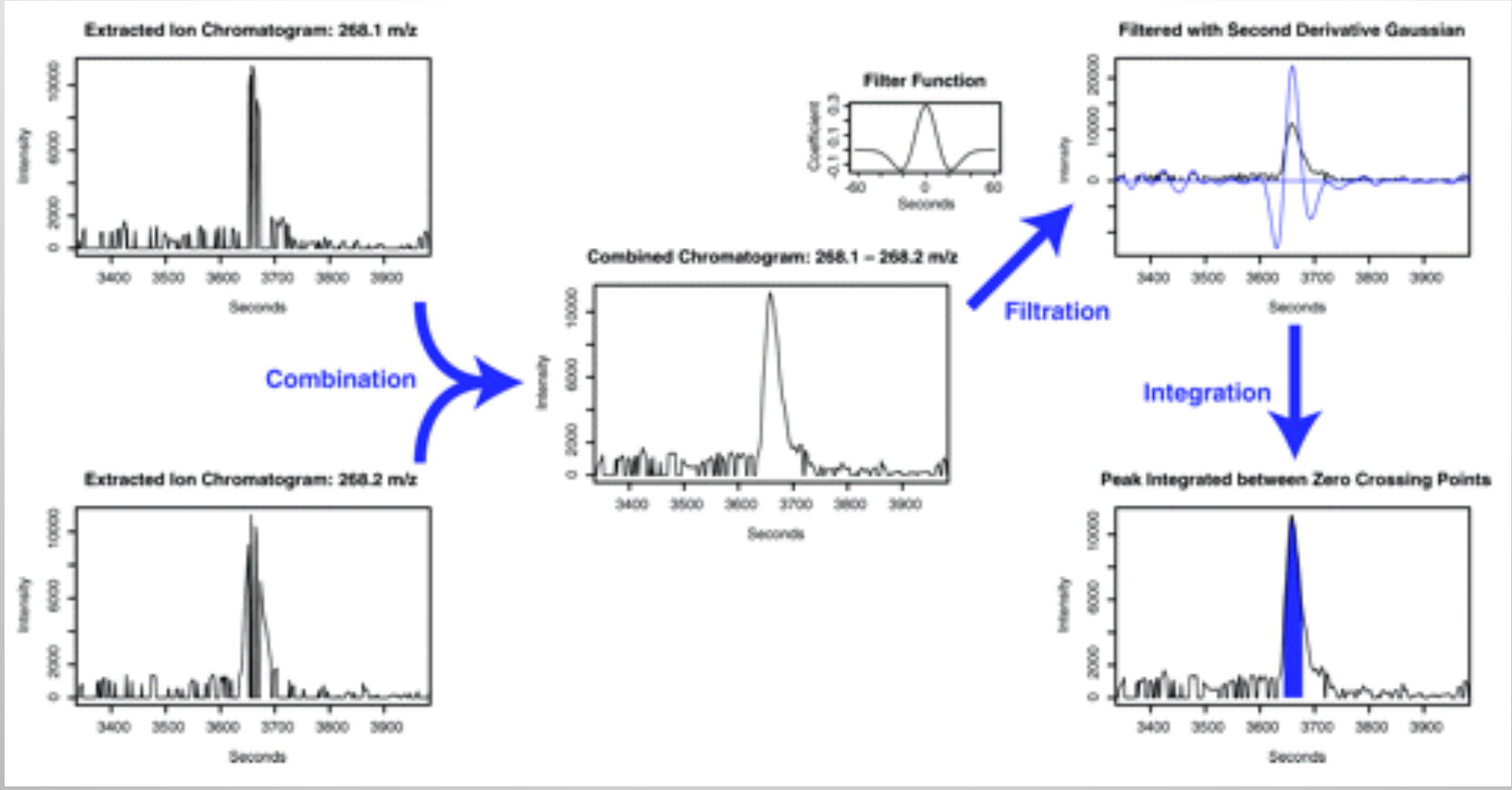
Peak Detection:



# LC/MS – some example methods

**Smith et.al. (2006) Analytical Chemistry. 78(3):779-87**

Matched Gaussian filter along the time axis after EIC extraction.



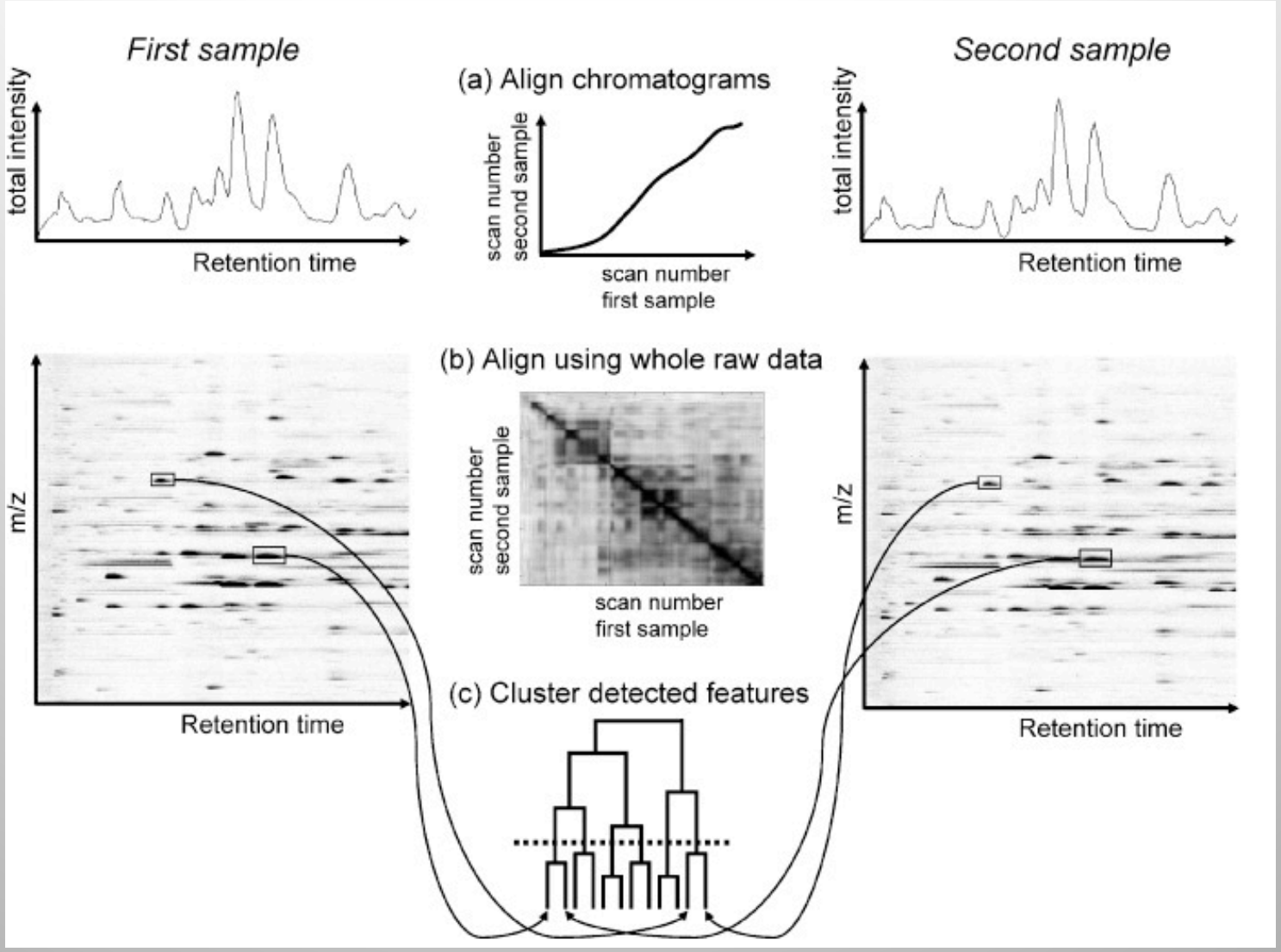

coefficients are equal to a second-derivative Gaussian function.

The filtered chromatogram crosses the *x*-axis roughly at the peak inflection points.

# LC/MS – some example methods

**Reviewed by Katajamaa&Oresic (2007) J Chr. A 1158:318**

Retention time alignment:

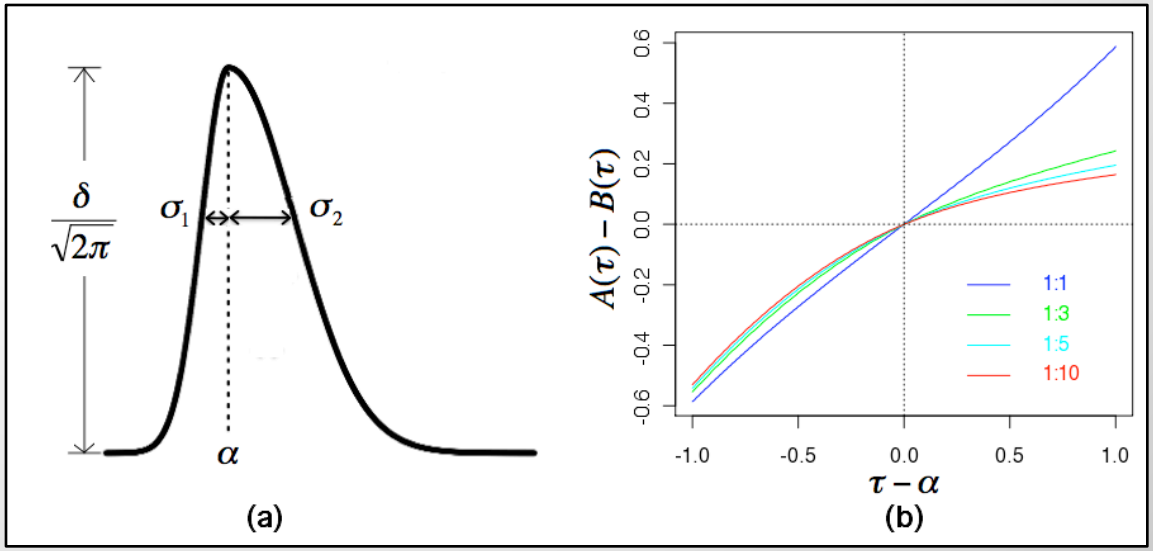


# LC/MS – some example methods

A model for assymetric peaks

Bi-Gaussian model:

$$g(t)=\begin{cases}\dfrac{\delta}{\sqrt{2\pi}}e^{-\frac{(t-\alpha)^2}{2\sigma_1^2}}, & t<\alpha\\ \dfrac{\delta}{\sqrt{2\pi}}e^{-\frac{(t-\alpha)^2}{2\sigma_2^2}}, & t\geq\alpha\end{cases}$$




Quantities used in peak location estimation:

$$A(\tau)=\log\left[\int_{-\infty}^{\tau}g(t)\,dt\right]-\log\left[\int_{\tau}^{\infty}g(t)\,dt\right]$$

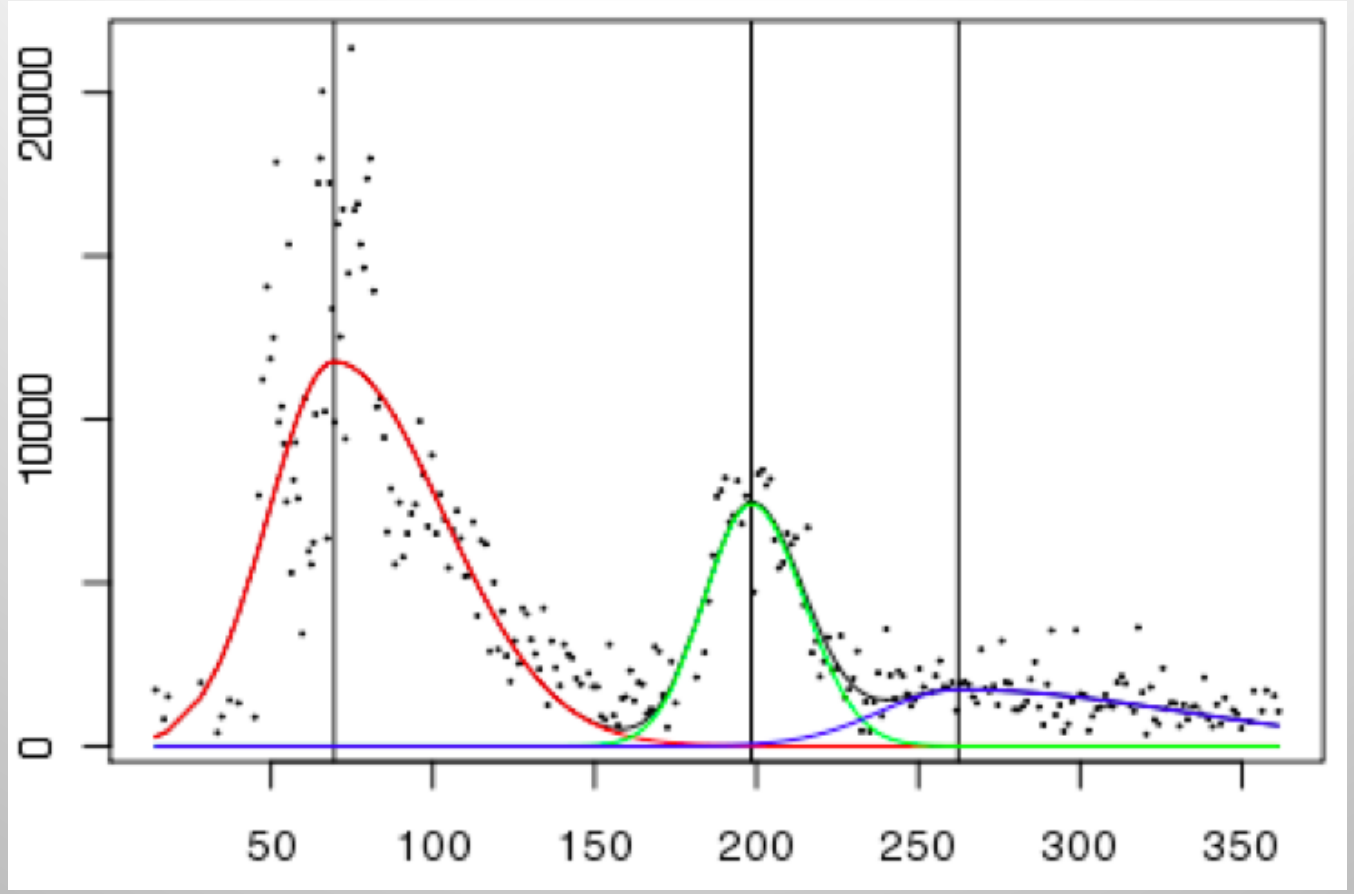
$$B(\tau)=\frac{1}{3}\log\left(\int_{-\infty}^{\tau}g(t)(t-\tau)^2\,dt\right)-\frac{1}{3}\log\left(\int_{\tau}^{\infty}g(t)(t-\tau)^2\,dt\right)$$

At the summit,

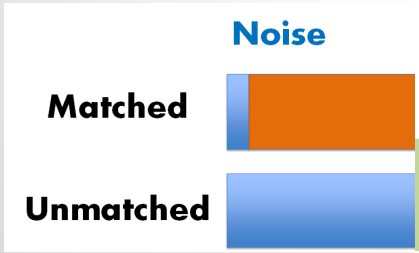
$$A(\alpha)=\log(\delta\sigma_1/2)-\log(\delta\sigma_2/2)=\frac{1}{3}\log\left(\frac{\delta\sigma_1^3}{2}\right)-\frac{1}{3}\log\left(\frac{\delta\sigma_2^3}{2}\right)=B(\alpha)$$

# LC/MS – some example methods

Fitting a mixture of bi-Gaussian curves using EM-like algorithms



# LC/MS – some example methods





LC/MS profile

Adaptive binning by iterative splitting (described in Yu et al. 2009)

Extracted Ion Chromatograms

Likely real peak locations based on database

Extract data features using different smoothing kernel and window sizes

Data features

Matching m/z at a certain tolerance level

Status of each EIC :
1: matched
0: unmatched

pROC selection and within-group selection

Selected data features

Variable selection, model comparison, and cross validation

EIC scores

Thresholding

Selected EICs

Peak detection, quantification and down-stream analyses

Data features:

RT span, RT peak location, m/z span, m/z standard deviation, the $0^{th}$, $25^{th}$, $50^{th}$,$75^{th}$ and $100^{th}$ percentile of signal intensity

Using each smoothing kernel (Gaussian or rectangular) and window size combination, the $0^{th}$, $25^{th}$, $50^{th}$,$75^{th}$ and $100^{th}$ percentile of:

point density along RT
point density weighted by signal intensity
point density weighted by log signal intensity

- Balance data by down-sampling unmatched
- Split data in half – training set and testing set
- Rank variable importance by Adaboost & Random Forest
- For each variable importance ranking, iterate from the top 1 variable to maximum allowable number of top variables, fit model using training set:
  - > Adaboost
  - > Logistic regression
  - > Support vector machine
  - > Random forest
- Based on the prediction of the testing set, choose the set of variables and predictive model that produce the best result.
- Compute the score for each EIC.

# Beyond pre-processing

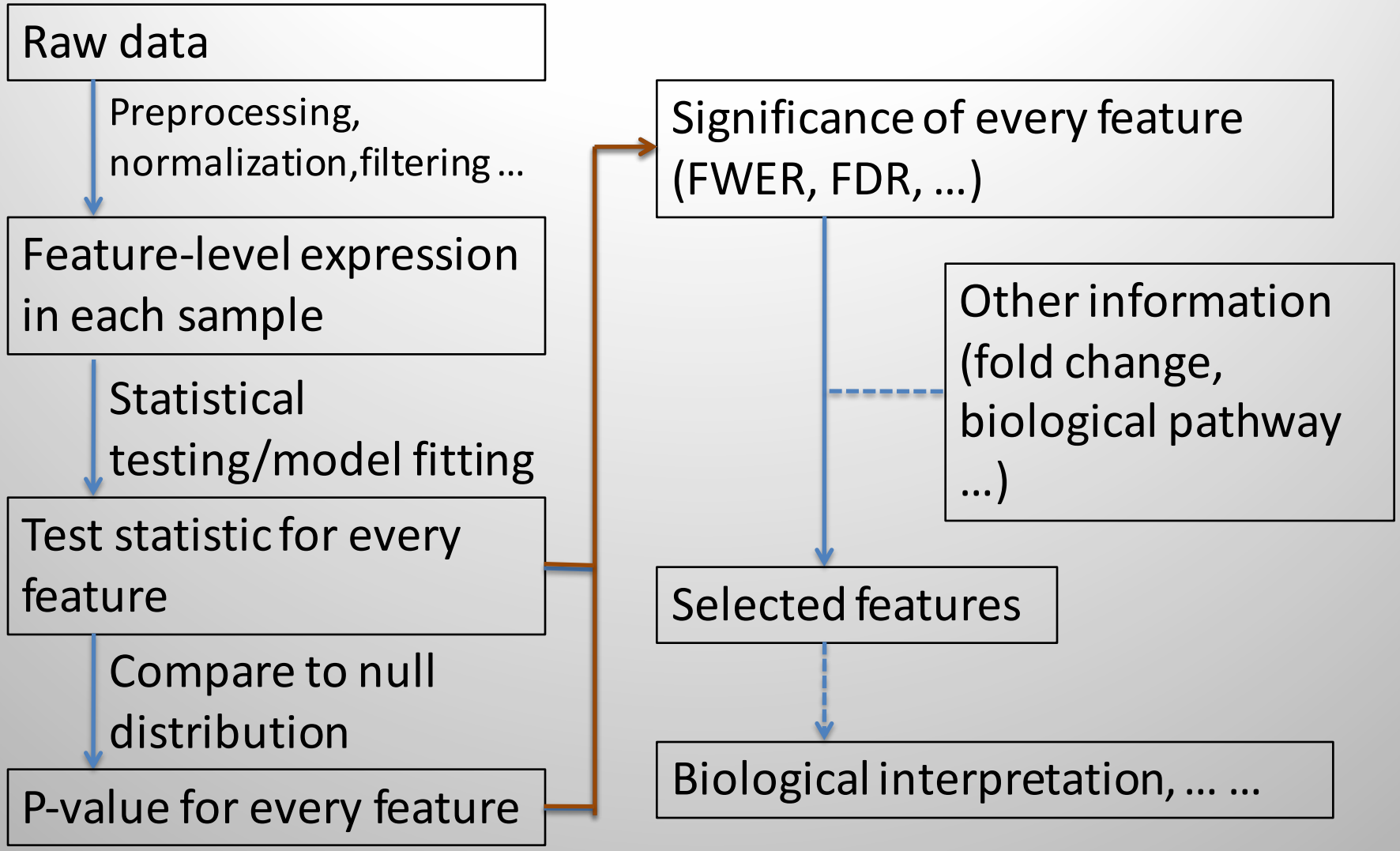
- Find features (genes/proteins/metabolites) significantly associated with a phenotype – “biomarkers”
- Find biological pathways or subnetworks associated with a phenotype
- Based on identified biomarkers, build predictive models for diagnosis, prognosis, or predict treatment response -> “personalized medicine”?
- Identify feature-feature associations and previously unknown structures in the data
- Find regulation patterns, both intrinsic and in relation to a phenotype

# Beyond pre-processing

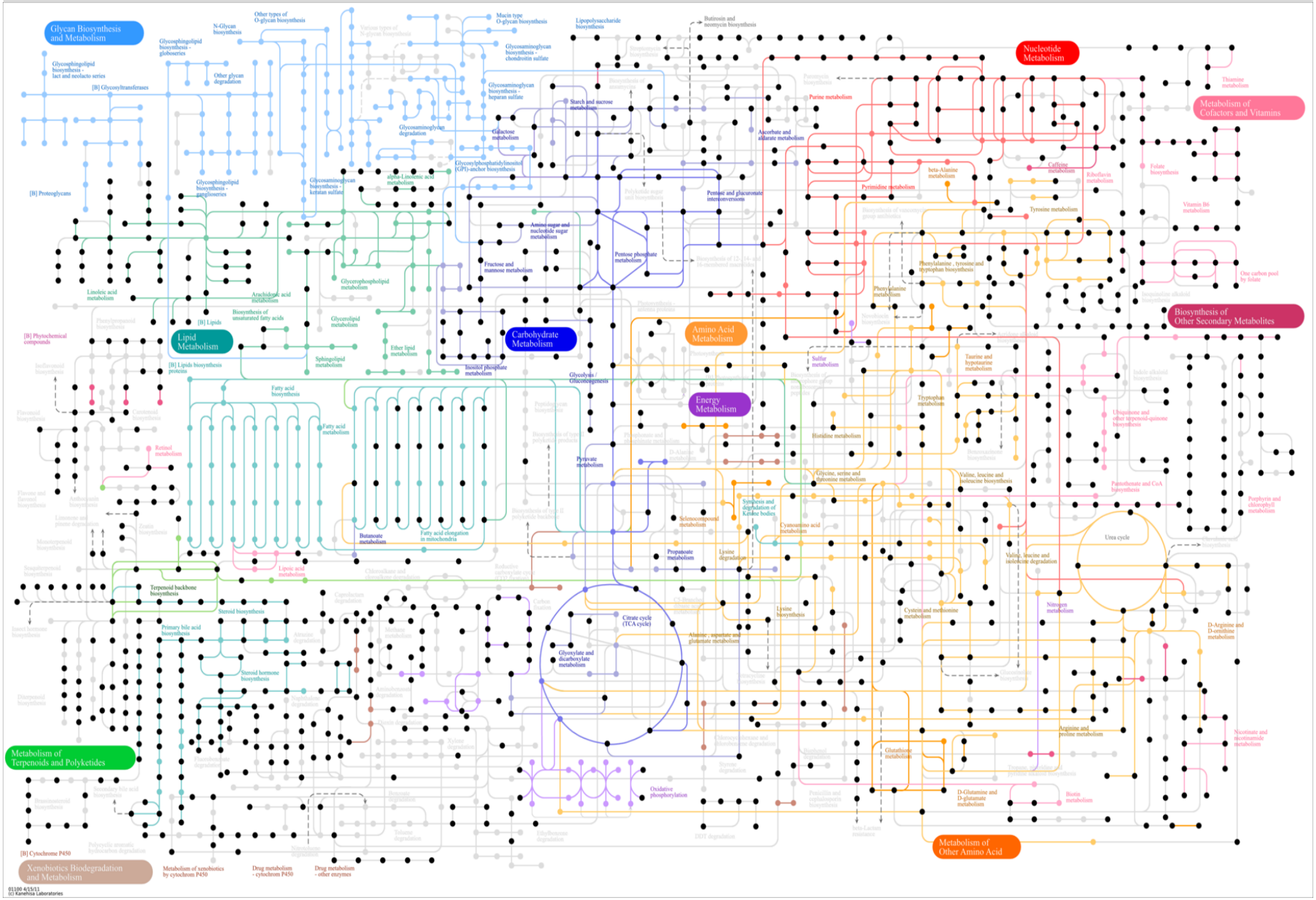
This is the common structure of microarray gene expression data from a simple cross-sectional case-control design.
Data from other high-throughput technology are often similar.

| | Control 1 | Control 2 | ..... | Control 25 | Disease 1 | Disease 2 | ..... | Disease 40 |
|---|---|---|---|---|---|---|---|---|
| Gene 1 | 9.25 | 9.77 | ..... | 9.4 | 8.58 | 5.62 | ..... | 6.88 |
| Gene 2 | 6.99 | 5.85 | ..... | 5 | 5.14 | 5.43 | ..... | 5.01 |
| Gene 3 | 4.55 | 5.3 | ..... | 4.73 | 3.66 | 4.27 | ..... | 4.11 |
| Gene 4 | 7.04 | 7.16 | ..... | 6.47 | 6.79 | 6.87 | ..... | 6.45 |
| Gene 5 | 2.84 | 3.21 | ..... | 3.2 | 3.06 | 3.26 | ..... | 3.15 |
| Gene 6 | 6.08 | 6.26 | ..... | 7.19 | 6.12 | 5.93 | ..... | 6.44 |
| Gene 7 | 4 | 4.41 | ..... | 4.22 | 4.42 | 4.09 | ..... | 4.26 |
| Gene 8 | 4.01 | 4.15 | ..... | 3.45 | 3.77 | 3.55 | ..... | 3.82 |
| Gene 9 | 6.37 | 7.2 | ..... | 8.14 | 5.13 | 7.06 | ..... | 7.27 |
| Gene 10 | 2.91 | 3.04 | ..... | 3.03 | 2.83 | 3.86 | ..... | 2.89 |
| Gene 11 | 3.71 | 3.79 | ..... | 3.39 | 5.15 | 6.23 | ..... | 4.44 |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| Gene 50000 | 3.65 | 3.73 | ..... | 3.8 | 3.87 | 3.76 | ..... | 3.62 |

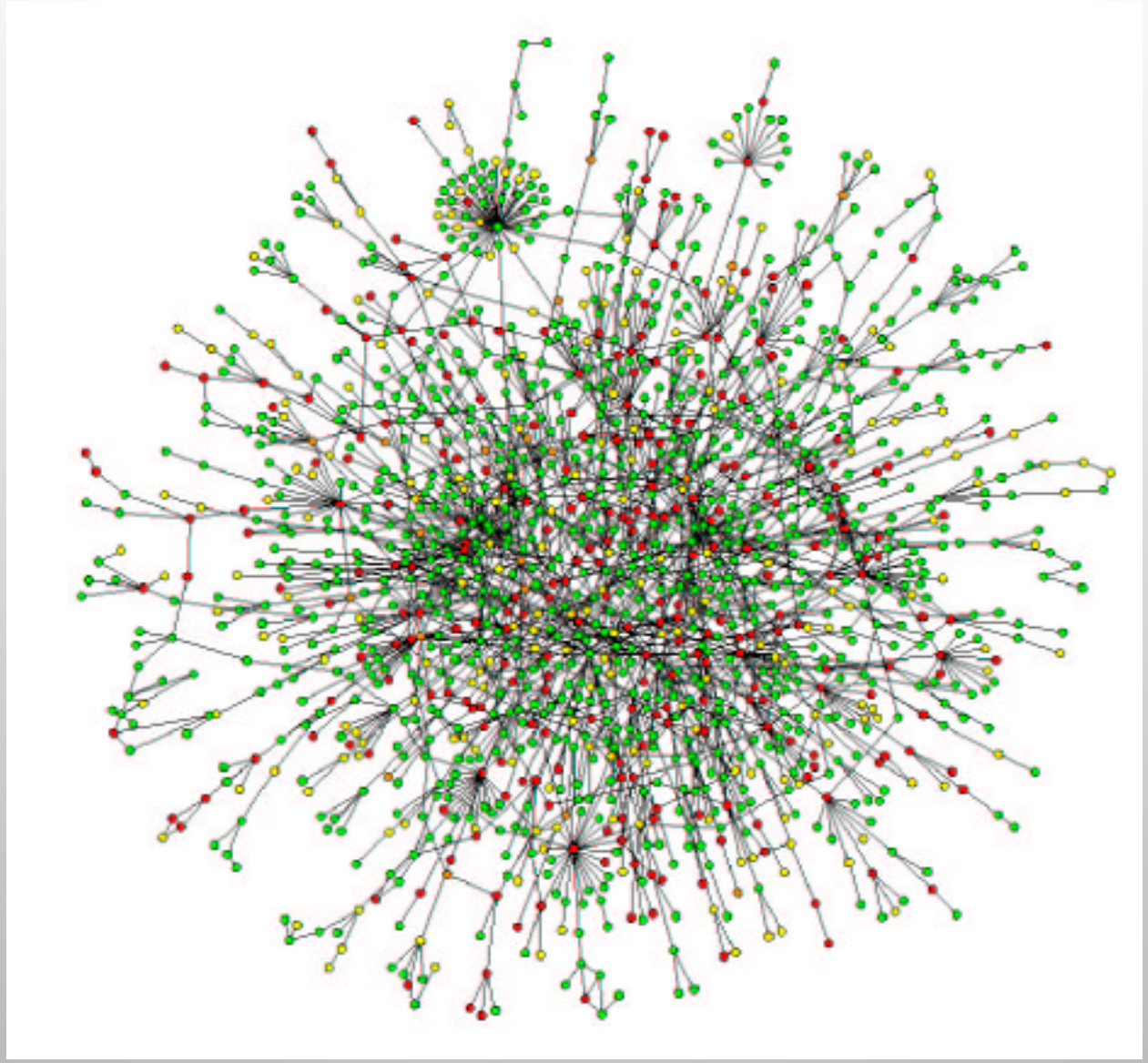
# Workflow of feature selection



# Genome-wide metabolic network



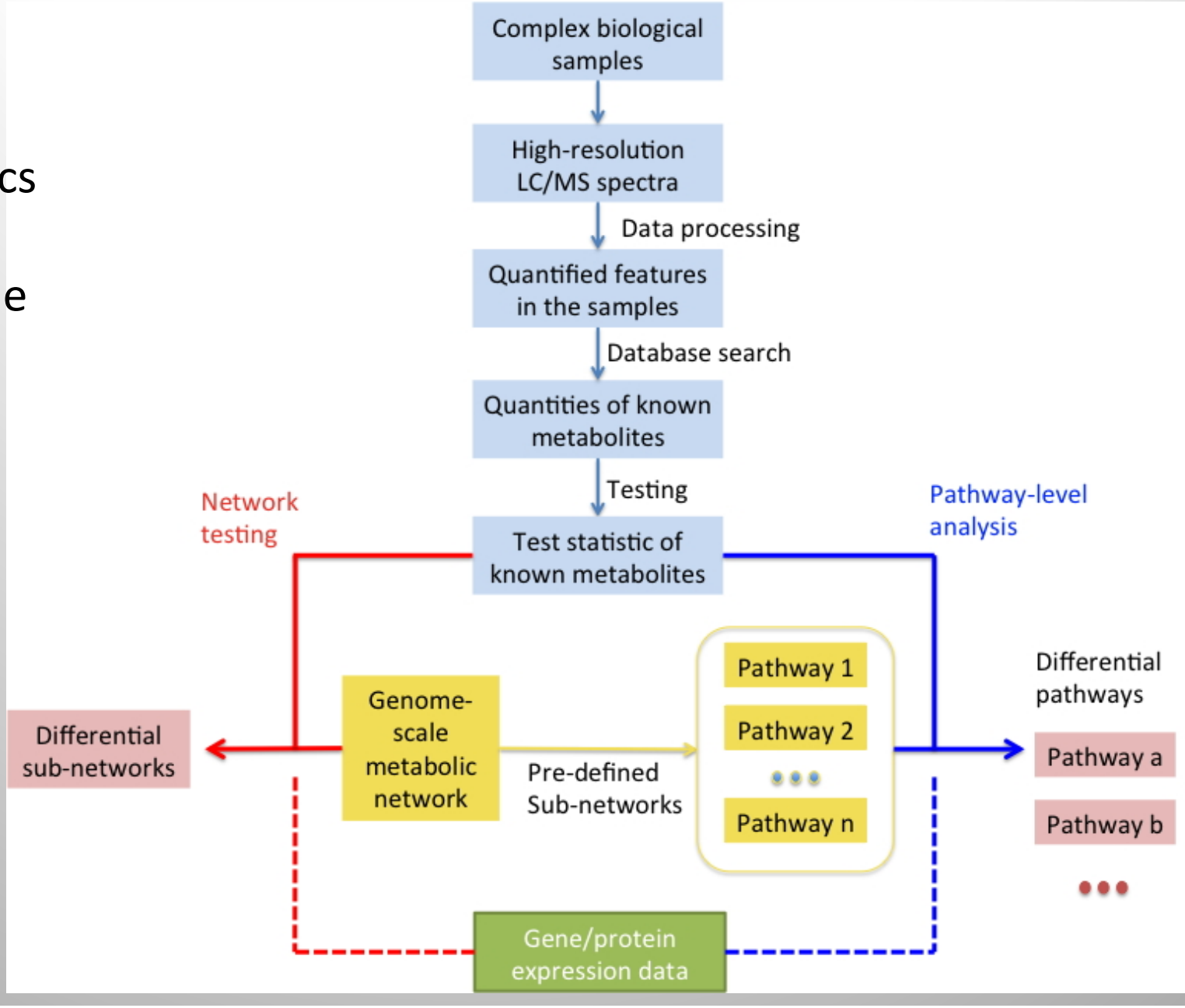
# Protein interaction network



S. Wuchty, E. Ravasz and A.-L. Barabasi: The Architecture of Biological Networks

# Knowledge-guided omics data analysis

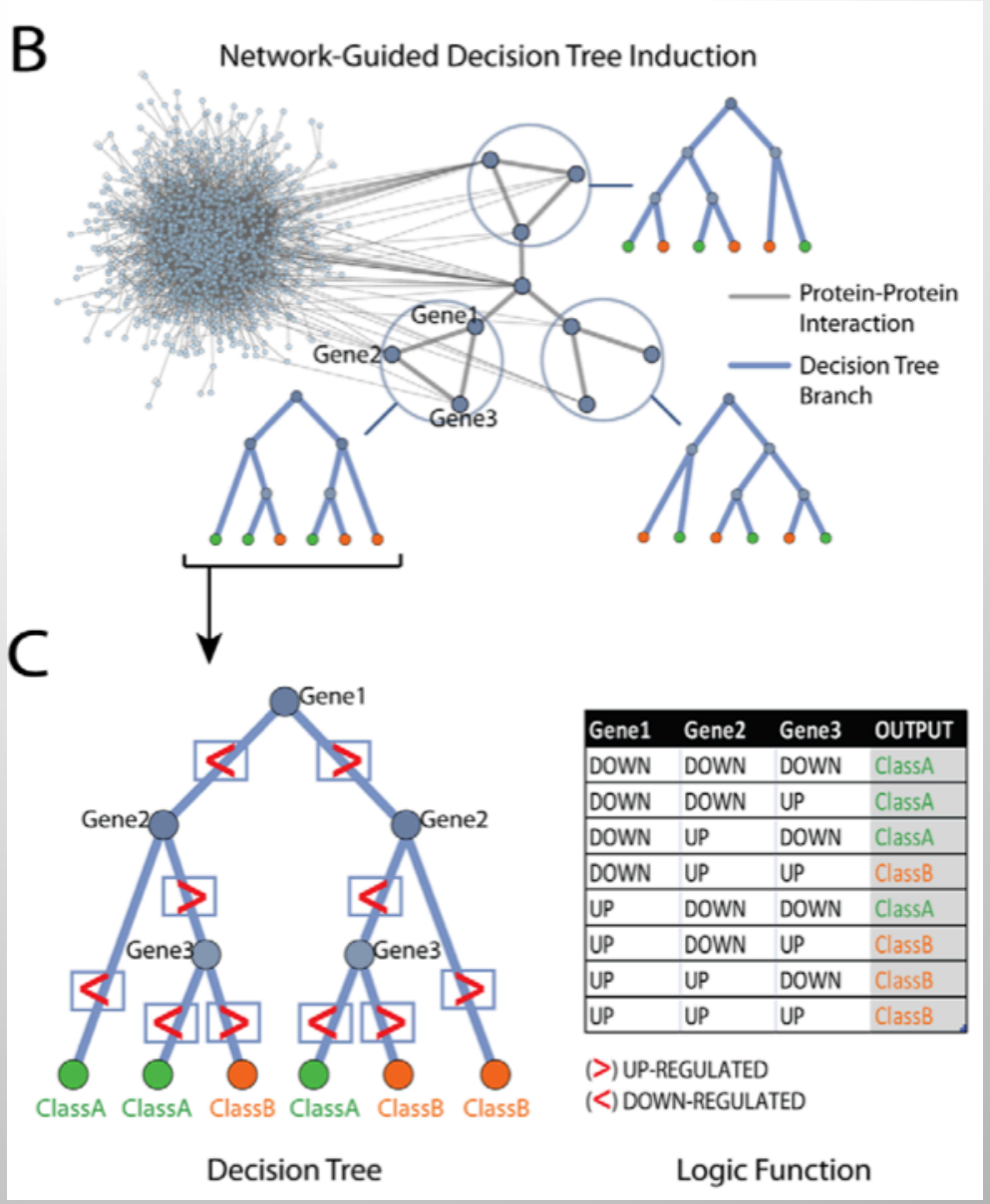
Example: analyzing metabolomics data with genome-scale metabolic network.



# Knowledge-guided omics data analysis

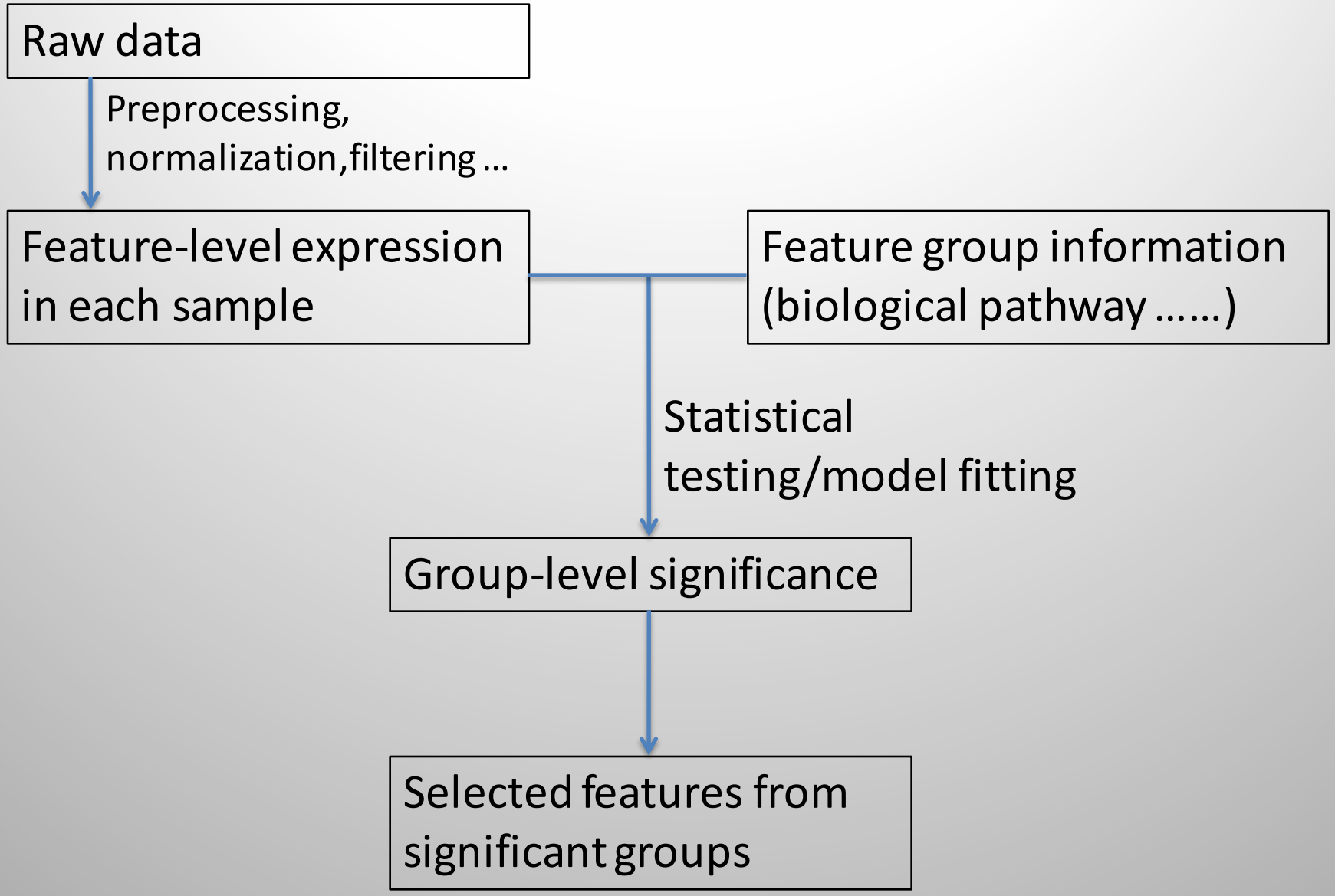
An example:

Selecting cancer-related sub-networks using Network-Guided Random Forests.



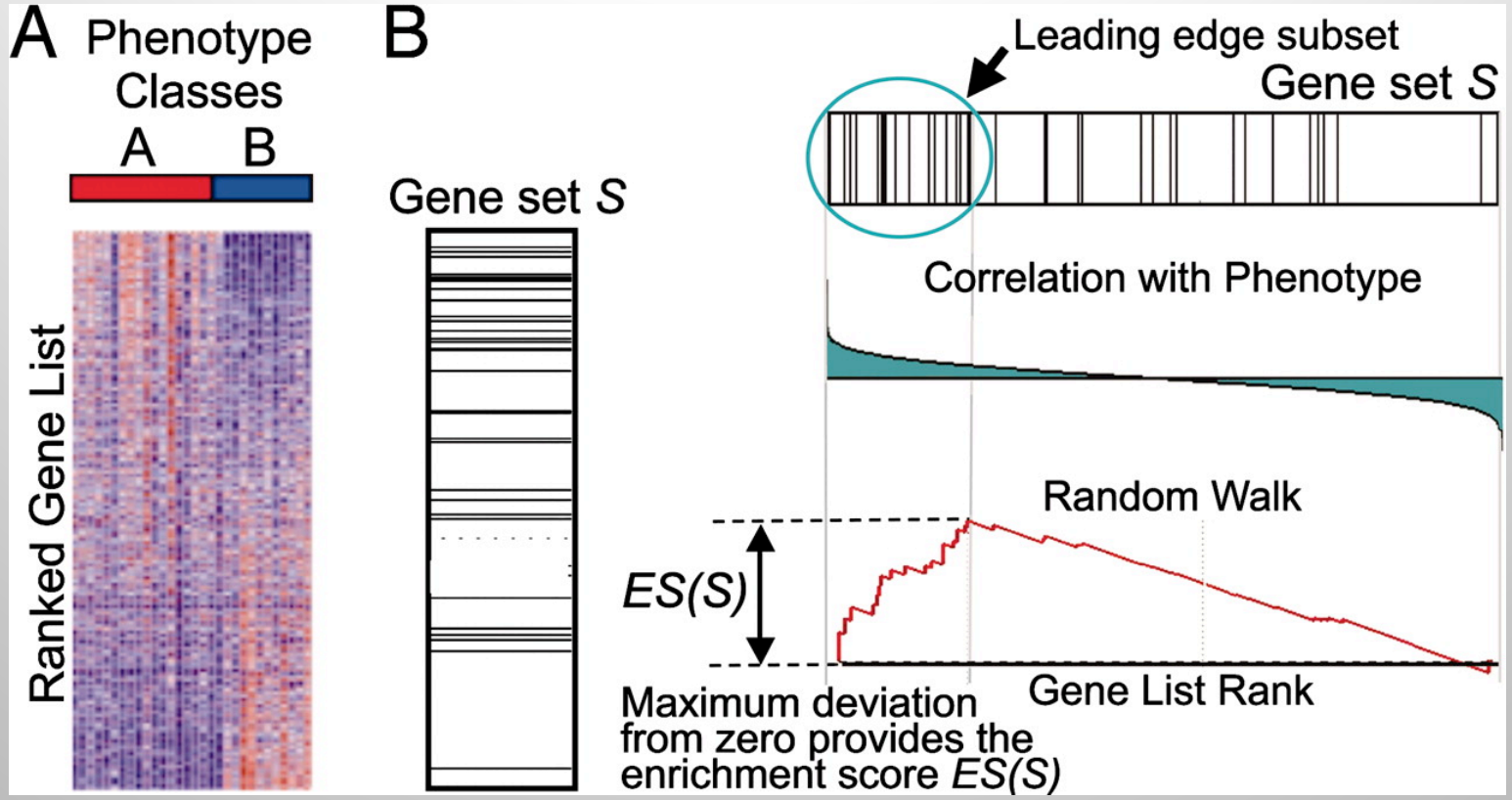

Dutkowski J, Ideker T (2011) PLoS Comput Biol 7(9): e1002180.

# Workflow of Gene Set Analysis



# Knowledge-guided omics data analysis

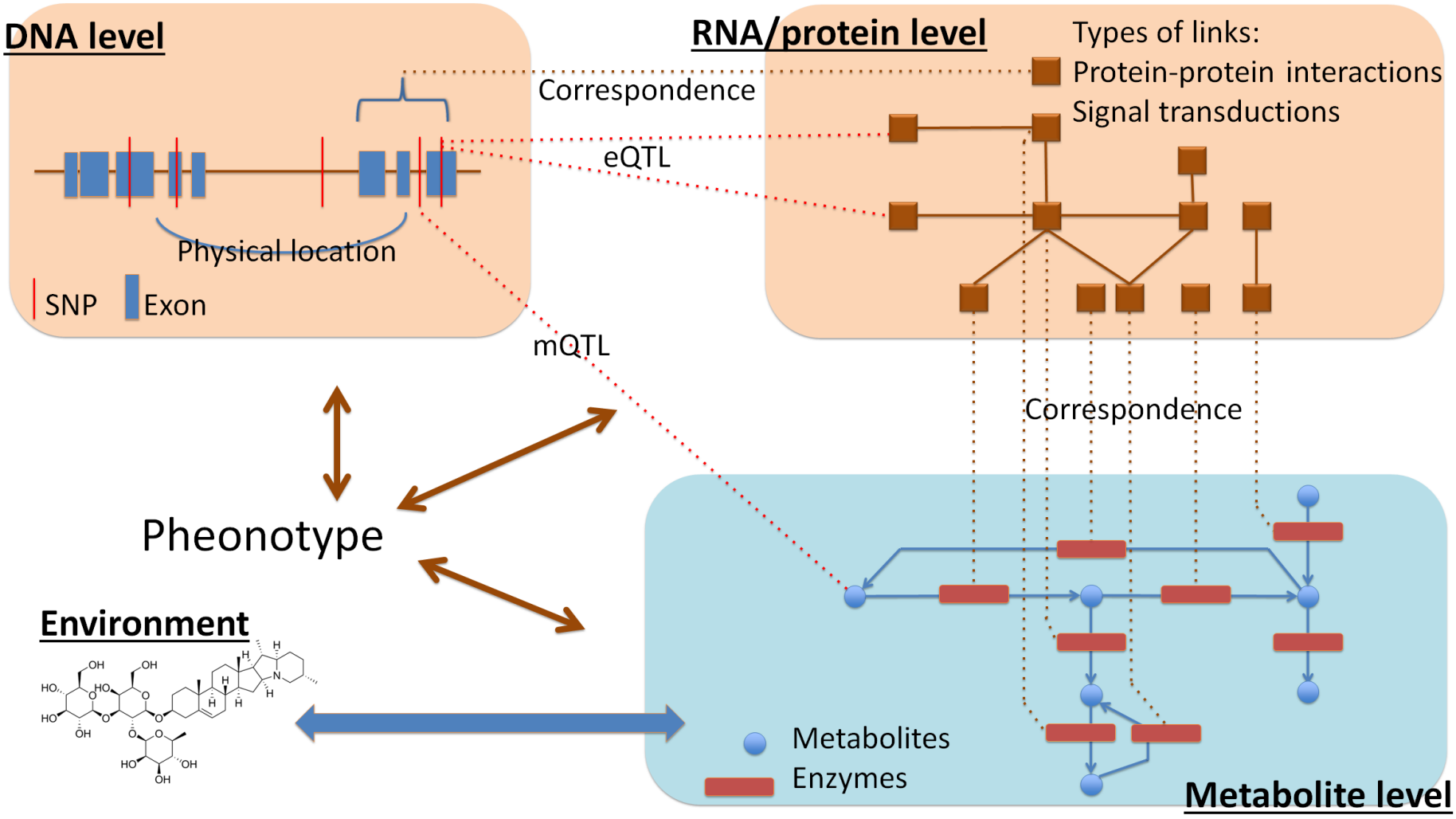
An example: Selecting gene sets (pre-defined gene groups) associated with phenotype.




PNAS, vol. 102, no. 43, 15545–15550

# Knowledge-guided omics data analysis

How to jointly model multi-layer omics data – an open question.



# Analyzing data together with biological network

An example application: pathways associated with vaccination response.