COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

1. Proteomics and Metabolomics

LU 1A – INTRODUCTION TO PROTEOMICS AND METABOLOMICS

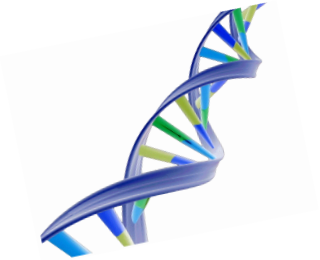
- Omics techniques and systems biology
- Difference between sequence-based techniques and MS-based techniques
- Applications of proteomics, metabolomics, lipidomics

Systems Biology – Definition

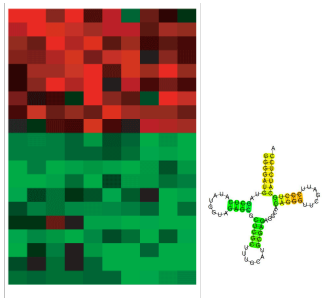
"**Systems biology** is a relatively new biological study field that focuses on the systematic study of complex interactions in biological systems, thus using a new perspective (**integration instead of reduction**) to study them. Particularly from year 2000 onwards, the term is used widely in the biosciences, and in a variety of contexts. Because the scientific method has been used primarily toward reductionism, one of the goals of systems biology is to discover new **emergent properties** that may arise from the **systemic view** used by this discipline in order to understand better the entirety of processes that happen in a biological system."

http://en.wikipedia.org/wiki/Systems_biology (06/06/2008)

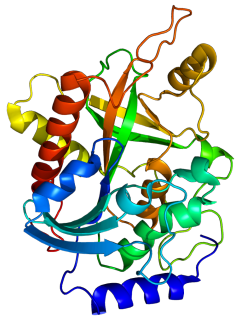
Technologies



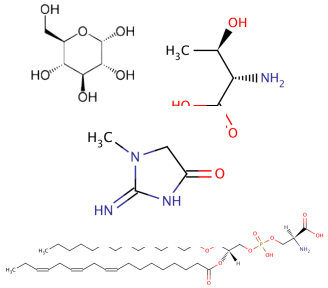
Genome
Epigenome



Transcriptome
RNOme



Proteome
Interactome



Metabolome
Lipidome

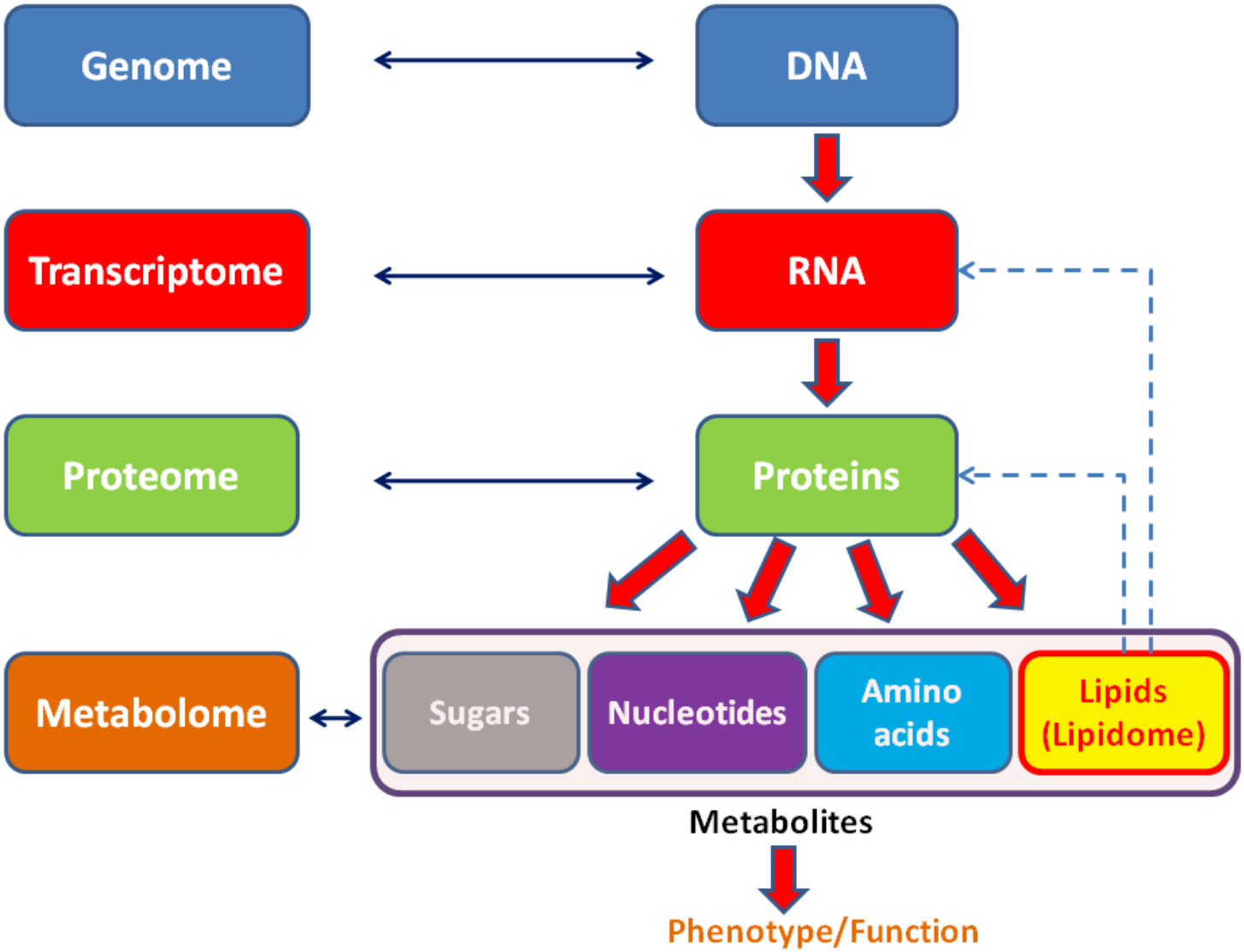
Next-Generation Sequencing

Mass Spectrometry

Amplification

- Sequencing-based methods have one massive advantage: **DNA can be amplified**
- **PCR** (polymerase chain reaction) can exponentially amplify existing DNA fragments with a low error rate
- 10 rounds of PCR increase the concentration of DNA in the sample by three orders of magnitude
- Metabolites and proteins cannot be amplified
- Methods for detecting and identifying metabolites and proteins thus need to be **more sensitive**

Omics Technologies



http://en.wikipedia.org/wiki/File:Metabolomics_schema.png , accessed 2014-03-10, 11:42:00 UTC

LU 1B - OVERVIEW OF SEPARATION TECHNIQUES

- Overview separation techniques (GE, LC, GC)
- Chromatographic techniques
- Separation principles (size, isoelectric point, hydrophobicity)

Sample Preparation Methods

- Samples for omics methods come from a wide range of sources: cell culture, primary tissue, body fluids
- Extraction of the required biomolecules is often difficult
  - Cells need to be broken up (mechanically, with detergents)
  - Proteins need to be denatured
  - Enzyme inhibitors, e.g., protease and phosphatase inhibitors, avoid enzymatic degradation
  - Small molecules are extracted by precipitating larger molecules (proteins) using strong organic reagents (e.g., methanol)
  - Metabolomics sample preparation must be very fast, since metabolites (intermediates of metabolism) can be rapidly degraded
  - Different solvents are required to extract/precipitate metabolites/ proteins
- Buffers and reagents should be compatible with MS!

Separation Methods

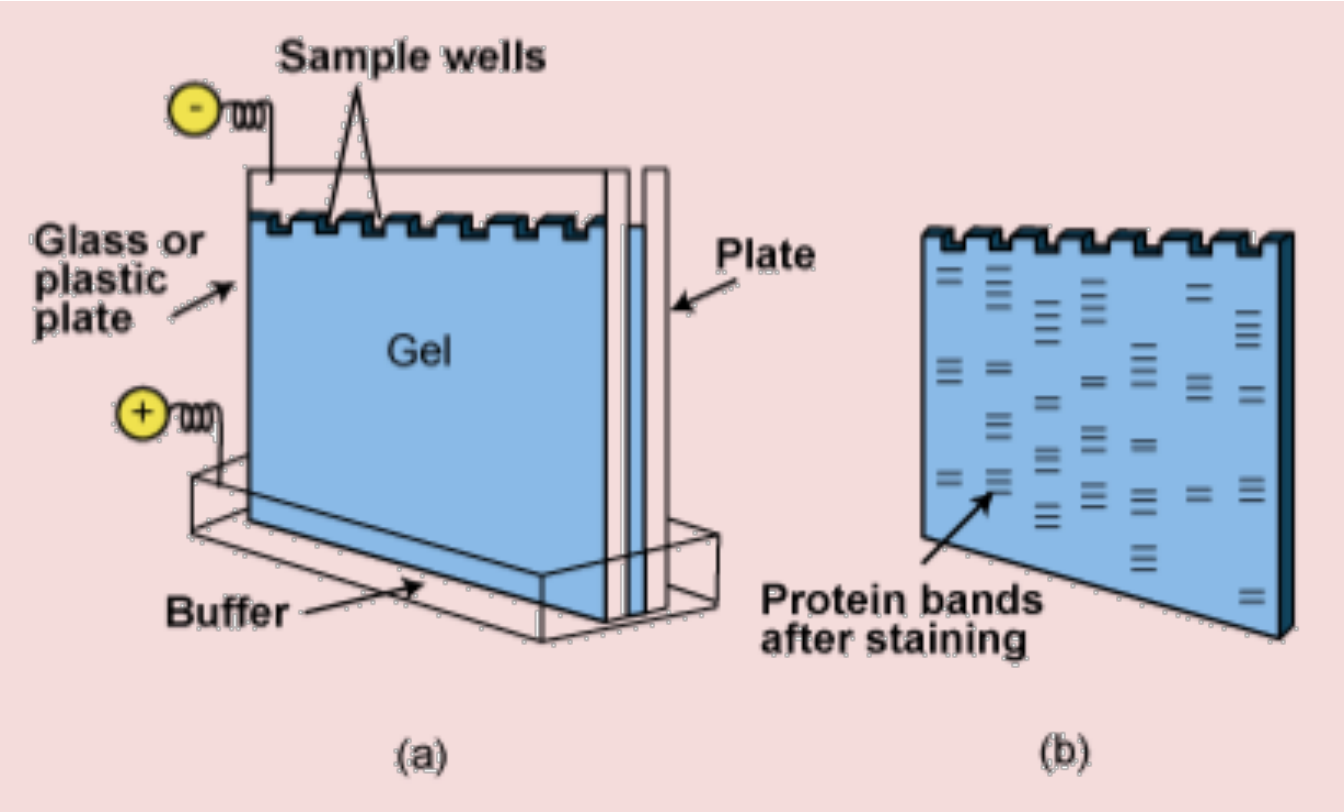
- Metabolomes and proteomes can be very complex (hundreds of thousands of analytes)
- Analyzing them at the same time reduces sensitivity and comprehensiveness of the analysis
- Idea:
  - Reduce the complexity
  - Split up the sample into smaller, less complex samples
- **Fractionation**
  - Separation is done before the analysis and results in a (small) number of new samples (usually dozens)
- **Online separation**
  - Separation happens simultaneously with the MS analysis

Overview Separation Methods

- Protein separation methods
  - 1D-PAGE (Polyacrylamide Gel Electrophoresis)
  - 2D-PAGE
  - (Capillary Electrophoresis)
- Peptide separation methods
  - Liquid chromatography
  - Isoelectric focusing of peptides
- Metabolite separation methods
  - Liquid chromatography
  - Gas chromatography

Gel Electrophoresis – Principles

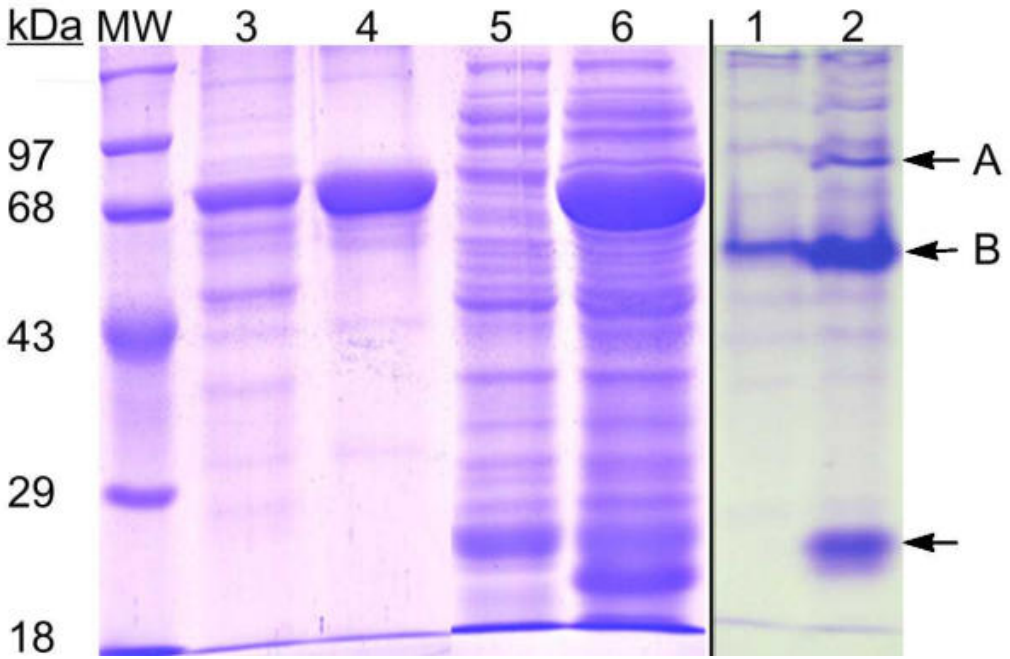
- Primarily used to separate proteins and DNA
- Proteins are charged (charge depends on pH, c.f. isoelectric point)
- Migrate through a gel if an external electrostatic field is applied
- Migration distance depends on charge (and/or size)



http://pharmaexposure.blogspot.de/2011/06/gel-electrophoresis.html

Gel Electrophoresis

- 1D-Polyacrylamide gel electrophoresis (PAGE)

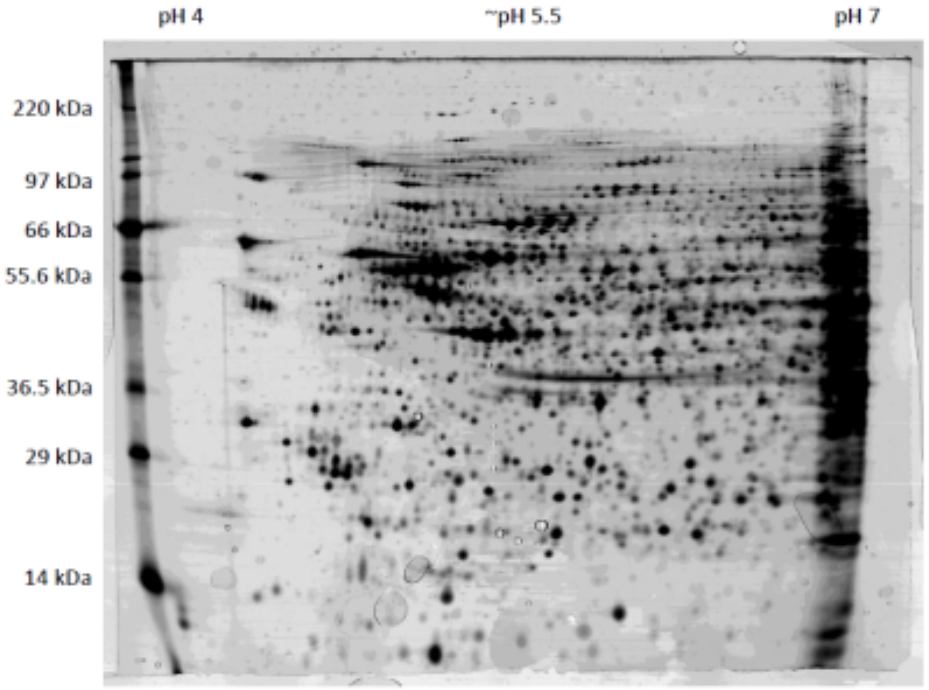


  - Cut gel into slices
  - Analyze slices separately

Senejani *et al*. *BMC Biochemistry*. 2001

Gel Electrophoresis

- 2D-Polyacrylamide gel electrophoresis (PAGE)



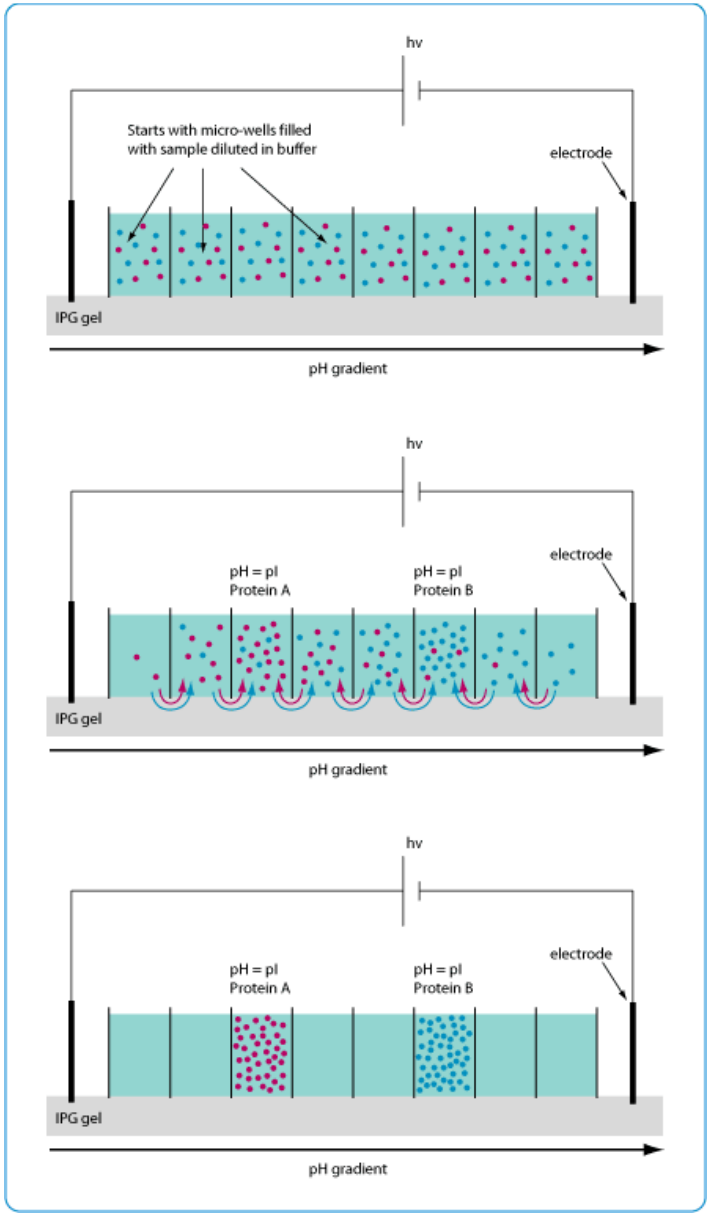
  - Initial separation on a pH gradient, then second separation (orthogonal to the first) based on size
  - Excise single protein spots
  - Analyze the protein spots separately

Off-Gel Separation

- Fractionation of peptides (or proteins) according to their *pI*)
- Reduce sample complexity: mixture will be split into several fraction
- Each fraction can be analyzed separately
- Analytes are kept in solution (they are kept off the gel)

Potential problems:

1. very basic or acidic peptides will not be captured
2. Measurement time is multiplied by the number of fractions
3. Protein quantification will have to include peptides from different fractions



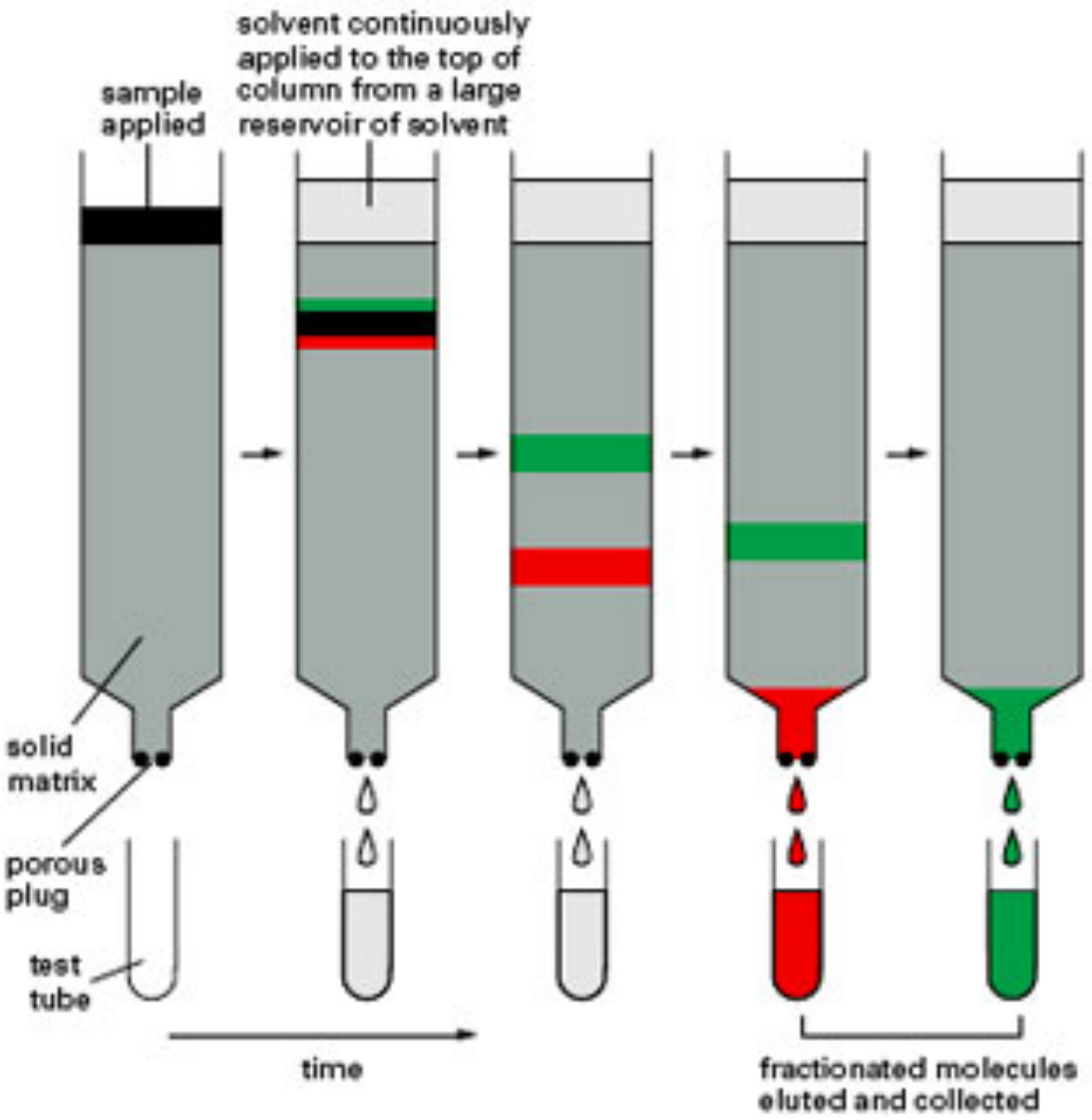
http://www.chem.agilent.com

Chromatography

- Chromatography is a separation technique
- From greek *chroma* and *graphein* – *color* and *to write*
- Initially developed by **Mikhail Semyonovich Tsvet**
- Simple fundamental idea:
  - Two phases: stationary and mobile
  - Analytes are separated while mobile phase passes along the stationary phase
- Various separation mechanisms, various choices for mobile/ stationary phases possible

M. S. Tsvet (1872-1919)

Column Chromatography



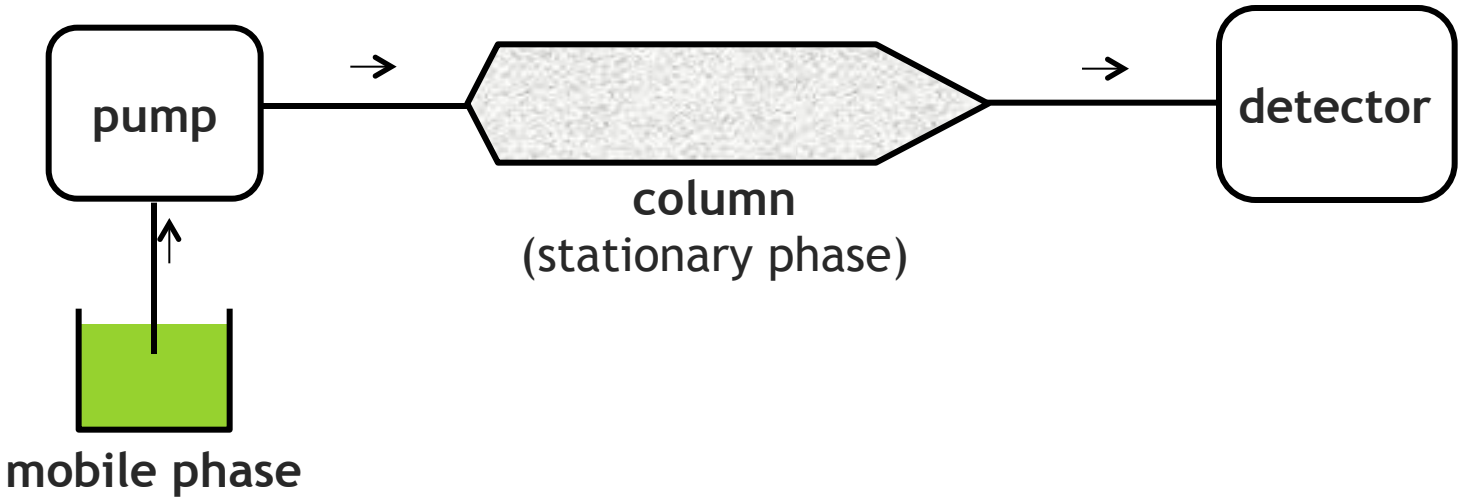
http://fig.cox.miami.edu/~cmallery/255/255hist/ecbxp4x3_chrom.jpg

Chromatography

- **Liquid chromatography** (LC)
  - Mobile phase liquid, stationary phase usually solid
  - Very versatile technique
  - High-Performance Liquid Chromatography (HPLC) for analytical purposes
- **Gas chromatography** (GC)
  - Mobile phase is a gas passing over the solid phase
  - Usually at higher temperatures
  - Limited to volatile compounds
- Others
  - Thin-Layer Chromatography (TLC)
  - Paper Chromatography (PC)
  - …

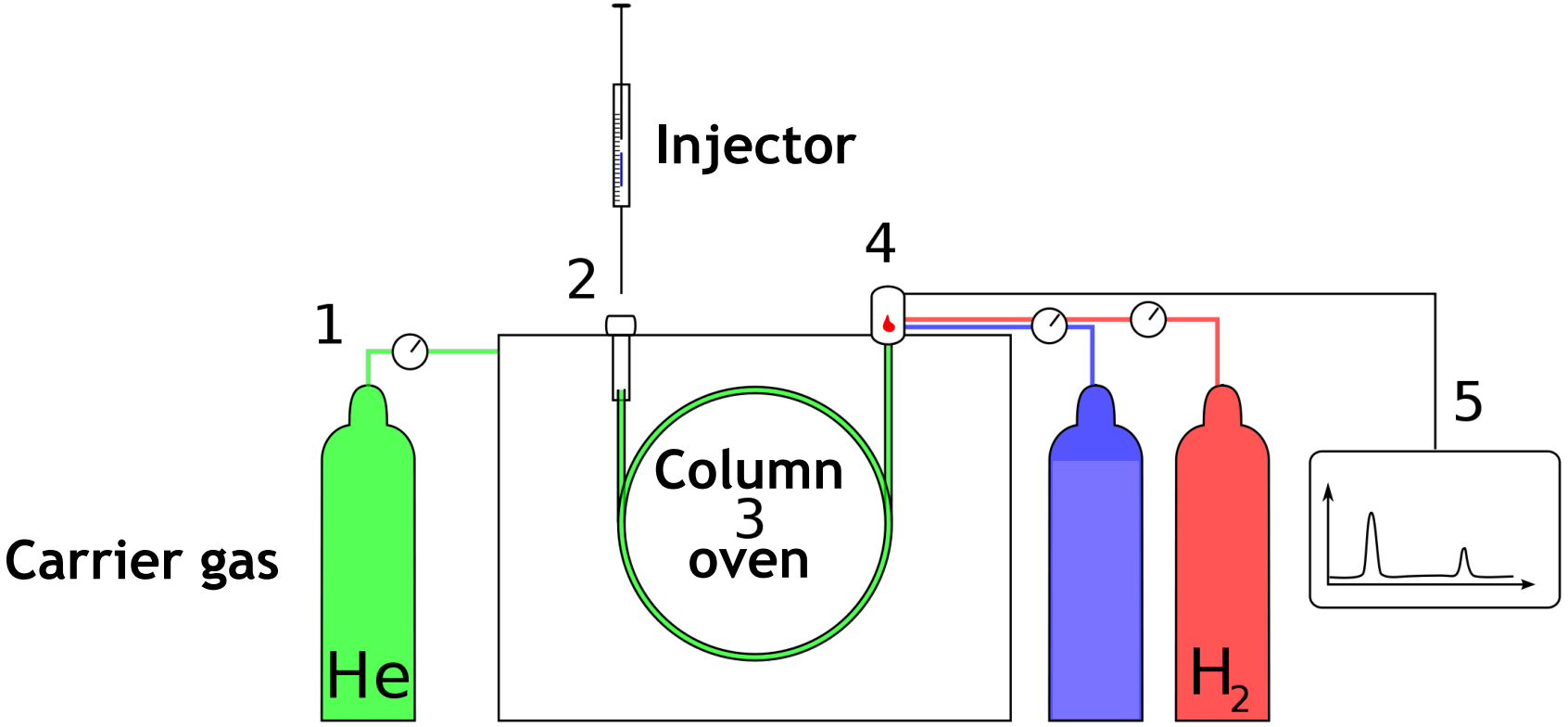
HPLC

- **High-performance liquid chromatography** (**HPLC**) uses small columns (µm inner diameter) and very high pressure (600 bar)
- **Reversed-phase (RP) chromatography** is the most common type: hydrophobic stationary phase, hydrophilic eluent (water/ acetonitrile) as mobile phase
- More hydrophobic analytes elute later than hydrophilic analytes



Gas chromatography

- Long column (10-200 m)
- Column is operated at very high temperatures (up to 450 °C)
- Requires analytes that are gaseous or evaporate easily
- **Derivatisation:** Convert non-volatile compounds to a volatile derivatives



http://de.wikipedia.org/wiki/Gaschromatographie

COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

1. Proteomics and Metabolomics

LU 1C - INTRODUCTION TO MASS SPECTROMETRY

- Definition of mass spectrometry, mass spectrum
- Overview of the three components of an MS (ion source, mass analyzer, detector)
- Molecular and atomic masses
- Isotope pattern/distribution, fine structure of isotope distribution

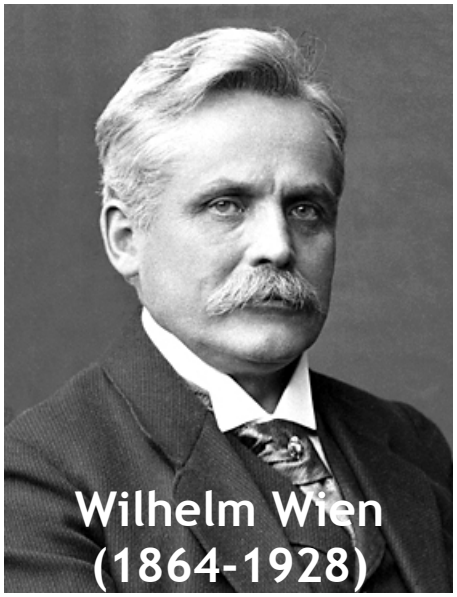
Mass Spectrometry

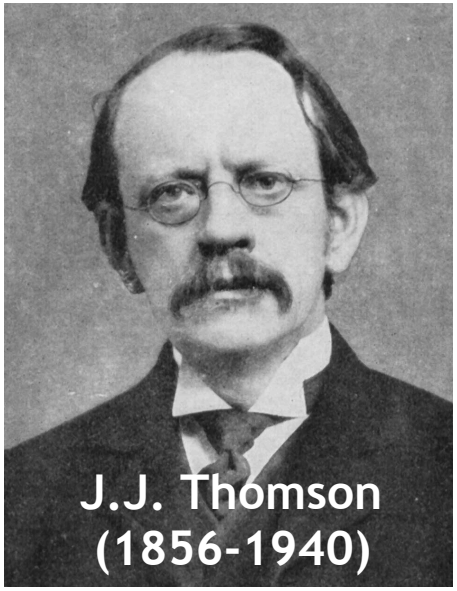
- **Definition:** **Mass spectrometry** is an analytical technique identifying type and amount of analytes present in a sample by measuring abundance and mass-to-charge ratio of analyte ions in the gas phase.

- Mass spectrometry is often abbreviated **mass spec** or **MS**
- The term **mass spectroscopy** is related, but its use is discouraged
- Mass spectrometry can cover a wide range of analytes and usually has very high sensitivity

Mass Spectrometry – Early History

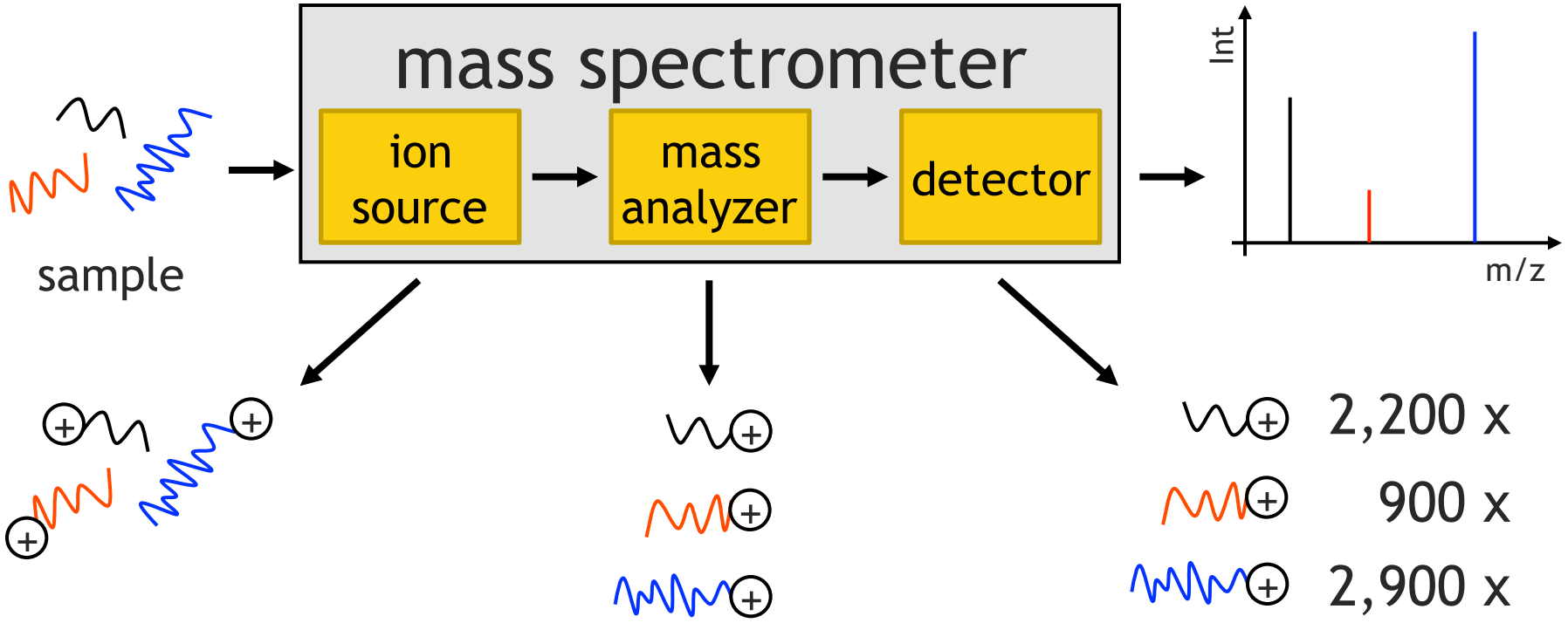
- **Wilhelm Wien** was the first to separate charged particles with magnetic and electrostatic fields in 1899
- **Sir Joseph J. Thomson** improved on these designs
- Sector mass spectrometers were used for separating uranium isotopes for the Manhattan project
- In the 1950s and 1960s **Hans Dehmelt** and **Wolfgang Paul** developed the ion trap





http://www.nobelprize.org/nobel_prizes/physics/laureates/1911/wien.jpg
http://en.wikipedia.org/wiki/J._J._Thomson#mediaviewer/File:J.J_Thomson.jpg

Components of a Mass Spectrometer

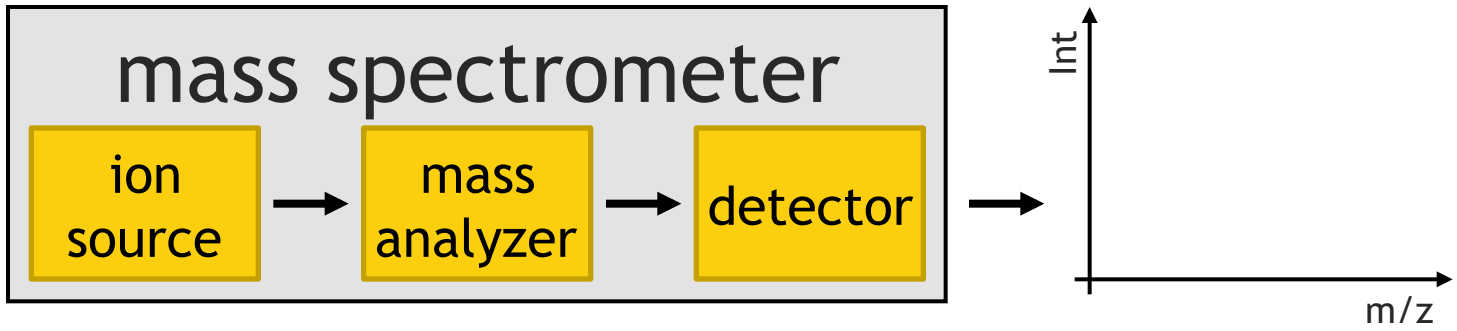


A mass spectrometer has three key components

- **Ion source** – converting the analytes into charged ions
- **Analyzer** – determining (and filtering by) mass-to-charge ratio
- **Detector** – detecting the ions and determining their abundance

Combining LC and Mass Spectrometry

- MS can be used as a very sensitive detector in chromatography
- It can detect hundreds of compounds (metabolites/peptides) simultaneously
- Coupling mass spectrometry to HPLC is then called **HPLCS-MS** (so-called 'hyphenated technique')
- **Idea:** analytes elute off the column and enter the MS more or less directly



Key Ideas in MS

- Ions are accelerated by electrostatic and electromagnetic fields
- Neutral molecules are unaffected
- Same idea: gel electrophoresis – but MS in vacuum/gas phase
- Force acting into a charged particle is governed by **Lorentz force**:

  $$F = q \cdot (E + v \times B)$$

  where
  - *q* is the charge of the particle, **v** is the velocity of the particle
  - **E** is the electric field, **B** is the magnetic field
  - **F** the force acting on the particle
- Together with **Newton's second law of motion** $F = m \cdot a$ we see that the acceleration **a** of the particle relates to the mass-to-charge ratio *m*/*q*:

  $$a = \frac{(E + v \times B)}{m/q}$$

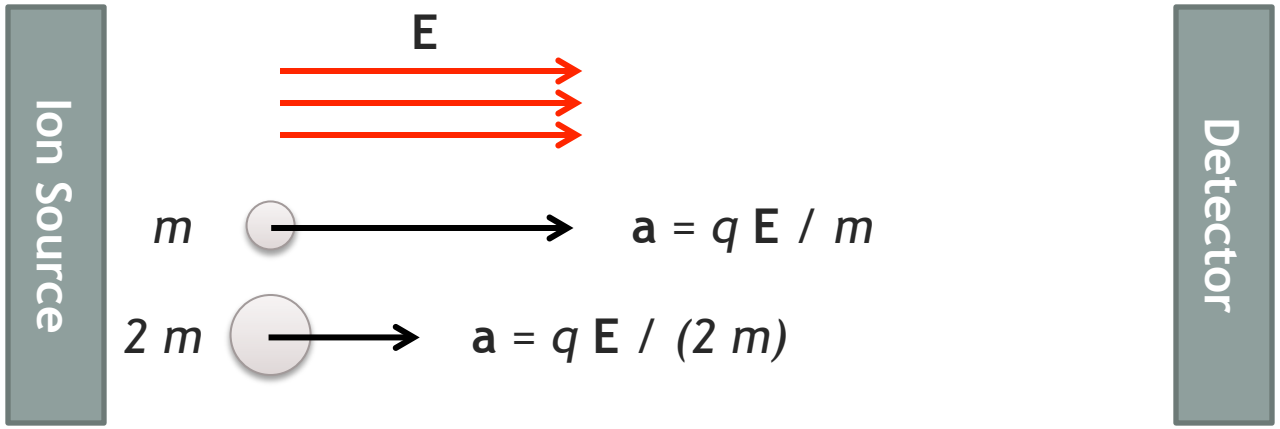
- Acceleration of the ions is then used to determine *m*/*q*

Key Ideas in MS

Acceleration: $a = \frac{(E + v \times B)}{m/q}$

Example:

- Applying the same **electrostatic field E** to different ions (e.g., different peptide ions) will result in a different acceleration, if they differ in the mass-to-charge ratio
- An ion with the twice the mass, but the same charge, will thus experience half the acceleration – and will hit the detector later!



Molecular Mass and Atomic Mass

- Atoms (and thus molecules) have a **mass**
- **Isotopes**: all chemical elements have naturally occurring isotopes that have the same atomic number but different masses
- Masses are generally given in units of kg (SI unit), however, there are a few conventions for atomic and molecular masses
- **Atomic mass** is the rest mass of an atom in its ground state
- Atomic mass is generally expressed in **unified atomic mass units**, which corresponds to 1/12 of the weight of ^{12}C ($1.6605402 \times 10^{-27}$ kg)
- Commonly used is also the non-SI unit 1 Dalton [Da], which is equivalent to the unified atomic mass unit. (1 Thompson [Th]) is equivalent)
- Another deprecated unit equivalent to Da still found in literature is *atomic mass unit (amu)*

IUPAC definition: doi:10.1351/goldbook.A00496

Molecular Mass

- Mass of a molecule is the sum of the masses of its atoms
- **Accurate mass** of a molecule is an experimentally determined mass
- **Exact mass** of a molecule is a theoretically calculated mass of a molecule with a specified isotopic composition
- **Molecular weight** or **relative molecular mass** is the ratio of a molecule's mass to the unified atomic mass unit
- For ions the mass of the missing/extra electron resp. proton needs to be included as well!

Note:

- Terms are not always used properly in the literature
- Be cautious with masses you google somewhere
- Reference: masses defined by IUPAC commission

Murray et al., Pure Appl. Chem., Vol. 85, No. 7, pp. 1515–1609, 2013
http://www.iupac.org/home/about/members-and-committees/db/division-committee.html?tx_wfqbe_pi1%5bpublicid%5d=210

Isotopes

- Isotopes are atom species of the same chemical element that have different masses
- Same number of protons and electrons, but different number of neutrons
- *For proteomics*: main elements occurring in proteins are C, H, N, O, P, S

Isotope	Mass [Da]	Nat. abundance [%]	Isotope	Mass [Da]	Nat. abundance [%]
^{1}H	1.007 825 0322(6)	99.985	^{16}O	15.994 914 620(2)	99.76
^{2}H	2.014 101 7781(8)	0.015	^{17}O	16.999 131 757(5)	0.038
^{12}C	12 (exact)	98.90	^{18}O	17.999 159 613(6)	0.2
^{13}C	13.003 354 835(2)	1.1	^{31}P	30.973 761 998(5)	100
^{14}N	14.003 074 004(2)	99.63	^{32}S	31.972 071 174(9)	95.02
^{15}N	15.000 108 899(4)	0.37	^{33}S	32.971 458 910(9)	0.75
			^{34}S	33.967 8670(3)	4.21

http://www.ciaaw.org/

Mass Number, Nominal, and Exact Mass

- The **mass number** is the sum of protons and neutrons in a molecule or ion
- The **nominal mass** of an ion or molecule is calculated using the most abundant isotope of each element rounded to the nearest integer
- The **exact mass** of an ion or molecule is calculated by assuming a single isotope (most frequently the lightest one) for each atom
- Exact mass is based on the (experimentally determined!) atomic masses for each isotope – numbers are regularly updated by IUPAC (International Union for Pure and Applied Chemistry)

Example:

Nominal mass of glycine ($C_2H_5NO_2$):

$$2 \times 12 + 5 \times 1 + 14 \times 1 + 16 \times 2 = 75$$

Exact mass of glycine ($C_2H_5NO_2$) using the lightest isotopes:

$$2 \times 12.0 + 5 \times 1.00782503226 + \ldots = 75.0320284\ldots$$

Monoisotopic Mass, Mass Defect

- **Monoisotopic mass** of a molecule corresponds to the exact mass for the most abundant isotope of each element of the molecule/ ion
- Note that for small elements (e.g., C,H,N,O,S) the most abundant isotope is also the lightest one
- **Mass defect** is the difference between the mass number and the monoisotopic mass
- **Mass excess** is the negative mass defect

Example

Monoisotopic mass of glycine ($C_2H_5NO_2$): 75.0320284…

Nominal mass of glycine: 75

Mass excess of glycine: 0.0320284…

Average Mass

- The **average mass** of a molecule is calculated using the average mass of each element weighted for its isotope abundance
- These average masses (weighted by natural abundance) are also the masses tabulated in most periodic tables

Example:

Average mass of glycine ($C_2H_5NO_2$):

$$
\begin{aligned}
& 2 \times (0.9890 \times M(^{12}C) + 0.0110 \times M(^{13}C)) \\
+ \; & 5 \times (0.99985 \times M(^{1}H) + 0.00015 \times M(^{2}H)) \\
+ \; & 1 \times (0.09963 \times M(^{14}N) + 0.00037 \times M(^{15}N)) \\
+ \; & 2 \times (0.9976 \times M(^{16}O) + 0.00038 \times M(^{17}O) + 0.002 \times M(^{18}O)) \\
= \; & \mathbf{75.0666\ Da}
\end{aligned}
$$

Simpler alternative: use average atomic weights from PTE!

Accurate Mass and Composition

- **Accurate mass** is an **experimentally determined** mass of an ion or molecule and it can be used to determine the **elemental formula**
- Accurate mass comes with a known accuracy or (relative) error, which is usually determined in ppm
  (10^{-6} = parts per million)
- Most mass spectrometers have a constant ***relative* mass accuracy** - **absolute mass error** often increases linearly with the measured mass

Example:

Measured accurate mass of valine ($C_5H_{11}NO_2$): 117.077 Da

Monoisotopic mass of valine: 117.078979 Da

Absolute mass error: -0.0178979 Da

Relative mass error: -0.0178979 Da/117.078979 Da = -16.9 ppm

IUPAC Terms

IUPAC (International union of pure and applied chemistry) defines the meaning of all the terms – so if you are unsure, look them up in the IUPAC Gold Book and in the IUPAC recommendations:

- **Exact mass**
- **Monoisotopic mass**
- **Average mass**
- **Mass number**
- **Nominal mass**
- **Mass defect**
- **Mass excess**
- **Accurate mass**

Murray et al., Pure Appl. Chem., Vol. 85, No. 7, pp. 1515–1609, 2013
http://goldbook.iupac.org/index.html

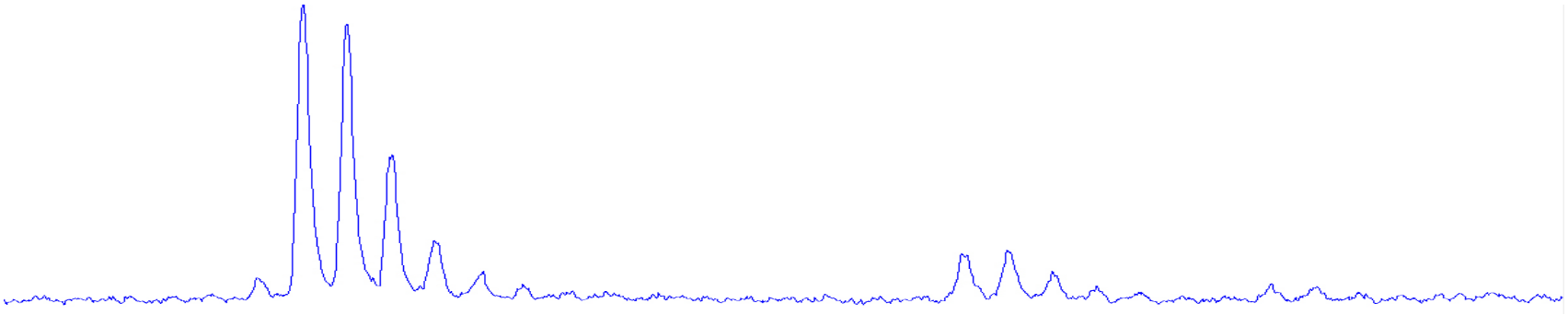
Isotope Patterns

- Molecule with one carbon atom
  - Two possibilities:
    - light variant ^{12}C
    - Heavy variant ^{13}C
  - 98.9% of all atoms will be light
  - 1.1% will be heavy

^{12}C	98.90%	^{13}C	1.10%		
^{14}N	99.63%	^{15}N	0.37%		
^{16}O	99.76%	^{17}O	0.04%	^{18}O	0.20%
^{1}H	99.98%	^{2}H	0.02%		

Isotope Patterns

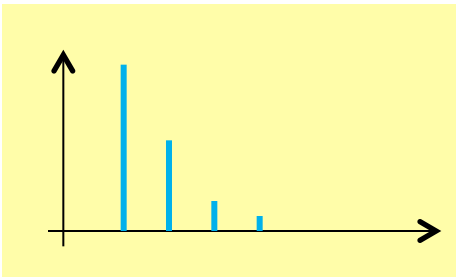
- Molecule with 10 carbon atoms
  - Lightest variant contains only ^{12}C
  - This is called ‘monoisotopic’
  - Others contain 1-10 ^{13}C atoms, these are heavier by 1-10 Da than the monoisotopic one
- In general, the relative intensities follow a binomial distribution, depending on the number of atoms
- For higher masses (i.e., a larger number of atoms), the **monoisotopic peak** will be no longer the most likely variant

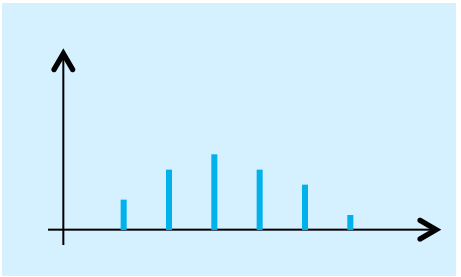


Isotope Patterns

- It is possible to compute approximate isotope patterns for any given *m*/*z*, by estimating the average number of atoms
- Heavier molecules have smaller monoisotopic peaks
- In the limit, the distribution approaches a normal distribution

m [Da]	P (k=0)	P (k=1)	P (k=2)	P (k=3)	P (k=4)
1,000	0.55	0.30	0.10	0.02	0.00
2,000	0.30	0.33	0.21	0.09	0.03
3,000	0.17	0.28	0.25	0.15	0.08
4,000	0.09	0.20	0.24	0.19	0.12



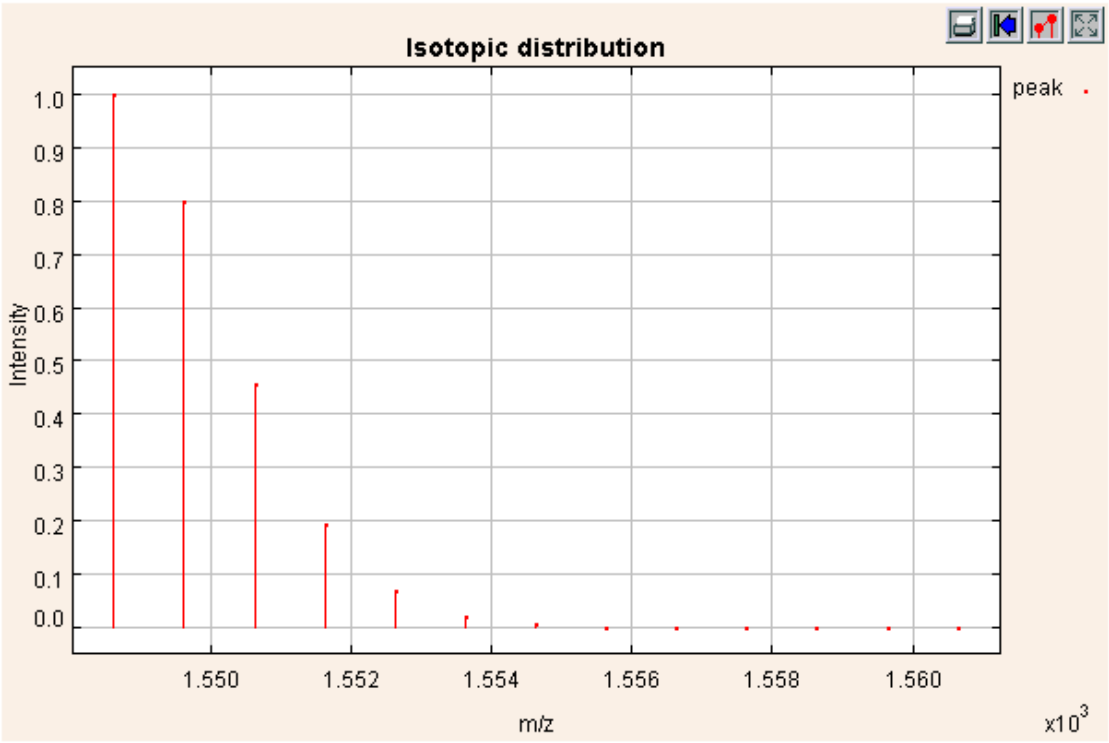


Online Calculator

The name of the query is: Unknown
The type of composition you've chosen is: Protein One-letter code [M]
You have entered:
TESTPEPTIDECPM

Element	Mass
C_{63}	756.674
H_{100}	100.794
N_{14}	196.094
O_{27}	431.984
S_2	64.130
Average mass:	1549.676

The monoisotopic mass is:
1548.632
Monoisotopic combination:
$^{12}C_{63}$
$^{1}H_{100}$
$^{14}N_{14}$
$^{16}O_{27}$
$^{32}S_2$

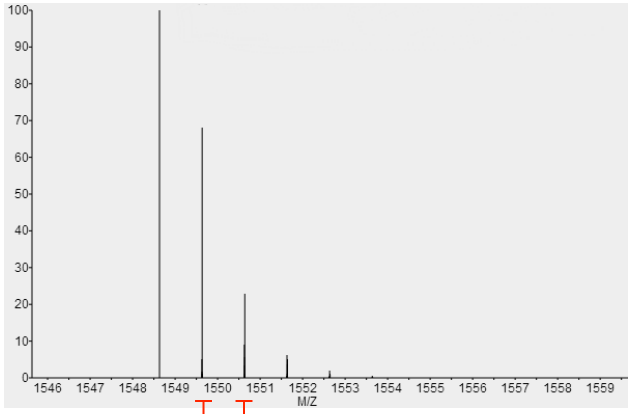


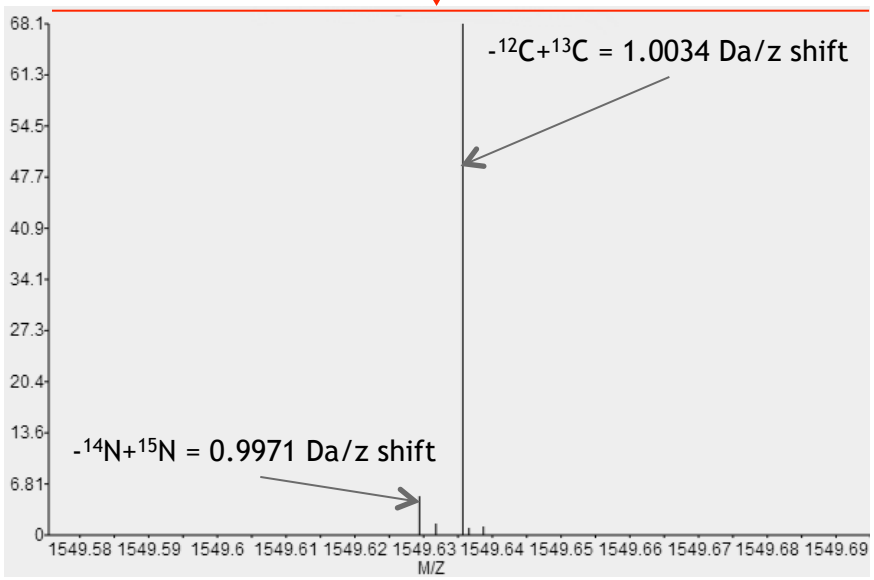
Most likely isotope combination:
$^{12}C_{63}$
$^{1}H_{100}$
$^{14}N_{14}$
$^{16}O_{27}$
$^{32}S_2$
Exact mass is 1548.632
Probability of combination is 39.255%
The most likely combination is 100.00% of those masses rounding to 1548 amu.

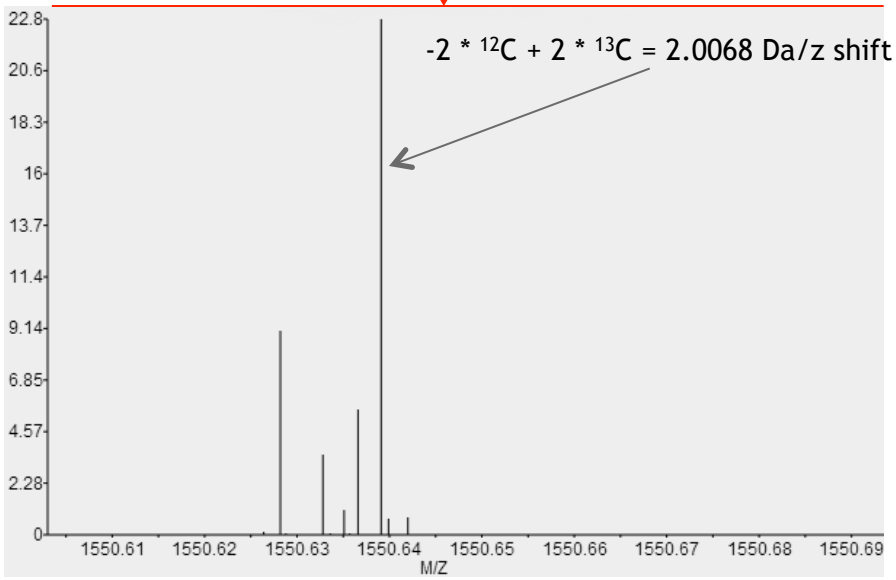
Download the data

Isotopic Fine Structure

- High-resolution MS reveals isotopic fine structure (Why?)







Computing the Isotopic Distribution

- For simplicity's sake, we will consider only nominal masses and no isotopic fine structure here
- Let *E* be a chemical element (e.g. H or N).
- Let $\pi_E[i]$ be the probability (i.e., natural abundance) of the isotope of *E* with *i* additional neutrons (*i* = 0 for the lightest isotope of *E*)
- Relative intensities of pure *E* are given by ($\pi_E[0]$, $\pi_E[1]$, … $\pi_E[k_E]$, where k_E = nominal mass shift of heaviest isotope of *E*)
- Given a molecule composed of two atoms of elements *E* and *E'*
- Probability for additional neutrons in the molecule is then the sum over all possible combinations and their respective probabilities

$$\pi_{EE'}[n] = \sum_{i=0}^{n} \pi_E[i]\pi_{E'}[n-i] \qquad \pi_{EE'}[l] = 0 \text{ for } l > k_E + k_{E'}$$

Computing the Isotopic Distribution

- This is known as a **convolution** and we can write

$$\pi_{EE'} = \pi_E * \pi_{E'}$$

with the convolution operator *

- **Convolution powers**

  Let $p^1 := p$ and $p^n :- p^{n-1} * p$ for any isotope distribution p

  p^0 with $p^0[0] = 1$, $p^0[l] = 0$ for $l > 0$ is the **neutral element** with respect to the operator *

Example: Compute the isotope distribution of CO

$$\pi_{CO}[0] = \pi_C[0]\ \pi_O[0]$$

$$\pi_{CO}[1] = \pi_C[1]\ \pi_O[0] + \pi_C[0]\ \pi_O[1]$$

$$\pi_{CO}[2] = \pi_C[2]\ \pi_O[0] + \pi_C[1]\ \pi_O[1] + \pi_C[0]\ \pi_O[2]$$

Computing the Isotopic Distribution

- The isotopic distribution for the chemical formula

$$E_{n_1}^1 \ldots E_{n_l}^l$$

consisting of n_i atoms of elements $E^1 \ldots E^l$ can be computed as

$$\pi_{E_{n_1}^1 \ldots E_{n_l}^l} = \pi_{E_1}^{n_1} * \ldots * \pi_{E_l}^{n_l}$$

- **Runtime:** quadratic in the number of atoms
  - Number of convolution operators is $n_1 + n_2 + \ldots n_l - 1$ and is thus linear in the number of atoms n
  - Convolution operator involves a summation for each $\pi[i]$
  - If the highest isotopic rank for E is k_E, then the highest isotopic rank of E_n is $n\ k_E$ – again, linear in the number of atoms

- There are several tricks and practical considerations to speed up these calculations

LU 1D – BASIC PROTEOMIC TECHNIQUES AND APPLICATIONS

- Definition and size of the proteome
- Protein databases
- Amino Acid masses, posttranslational modifications, protein isoforms
- Top-down proteomics
- Shotgun proteomics, tryptic digest
- Applications: clinical proteomics, signaling

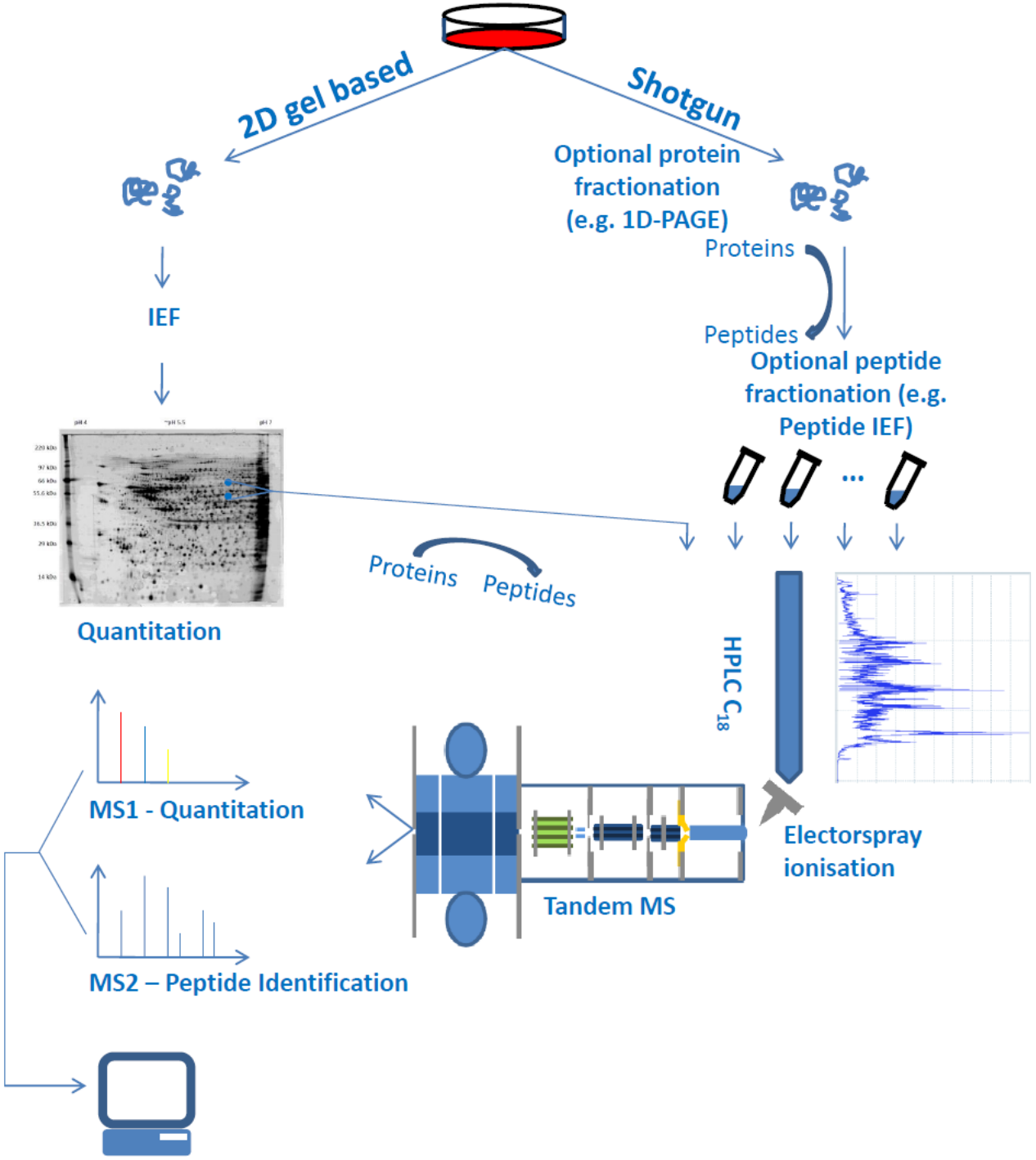
Top-Down vs. Bottom-Up Proteomics

- Two fundamentally different approaches in proteomics
  - **Top-down proteomics**: intact proteins are analyzed
  - **Bottom-up proteomics** (**shotgun proteomics**): proteins are digested to peptides, peptides are analyzed
- Bottom-up approaches are currently more popular
  - Absolute mass error increases with measured mass to charge (m/z) value
  - Hard to determine mass for a protein – broad mass distribution
  - The sensitivity of mass measurements at a protein range is significantly worse than at peptide level
  - The existence of modifications complicates the analysis of complete proteins
  - Peptides are easier to separate using HPLC than proteins

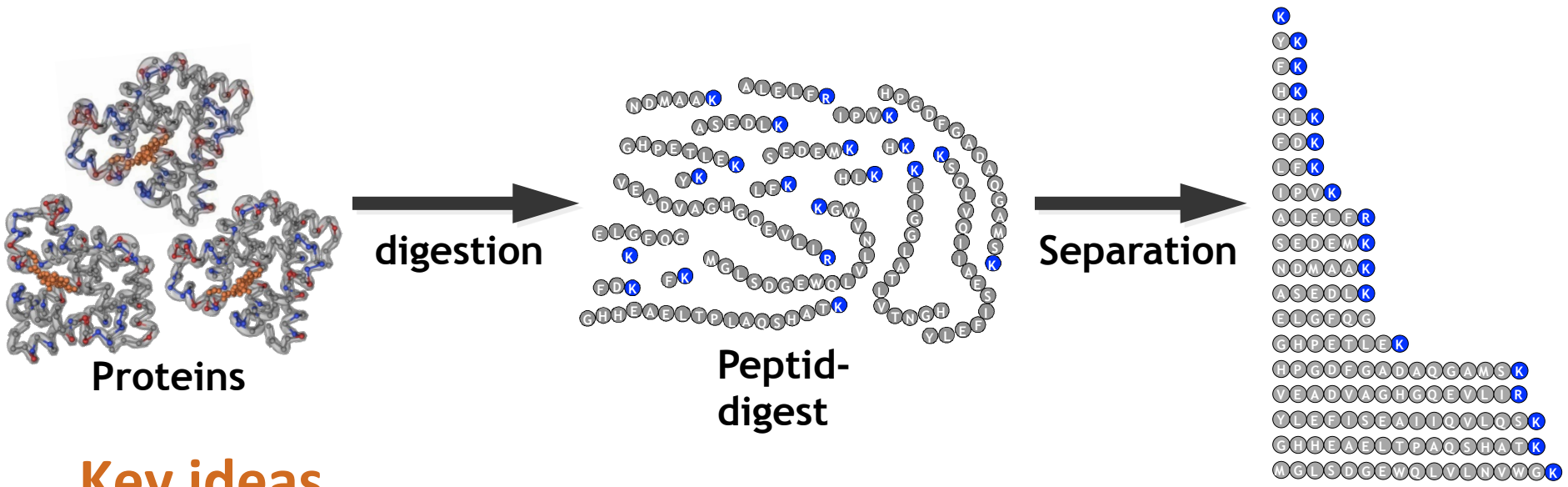
Bottom-Up Proteomics

- Note that bottom-up and shotgun proteomics are used equivalently most of the time
- There are two conceptually different approaches in bottom-up proteomics

Peptide mass fingerprinting	Peptide sequencing
• Peptide masses are used to identify the protein • Often used in combination with 2D gels	• Peptides are fragmented • Fragments are used to interfere the sequence • Shotgun proteomics


2D gel based
Shotgun
Optional protein fractionation (e.g. 1D-PAGE)
Proteins
Peptides
IEF
Optional peptide fractionation (e.g. Peptide IEF)
Quantitation
Proteins
Peptides
HPLC C_{18}
MS1 - Quantitation
Tandem MS
Electorspray ionisation
MS2 – Peptide Identification

Shotgun Proteomics



Key ideas

- Separation of whole proteins possible but difficult, hence digestion preferred
- Usually: trypsin – cuts after K and R and ensures peptides suitable for MS (positive charge at the end)
- Separate peptides; this is easier than separating proteins
- Identify proteins through peptides

Amino Acid Masses

AA	Chemical formula	Mono-isotopic [Da]	Average [Da]
Ala	C_3H_5ON	71.03711	71.0788
Arg	$C_6H_{12}ON_4$	156.10111	156.1875
Asn	$C_4H_6O_2N_2$	114.04293	114.1038
Asp	$C_4H_5O_3N$	115.02694	115.0886
Cys	C_3H_5ONS	103.00919	103.1388
Glu	$C_5H_7O_3N$	129.04259	129.1155
Gln	$C_5H_8O_2N_2$	128.05858	128.1307
Gly	C_2H_3ON	57.02146	57.0519
His	$C_6H_7ON_3$	137.05891	137.1411
Ile	$C_6H_{11}ON$	113.08406	113.1594

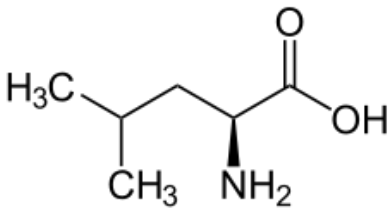
AA	Chemical formula	Mono-isotopic [Da]	Average [Da]
Leu	$C_6H_{11}ON$	113.08406	113.1594
Lys	$C_6H_{12}ON_2$	128.09496	128.1741
Met	C_5H_9ONS	131.04049	131.1926
Phe	C_9H_9ON	147.06841	147.1766
Pro	C_5H_7ON	97.05276	97.1167
Ser	$C_3H_5O_2N$	87.03203	87.0782
Thr	$C_4H_7O_2N$	101.04768	101.1051
Trp	$C_{11}H_{10}ON_2$	186.07931	186.2132
Tyr	$C_9H_9O_2N$	163.06333	163.1760
Val	C_5H_9ON	99.06841	99.1326

Note: these masses are for **amino acid residues** – HN-CHR-CO, not the full amino acid! It is thus the mass by which a protein mass increases, if this amino acid is inserted in the sequence.

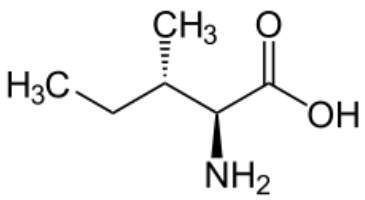
Amino Acid Masses

- Leu and Ile (L/I) are structural isomers
- They thus have identical mass and cannot be distinguished by their mass alone!
- Fragments with same mass are called isobaric
- Gln and Lys (Q/K) have nearly identical masses: **128**.09496 Da and **128**.05858 Da
- For low-resolution instruments they are indistinguishable, too

AA	Chemical formula	Mono-isotopic [Da]	Average [Da]
Leu	$C_6H_{11}ON$	113.08406	113.1594
Ile	$C_6H_{11}ON$	113.08406	113.1594
Gln	$C_5H_8O_2N_2$	128.05858	128.1307
Lys	$C_6H_{12}ON_2$	128.09496	128.1741



Leu



Ile

Post-Translational Modifications

- Alterations to the chemical structure of proteins after the translation are called **post-translational modifications** (**PTMs**)
- **Chemical modifications** (e.g., isotopic labels) are not PTMs
- The UniMod database (www.unimod.org) contains a wide range of potential modifications to
- PTMs play very important roles in cellular signaling
- Best known example: **phosphorylation**
  - Phosphorylation of amino acids (primarily Ser, Thr, Tyr) can activate or inactivate protein function
  - Example: MAP kinase pathway

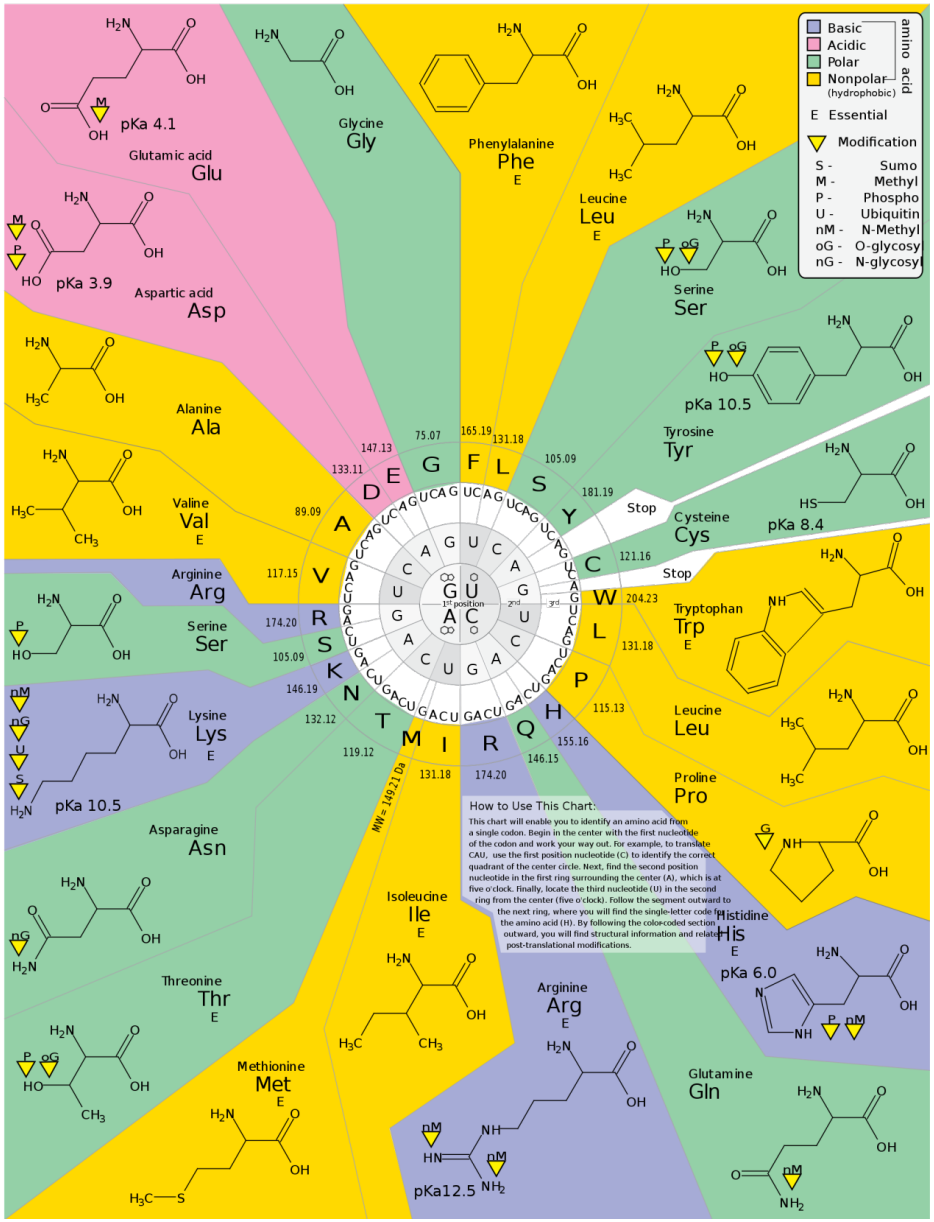
Post-Translational Modifications

Most common *in vivo* PTMs

- Phosphorylation
- Acetylation
- Oxidation
- Methylation
- Glycosylation
- …

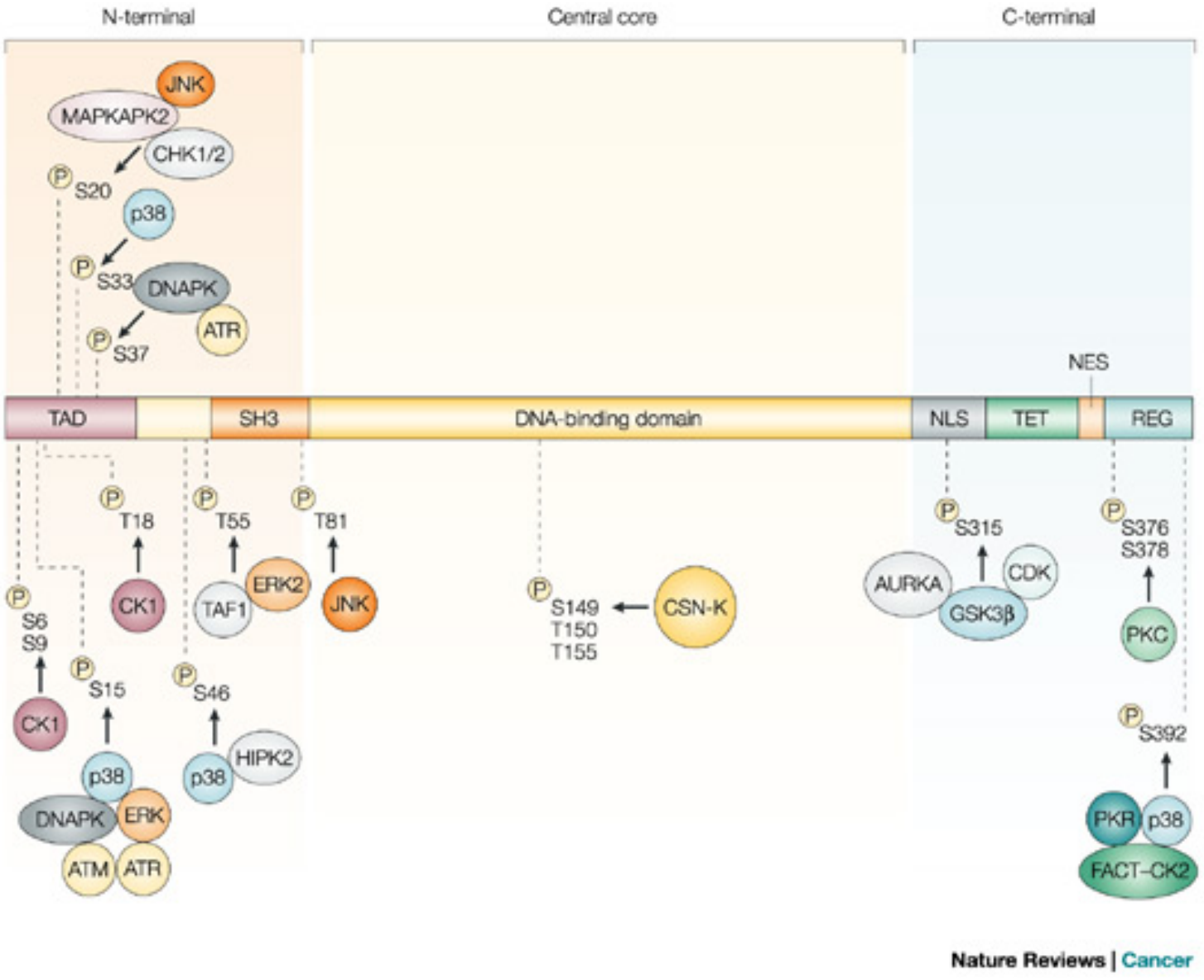
Mechanisms inducing PTMs

- Enzymes
- Covalent linking to other proteins
- Change of cellular conditions



http://www.enotes.com/topic/Posttranslational_modification

P53 Phosphorylation Sites



Protein Sequence Databases

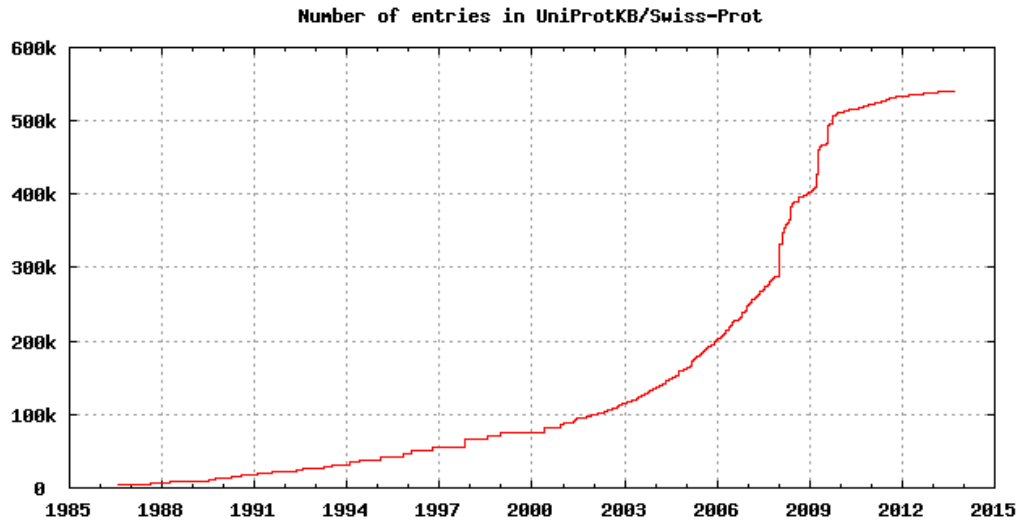
- Protein sequence databases are important to link mass spectra to proteins
- These databases do not only provide sequence information, but also:
  - Names
  - Taxonomy
  - Polymorphisms, isoforms, PTMs, etc.
- Important databases
  - **UniProt** Knowledgebase (SwissProt/ TrEMBL, PIR)
  - **NCBI** non redundant database
  - The International Protein Index (**IPI**)
  - **NextProt**

UniProtKB

- Is built on three established databases: SwissProt, TrEMBL and PIR (Protein Information Resource)
- It contains:
  - Accession number that serves as an unique identifier for the sequence
  - Sequence
  - Molecular mass
  - Observed and predicted modifications

UniProtKB/SwissProt

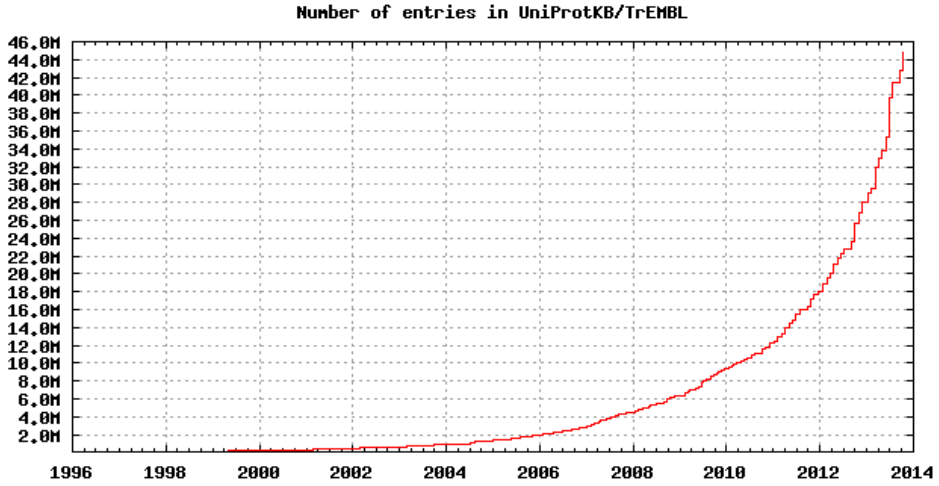
- http://www.uniprot.org/
- SwissProt is the manually curated section of the **UniProt Knowledgebase (UniProtKB)**
- Curated by ExPASy (Expert Protein Analysis System)
- manually annotated with minimal redundancy
- > 500k entries and 20,271 human proteins (Note: 20,248 in 2011)



http://web.expasy.org/docs/relnotes/relstat.html

TrEMBL

- http://www.uniprot.org/
- Translated EMBL nucleotide sequences
  - European Molecular Biology Laboratory / European Bioinformatics Institute (EBI)
- Computer annotated section of the **UniProt Knowledgebase** (UniProtKB)
  - 42,821,879 entries among them: 113,507 human proteins
  - 2011: 16,886,838 and 794,190 -> no saturation!



http://web.expasy.org/docs/relnotes/relstat.html

NCBI NR

- NCBI: National Center for Biotechnology Information
- Groups different information: SwissProt, TrEMBL and RefSeq
- RefSeq consists of XP and NP entries. For NP entries there is experimental evidence and XP entries are purely predicted
- NCBI is non-redundant at the absolute protein level -> no two sequences are identical
- History management is provided via the Entrez web interface

NextProt

- *http://www.nextprot.org*
- **neXtProt** is an on-line knowledge platform on human proteins
- Integrates various sources of information, such as UniProt, GeneOntology, ENZYME and PubMed
- Potentially best curated knowledgebase for human proteins: Oct. 2013: 20,133 human proteins

Other Databases

- MSDB (Mass Spectrometry DataBase): combination of different databases
- HPRD (Human Protein Reference Database): manually curated from literature
- PDB (Protein Data Bank): protein structure database

LU 1E – BASIC METABOLOMIC TECHNIQUES AND APPLICATIONS

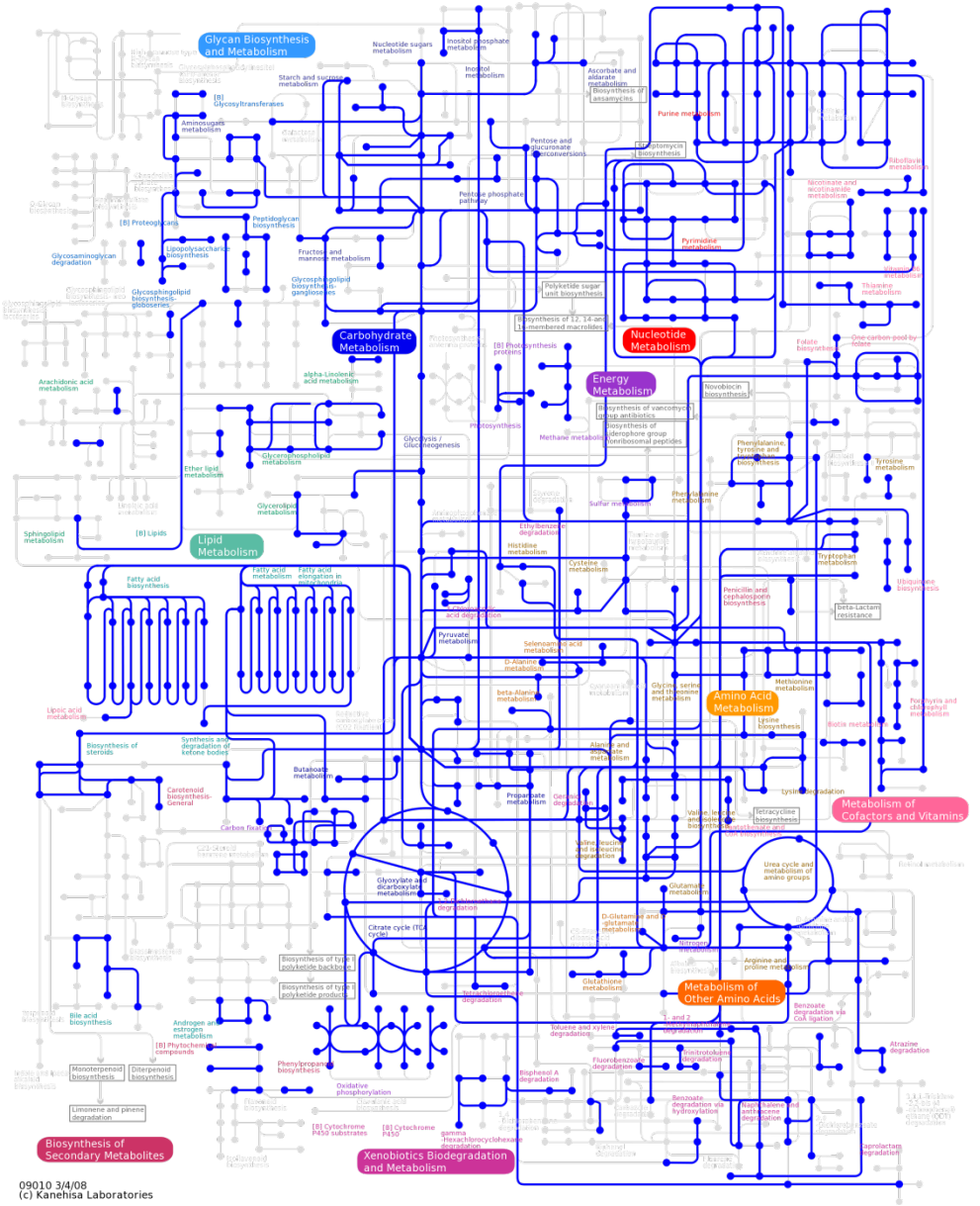
- Metabolome - differences, similarities to proteome, MW distribution
- Metabolic pathways, connection between proteome and metabolome
- Metabolomics databases
- Metabolomics techniques (targeted, non-targeted; LC, GC, NMR
- Metabolomics applications: biomarker discovery

Metabolism

- **Metabolism** = sum of all the chemical processes occurring in an organism at one time
- Concerned with the management of material and energy resources within the cell
- Two types of metabolic processes
  - **Anabolic processes** – processes constructing larger molecules from smaller units (building up)
  - **Catabolic processes** – processes breaking down larger units (degradation or energy generation)
- Metabolites are both educts and products of metabolic processes
- Enzymes (proteins) usually catalyze these metabolic processes (reactions)
- A sequence of several coupled metabolic processes is called a **metabolic pathway**

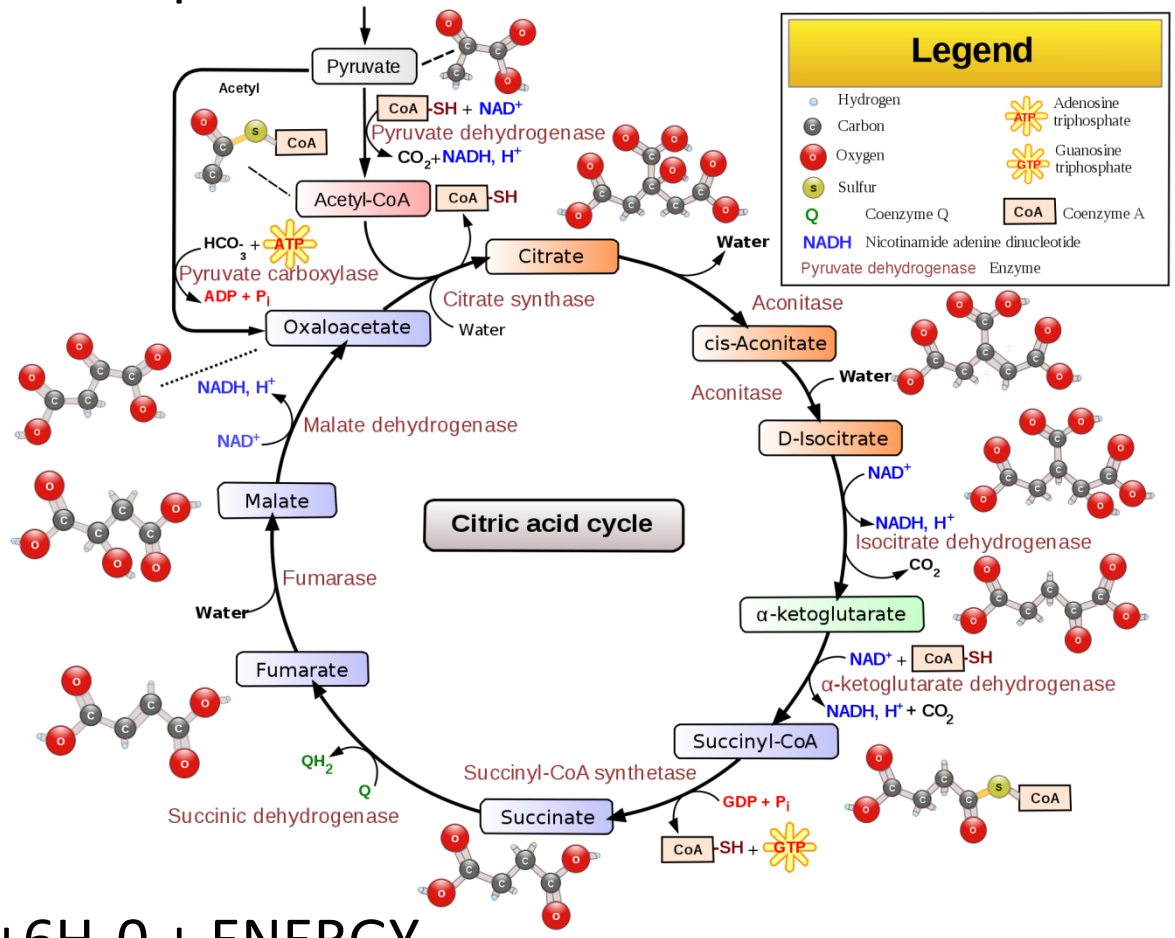
Metabolism

- Metabolic map *Rhodobacter capsulatus*
- Highly complex network structure



Catabolic Pathways

- Pathways that release energy by reaction that *catabolize* complex molecules to simpler compounds



Example:

Krebs cycle

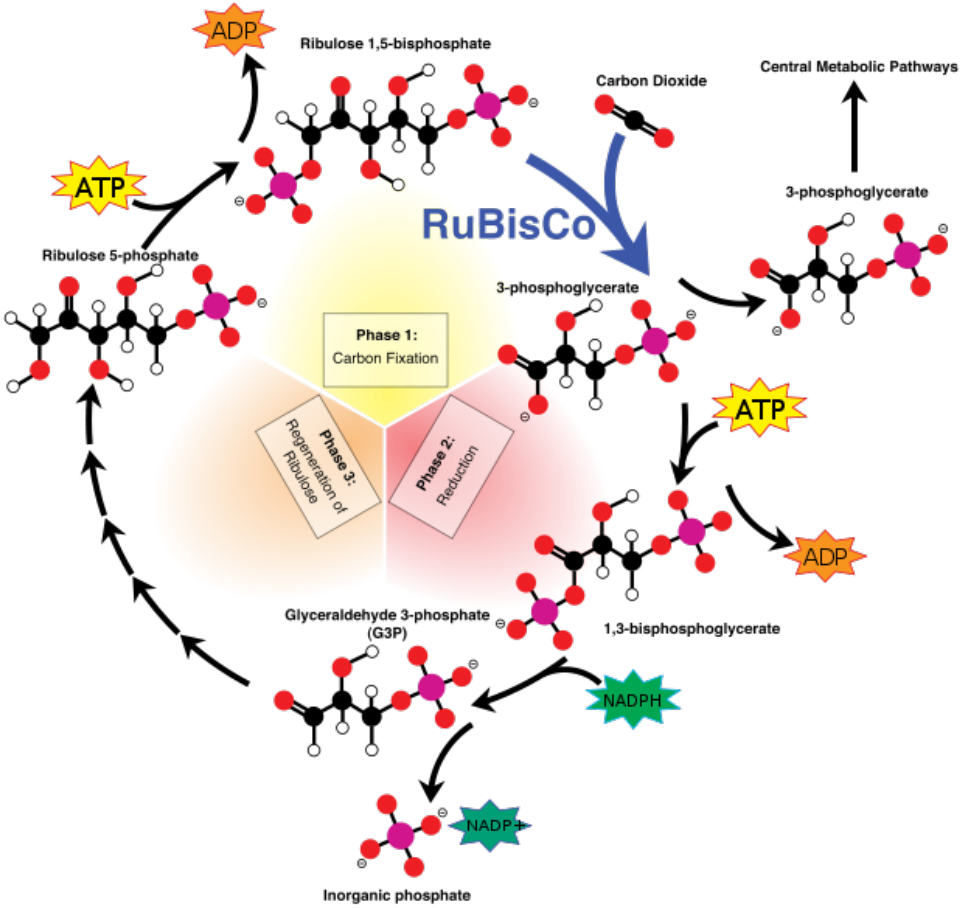
Cellular respiration:

$C_6H_{12}0_6 + 6O_2 \rightarrow 6CO_2 + 6H_20 + ENERGY$

http://en.wikipedia.org/wiki/Citric_acid_cycle

Anabolic Pathways

- Pathways that consume energy to *anabolize more* complex molecules from simpler compounds



Example:

Calvin cycle:

$$3\ CO_2 + 9\ ATP + 6\ NADPH + 6\ H^+ \rightarrow C_3H_6O_3\text{-phosphate} + 9\ ADP + 8\ P_i + 6\ NADP^+ + 3\ H_2O$$

http://no.wikipedia.org/wiki/Fil:Calvin-cycle3.png

Metabolites

- Metabolites comprise a heterogeneous set of biomolecules: all small molecules in a system excepting salts and macromolecules (proteins, long peptides, RNA, DNA)
- Lipids and sugars are metabolites as well
- There are separate fields dealing with lipids and sugars (lipidomics, glycomics), techniques are very similar

Examples:

Metabolite	mol l^{-1}	Metabolite	mol l^{-1}	Metabolite	mol l^{-1}
Glutamate	9.6×10^{-2}	UDP-glucuronate (51)	5.7×10^{-4}	*N*-Acetyl-ornithine (79)	4.3×10^{-5}
Glutathione	1.7×10^{-2}	ADP	5.6×10^{-4}	Gluconate (80)	4.2×10^{-5}
Fructose-1,6-bisphosphate	1.5×10^{-2}	Asparagine (52)	5.1×10^{-4}	Malonyl-CoA (81)	3.5×10^{-5}
ATP	9.6×10^{-3}	α-Ketoglutarate	4.4×10^{-4}	Cyclic AMP (82)	3.5×10^{-5}

Extracted from Bennett et al.: some of the most abundant small molecules in *E. coli*

Bennett et al. *Nature Chemical Biology.*
Bennett et al. *Nature Chemical Biology.* 2009

Metabolome vs. Proteome

- Size and complexity of the metabolome still largely unknown
- Similar to protein sequence databases, there are also metabolite databases listing all known metabolites (usually contains tens of thousands of metabolites)
- Differences between proteome and metabolome:
  - Metabolites belong to wider ranger of chemical compound classes (lipids, sugars, amino acids)
  - Proteins have a more homogenous chemistry (20 proteinogenic amino acids)
  - Metabolites can have complex structures that require a structural formula for a comprehensive description
  - Proteins have a simple, linear structure that can be represented by a sequence
  - Metabolites are **light**: average metabolite mass a 100-300 Da
  - Proteins are **heavy**: median protein length around 300-500 aa, about 40,000 Da molecular weight

Metabolomics Techniques

- Fundamentally two types of approaches
  - **Targeted metabolomics**
    - Identify only a well-defined subset of metabolites, but those with higher accuracy (hundreds?)
    - All metabolites can be identified
  - **Non-targeted metabolomics (metabolic profiling)**
    - Try to see as much of the metabolome as possible (thousands and more)
    - Majority of metabolites can be seen
    - Only a small fraction will be identified
- Similarly, there is also targeted and non-targeted proteomics
- In proteomics, the identification problem is less difficult, though, which is why this distinction is more relevant in metabolomics (where identification is much harder)

KEGG

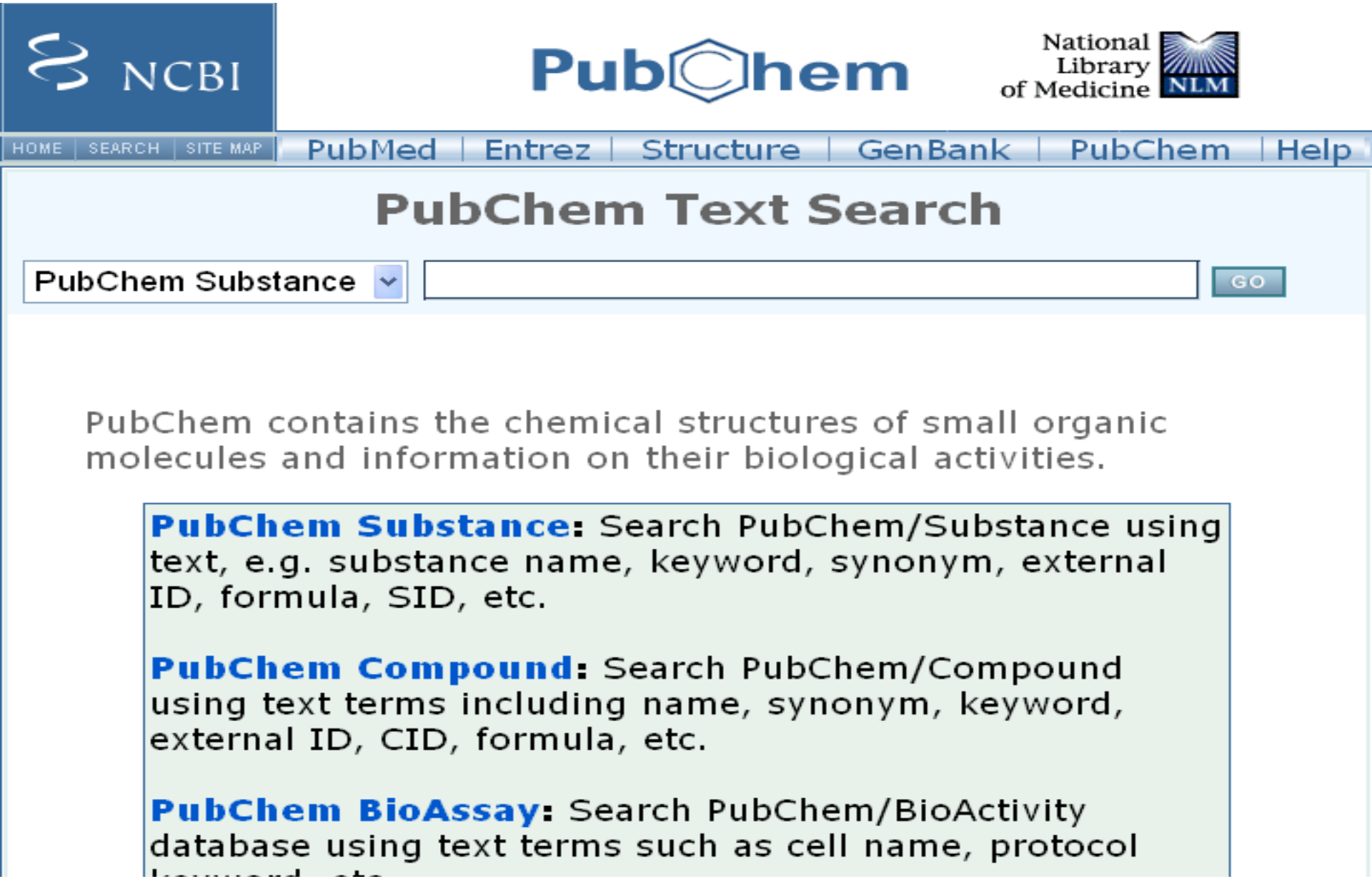
KEGG - Table of Contents

KEGG2 PATHWAY GENES LIGAND KO SSDB EXPRESSION BRITE XML API DBGET

1. KEGG Databases

Category	Database		Search & Compute	DBGET Search	
Pathway information	**KEGG PATHWAY Database**	XML	Search objects in KEGG pathways Color objects in KEGG pathways	PATHWAY	
Genomic information	**KEGG GENES Database**	KO	Search orthologs or gene clusters in SSDB	KO	
			Search similar GENES sequences	GENES	
			Search similar GENOME sequences	GENOME	
Chemical information	**KEGG LIGAND Database**	RC	Search similar compound structures	COMPOUND	LIGAND
			Search similar glycan structures	GLYCAN	
			Predict reactions and assign EC numbers	REACTION	
				RPAIR	

PubChem


NCBI
PubChem
National Library of Medicine
NLM
HOME
SEARCH
SITE MAP
PubMed
Entrez
Structure
GenBank
PubChem
Help
PubChem Text Search
PubChem Substance
GO
PubChem contains the chemical structures of small organic molecules and information on their biological activities.
PubChem Substance: Search PubChem/Substance using text, e.g. substance name, keyword, synonym, external ID, formula, SID, etc.
PubChem Compound: Search PubChem/Compound using text terms including name, synonym, keyword, external ID, CID, formula, etc.
PubChem BioAssay: Search PubChem/BioActivity database using text terms such as cell name, protocol

HumanCyc

HumanCyc

Encyclopedia of Human Genes and Metabolism

HumanCyc Quick S

BioCyc Home

Search

Database Search
Advanced Database Search
Help

News

Nov 09 BioCyc 8.6 released
Sep 17 BioCyc 8.5 released
Sep 17 Online Licensing

Services

Software/Data Download
User Support
Subscribe to Mailing List
EcoCyc T-shirts

Information

Introduction to BioCyc
Guided Tour
Pathway Tools Software

HumanCyc: Encyclopedia of *Homo sapiens* Genes and Metabolism

- Query the HumanCyc database

Authors

Pedro Romero, Markus Krummenacker and Peter D. Karp, SRI International.

Project Overview

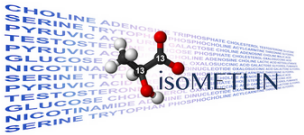
HumanCyc is a bioinformatics database that describes the human metabolic pathways and the human genome. By presenting metabolic pathways as an organizing framework for the human genome, HumanC provides the user with an extended dimension for functional analysis of

METLIN

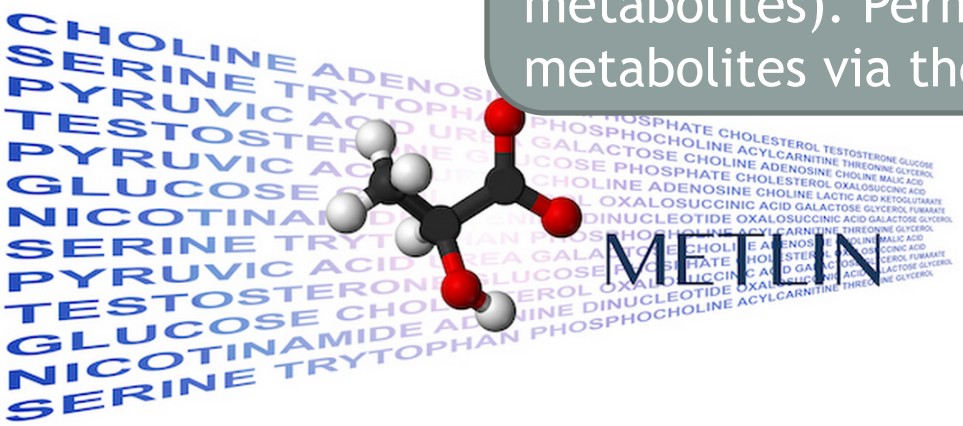
Scripps Center For Metabolomics

METLIN: Metabolite and Tandem MS Database

MS HOME | Overview | Search | XCMSOnline | Software/Services | Metabolomics Science | Publications



Database containing a large number of metabolites (240,000+) and spectra for those (12,000 metabolites). Permits search of metabolites via their mass spectra.



Statistics

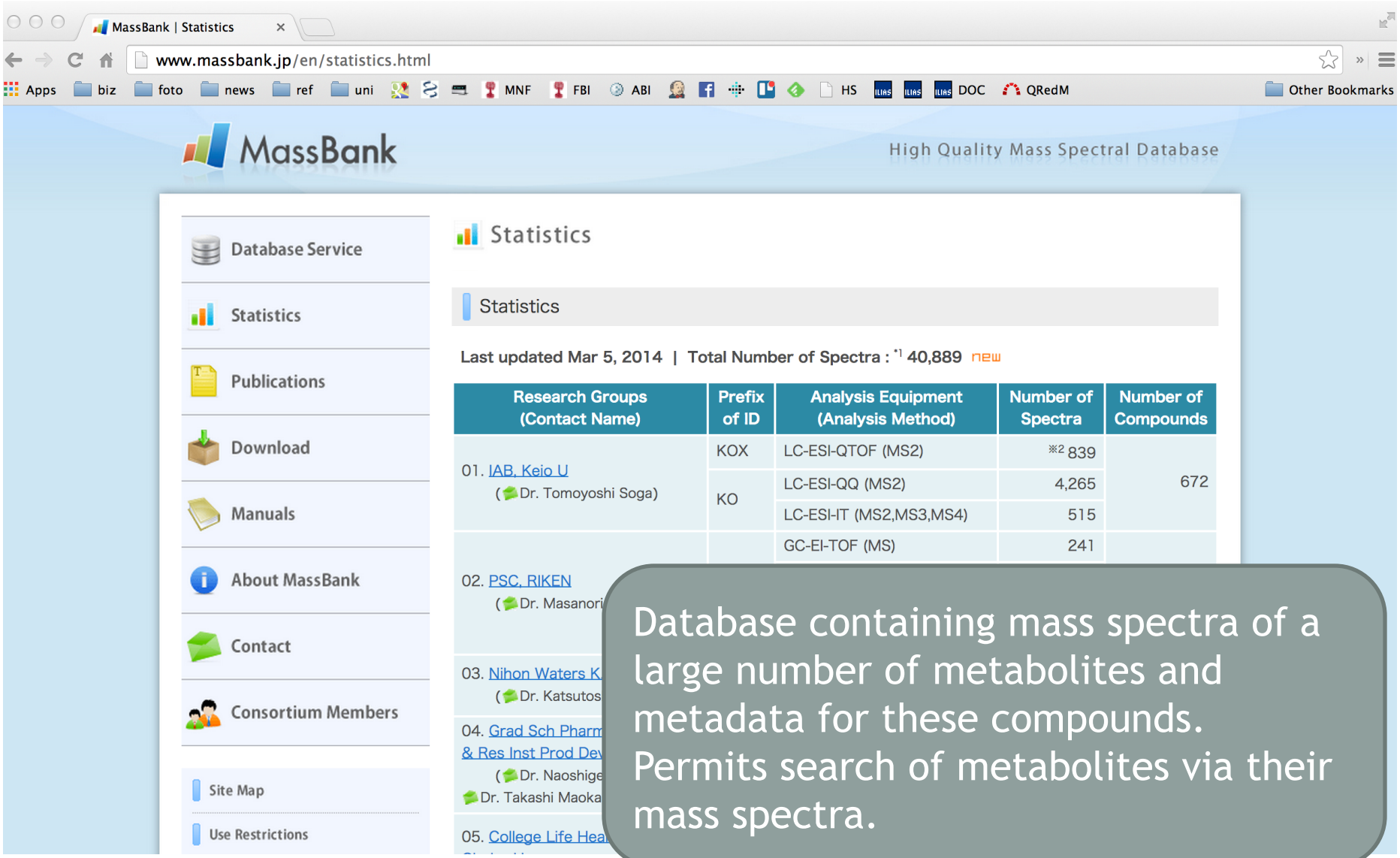
- # Metabolites: 240,516
- # High Resolution MS/MS Spectra: 61,872
- # Metabolites w/ High Resolution MS/MS: 12,057

Functionality

- **Single** & **Batch** Precursor Ion *(m/z)* searching
- **Single** & **Multiple** Fragment Ion *(m/z)* searching

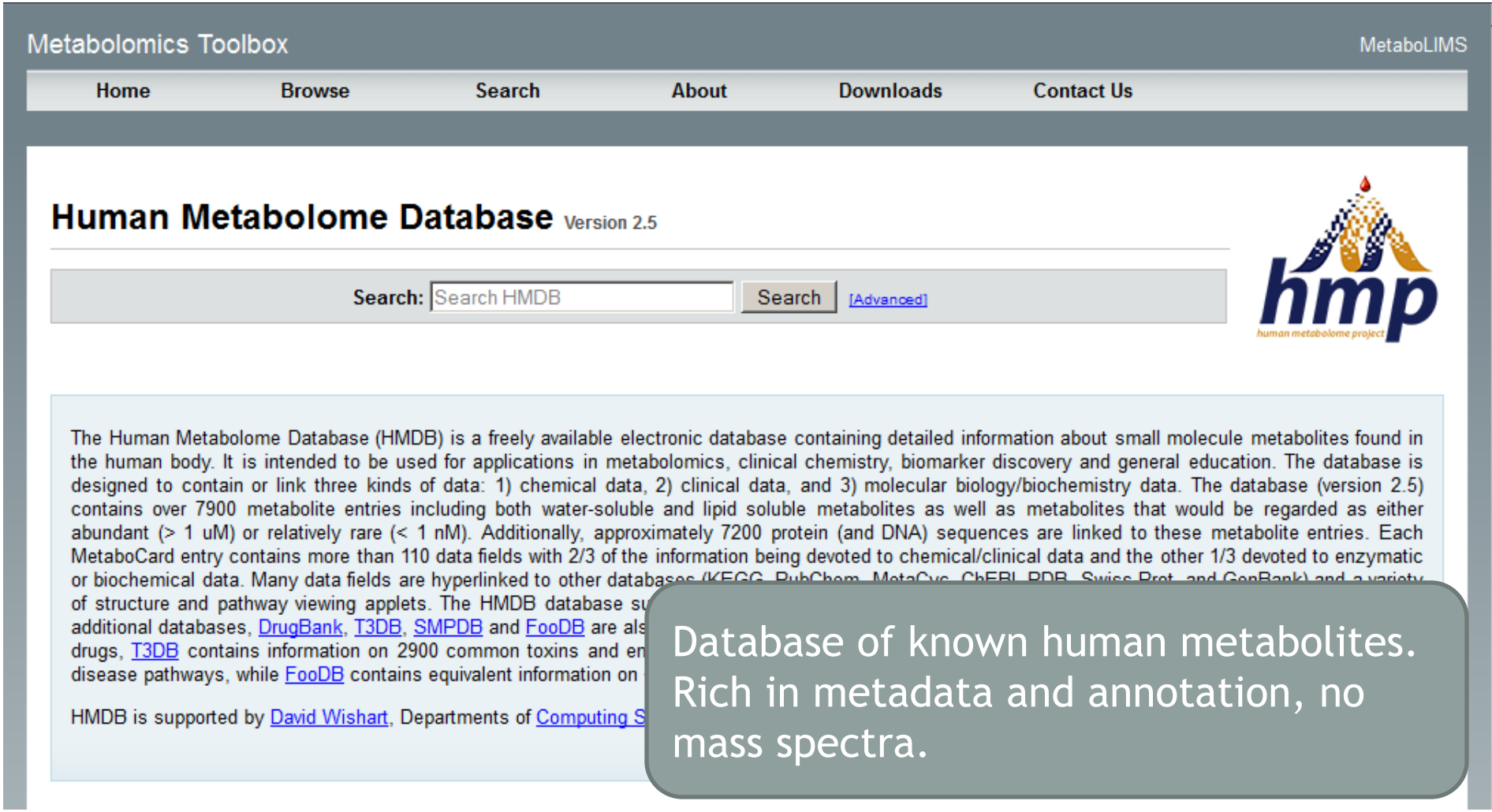
http://metlin.scripps.edu/

MassBank



http://www.massbank.jp/

HMDB


Metabolomics Toolbox
MetaboLIMS
Home
Browse
Search
About
Downloads
Contact Us
Human Metabolome Database Version 2.5
Search: Search HMDB
Search
[Advanced]
hmp
human metabolome project
The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education. The database is designed to contain or link three kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. The database (version 2.5) contains over 7900 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 uM) or relatively rare (< 1 nM). Additionally, approximately 7200 protein (and DNA) sequences are linked to these metabolite entries. Each MetaboCard entry contains more than 110 data fields with 2/3 of the information being devoted to chemical/clinical data and the other 1/3 devoted to enzymatic or biochemical data. Many data fields are hyperlinked to other databases
of structure and pathway viewing applets. The HMDB database su
additional databases, DrugBank, T3DB, SMPDB and FooDB are als
drugs, T3DB contains information on 2900 common toxins and en
disease pathways, while FooDB contains equivalent information on
HMDB is supported by David Wishart, Departments of Computing S
Database of known human metabolites. Rich in metadata and annotation, no mass spectra.

Materials

- Learning units 1A-E

COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

02. Chromatography and Mass Spectrometry

LU 2A - CHROMATOGRAPHY

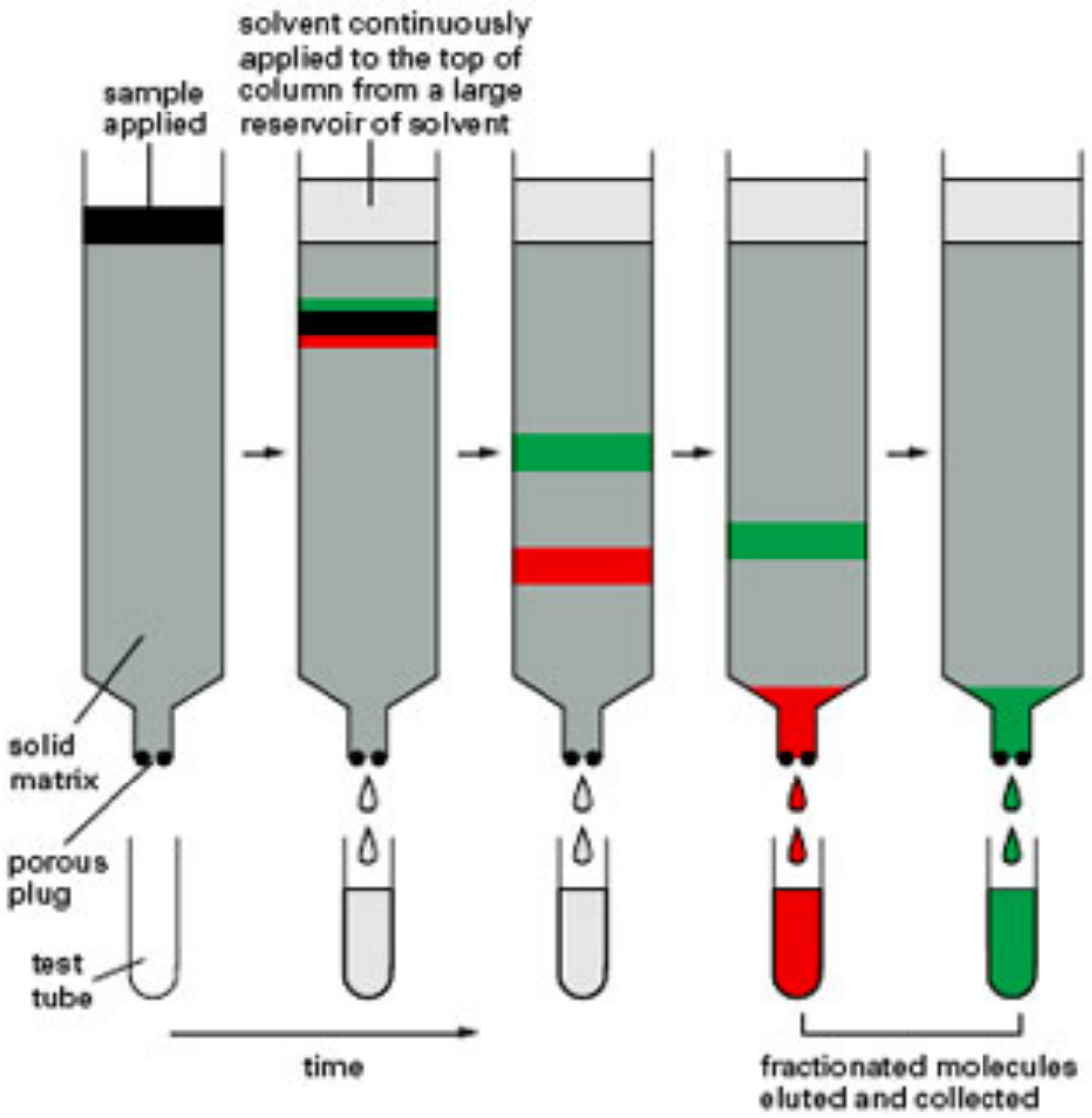
- History of chromatography
- Add description and experiments (chalk chrom. video)?
- Types of chromatography (TLC, LC, GC)
- Separation principles (RP, SAX, SEC)
- Model of theoretical plates, simulation, plate number, capacity
- Peak shapes and properties, asymmetry

Chromatography

- Chromatography is a separation technique
- From greek *chroma* and *graphein* – *color* and *to write*
- Initially developed by **Mikhail Semyonovich Tsvet**
- Simple fundamental idea:
  - Two phases: stationary and mobile
  - Analytes are separated while mobile phase passes along the stationary phase
- Various separation mechanisms, various choices for mobile/ stationary phases possible

M. S. Tsvet (1872-1919)

Column Chromatography



http://fig.cox.miami.edu/~cmallery/255/255hist/ecbxp4x3_chrom.jpg

Chromatography

- Family of techniques used to separate a mixture into individual components
- Separation by passing the mixture through immobilized porous substance
- Individual components interact to different degrees
- **Retention time** ≔ time an individual component takes to pass through the system
- Chromatography has a long tradition and is a scientific field in itself
- In Proteomics/ Metabolomics: used for separation

Chromatography

- The mobile phase (containing the sample) moves through a stationary phase
- At any time a component is interacting either with the stationary or with the (moving) mobile phase
- The more a component interacts with the stationary phase relative to the mobile phase, the longer the migration takes
- This leads to different retention times for different components

Column chromatography

- Most commonly used type of chromatography in proteomics/ metabolomics
- Consists of column (glass, metal, synthetic material) containing the stationary phase through which the mobile phase is passing
- Types of column chromatography used for organic components:
  - Liquid chromatography (dominates the field of proteomics)
  - Gas chromatography (principle see last lecture)
- Other chromatography types:
  - Paper chromatography
  - Thin-layer chromatography

High-pressure liquid chromatography (HPLC)

- Liquid as the mobile phase and a porous solid as the stationary phase
- Surface of the solid can be shaped for specific properties
- *High-pressure* pumps are used to pump liquid through the system

Some further definitions:

- The stationary phase is also called packing
- The mobile phase is also called solvent or eluent

Technical information:

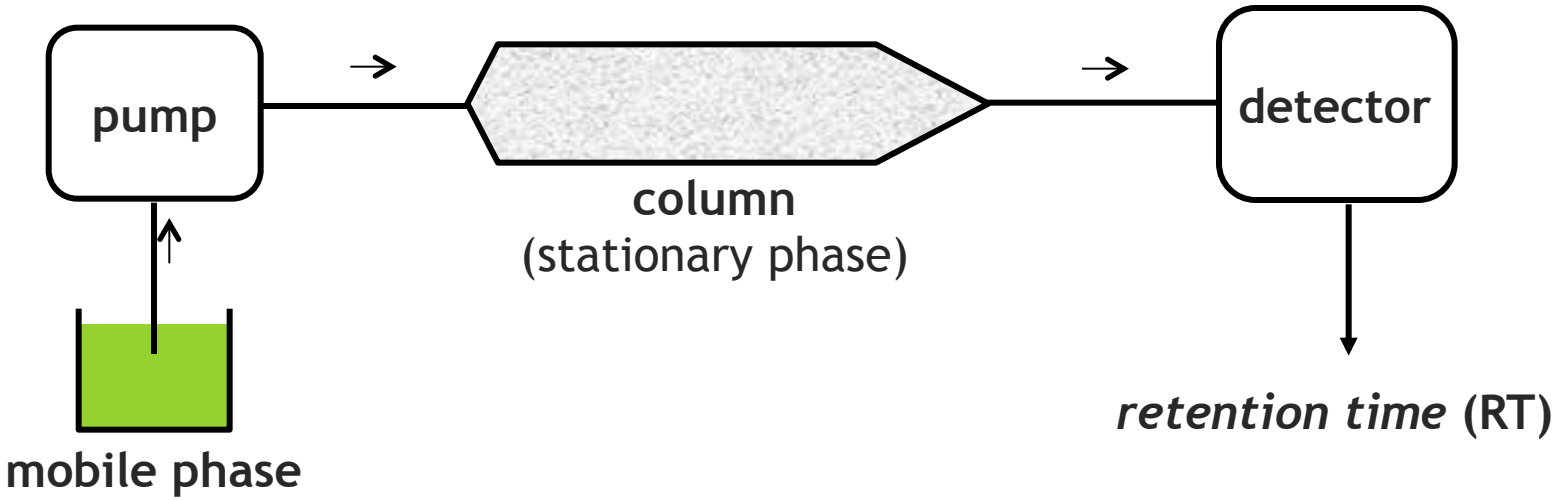
- Columns are typically 10-25 cm long (can be up to several meters)
  - µ-LC columns have an inner diameter (ID) of approx. 1 mm (preparative use)
  - Capillary LC columns: ID < 300 µM
  - nanoLC columns: ID 50-100 µm

Columns

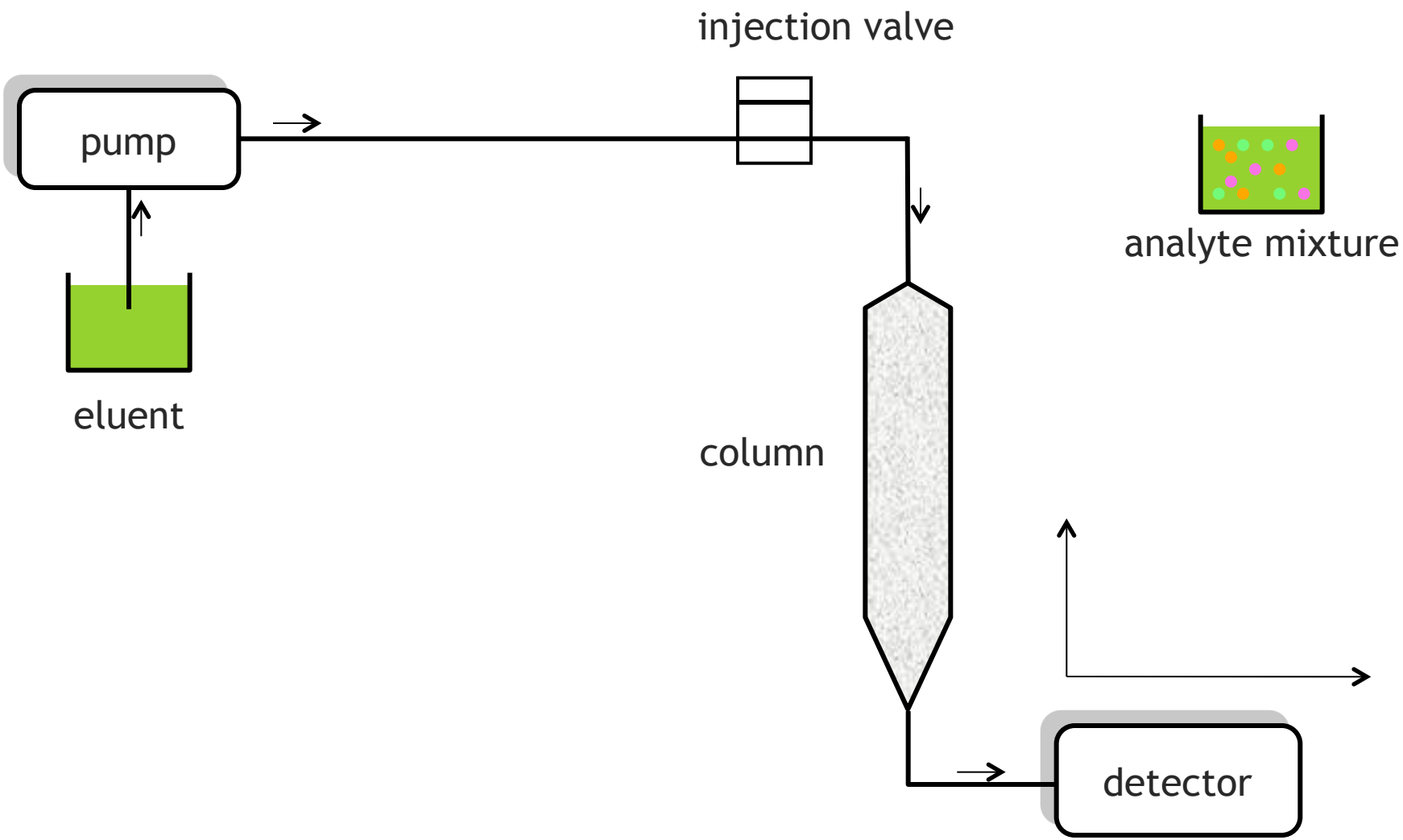
- Tsvet used an open glass column
- Better separation and low sample amounts require columns with smaller inner diameter
- Smaller, tightly packed columns require higher eluent pressures (Tsvet: hydrostatic pressure only) to achieve rapid separations

HPLC steel column (internal diameter 4.6 mm, length 150 mm)

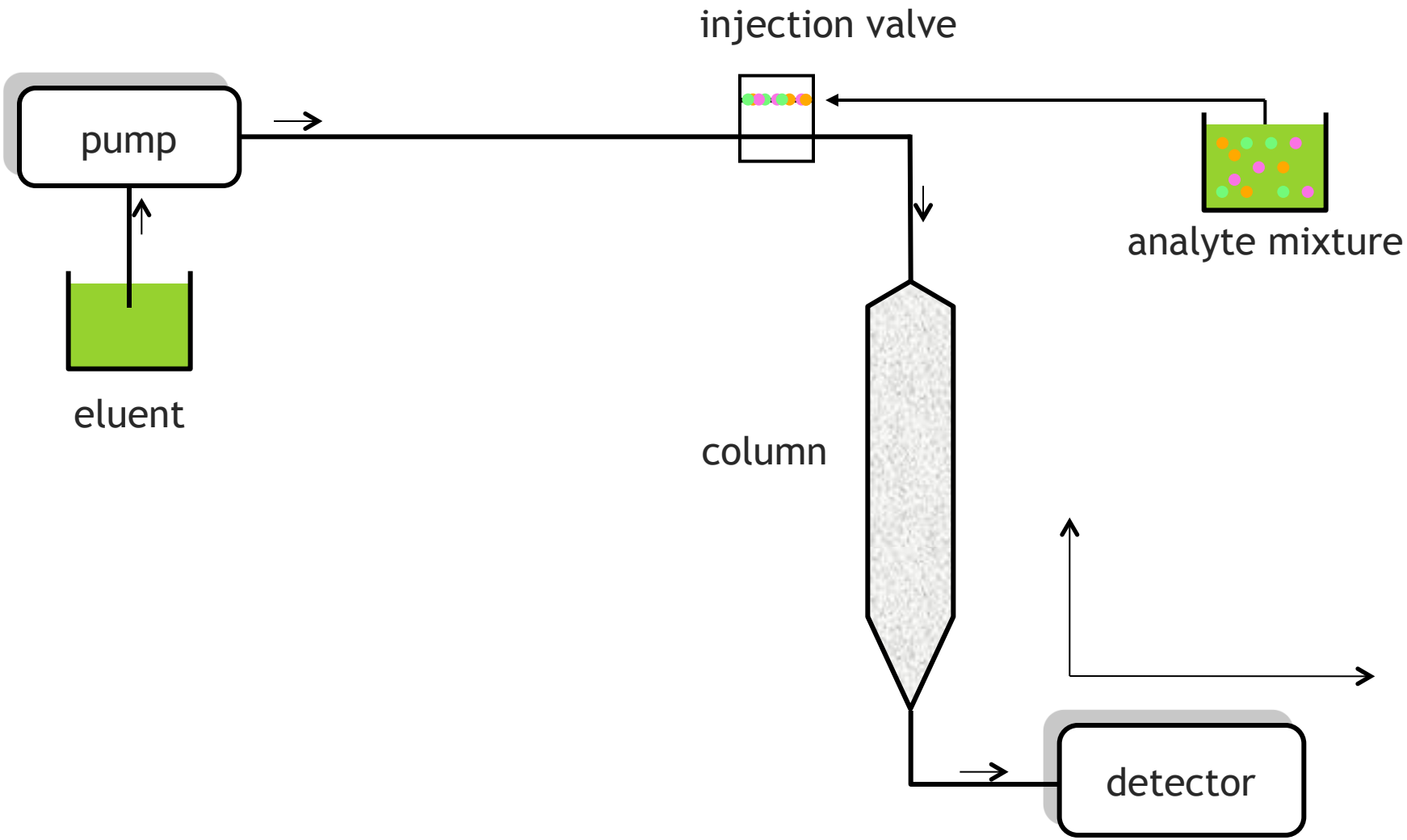
HPLC (High Performance Liquid Chromatography)



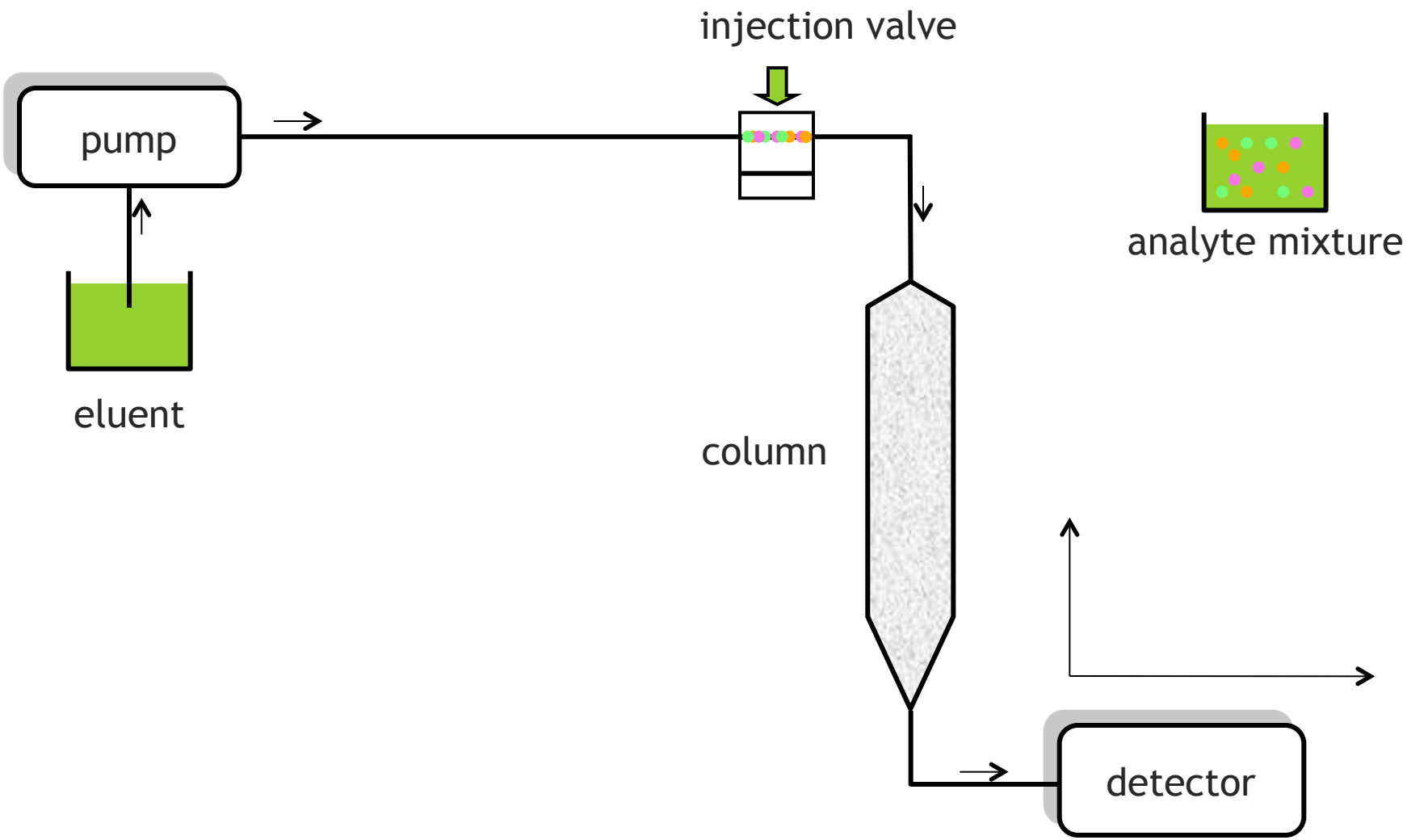
HPLC (High Performance Liquid Chromatography)



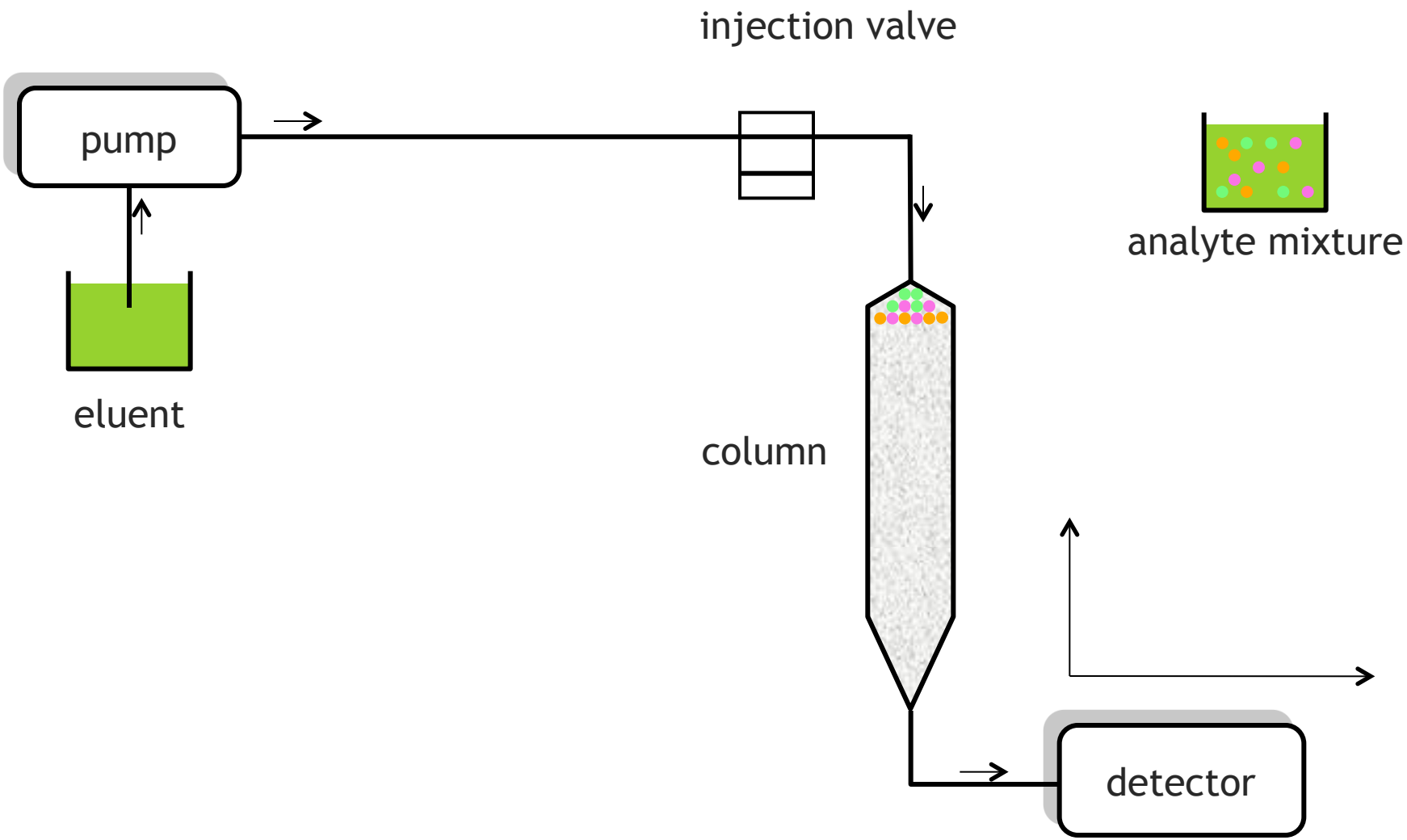
HPLC (High Performance Liquid Chromatography)



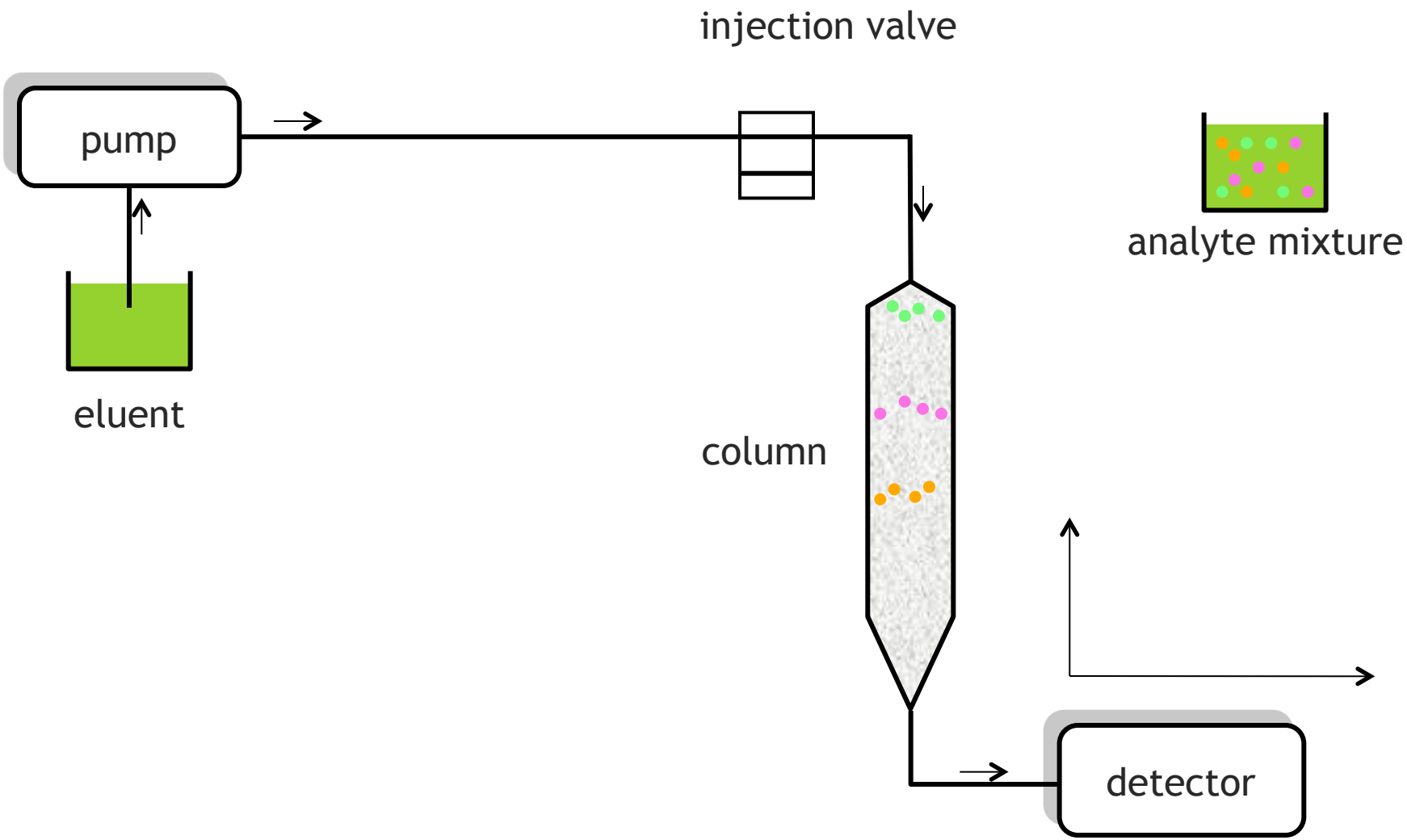
HPLC (High Performance Liquid Chromatography)



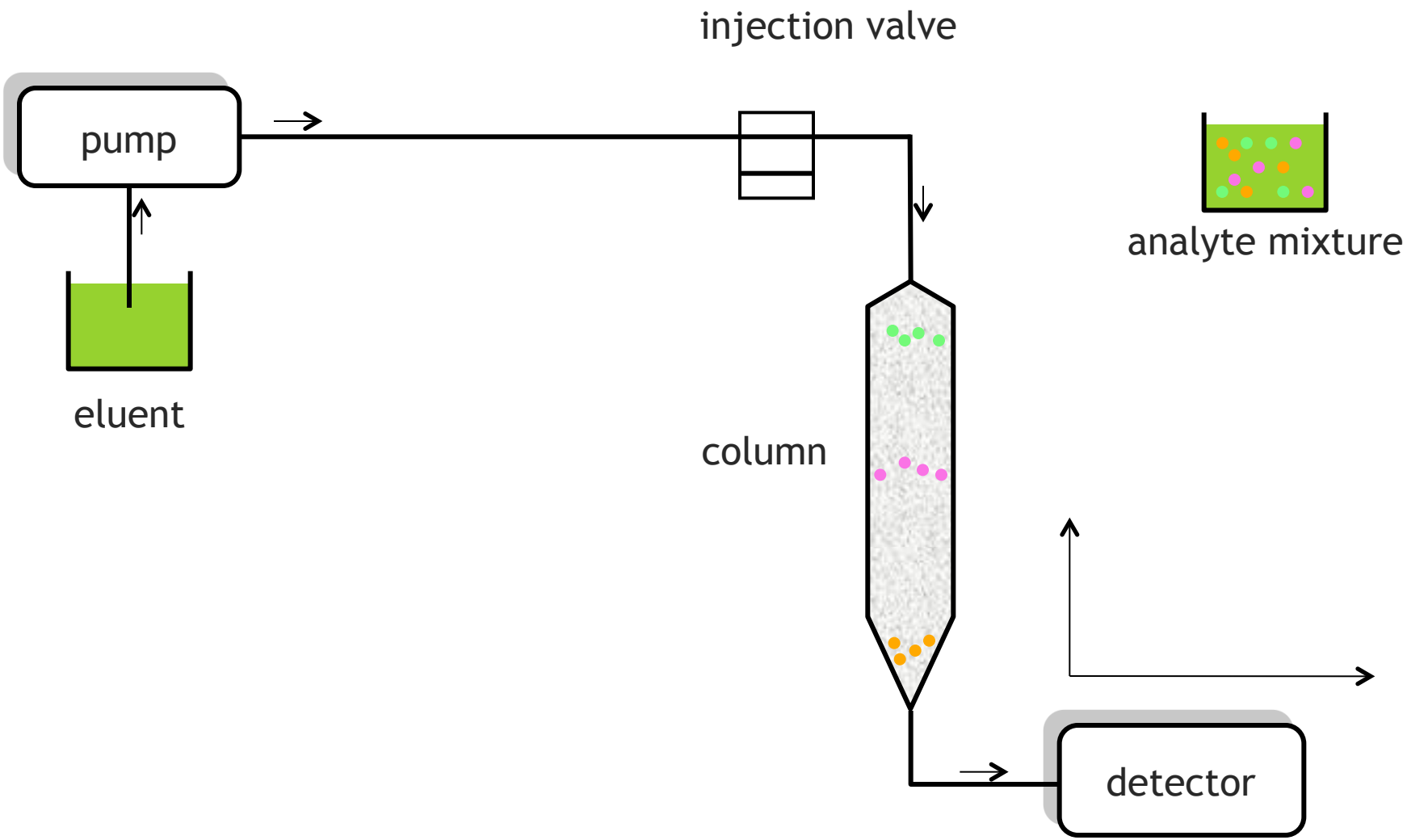
HPLC (High Performance Liquid Chromatography)



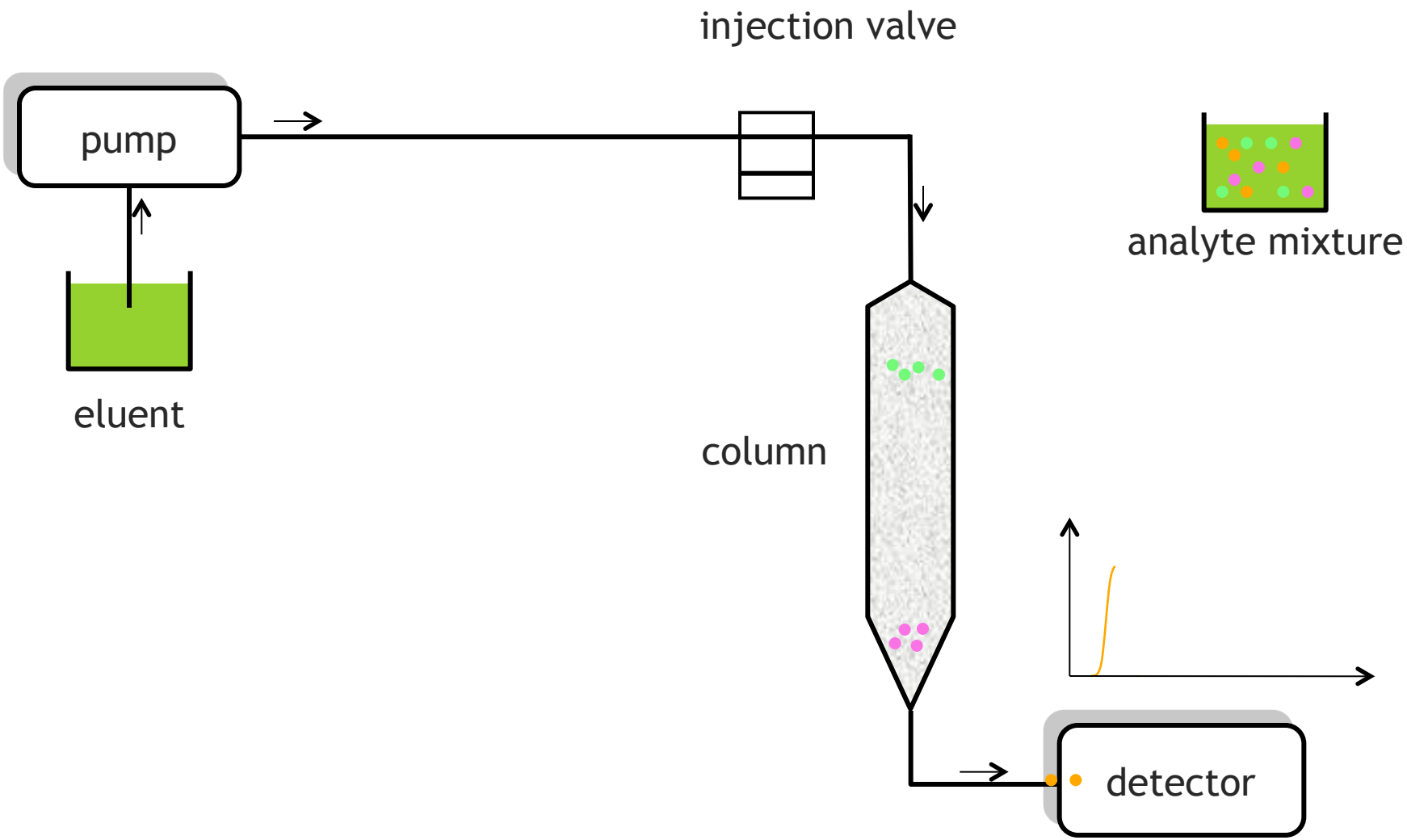
HPLC (High Performance Liquid Chromatography)



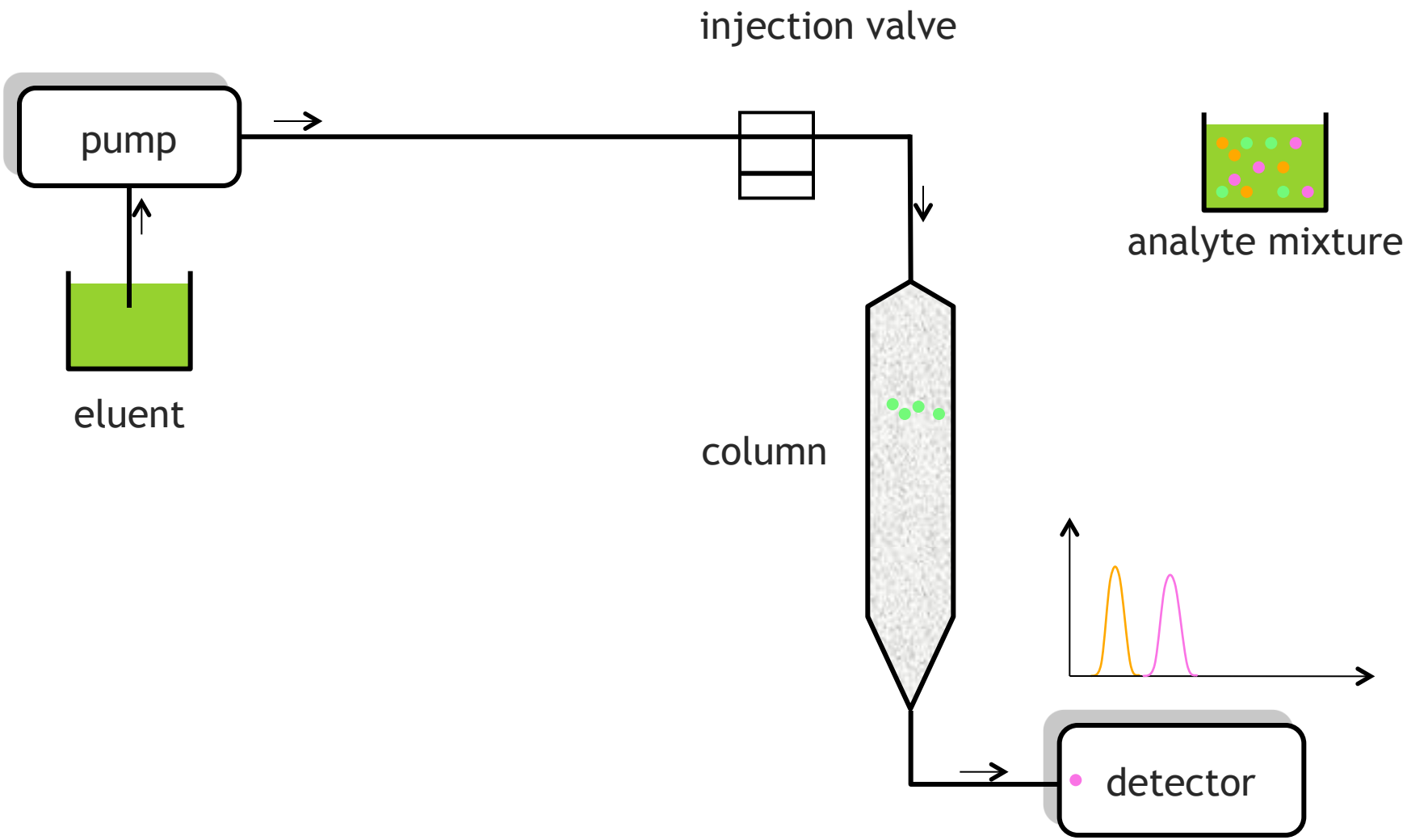
HPLC (High Performance Liquid Chromatography)



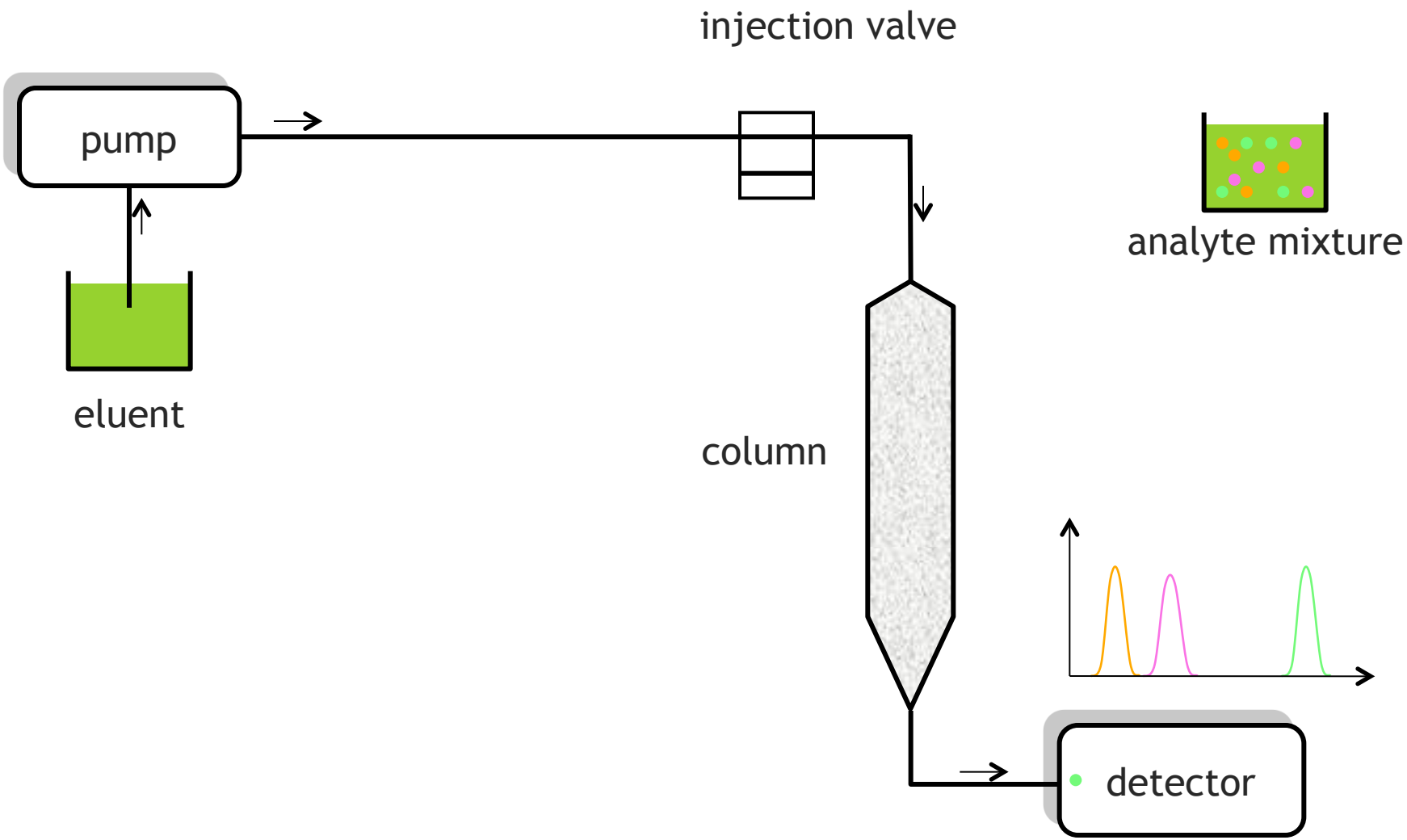
HPLC (High Performance Liquid Chromatography)



HPLC (High Performance Liquid Chromatography)

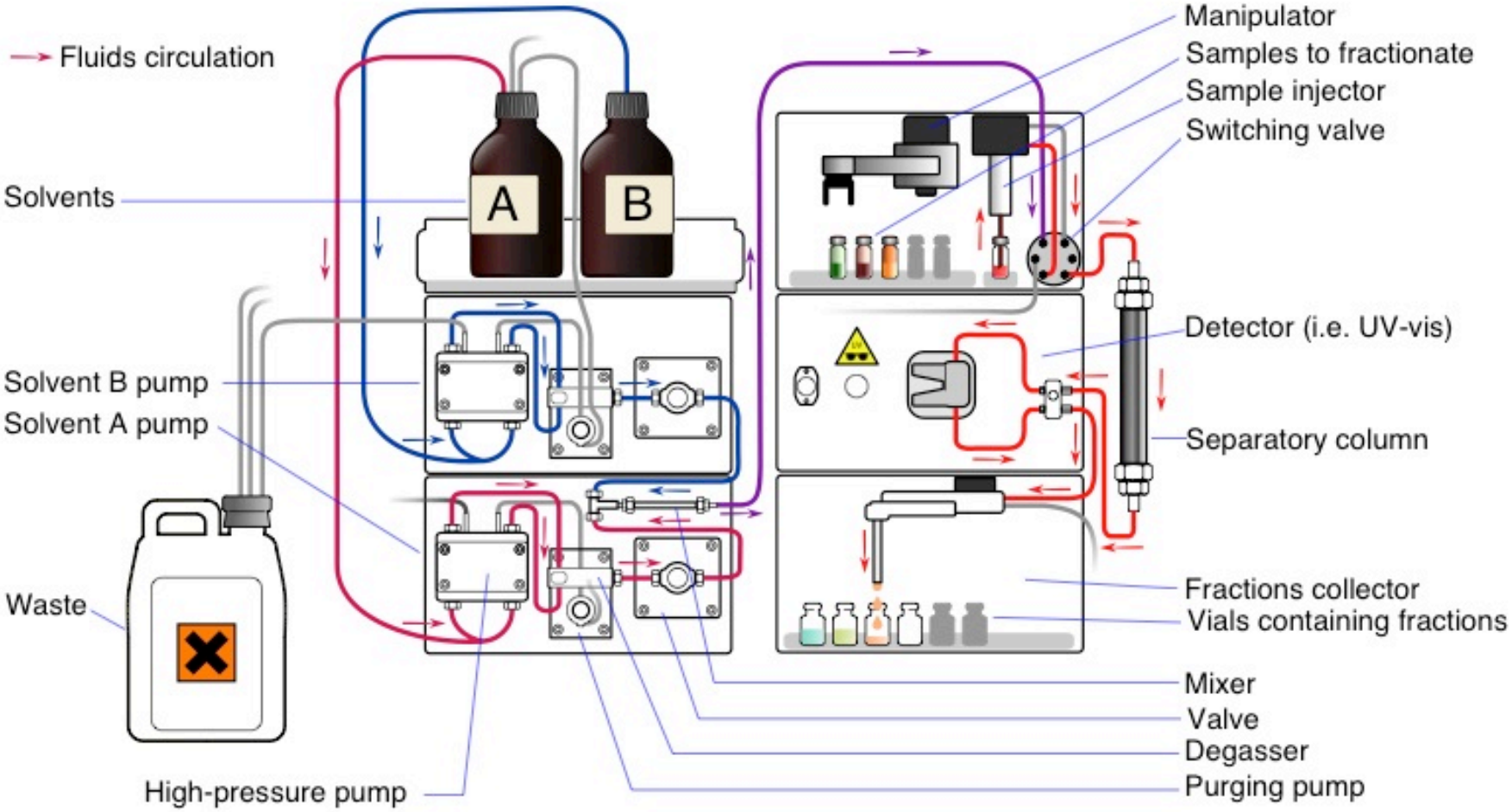


HPLC (High Performance Liquid Chromatography)



Modern HPLC

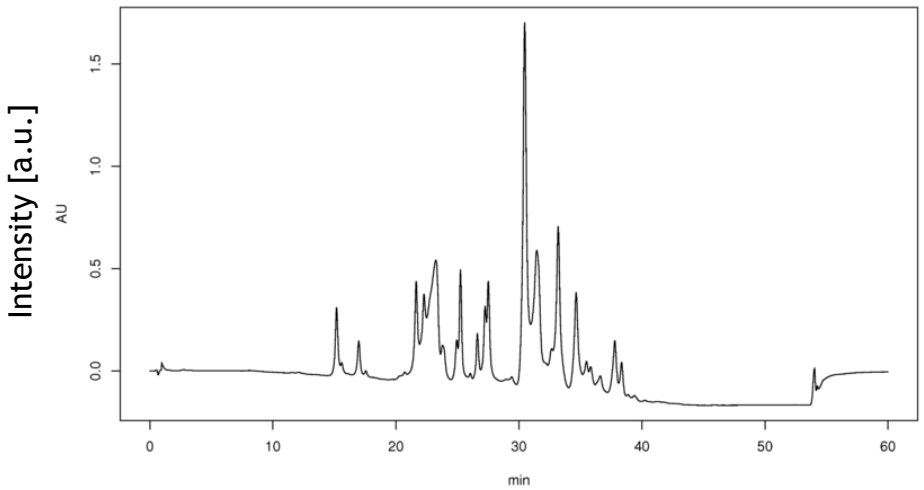
Schematic representation of a modern HPLC system



Yassine Mrabet; http://en.wikipedia.org/wiki/Chromatography#mediaviewer/File:Preparative_HPLC.svg

Detectors Used for HPLC

- Detectors registers the components as they elute off the column
- Common detectors use
  - Light absorption (photometric detector)
  - Fluorescence
  - Change in diffraction index
  - Mass spectrometers
- Detector registers some sort of intensity as function of time
- Detector response over time is called a **chromatogram**



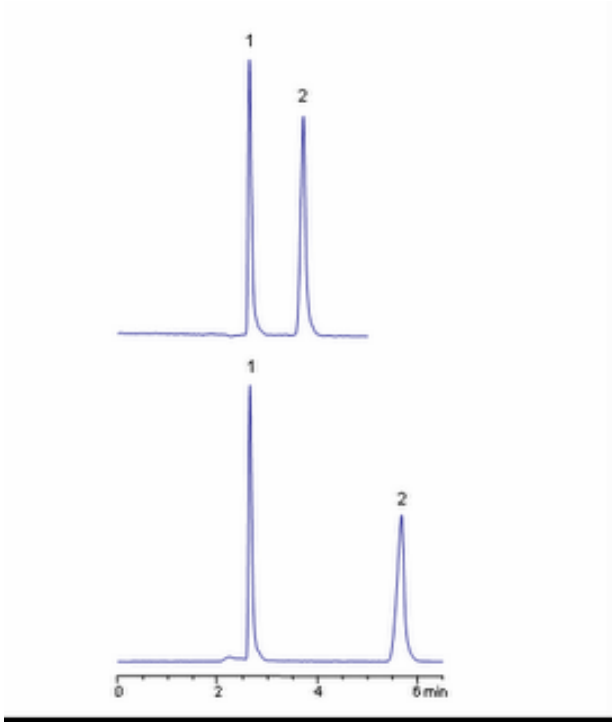
Good HPLC performance

- Find an optimal balance between the components' affinity for the stationary phase and the solubility of the components in the mobile phase
  - Different components should migrate at different rates
  - Narrow elution peak of different components
  - Ideally elution peaks of different components should not overlap
- **Challenge**
  - Achieve different rates of migration for the different components
  - Narrow elution peaks
- **Difficulties**
  - Noise. *Signal-to-noise* ratios can be used to quantify how well a real signal can be differentiated from background signal
  - Baseline drift. The baseline is recorded when only the mobile phase elutes. This can vary over time.

HPLC performance

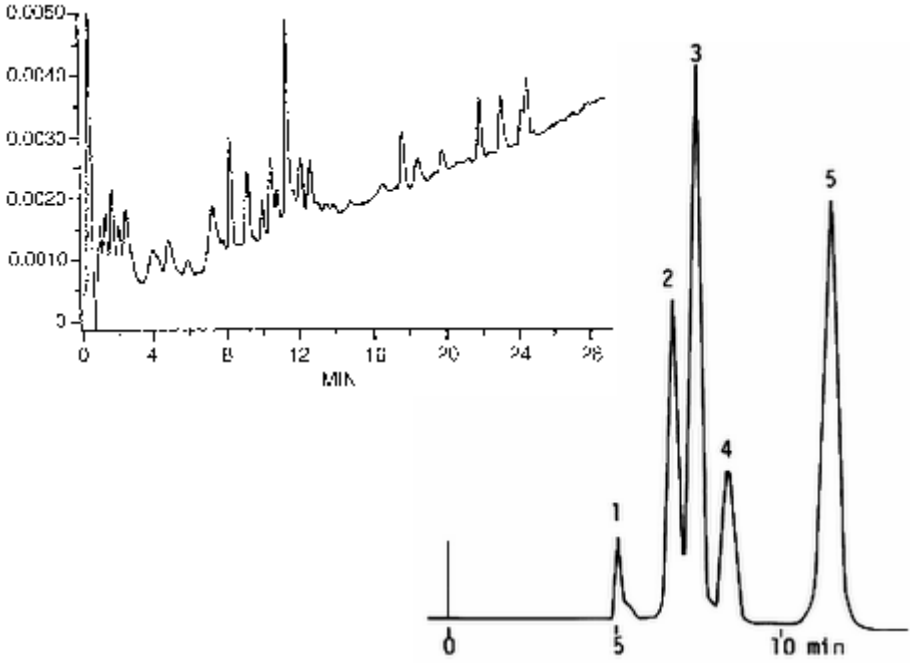
Good separation

- Peaks are well separated
- Peaks are sharp
- Peaks come down to baseline



Poor separation

- Peaks are not well resolved
- Baseline is drifting upwards



epharmacist.blogspot.com/2008_04_01_archive.html

HPLC methods

- Essential components are
  - **Stationary phase**
    - Interaction of stationary phase with components
  - **Mobile phase**
    - Solubility of components in mobile phase

- Most common HPLC methods
  - Reversed-phase (RP) chromatography
  - Strong cation/anion exchange (SCX/SAX) chromatography
  - Affinity chromatography
  - Size exclusion chromatography

Reversed-Phase Chromatography

- **Stationary phase**
  - surface-modified silica (most commonly alkyl chains: (C_4, C_8 or C_{18}); comparable to fatty acid chains)
  - 'Reversed phase' – while silica is generally hydrophilic, the hydrophobic modifications turn it into a hydrophobic phase
- **Mobile phase (eluent)**
  - Usually a mixture of water and an organic solvent (e.g., acetonitrile [ACN])
  - Composition of the eluent usually changes ('**gradient**') with time
  - Start with hydrophilic eluent (mostly water)
  - Higher ACN content towards the end to elute hydrophobic peptides 'sticking' to the column)

Strong Cation Exchange (SCX) Chromatography

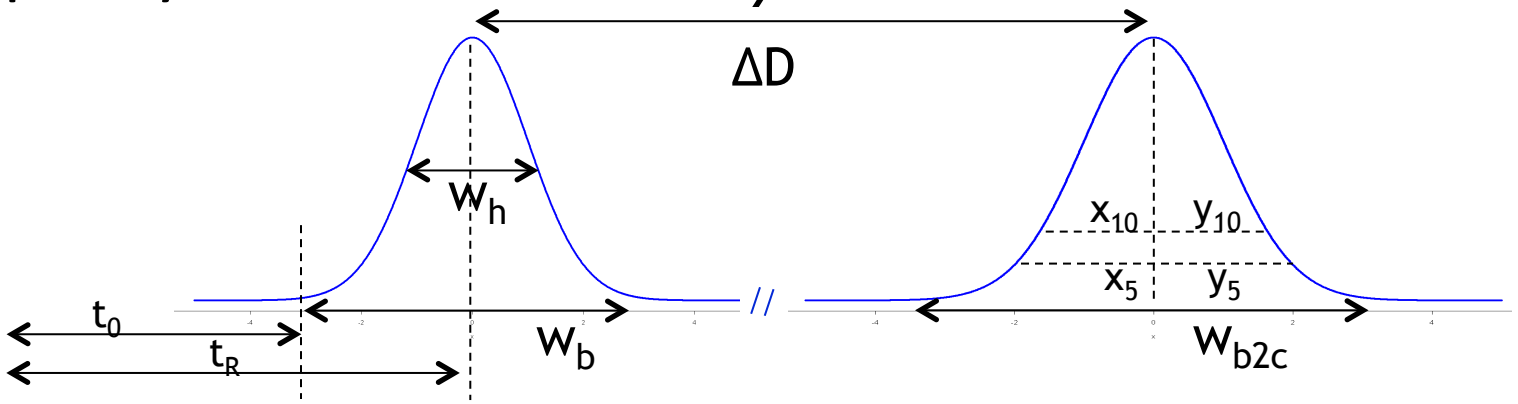
- Opposite charges attract each other
- Net charge of a peptide depends on pH
- Stationary phase is an SCX
  - Surface modified by sulfonic acid groups (neg. charged at pH above 2-3)
- Peptides injected at low pH (~3), thus positively charged (cations)
- The more positive the charges – the stronger the interaction
- For elution, increase ionic strength within solution B (using salt)

Multidimensional Chromatography

- To increase the separation power two or more different separation techniques can be used in series
- For proteomics this is called MudPIT (multidimensional protein identification technology)
  - Wolters DA, Washburn MP, Yates JR, 73[23]: 5683-90. Analytical Chemistry. 2001

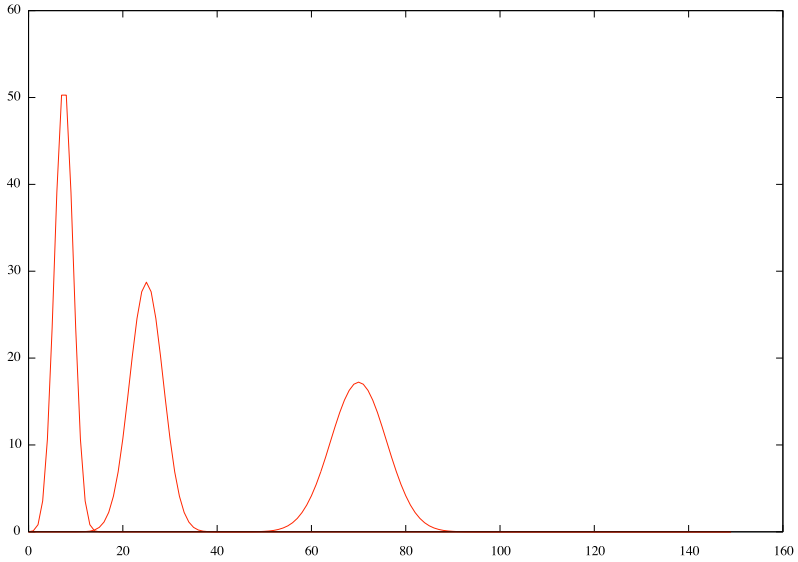
Component Migration

- Retention time (t_R):=
  t_R is the time a component takes from injection until the elution peak maximum
- Dead time (t_0):=
  t_0 is the time a component takes from injection until the elution peak maximum, assuming no interaction with the stationary phase
- Capacity factor $k' := tR - t0 / t0$



Peak Width in Chromatography

- Peaks broaden with retention time
- Early peaks are sharp and narrow
- Later peaks tend to be broader
  - Diffusion along the column during the separation
  - Analytes eluting later had more time to diffuse
- Peak area remains constant, though
  - No analyte is getting lost
  - Increased width is compensated by reduced height

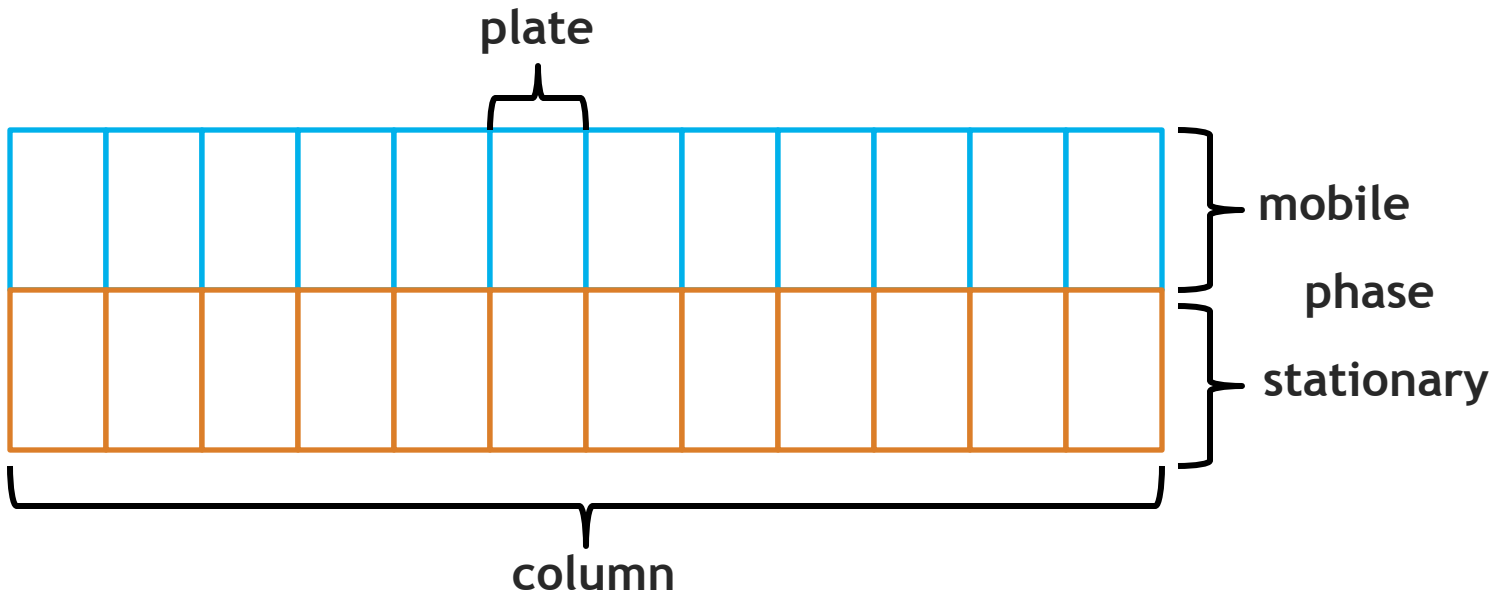


Peak Shape

- Baseline width:= w_b is the width of the baseline
- Half-height width := w_h is the width at 50% at the half peak height. This is also called FWHM (full width at half maximum)
- Peaks should be narrow and symetrical, but peak width generally increases with the retention time
  - **Plate number** is a better indicator (than peak width) to show how good an LC performs in producing narrow eluent peaks
  - **Symmetry** is very rarely observed. Tailing is frequently observed
  - **Resolution** is a measure of how well two adjacent peaks are separated

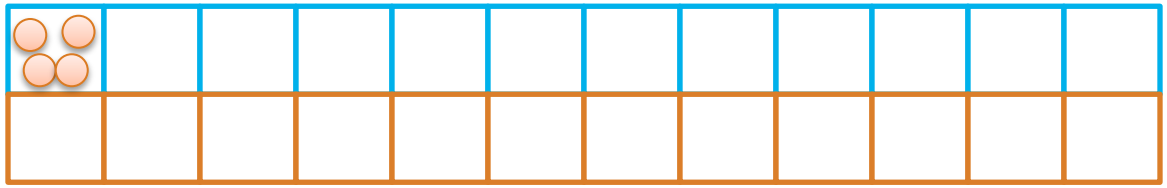
Theoretical Plates

- Conceptual model of the chromatographic process
- Term ‘plates’ originates from fractionated distillation
- Same mathematical formalism is used
- A **plate** represents a stage of **discrete equilibration** between two phases (in chromatography there are no physical plates, but only “theoretical plates”)

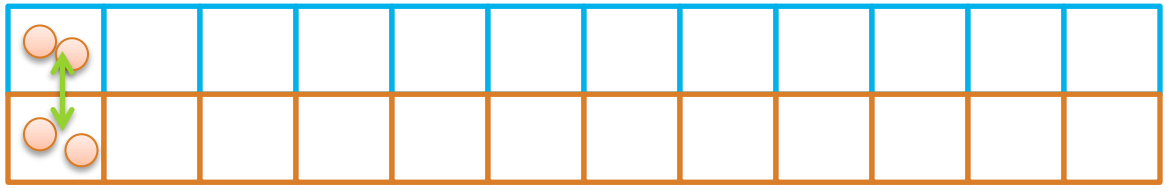


Theoretical Plates - Model

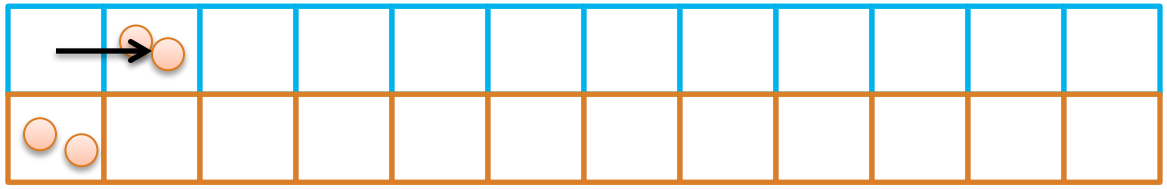
- Analyte enters the head of the column dissolved in the eluent



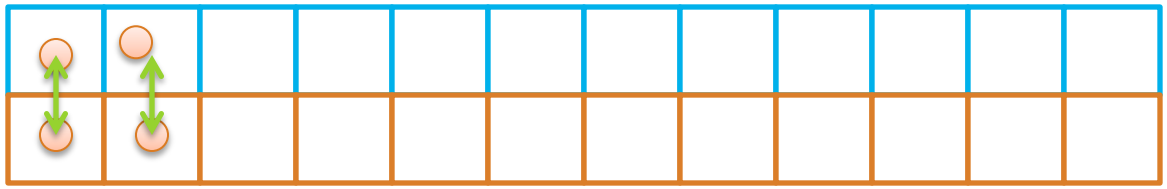
- Analyte equilibrates between both phases in each plate



- Mobile phases moves on in a discrete step and brings mobile phases into contact with the next plate



- Plates equilibrate, process continues as above

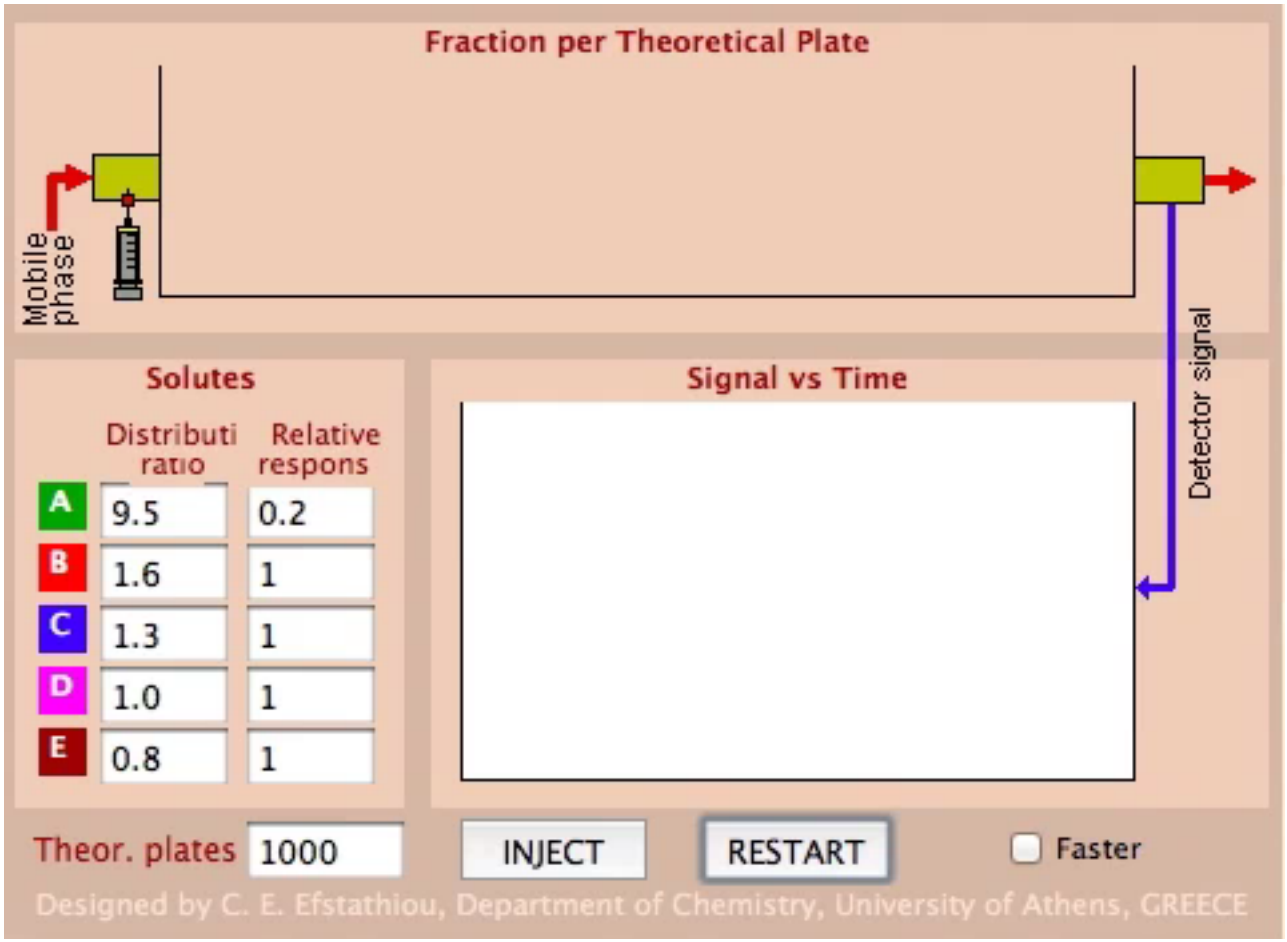


Theoretical Plates – Example

- Analytes **distribute between mobile phase and stationary phase** (here: equally)
- Mobile phase moves again, equilibrates, etc.
- Numbers express the concentration in each plate
- This simple model yields a **Gaussian** shape for an infinite number of plates
- Note that the sum over all plates always remains the same (**no analyte is lost**) until it starts eluting from the column

1	16	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
2	8	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0
3	0	8	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0
4	4	4	0	0	0	0	0	0	0	0
	4	4	0	0	0	0	0	0	0	0
5	0	4	4	0	0	0	0	0	0	0
	4	4	0	0	0	0	0	0	0	0
6	2	4	2	0	0	0	0	0	0	0
	2	4	2	0	0	0	0	0	0	0
7	0	2	4	2	0	0	0	0	0	0
	2	4	2	0	0	0	0	0	0	0
8	1	3	3	1	0	0	0	0	0	0
	1	3	3	1	0	0	0	0	0	0
9	0	1	3	3	1	0	0	0	0	0
	1	3	3	1	0	0	0	0	0	0

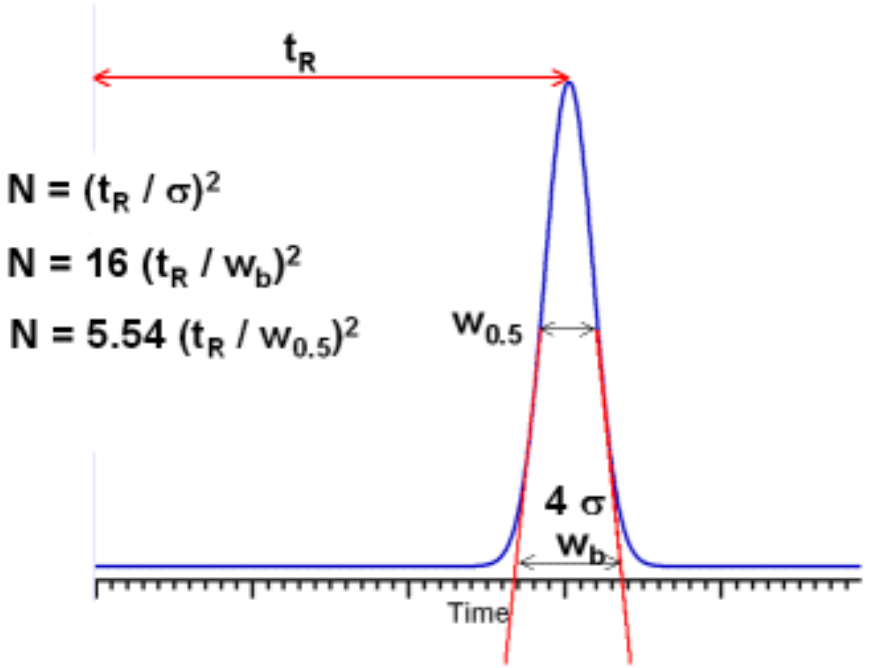
Theoretical Plates – Simulation



http://www.chem.uoa.gr/applets/appletchrom/appl_chrom2.html

Plate Number

- Model of theoretical plates gives (asymptotically) rise to a **Gaussian peak shape**
- Ratio of retention time and peak width
- Three different ways to calculate the *plate number N*
  - *From the retention time*
  - *From the baseline width*
  - *From the half-height width*

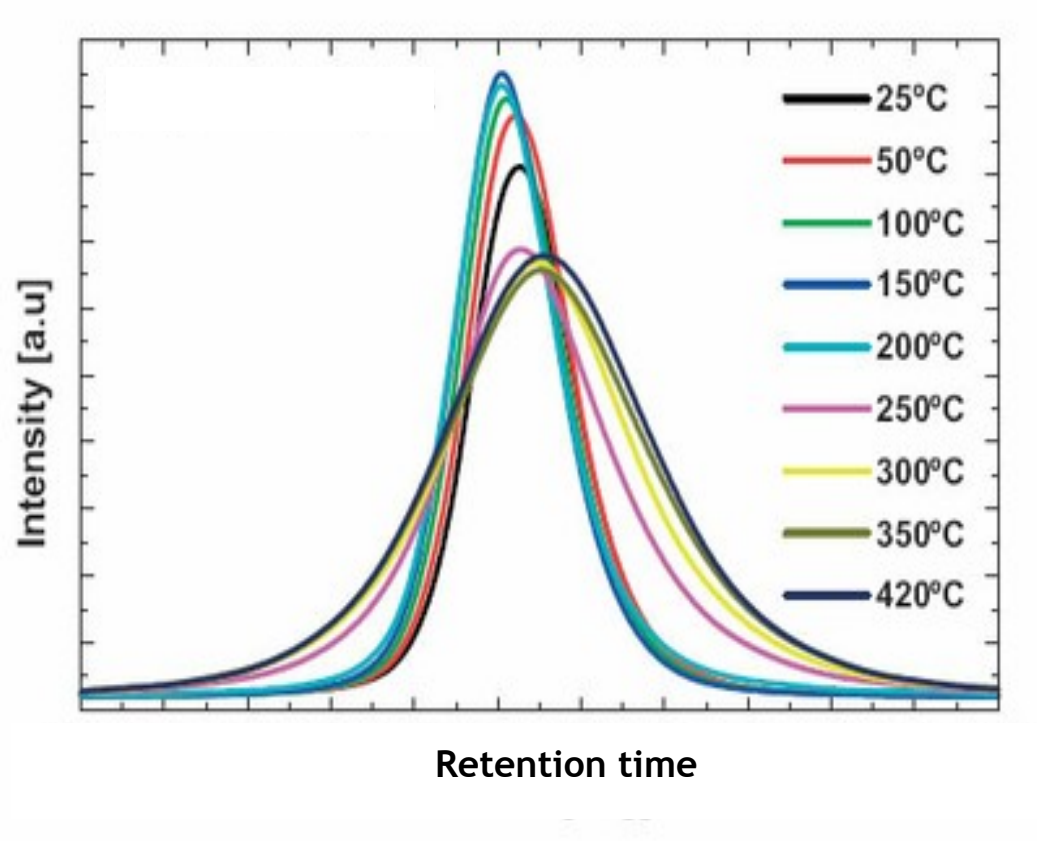


http://www.lcresources.com/resources/TSWiz/PlateCalc.gif

Peak Broadening and Asymmetry

Many factors affect peak shape and symmetry

- Column "secondary interactions"
- Column packing voids
- Column contamination
- Column aging
- Column loading
- Extra-column effects
- Temperature (column and environment)

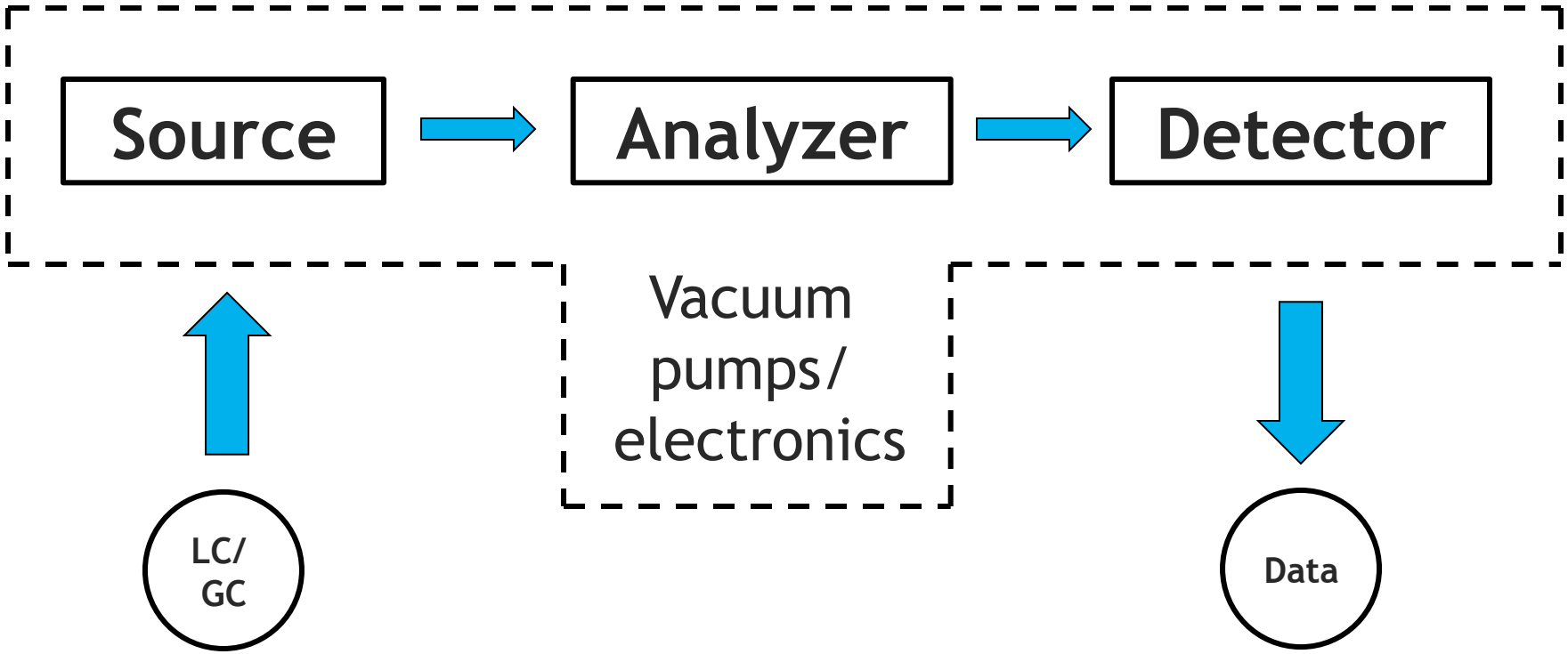


http://www.esrf.eu/UsersAndScience/Experiments/StructMaterials/ID31/Applications/calcite.jpg/image_preview

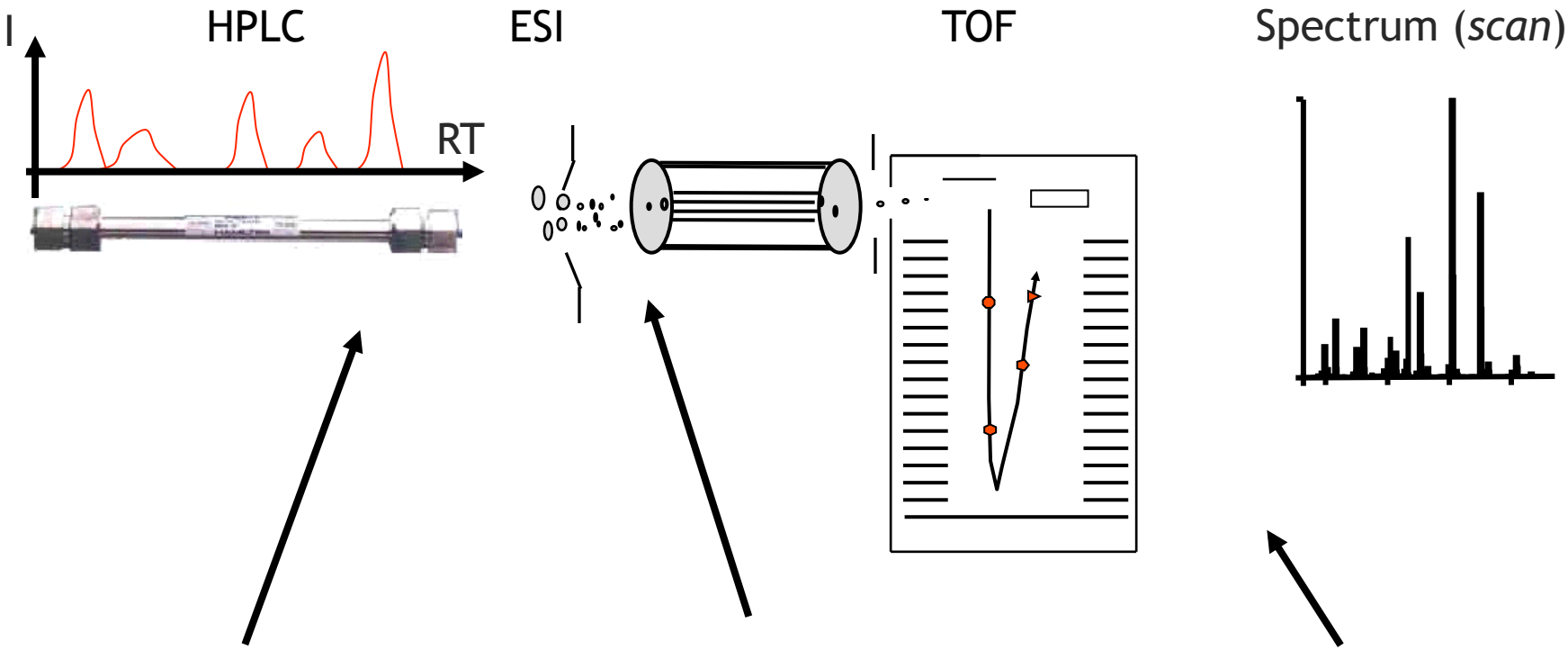
LU2B – MASS SPECTROMETRY

- Basic components of a mass spectrometer
- Ionization methods: MALDI, ESI
- Mass analyzers (TOF, quadrupole, ion trap, orbitrap)
- Detectors
- Tandem MS, MS/MS fragmentation methods
- Product ion generation, ion types, charge states
- Mass accuracy and resolution
- Technical characteristics of typical instruments
- DDA and DIA

Schema Mass spectrometer



HPLC-MS



Separation 1
separate peptides by their retention time on column

Ionization
electrospray, transfers charge to the peptides

Separation 2
MS separates by mass-to-charge ratio (m/z)

Ionization methods

- ElectroSpray Ionization (ESI) *-soft ionization-*
- Matrix Assisted Laser Desorption/ Ionization (MALDI)
  -soft ionization-
- Electron Impact (EI) *-hard ionization-*
- Other methods:
  - Particle bombardment; Field Desorption; Field Ionization;

The following (ionization) is partly based on education material from the Genetics department, Wisconsin University, USA, http://skop.genetics.wisc.edu/AhnaMassSpecMethodsTheory.ppt

Electron impact ionization

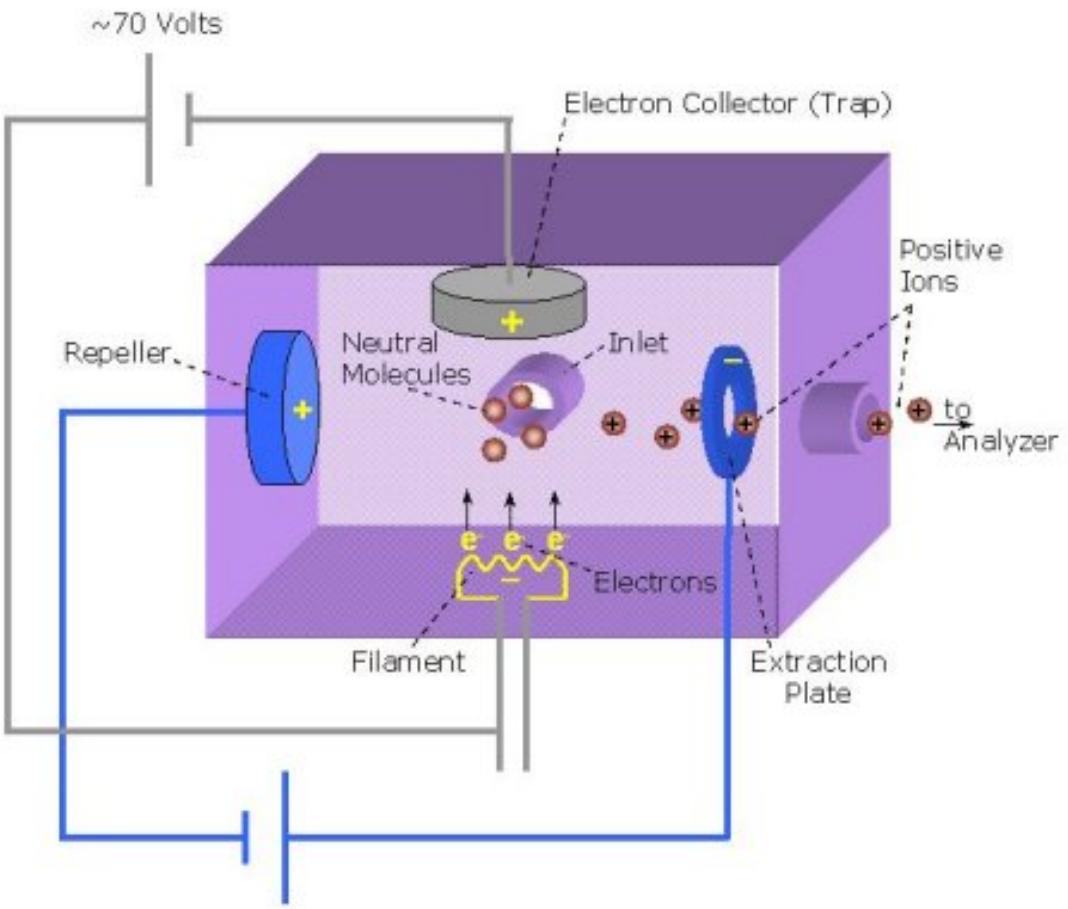
- Sample introduced into instrument by heating it until it evaporates

- Gas phase sample is bombarded with electrons coming from a filament, e.g., rhenium (energy = 70 eV)
  - Electron volt is a unit of energy:
    - $1\ eV = 1.602 \times 10\_19\ Joules$

- Molecule is "shattered" into fragments (70 eV >> 5 eV bonds)

- Fragments sent to mass analyzer

Electron impact ion source



EI ionization of CH_3OH (Methanol)

$$CH_3OH \longrightarrow CH_3OH^+$$

$$CH_3OH \longrightarrow CH_2O{=}H^+ + H$$

$$CH_3OH \longrightarrow {}^+CH_3 + OH$$

$$CH_2O{=}H^+ \longrightarrow CHO{=}H^+ + H$$

Why wouldn't Electron Impact be suitable for analyzing proteins?

EI is not for proteomics

- EI shatters chemical bonds
- A peptide/protein contains (20 different) amino acids
- EI would shatter the peptide not only into amino acids but also amino acid sub-fragments
- Result is tens of thousands of different signals from a single peptide -- too complex to interpret
- EI might be well suited for some applications in metabolomics

Soft ionization methods

- Soft ionization techniques keep the molecule of interest fully intact
- Electro-spray ionization first conceived in 1960's by Malcolm Dole but put into practice in 1980's by John Fenn (Yale)
- MALDI first introduced in 1985 by Franz Hillenkamp and Michael Karas (Frankfurt)
- Made it possible to analyze large molecules via inexpensive mass analyzers such as quadrupole, ion trap and TOF

The Nobel Prize in Chemistry 2002

"for the development of methods for identification and structure analyses of biological macromolecules"

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

"for his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution"

John B. Fenn

1/4 of the prize

USA

Virginia Commonwealth

Koichi Tanaka

1/4 of the prize

Japan

Shimadzu Corp.
Kyoto, Japan

Kurt Wüthrich

1/2 of the prize

Switzerland

Eidgenössische Technische

The Nobel Prize in Chemistry 2002
Press Release
Advanced Information
Information for the Public
Presentation Speech
Illustrated Presentation

John B. Fenn
Nobel Lecture
Banquet Speech
Nobel Diploma
Prize Award Photo
Other Resources

Koichi Tanaka
Nobel Lecture
Interview
Nobel Diploma
Prize Award Photo
Other Resources

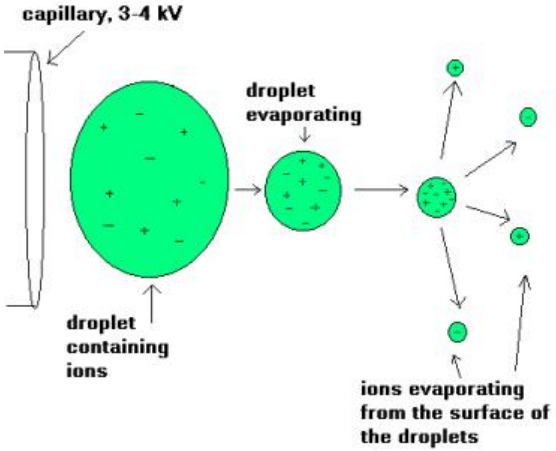
Kurt Wüthrich
Nobel Lecture
Interview
Nobel Diploma
Prize Award Photo
Educational
Other Resources

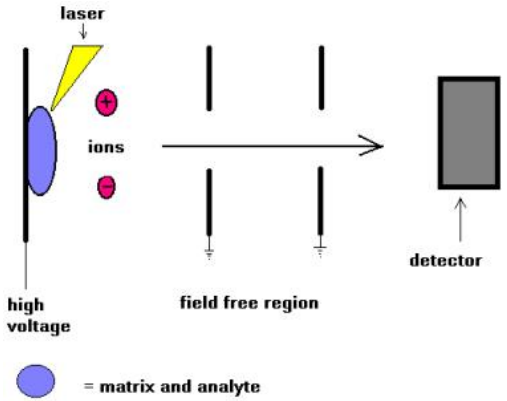
2001

The 2002 Prize in:
Physics
Chemistry
Physiology or Medicine

Soft ionization methods

- **Electrospray mass spectrometry (ESI-MS)**
  - Liquid containing analyte is forced through a steel capillary at high voltage to electrostatically disperse analyte. This induces charged droplets. Ions are formed by extensive evaporation





- **Matrix-assisted laser desorption ionization (MALDI)**
  - Analyte (protein) is mixed with large excess of matrix (small organic molecule)
  - Irradiated with short pulse of laser light. Wavelength of laser is the same as absorbance maximum of matrix.

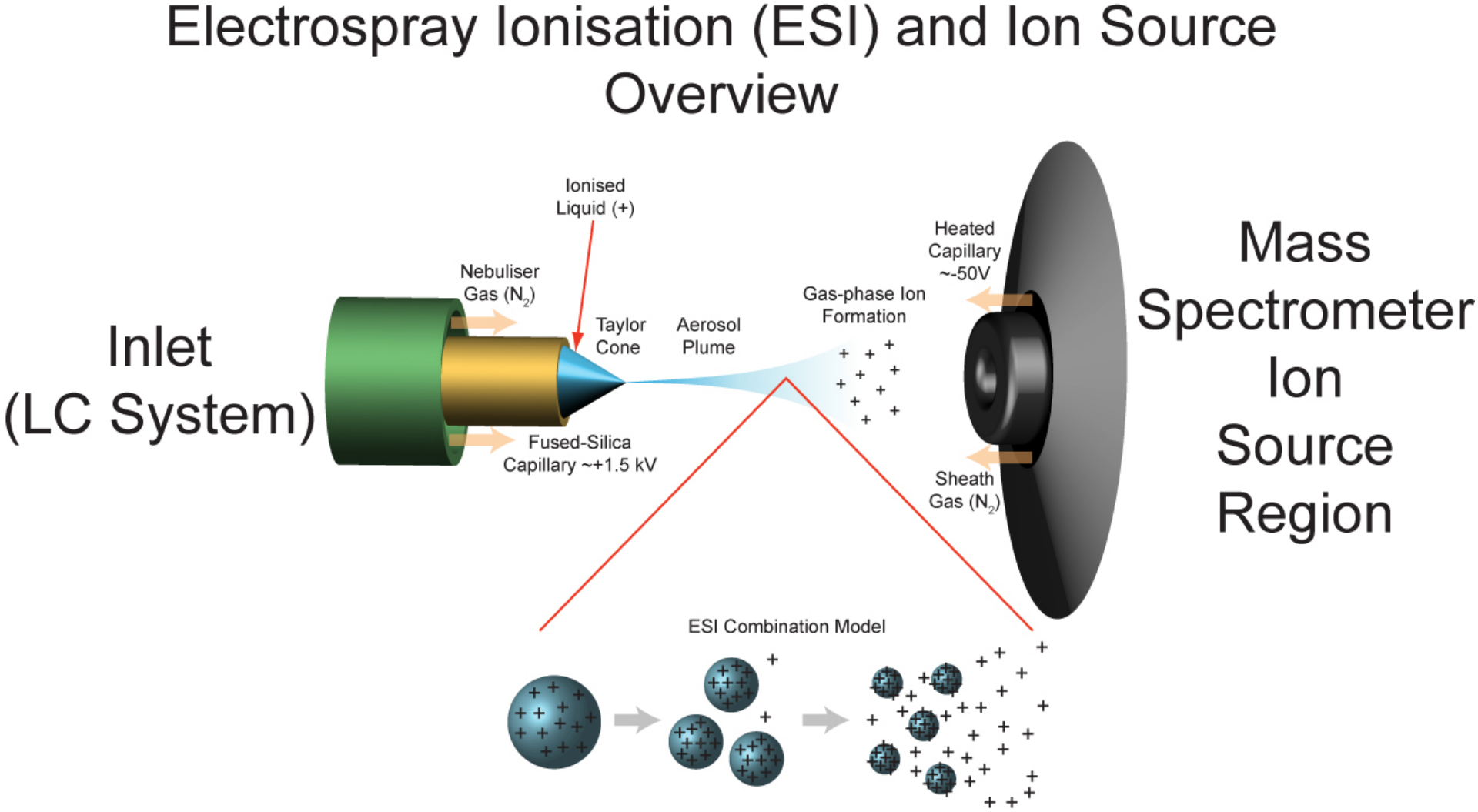
Electrospray ionization

- Sample dissolved in polar, volatile buffer (no salts) and pumped through a stainless steel capillary (70 - 150 μm) at a rate of 10-100 μL/min (for nano spray this can also be at 200 nl/min)

- High voltage (3-4 kV) applied at tip along with flow of nebulizing gas causes the sample to "nebulize" or aerosolize

- Aerosol is directed through regions of higher vacuum until droplets evaporate to near atomic size (still carrying charges)

ESI



http://www.lamondlab.com/MSResource/images/lcms/ESI.jpg

ESI

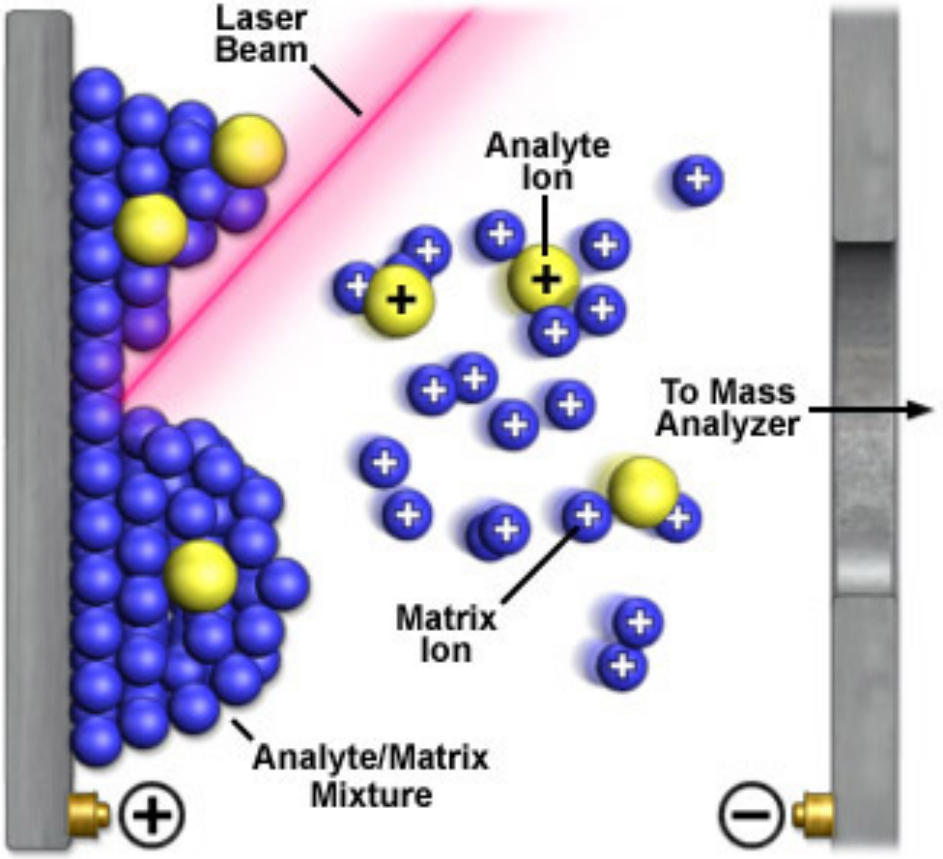
- Very sensitive technique, requires less than a picomole of material
- Strongly affected by salts and detergents
- Positive ion mode measures $(M + H)^+$ (add formic acid to solvent)
- Negative ion mode measures $(M - H)^-$ (add ammonia to solvent)

Positive or negative mode

- Functional groups that readily accept H^+ (such as amide and amino groups found in peptides and proteins) can be ionized using positive mode ESI.

- Functional groups that readily lose a proton (such as carboxylic acids and hydroxyls as found in nucleic acids and sugars) should be ionized using negative mode ESI

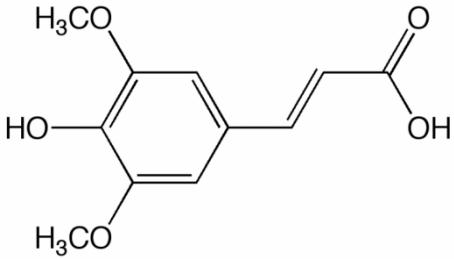
MALDI


Laser
Beam
Analyte
Ion
To Mass
Analyzer
Matrix
Ion
Analyte/Matrix
Mixture

MALDI

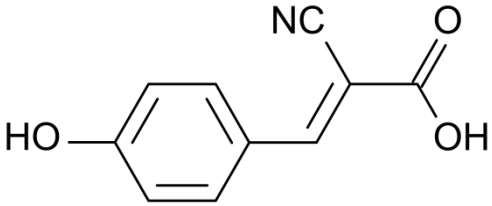
- Sample is ionized by bombarding sample with laser light
- Sample is mixed with a UV absorbant matrix (sinapinic acid for proteins, 4-hydroxycinnamic acid for peptides)
- Light wavelength matches that of absorbance maximum of matrix so that the matrix transfers some of its energy to the analyte (leads to ion sputtering)

Sinapinic acid



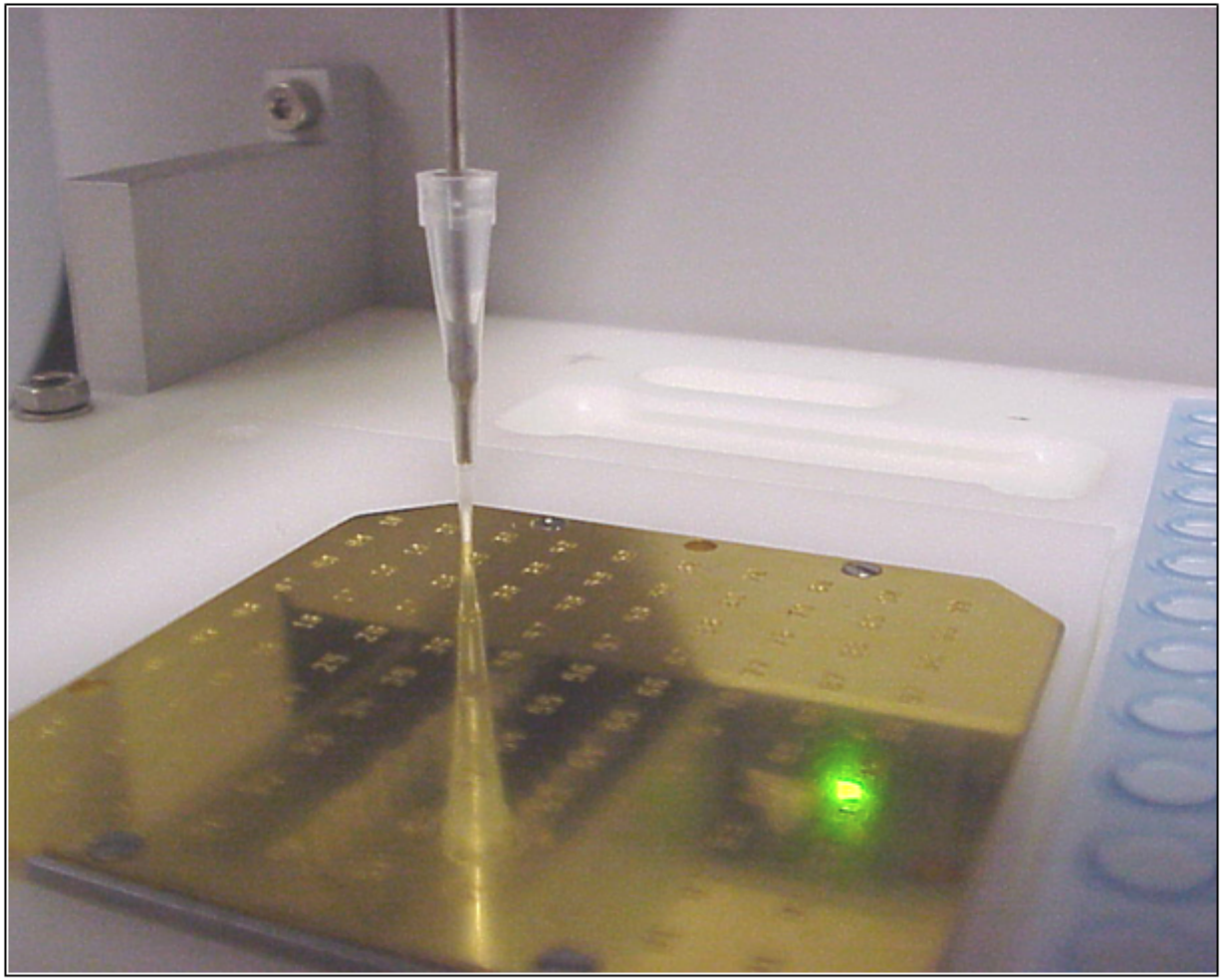
http://upload.wikimedia.org/wikipedia/commons/6/6f/Sinapinic_acid.gif

4-hydroxycinnamic acid

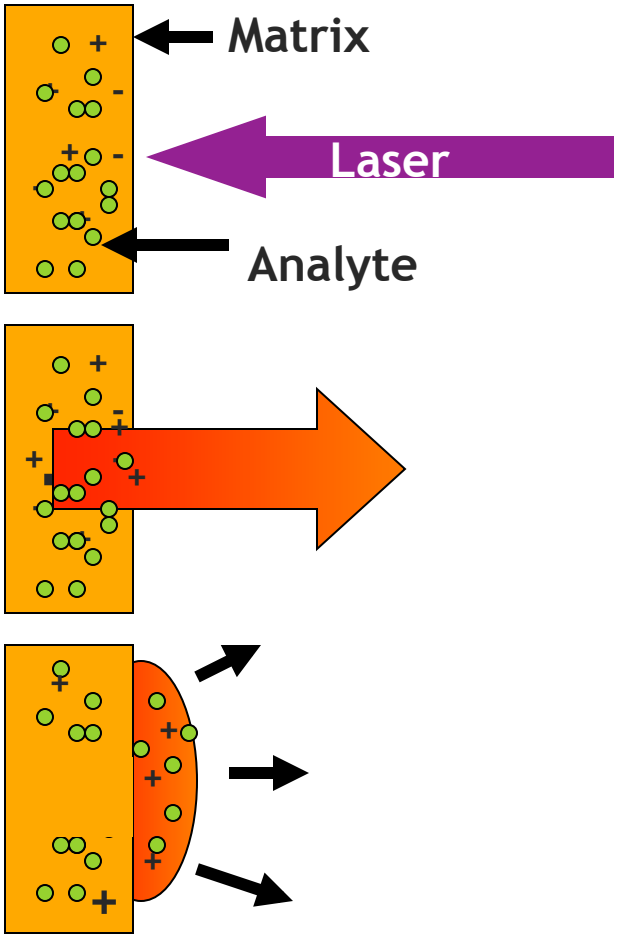


https://commons.wikimedia.org/wiki/File:%CE%91-cyano-4-hydroxycinnamic_acid.svg

Spotting on a MALDI plate



MALDI ionization



- Absorption of UV radiation by chromophoric matrix and ionization of matrix
- Dissociation of matrix, phase change to super-compressed gas, charge transfer to analyte
- Expansion of matrix, analyte trapped in expanding matrix plume (explosion/"popping")

MALDI

- Unlike ESI, MALDI generates spectra that have just a singly charged ion
- Positive mode generates ions of $(M + H)^+$
- Negative mode generates ions of $(M – H)^-$
- Generally more robust that ESI (tolerates salts and nonvolatile components)
- Easier to use and maintain, capable of higher throughput
- Requires 10 μL of 1 pmol/μL sample

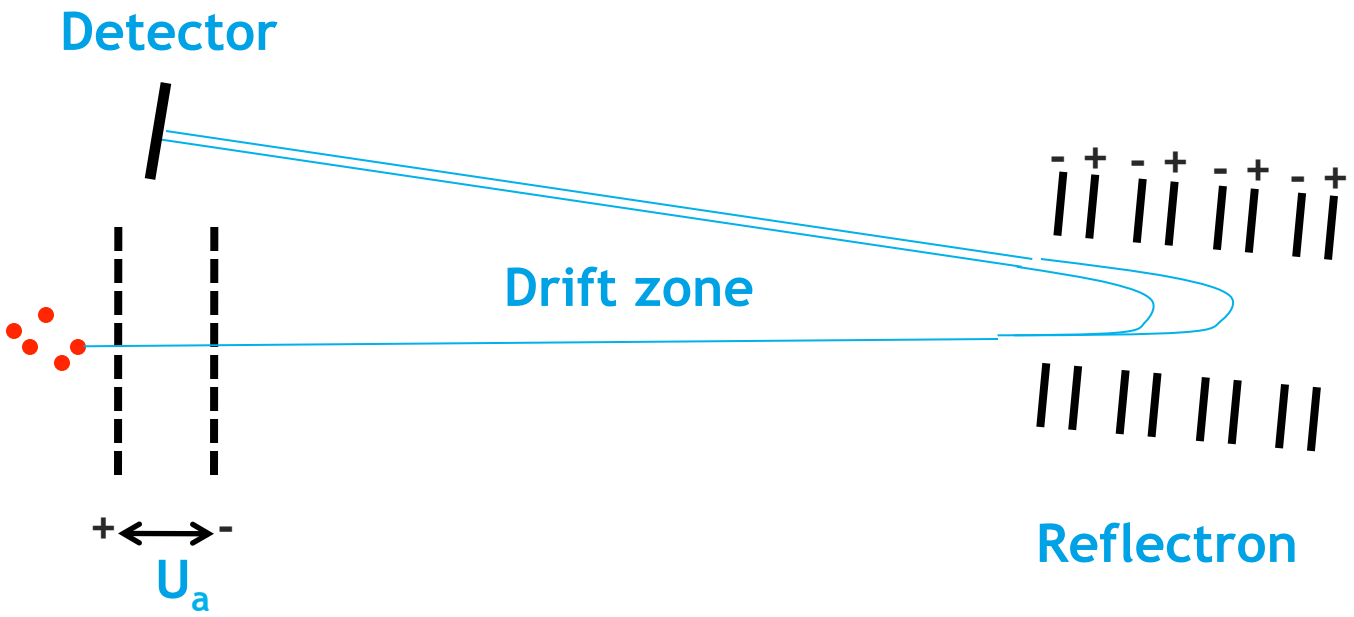
Mass analyzers

- Main operations:
  - Separate peptides
  - Selection of ions (within appropriate m/z)
  - (Fragmentation of selected precursor ions)
  - Measure the m/z of ions

Mass analyzers

- TOF
- Ion trap
- Quadrupole
- Orbitrap
- …

Mass Analyzer: Time of Flight

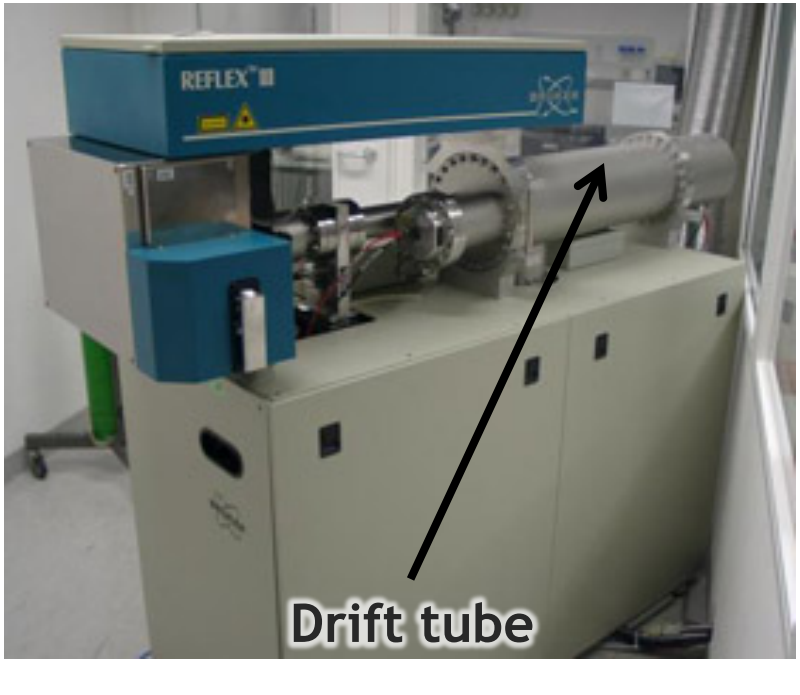


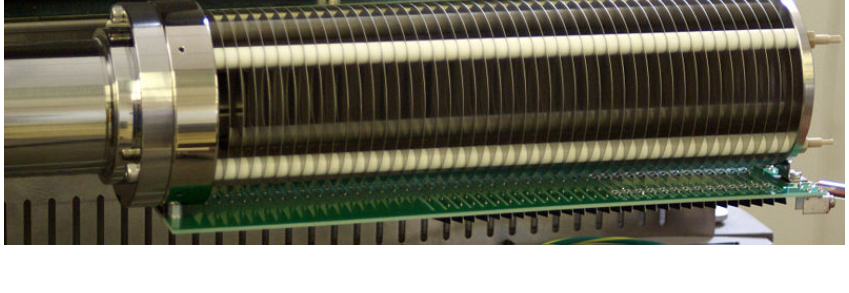
Time-of-flight mass analyzer (TOF):

- Ions are extracted from the ion source through an electrostatic field in pulses in a field-free drift zone
- An 'electrostatic mirror' (reflectron) reflects the ions back onto the detector
- Detector counts the particles and records the time of flight between extraction pulse and a particle hitting the detector

Mass Analyzer: Time of Flight

- **Drift tubes** have sizes of over a meter in real-world instruments
- A **reflectron** doubles the drift length, and thus the instrument's resolution
- It also focuses the ions onto the detector

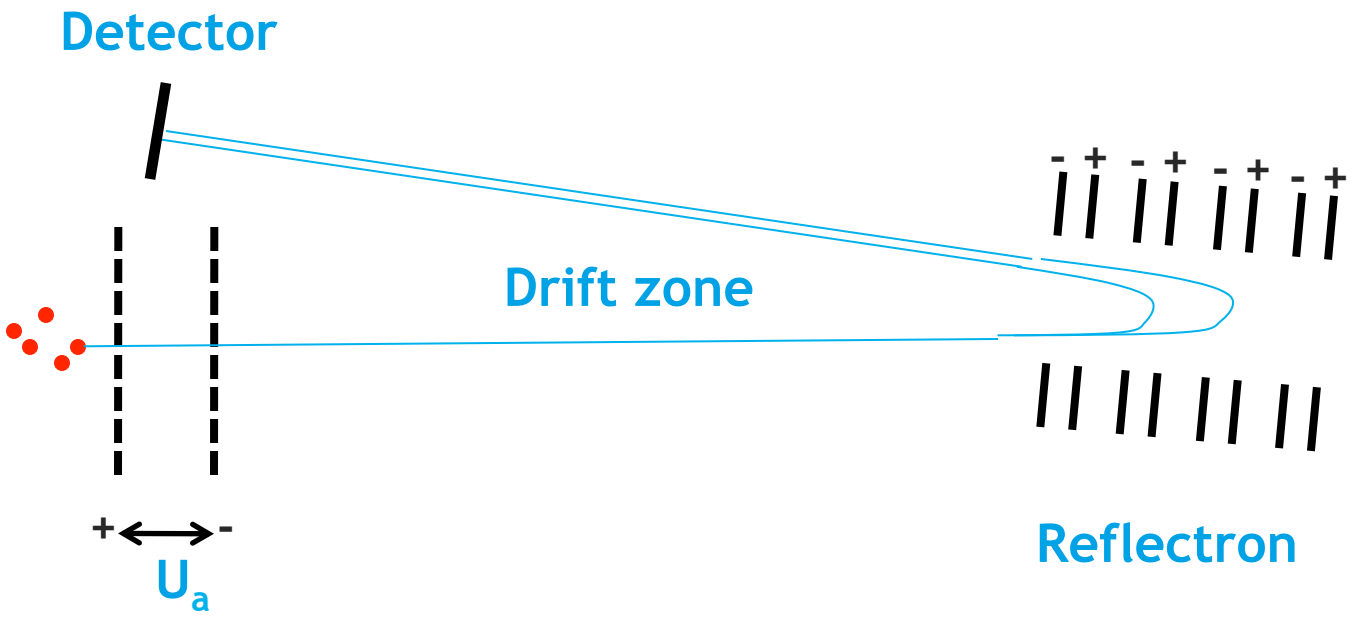




Reflectron

http://www.biochem.mpg.de/nigg/research/koerner/instruments/absatz_pic_reflexIII.jp
http://upload.wikimedia.org/wikipedia/commons/thumb/d/d8/Reflectron.jpg/800px-Reflectron.jp

Mass Analyzer: Time of Flight



- The kinetic energy transferred to the ions depends on the acceleration voltage U_a and the particle's charge
- Lighter particles fly faster that heavier particles of the same charge
- Hence, they arrive later at the detector
- The time of flight is thus a measure of the particle's mass

Mass Analyzer: Time of Flight

- Energy transferred to an ion with charge q accelerated by an electrostatic field with acceleration voltage U_a:

$$E_{pot} = qU_a$$

- This energy is obviously converted into kinetic energy as the ion accelerates:

$$E_{kin} = \tfrac{1}{2}\, mv^2 = qU_a$$

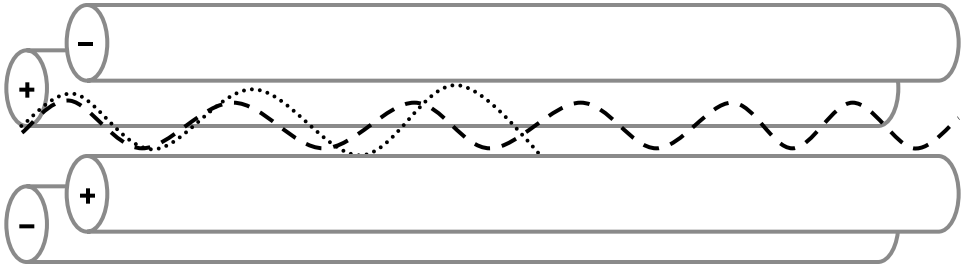
- For a given path length s from extraction to detector, the time of flight t is thus

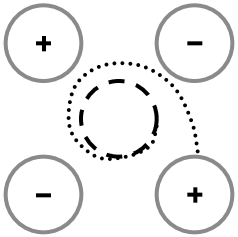
$$t = s / v$$

- Time of flight for a given path length and acceleration voltage, which are instrument parameters, depends on the ion's charge and mass only

Mass Analyzer: Quadrupole

- Oscillating electrostatic fields stabilize the flight path for a specific mass-to-charge ratio – these ions will pass through the quadrupole
- Ions with different *m/z* will be accelerated out of the quadrupole
- Changing the frequency allows the selection of a different *m/z*



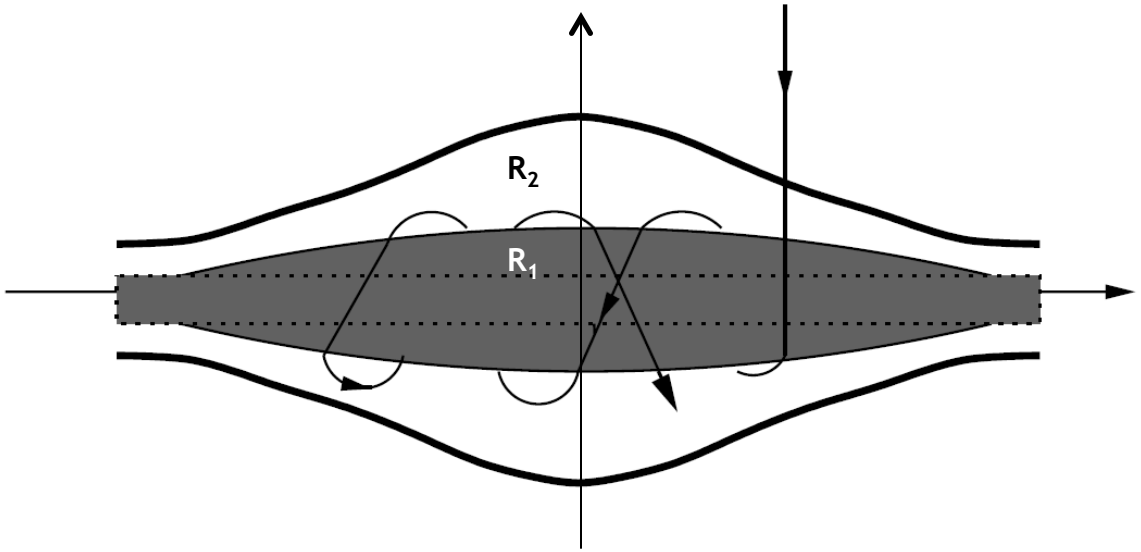


--- stable ion path
····· unstable ion path

Ion Trap

- Ions are captured in a region of a vacuum system or tube
- Trapping of ions is based on a combination of magnetic and electric fields
- There is a long history of ion trapping and a variety of different technologies have emerged over the years
  - Penning trap
  - Paul trap
  - Kingdon trap

Orbitrap (based on Kingdon trap)



- Outer and inner coaxial electrode with radii R_2 and R_1, respectively
- Electrostatic field
- Ions form harmonic oscillation along the axis of the electrostatic field
- The harmonic oscillator with frequency ω is used to determine m/z with $\omega = \sqrt{kz/m}$, where *k* is a constant

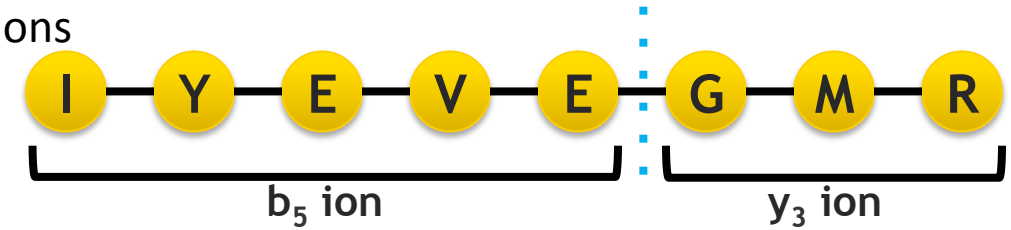
Tandem MS (MS/MS)

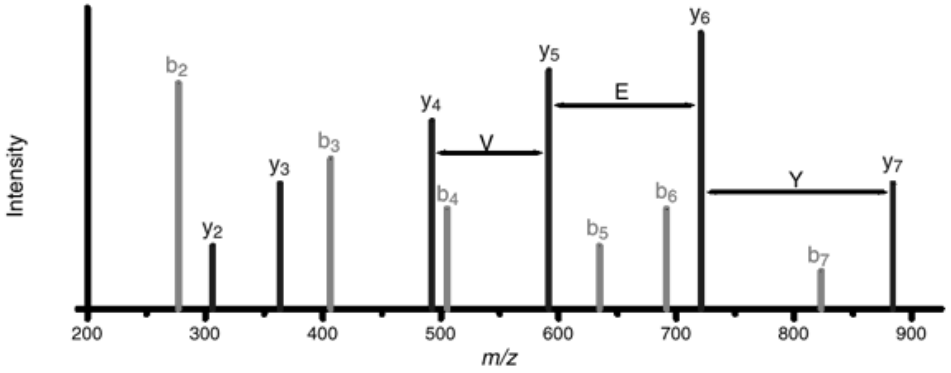
- Uses two mass-to-charge measurements to analyze *precursor* and *product* ions
- Fragmentation is used to dissociate the analytes into smaller fragments
- MS/MS capable instruments
  - Same mass analyzers are used: *in-time set-up*
  - Different analyzers are used (hybrid instruments): *in-space set-up*

Peptide Identification via MS/MS

Why can we identify peptides from tandem MS spectra?

- **Goal: identify sequence**
- Tandem MS
  - Sequence consists of the **same 20 building blocks** (amino acids)
  - CID: peptide breaks preferentially along the **backbone**
  - Peptide **fragment ions correspond to prefixes and suffixes** of the whole peptide sequences
  - Complete ion series (ladders) reveal the sequence via mass differences of adjacent fragment ions

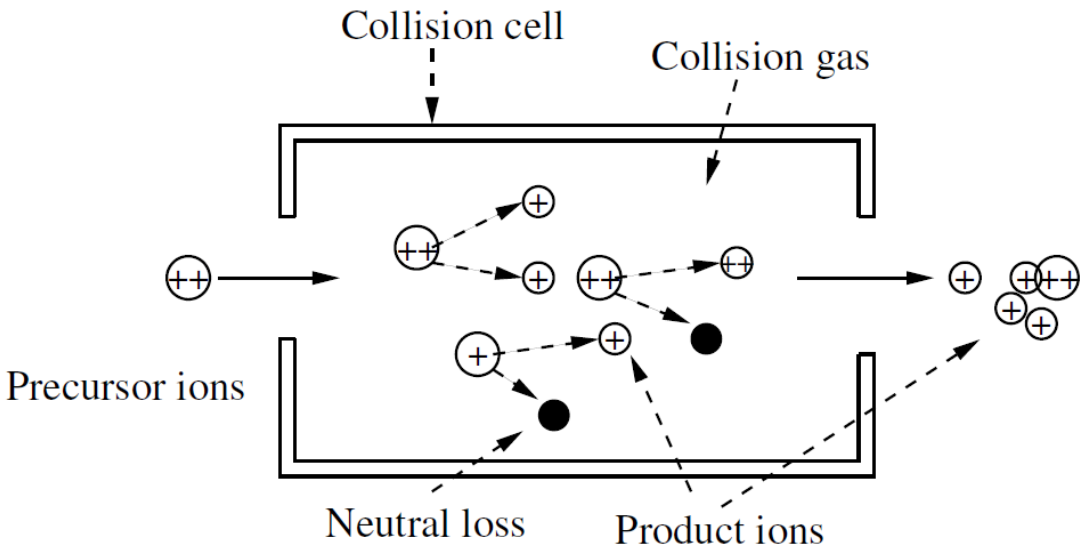




Tandem MS fragmentation methods

- Different fragmentation techniques
  - **Collision-Induced-Dissociation (CID)**
  - Pulsed Q Dissociation (PQD)
  - Electron transfer dissociation (ETD)
  - Electron capture dissociation (ECD)
  - Higher energy collisional dissociation (HCD)

Collision-induced dissociation



- Two colliding molecules
- Fragmentation is performed in collision cell
- Inert collision gas (e.g., Ar, He) is used for collision
- Precursor that reaches energy threshold will fragment into products and/or neutral losses
- Typical settings: high (>1000 eV) or low energy (<100 eV) CID
- Peptides are fragmented at the peptide bond !

Hybrid mass spectrometer

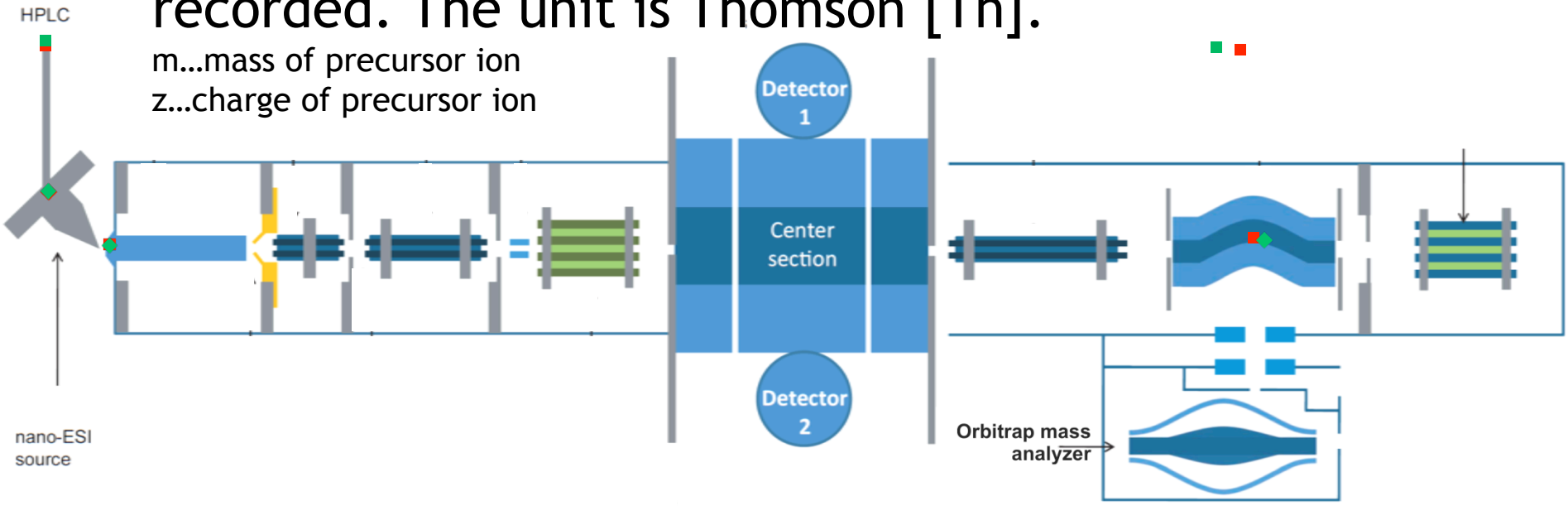
Different hybrid mass spectrometers are used for different applications. The most frequently used combinations are

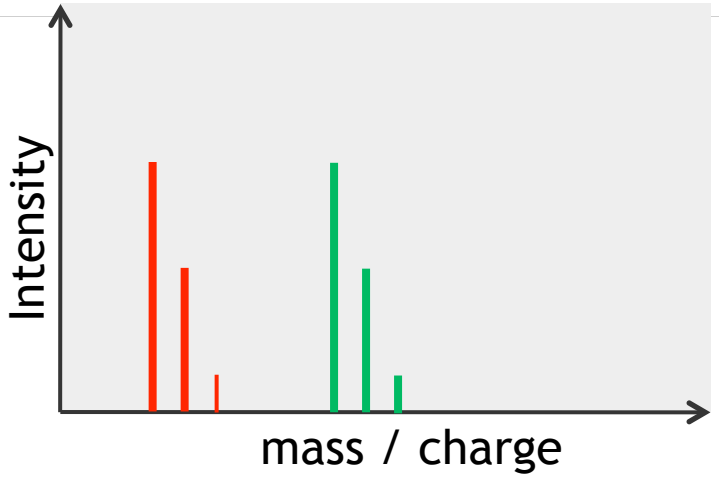
- Q-TOF
- Q-Trap
- **LTQ (linear ion trap)- Orbitrap**

LTQ-Orbitrap – MS

Mass-to-charge ratios (**m/z**) are recorded. The unit is Thomson [Th].

m…mass of precursor ion
z…charge of precursor ion

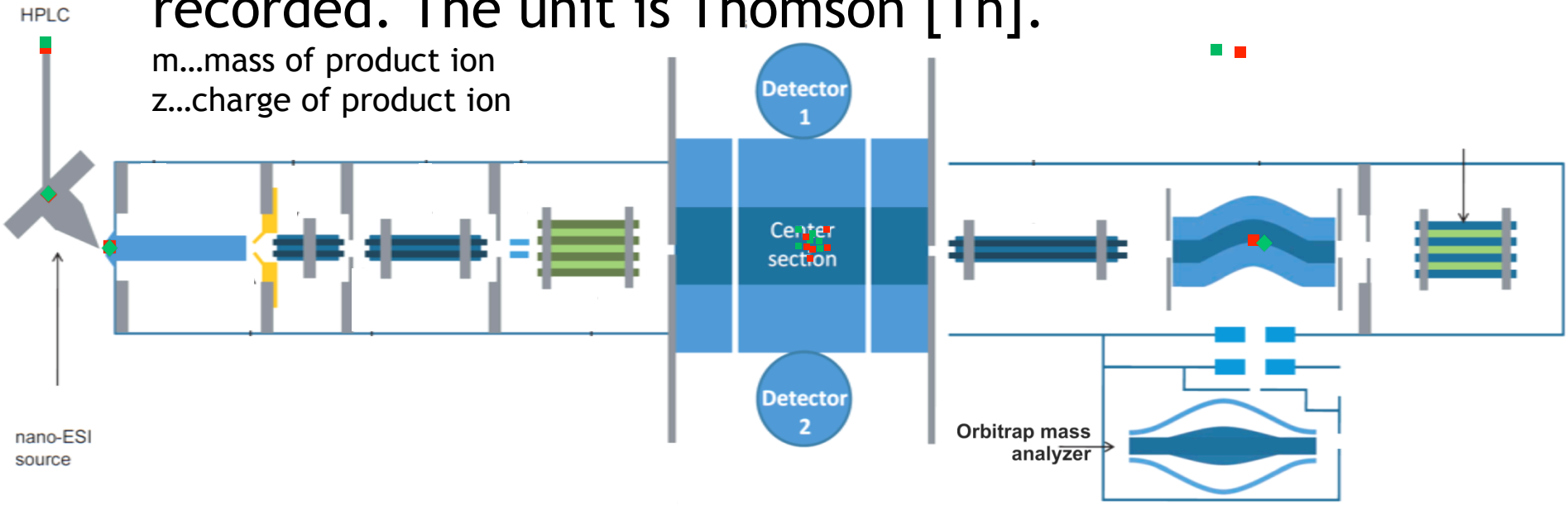


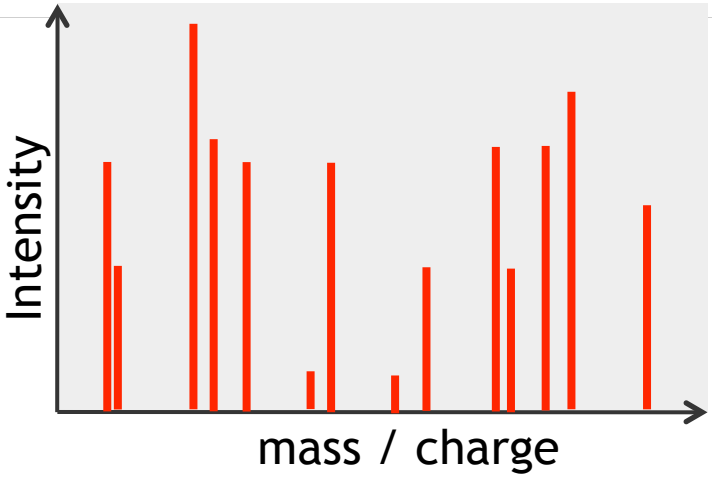


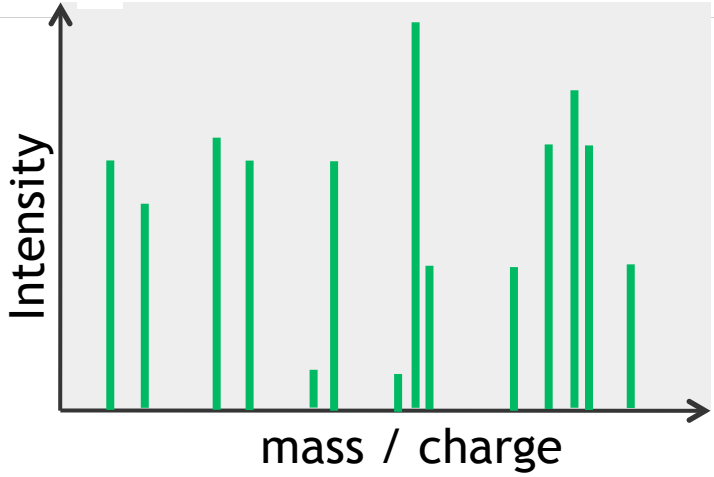
LTQ-Orbitrap – MS/MS

Mass-to-charge ratios (**m/z**) are recorded. The unit is Thomson [Th].

m…mass of product ion
z…charge of product ion

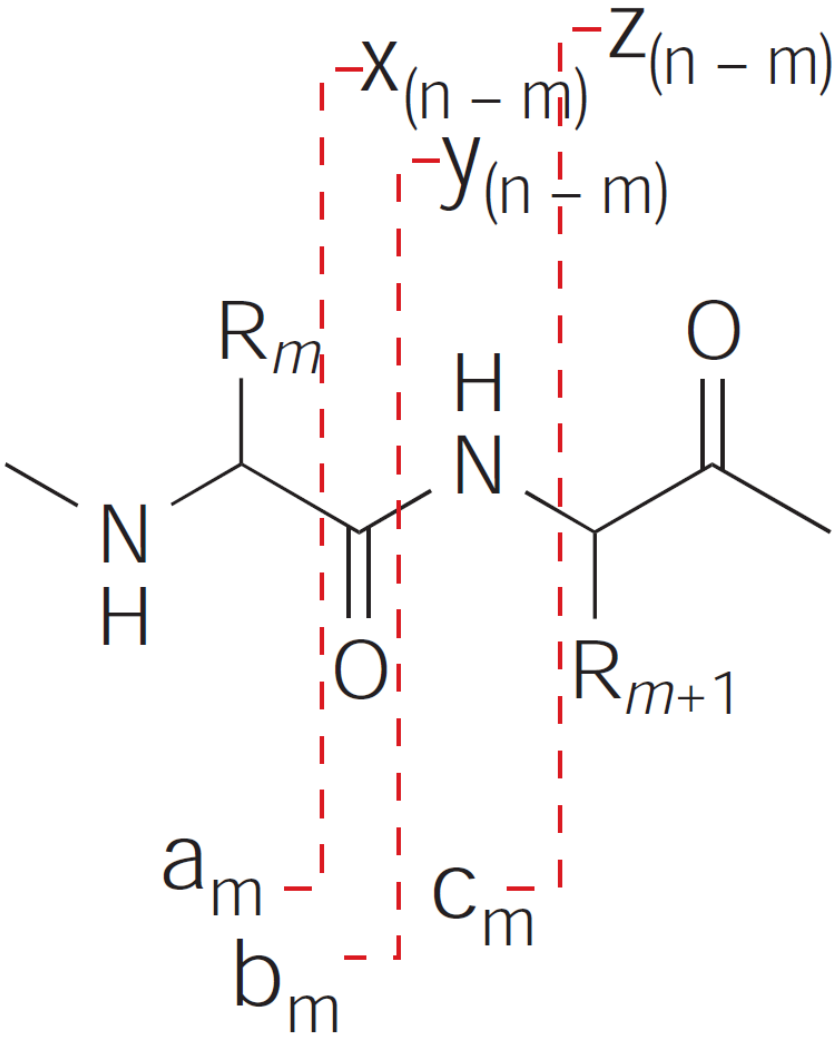






Product ion generation

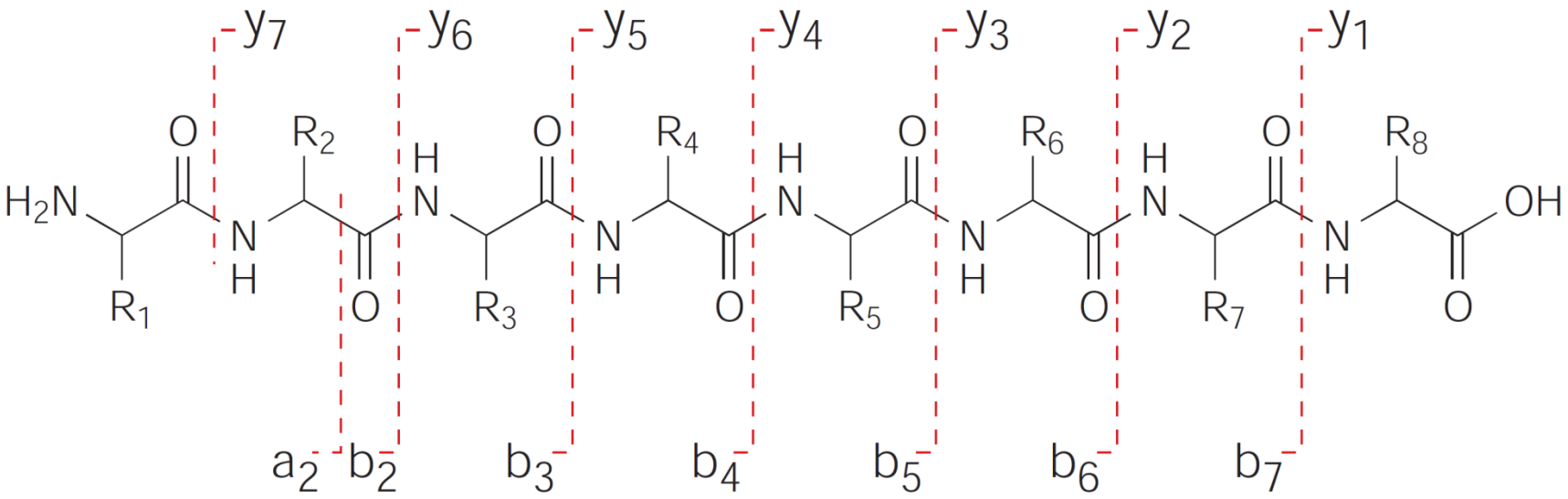
A peptide of length *n* can potentially give rise to a,b,c and x,y,z ions. This example shows the fragments that can be produced between amino acids R_m and R_{m+1}



Steen and Mann. Nature Reviews, Molecular Cell Biology, Vol. 5 2004

b/y Ions in CID

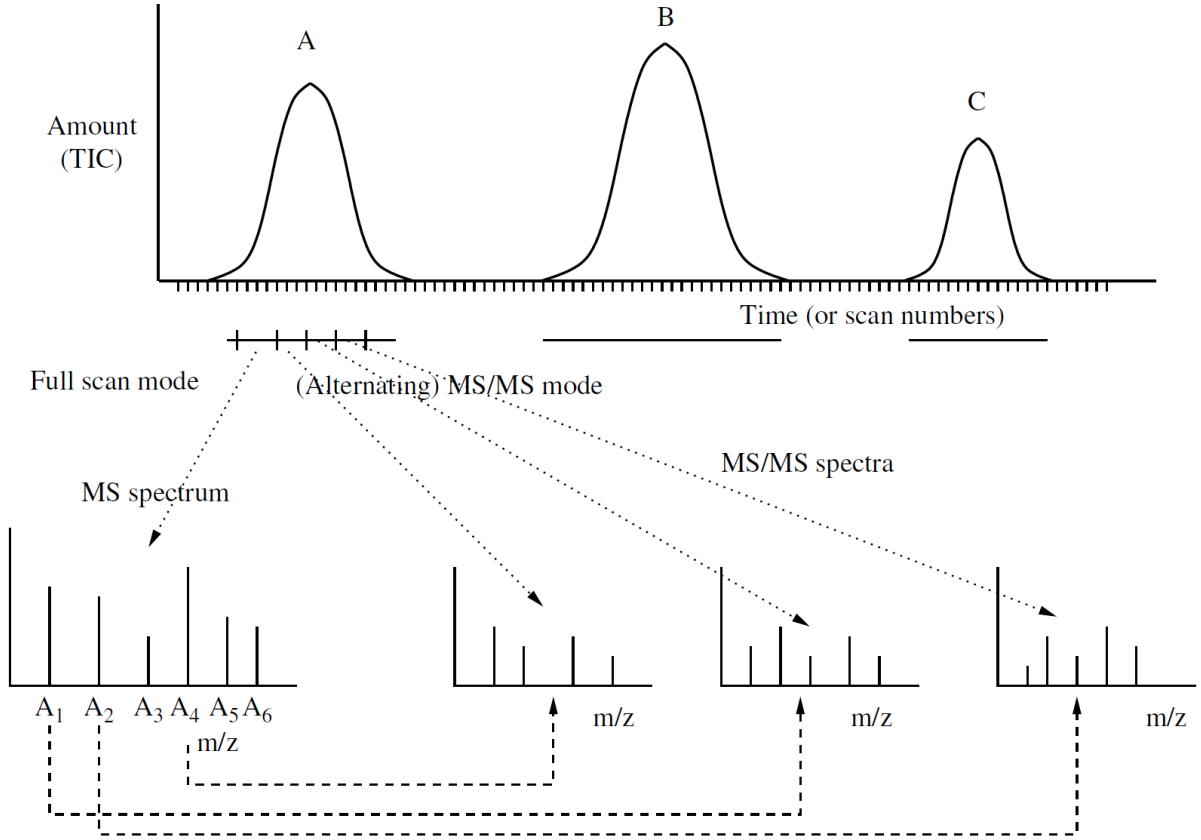
CID fragmentation predominately produces b and y ions



Steen and Mann. Nature Reviews, Molecular Cell Biology, Vol. 5 2004

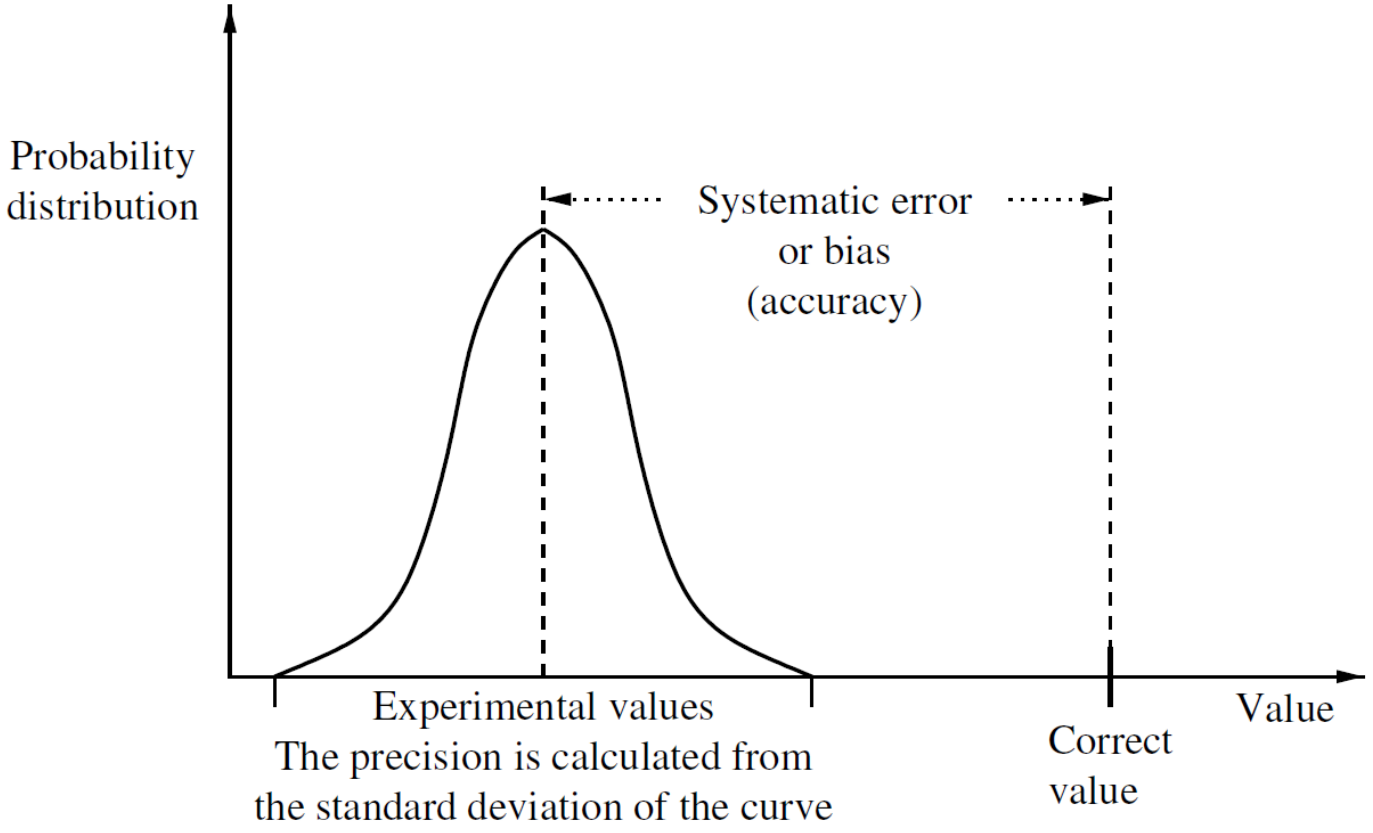
Processes for MS/MS recording

- Select *n* most abundant peaks (usually $3 \leq n \leq 20$) are selected for fragmentation
- Select for specific charge states
- Inclusion list to specify m/z values for fragmentation



Accuracy vs. precision …

… (of a mass measurement)

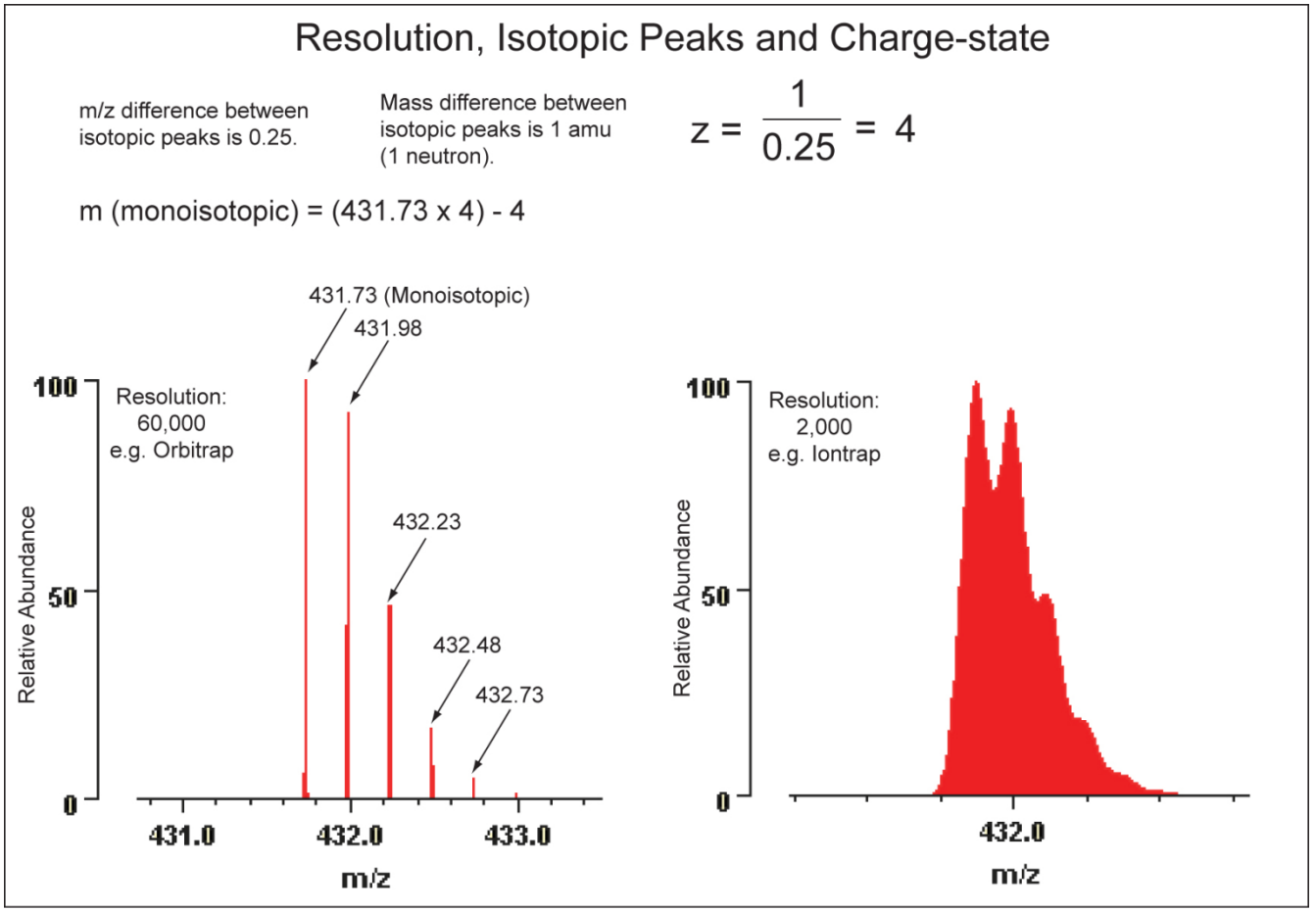


Resolution

(peak width definition)

Resolution:= For a single peak made up of singly charged ions at mass *m* in a mass spectrum, the resolution may be expressed as $m/\Delta m$, where Δm is the width of the peak at a height which is a specified fraction of the maximum peak height. It has been standardized to use 50% of the maximum peak height. FMWH (Full Width at Half Maximum) is commonly used. Note that resolution is dimensionless. Furthermore, in proteomics it has become common to report the resolution for ions at 400 Th.

Charge states



- Charge state determination is easy if the resolution is high enough
- For low resolution data this can become difficult

http://www.lamondlab.com/MSResource/LCMS/MassSpectrometry/chargeState.php

Materials

- Learning Units 2A and 2B

- Video on the separation of plant pigments (Tsvet's experiment) on YouTube

  http://www.youtube.com/watch?v=3N1Rt6nWczY