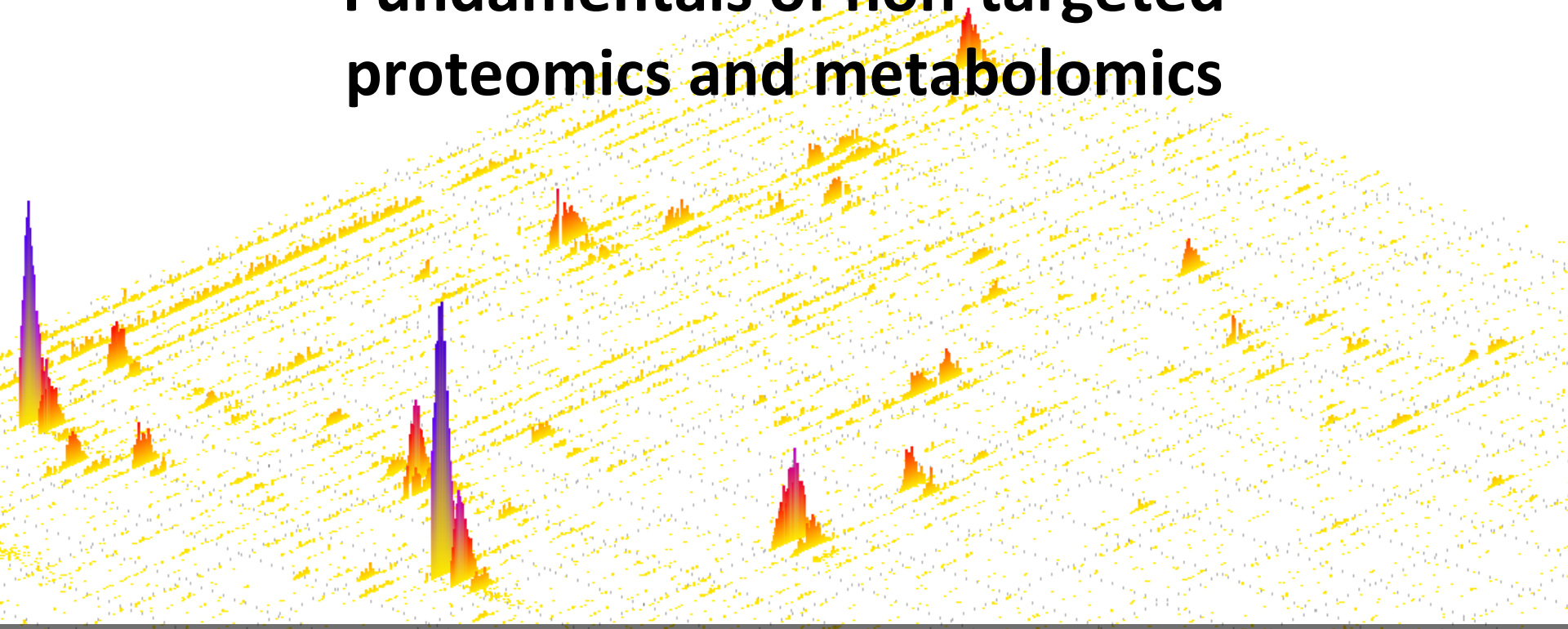
May Institute 2017

Computation and statistics for mass spectrometry and proteomics

Fundamentals of non-targeted proteomics and metabolomics



Oliver Kohlbacher

University of Tübingen and

MPI for Developmental Biology

KohlbacherLab.org | @okohlbacher

MAX-PLANCK-GESELLSCHAFT

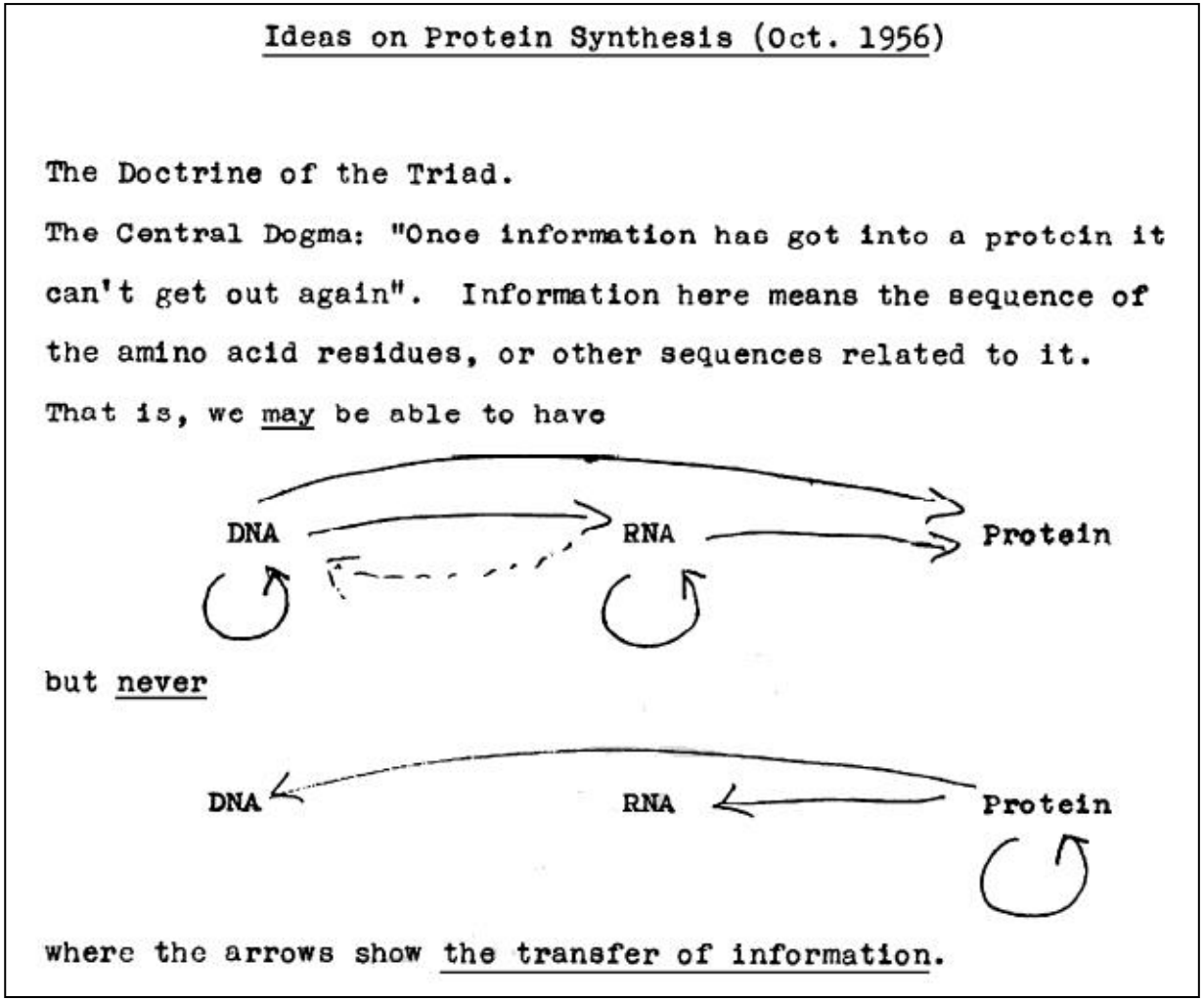
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Overview

- **Introduction to proteomics and metabolomics**
  - Systems biology and omics technologies
  - Key ideas in proteomics and metabolomics
  - Basics of LC-MS
- **OpenMS**
  - Philosophy – Workflows and reproducible science
  - OpenMS – Key ideas and concepts
  - KNIME – Workflow construction and execution

Central Dogma of Molecular Biology

- First described by Francis Crick in 1956
- Published in Nature in 1970



Ideas on Protein Synthesis (Oct. 1956)

The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have

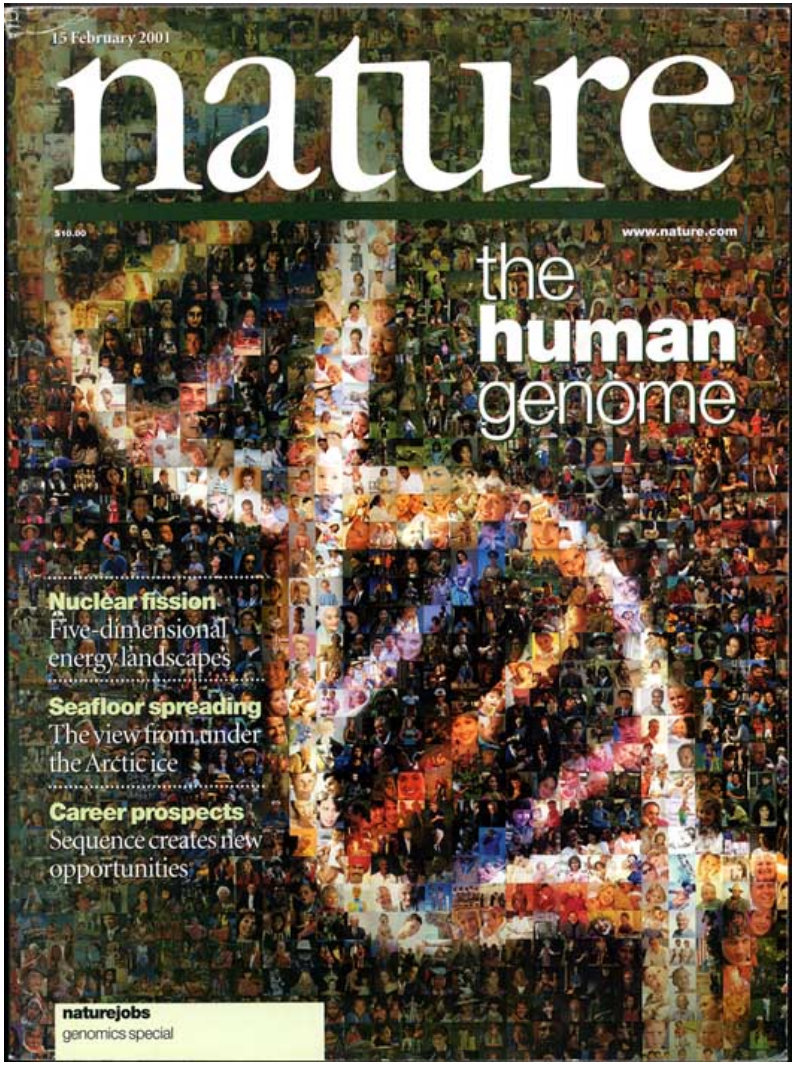
but never

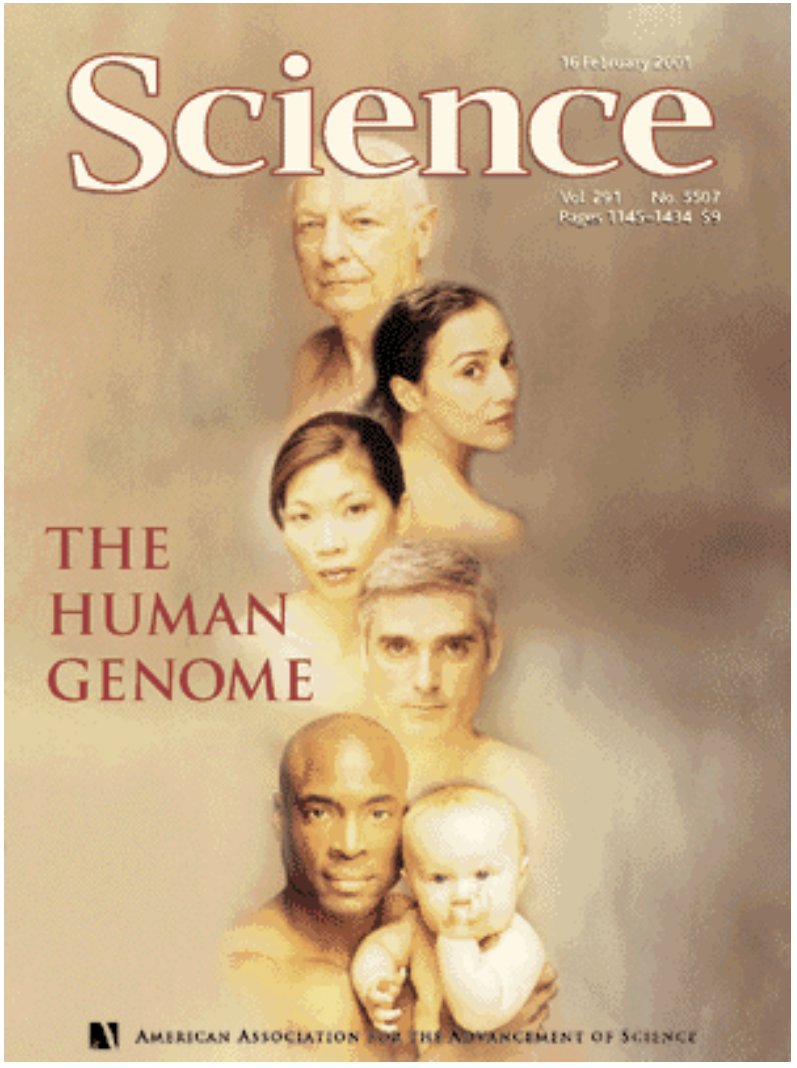
where the arrows show the transfer of information.

Origin of the "Central Dogma of Molecular Biology" (Francis Crick, 1956)

Genome sequencing

February 2001 – Publication of the first draft of the human genome





‘Postgenomics’ – The Age of Omes

-ome, *comb. form*

[…]

3. *Cell Biol. and Molecular Biol.* Forming nouns with the sense ‘all of the specified constituents of a cell, considered collectively or in total’, as plastidome n., plastome n., vacuome n.

(Oxford English Dictionary online)

Ever since the rise of genomics, the suffix "-omics" has been added to many fields to denote studies undertaken on a large or genome-wide scale. While not everyone agrees with this change of terms, we felt that the terms are sufficiently widely used to serve as pointers to our published papers in the area.

(Website of ‘Nature’)

http://www.oed.com
http://www.nature.com/omics/about/index.html

OMICS Mania

Alphabetically ordered list of omes and omics

Alphabetically ordered Omes and Omics. You can freely add and edit the entries.

--A--

Alignmentome: conceived before 2003. The whole set of multiple sequence and structure alignments in bioinformatics. Alignments are the most important representation in bioinformatics especially for homology and evolution study. (Alignmentome.org)

Alignmentomics: conceived before 2003. The study of aligning strings and sequences especially in bioinformatics. (Alignmentomics.org)

Alignome: 2003 . The whole set of string alignment algorithms such as FASTA, BLAST and HMMER. (Alignome.org)

Alignomics: The omics approach research of Alignomics (Alignomics.org) in biology

Alternatome: 2006. The totality of alternative spliceable elements. Suggested by people in KOBIC and UCSC. (Alternatome.org)

Alternatomics: The omics approach research of Alternatomics (Alternatomics.org) in biology

Animalome: 2000 . The whole set of animals and their genetic components on Earth. While animal kingdom traditionally means the totality of animals, animalome indicates the system of animals, animal genes, animality, and complex network of animal genes and proteins. Animals contain proteins that are special. (Animalome.org)

Animalomics: The omics approach research of Animalomics (Animalomics.org) in biology

Aniome: 2003 . The whole set of any biologically relevant things in the universe. (Aniome.org)

Antibodyome: conceived around 2003 in association with immunolome in artificial immune system as computational system (Jong). (Antibodyome.org)

Antibodyomics: The omics approach research of Antibodyome (Antibodyomics.org) in biology

Antiome: The totality of people who object the propagation of omes.

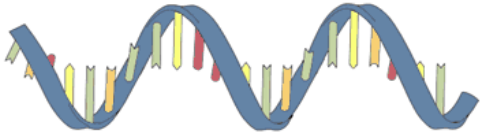
Antiomics: The omics study of analyzing the trend of attaching omics suffix to debunk it.

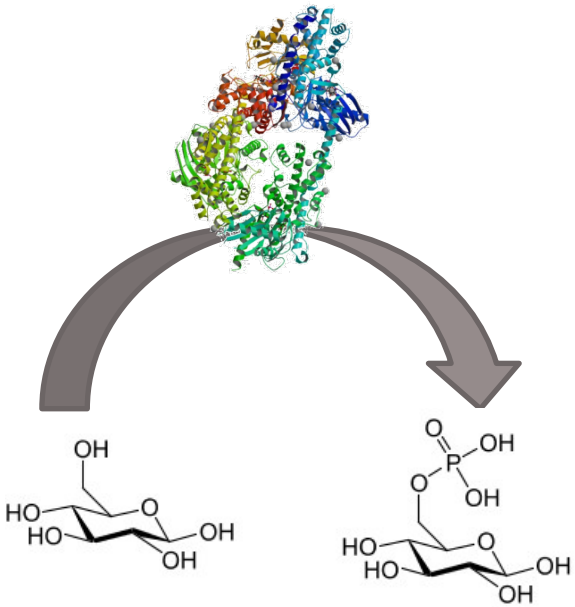
Archaeome: 2002 . All the species of archae and their proteins especially. (Archaeome.org)

Archaeomics: The omics approach research of Archaeomics (Archaeomics.org) in biology

Archiome: 2002 . The same as archaeome. (Archiome.org)

The World of Omes





DNA

mRNA

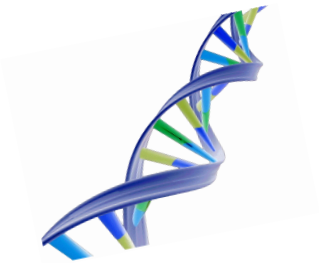
Protein

Metabolites

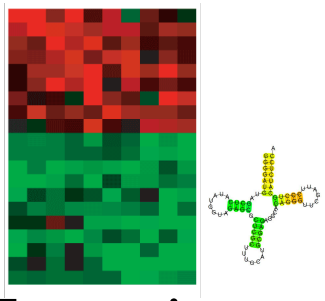
- Genome
- Transcriptome
- Proteome
- Metabolome

http://www.pdb.org/pdb/images/2nzt_bio_r_500.jpg

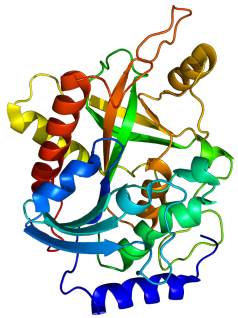
Technologies



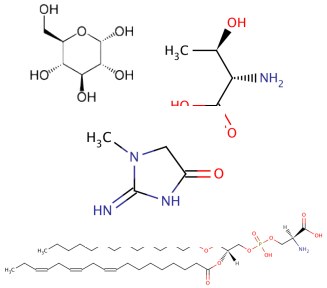
Genome
Epigenome



Transcriptome
RNOme



Proteome
Interactome

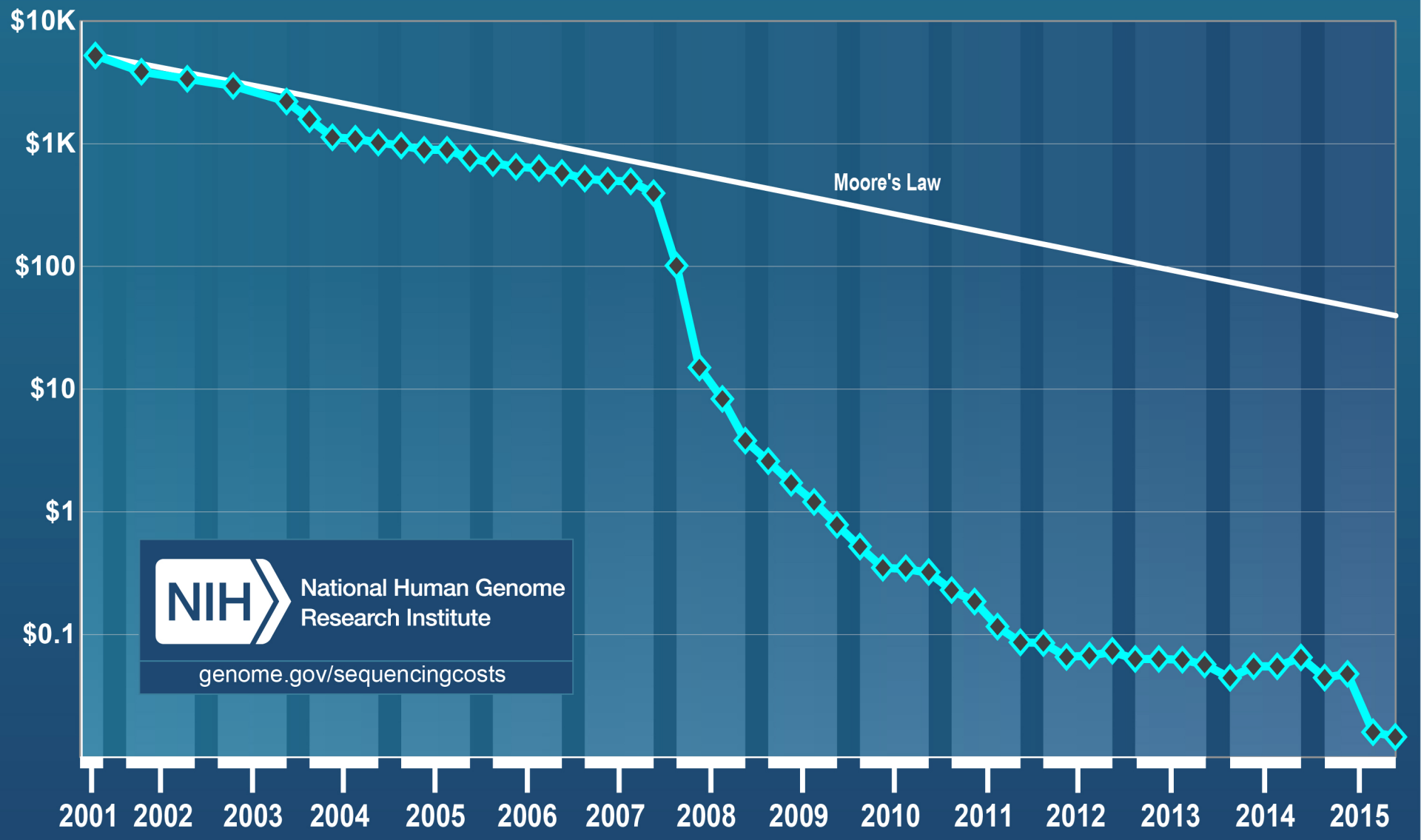


Metabolome
Lipidome

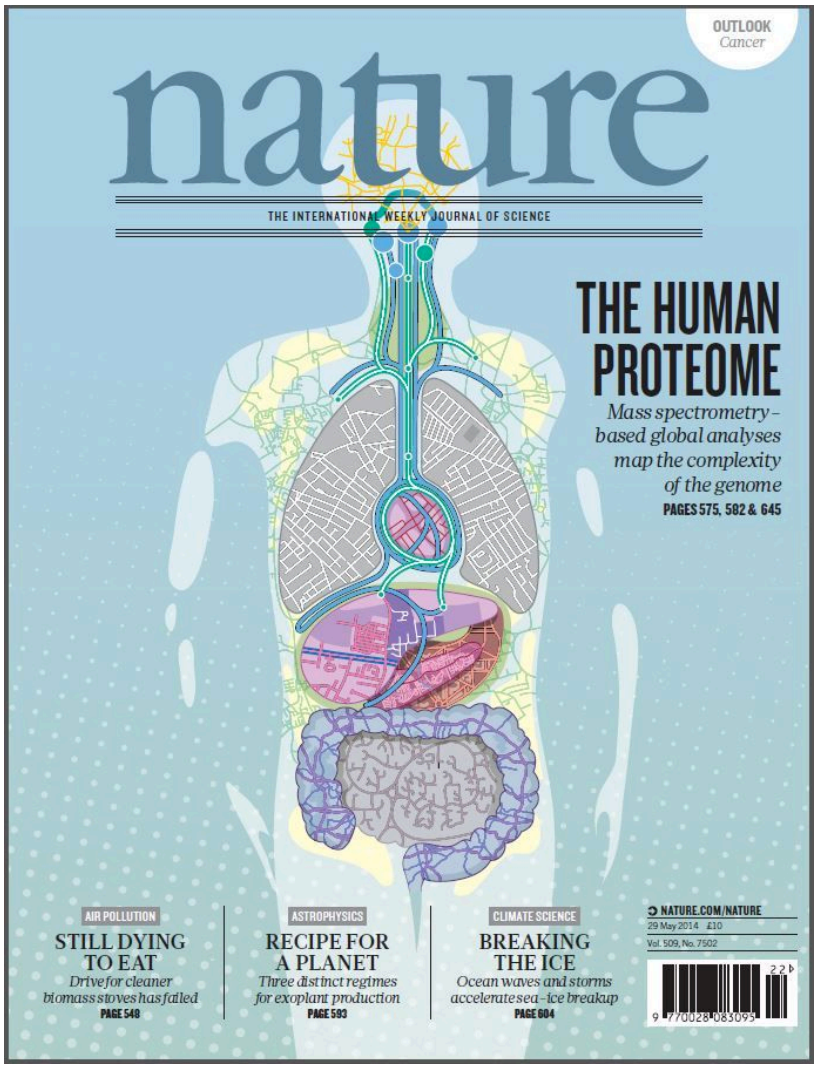
Next-Generation Sequencing

Mass Spectrometry

Cost per Raw Megabase of DNA Sequence



Human Proteome



Nature cover May 2014

- Two draft versions of the human proteome (for various) tissues
- Claim ~90% coverage of the proteome

Nature, 509 (2014) [cover]

OMICS Data

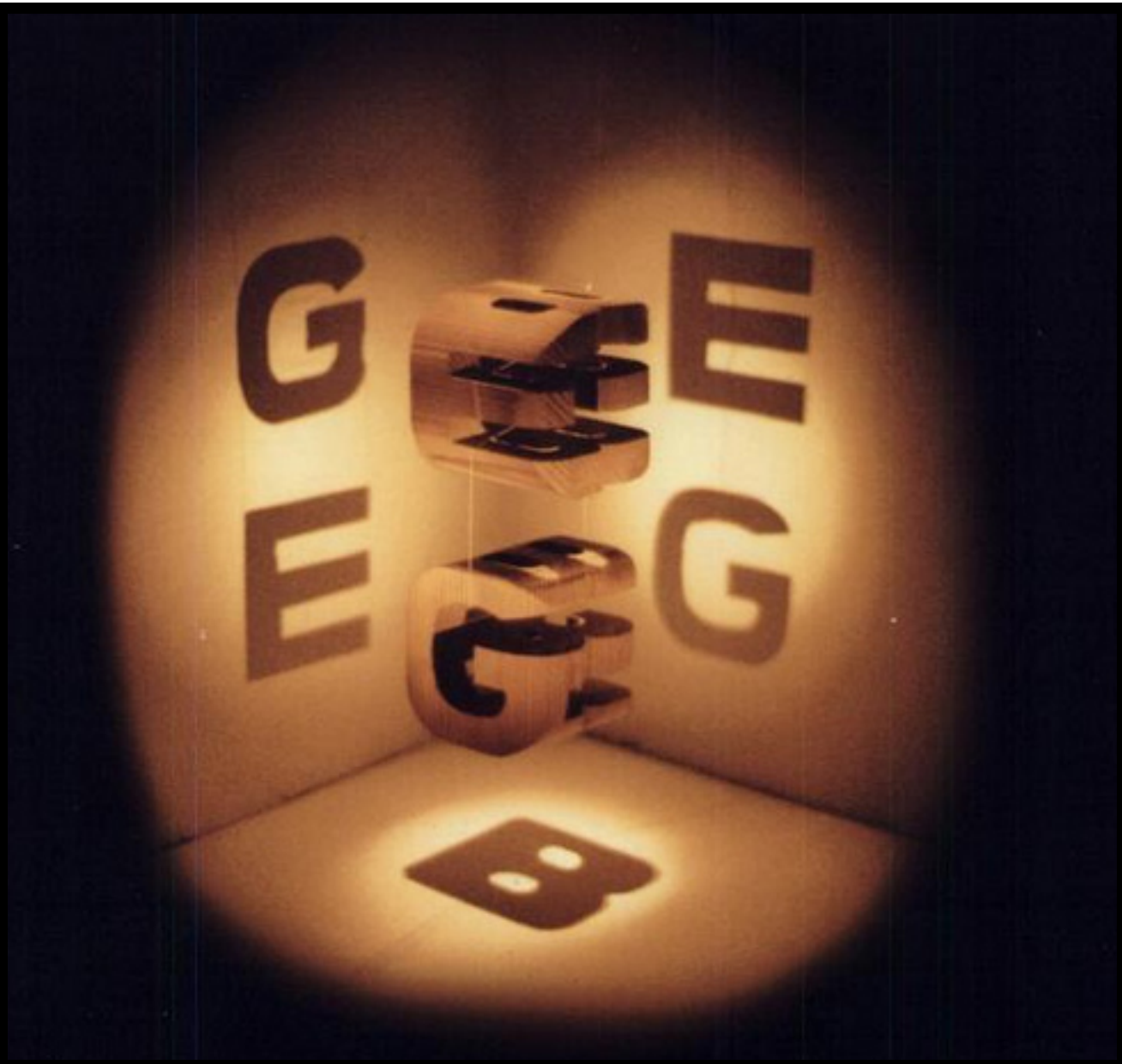
- **High-throughput techniques** provide data for one specific type of relationship
  - **Genomics**: DNA sequence data
  - **Transcriptomics**: mRNA concentration
  - **Proteomics**: protein concentrations/sequence
  - **Metabolomics**: metabolite concentrations
  - **Interactomics**: protein-protein interaction data

- OMICS data is reductionist, but at a very large scale
- OMICS data is often voluminous, but of low quality/noisy

Classical Data vs. Omics Data

Classical	Omics
• Low-throughput	• High-throughput
• Low-dimensional, often single facts	• High-dimensional, measuring many parameters in parallel
• High accuracy, every data point supported by multiple experiments	• Often low accuracy, lots of noise
• Analysis of experiments simple (small data volume!)	• Often not interpretable without statistics/bioinformatics

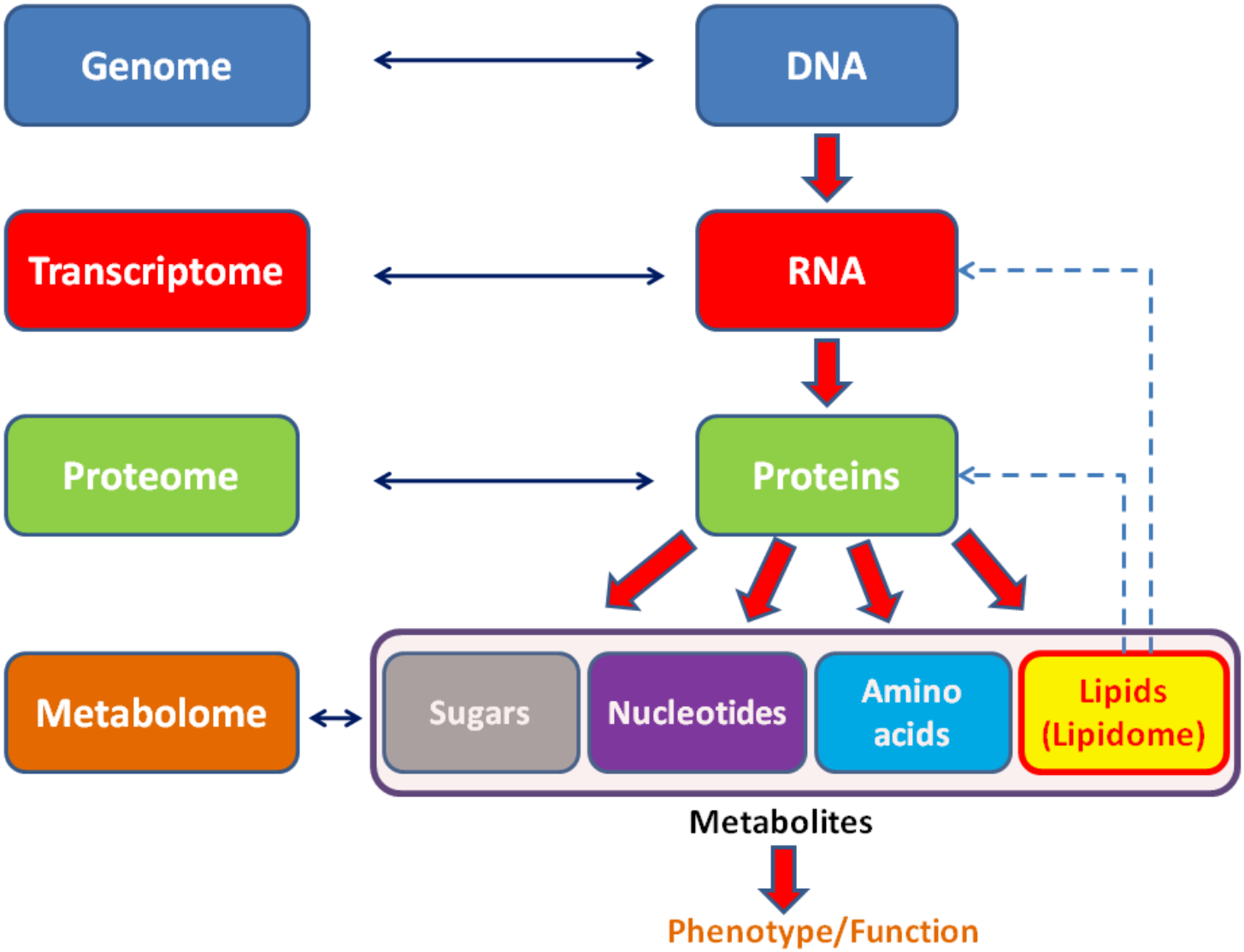
Omics is a Matter of Perspective!



Omics is a Matter of Perspective

- Each omics technology/level provides a cross-section of one **particular type of biomolecules**
- Different levels correlate (roughly) to distinct functional levels as well:
  - **Genomics**: what can the cell potentially do?
  - **Transcriptomics**: what is currently being turned on?
  - **Proteomics**: what enzymes are currently active? which signals are being transduced?
  - **Metabolomics**: what is being produced/consumed?

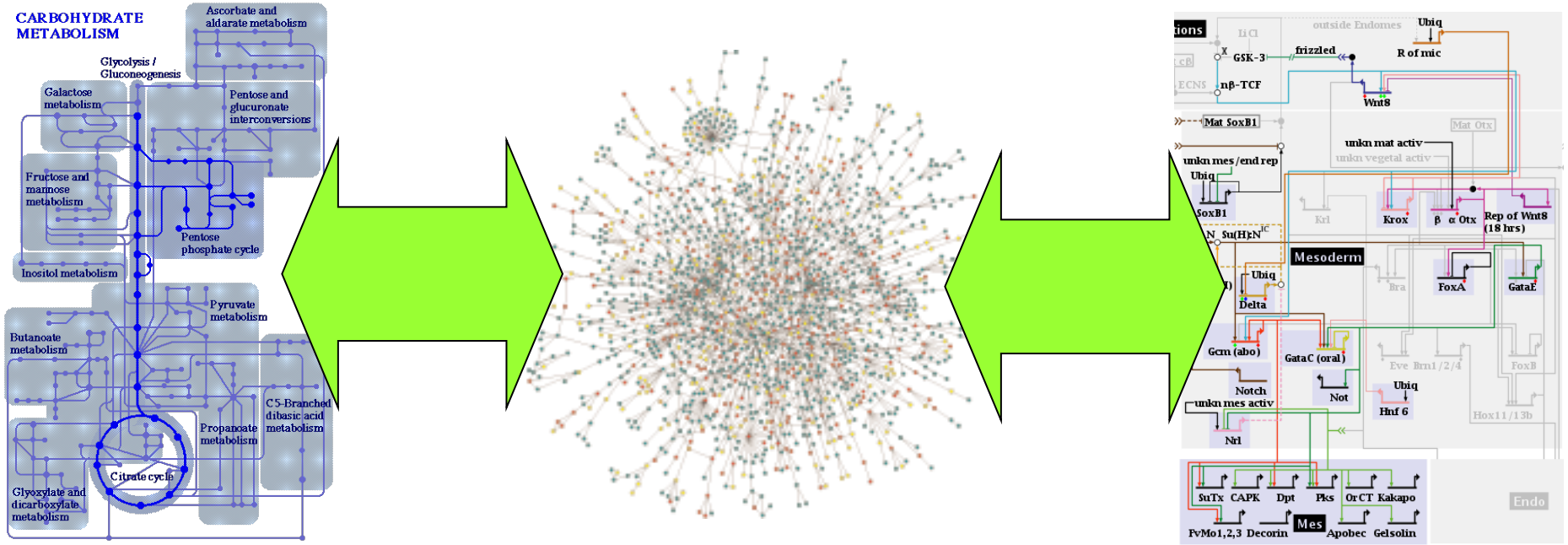
Omics Technologies



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

Integrative Analysis

- Analyzing individual data set is trivial
- **Simultaneous integrated analysis** of data from multiple layers/types of data is currently still the major challenge!

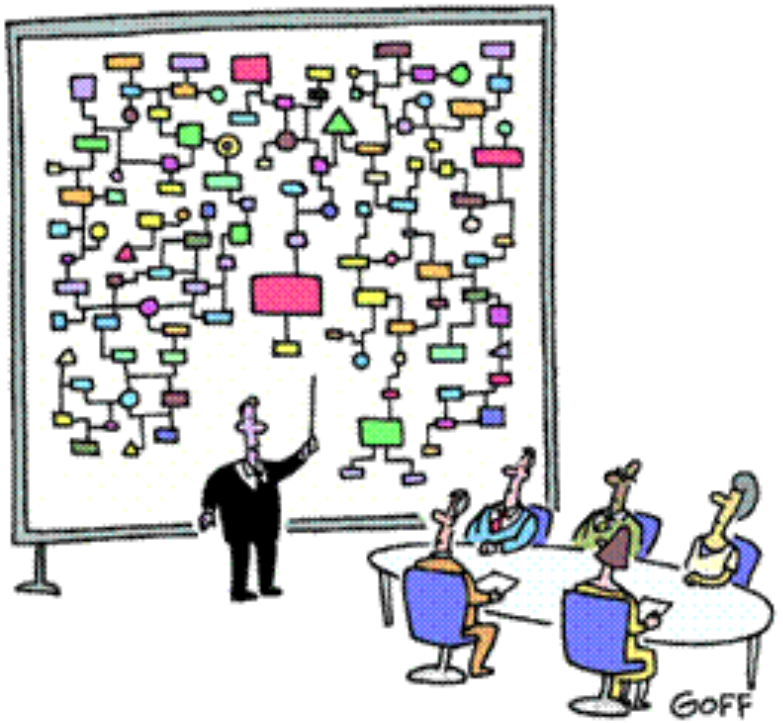


Computational Systems Biology

- The complexity and also the sheer amount of data produced with high-throughput techniques makes manual analysis difficult

- Systems biology thus requires a strong computational component:

  Computational Systems Biology



"And that's why we need a computer."

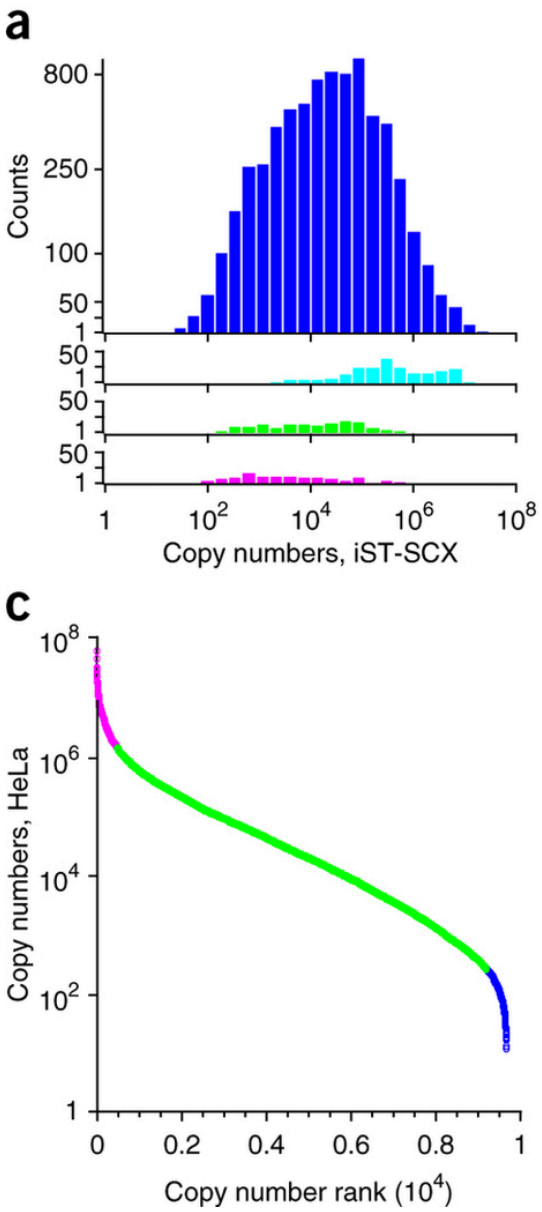
http://www.sys-bio.org/contentimages/WhyWeNeedComputer.png

Challenges in Data Integration

- **Semantic integration** of data from different sources
  - Different data formats
  - Ambiguities, nomenclature
- **Lack of data**
  - We do not know everything!
  - High-throughput methods show only a fraction of ‘everything’ (detection limits!)
- **Different scales**
  - Time scales different, length scales different
  - How to model different resolutions simultaneously?

Proteomics

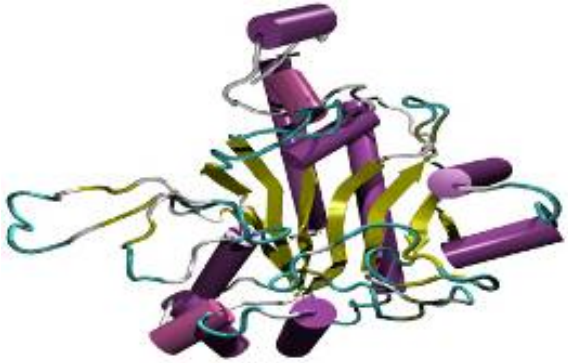
- **Proteomics**: study of a proteome
- **Proteome**: sum of all proteins in a given sample (e.g., tissue, cell, time-point)
- Proteomics typically tries to
  - Catalog the proteins in a sample (**qualitative proteomics**)
  - Quantify the proteins in a sample, i.e., determine the concentrations of all proteins (**quantitative proteomics**)
- Concentrations in a sample vary drastically – large dynamic range required (see figure on the right)



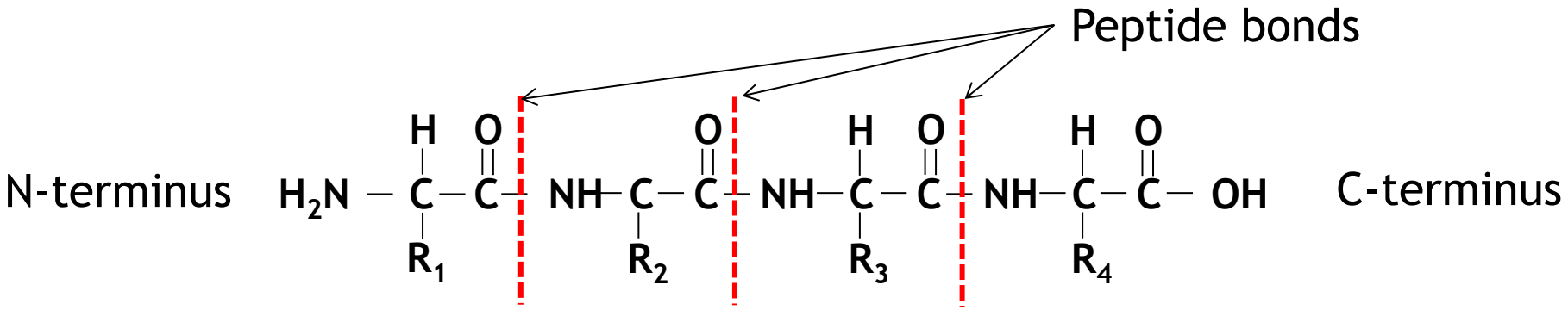
Kulak et al., Nat. Methods 11, 319-324 (2014) doi:10.1038/nmeth.2834

Protein

- A protein or polypeptide consists of a linear chain of amino acids that form three-dimensional structures



- Amino acids are connected via peptide bonds



Proteomics – Typical Questions

- There are some problematic issues on defining a protein
  - Protein identity: unique amino acid sequence and single source of origin?
    - There may be different genes encoding the identical amino acid sequence
    - Different organisms may encode identical proteins
  - Splice variants: A gene can give rise to different mRNAs
  - Polymorphisms: many genes occur in allelic variants encoding sequence variations
  - Post-translational modifications (PTMs): PTMs are very hetero-geneous and significantly alter the function of the protein

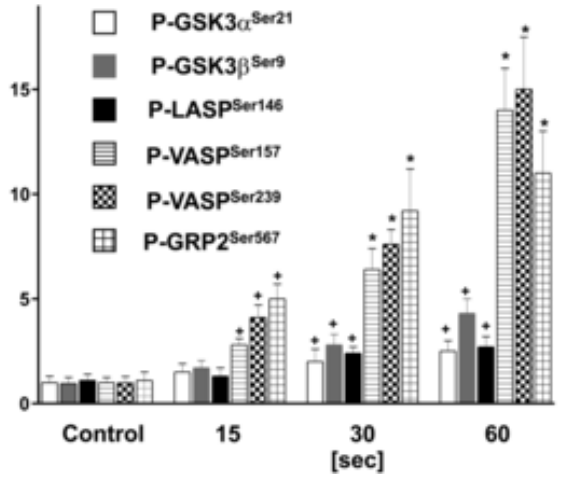
Proteomics - Examples

Understanding phenotypes:
Genome remains the same…

…proteome changes

Understanding signaling:
Platelets are non-nucleated cells – to understand their behavior (blood clotting) phosphoproteomics is required. It reveals time-resolved activation of kinases.

Activated platelets



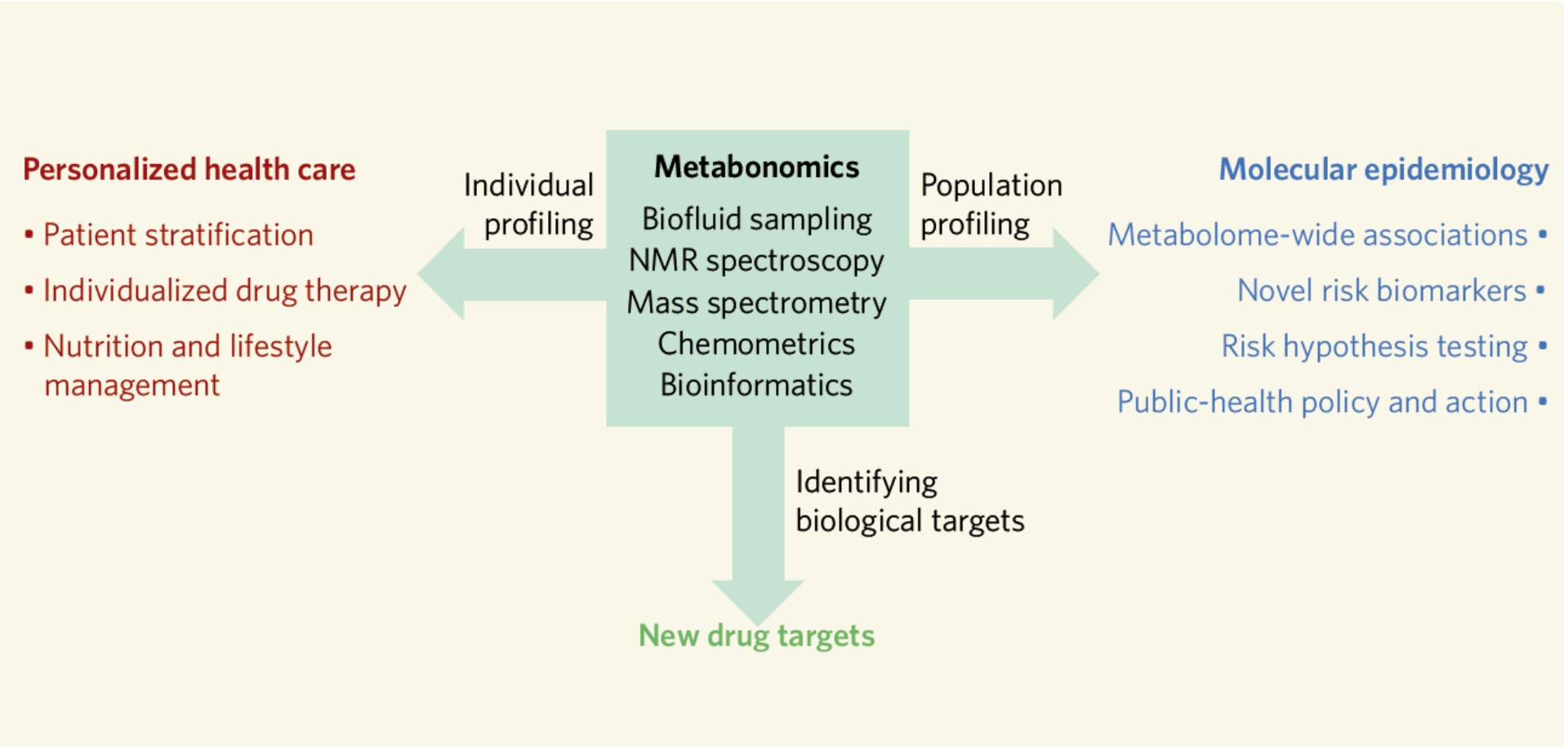
Time course of selected phosphopeptides (Beck et al., 2014)

http://www.iamashcash.com/wp-content/uploads/2011/03/caterpillar-to-butterfly1.jpg, accessed: 14/10/2013 6 PM
http://www.ufrgs.br/laprotox/en/what-we-do/research-lines/ureases-non-enzymatic-properties/ureases-induce-platelet-secretion-and-aggregation
Beck et al., Blood (2014), 123(5):e1-e10. doi: 10.1182/blood-2013-07-512384

Metabolites

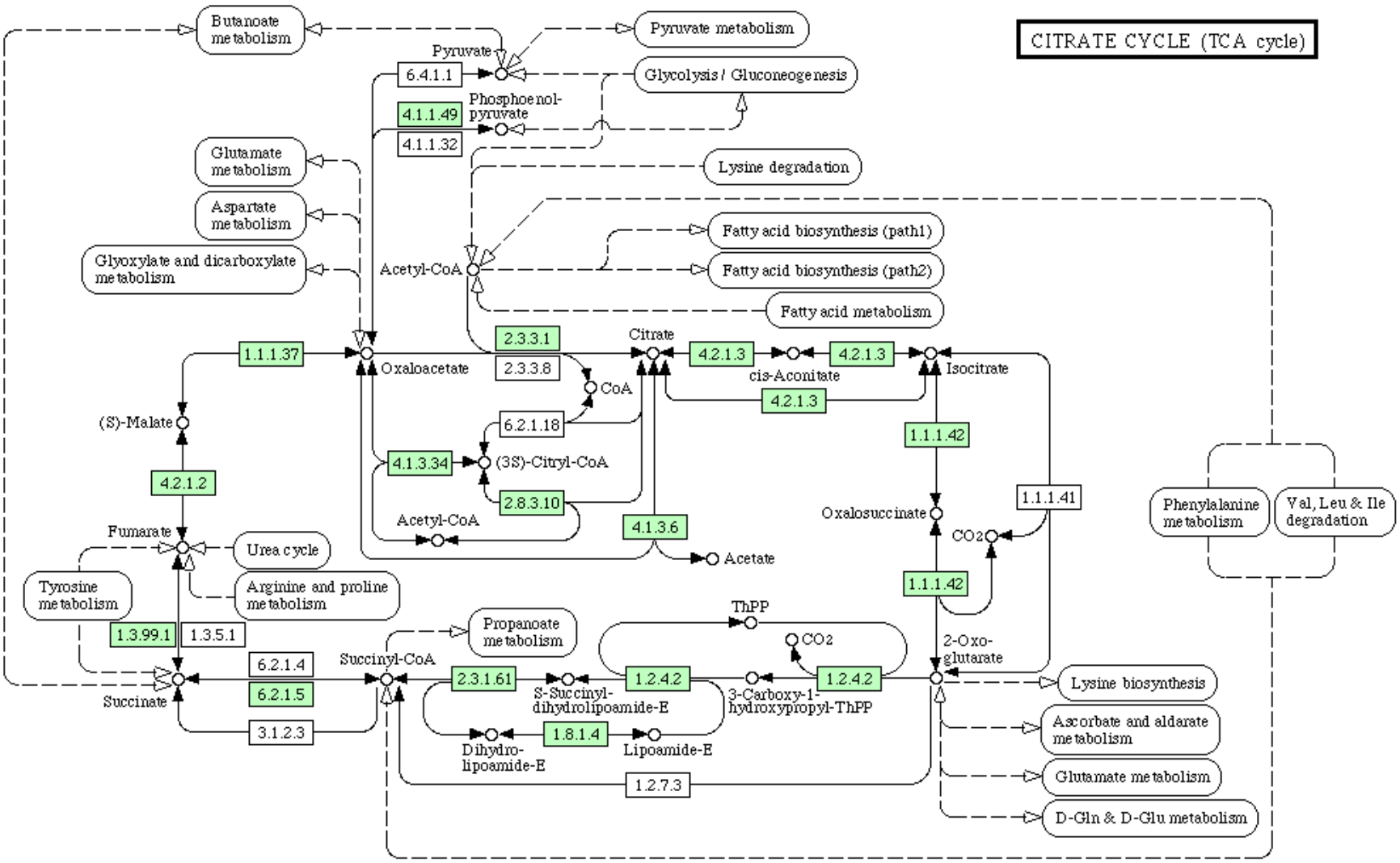
- **Metabolites** are intermediates and products of metabolic processes – everything that biochemistry can create
- Technically speaking also DNA, RNA and proteins could be considered metabolites
- The term is usually restricted to small molecules
- Spans a variety of substance classes (not complete):
  - Amino acids
  - Lipids
  - Sugars
  - …
- Chemically much more diverse than proteome!

Metabolomics – The Big Picture



Nicholson and Lindon. Nature 2008, 455, 1054-1056

Metabolic Networks



http://www.genome.jp/dbget-bin/www_bget?pathway+ecj00020

Technologies

Modern proteomics and metabolomics studies primarily driven by

Chromatography coupled to mass spectrometry (MS)

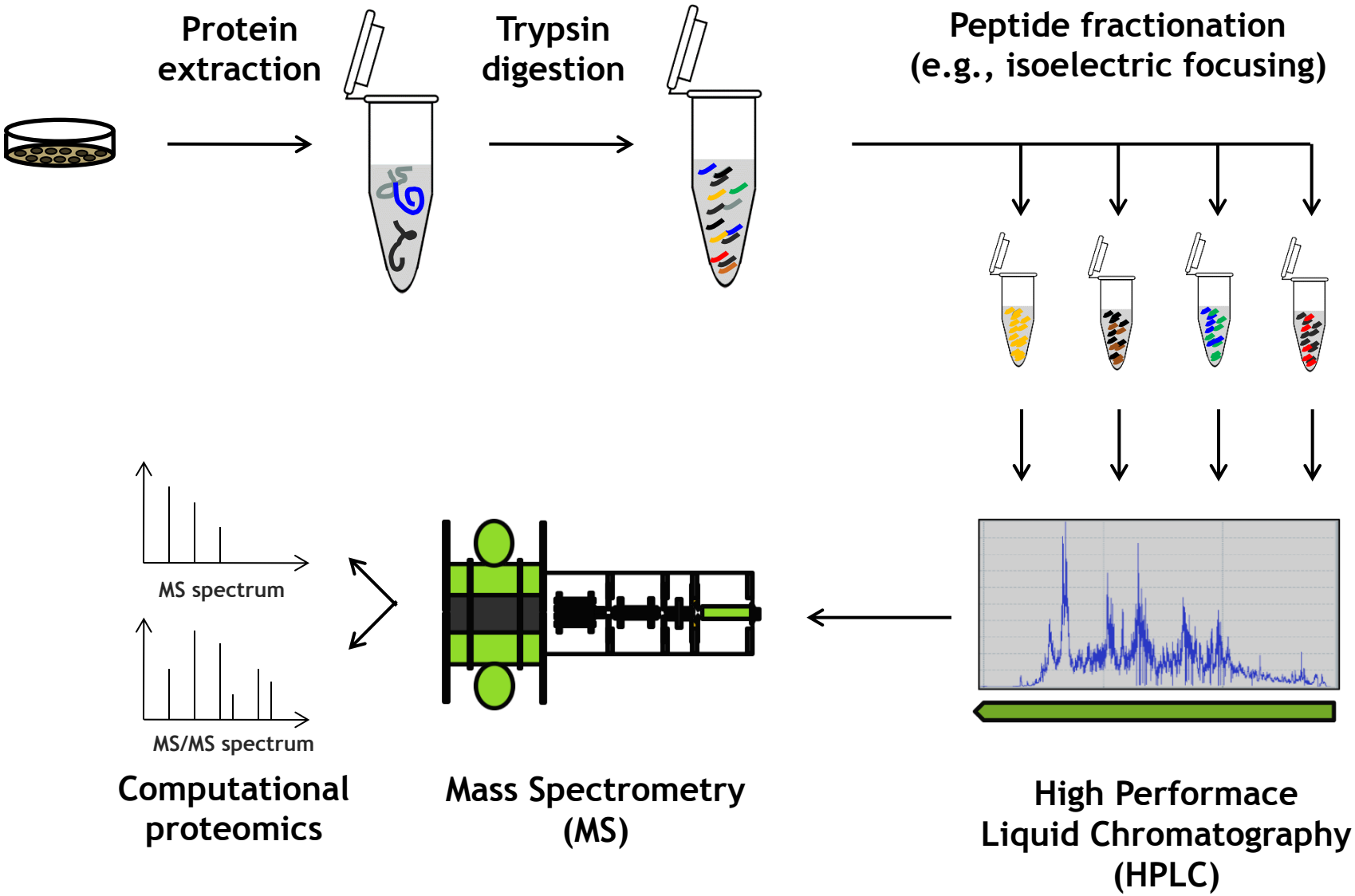
http://en.wikipedia.org/wiki/High-performance_liquid_chromatography.
Access 14/10/2013, 5 PM

www.planetorbitrap.com
Access 14/10/2013, 5 PM

Technologies

- **Chromatography (GC/LC)**
  - Chromatography separates proteins/peptides or metabolites
  - Reduces complexity of samples
- **Mass spectrometry (MS)**
  - Identifies the biomolecules (mass spectrum often used similar to a ‘fingerprint’ of the molecule)
  - Signal intensity is proportional to concentration of the molecule in the sample

Shotgun proteomics



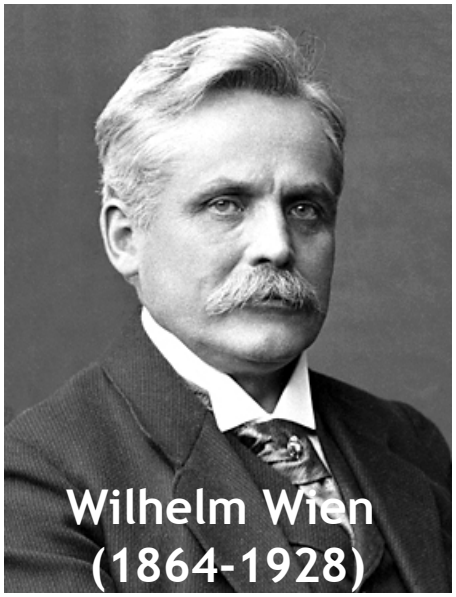
Mass Spectrometry

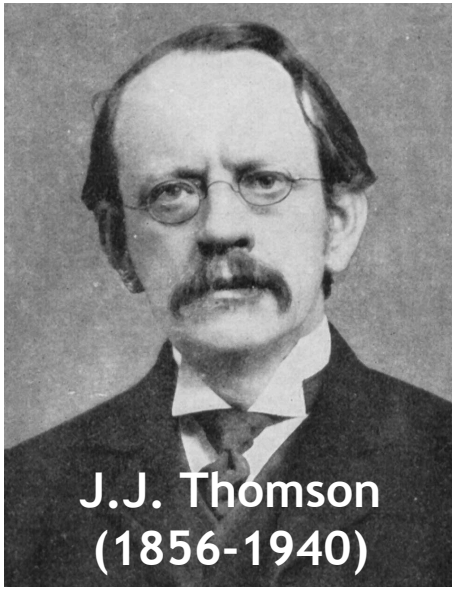
- **Definition:** **Mass spectrometry** is an analytical technique identifying type and amount of analytes present in a sample by measuring abundance and mass-to-charge ratio of analyte ions in the gas phase.

- 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

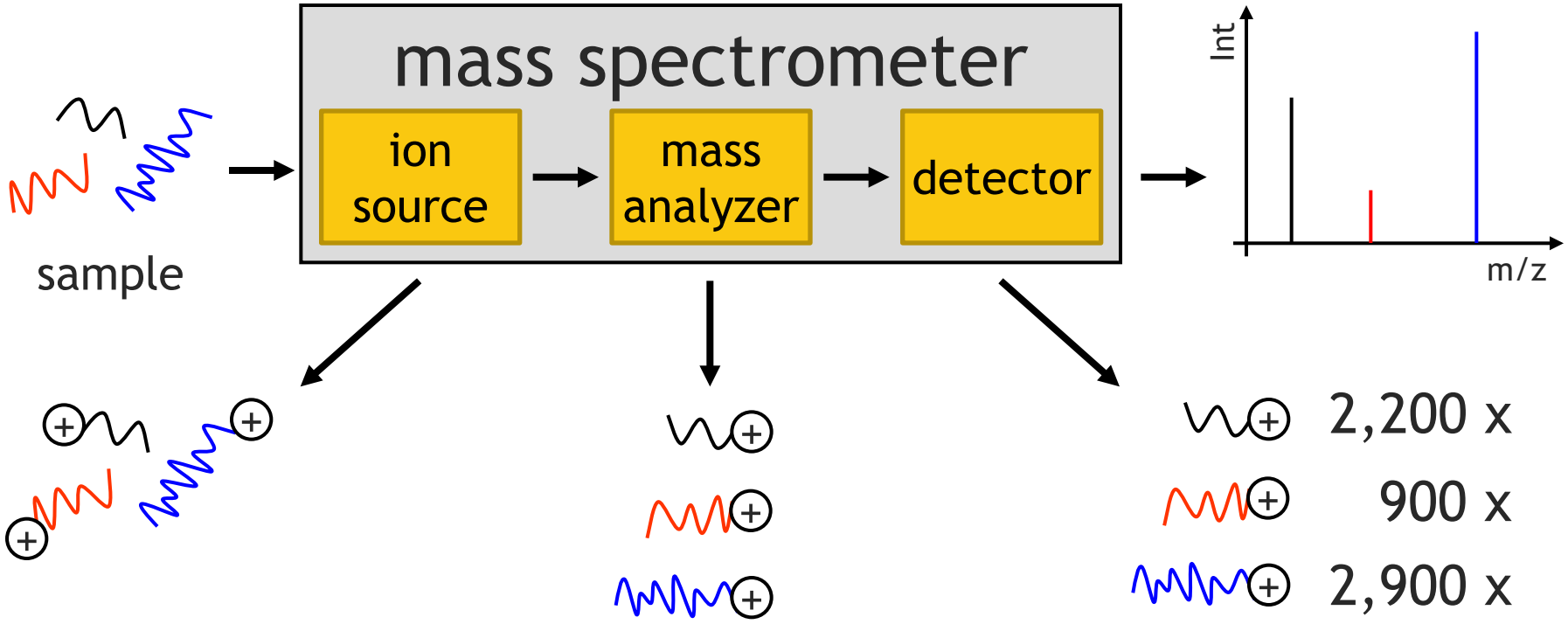
- **Wilhelm Wien** was the first to separate charged particles with magnetic and electrostatic fields in 1899
- **Sir Joseph J. Thomson** improved on these designs
- Sector mass spectrometers were used for separating uranium isotopes for the Manhattan project
- In the 1950s and 1960s **Hans Dehmelt**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


Wilhelm Wien
(1864-1928)


J.J. Thomson
(1856-1940)

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

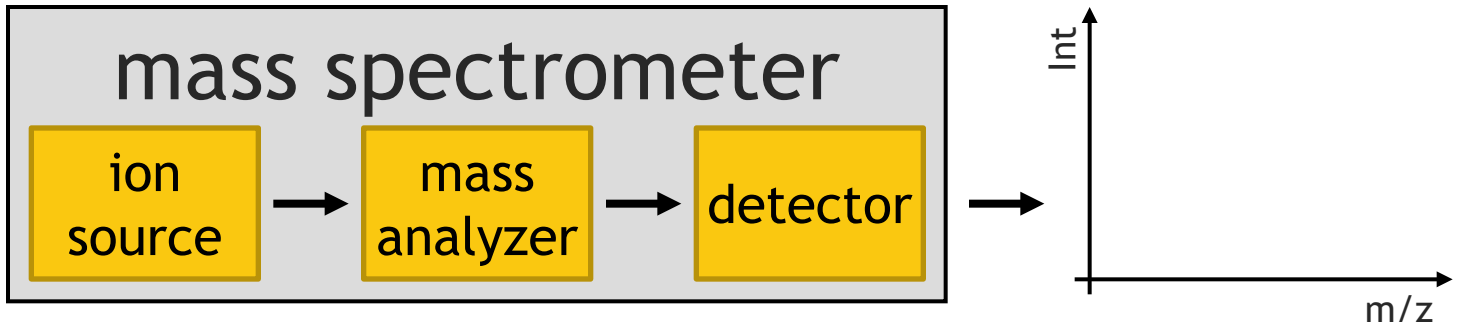


A mass spectrometer has three key components

- **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 onto 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:

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

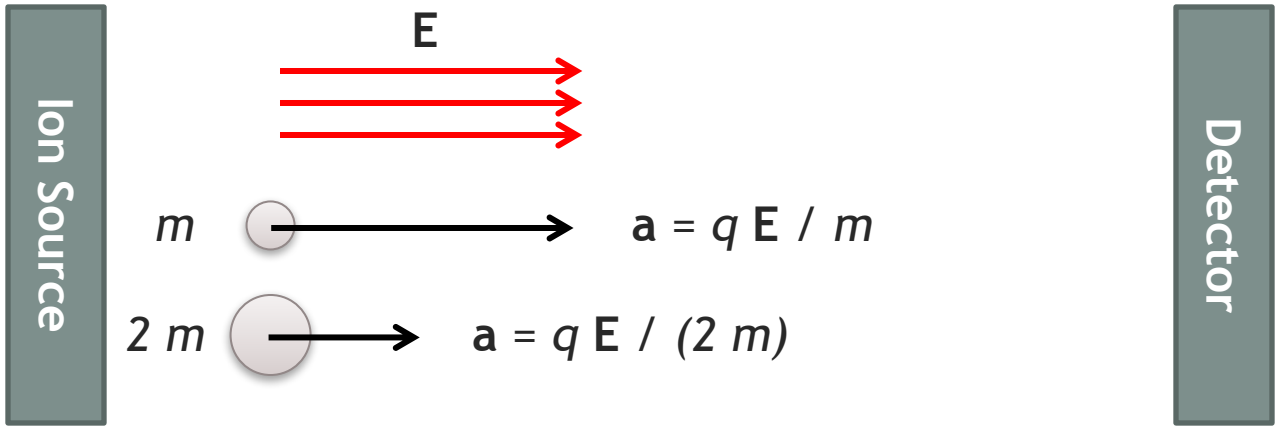
- Acceleration of the ions is then used to determine m/q

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 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 generally given in units of kg (SI unit), however, there are different 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
- 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	Atomic Weight	Nat. abun-dance [%]	Isotope	Atomic Weight	Nat. abun-dance [%]
^{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 light 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 ($C_2H_5NO_2$):

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

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

Monoisotopic mass of valine: 117.078979

Absolute mass error: -0.0178979

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

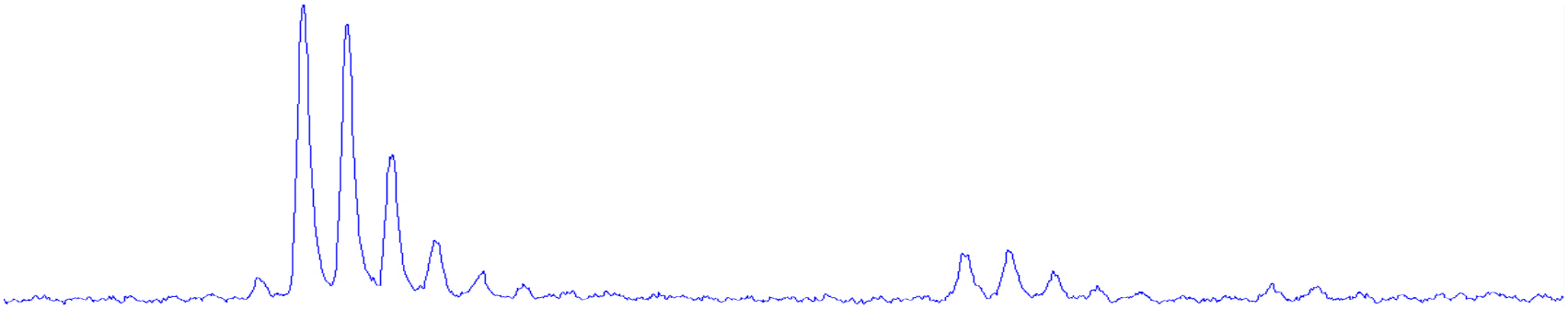
Isotope Patterns

- 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

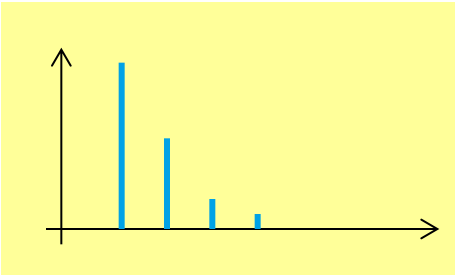
- Molecule with 10 carbon atoms
  - Lightest variant contains only ^{12}C
  - This is called ‘monoisotopic’
  - Others contain 1-10 ^{13}C atoms, these are heavier by 1-10 Da than the monoisotopic one
- In general, the relative intensities follow a binomial distribution, depending on the number of atoms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

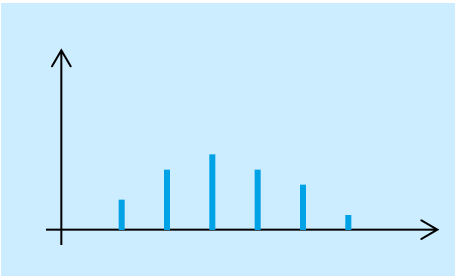


Isotope Patterns

- 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



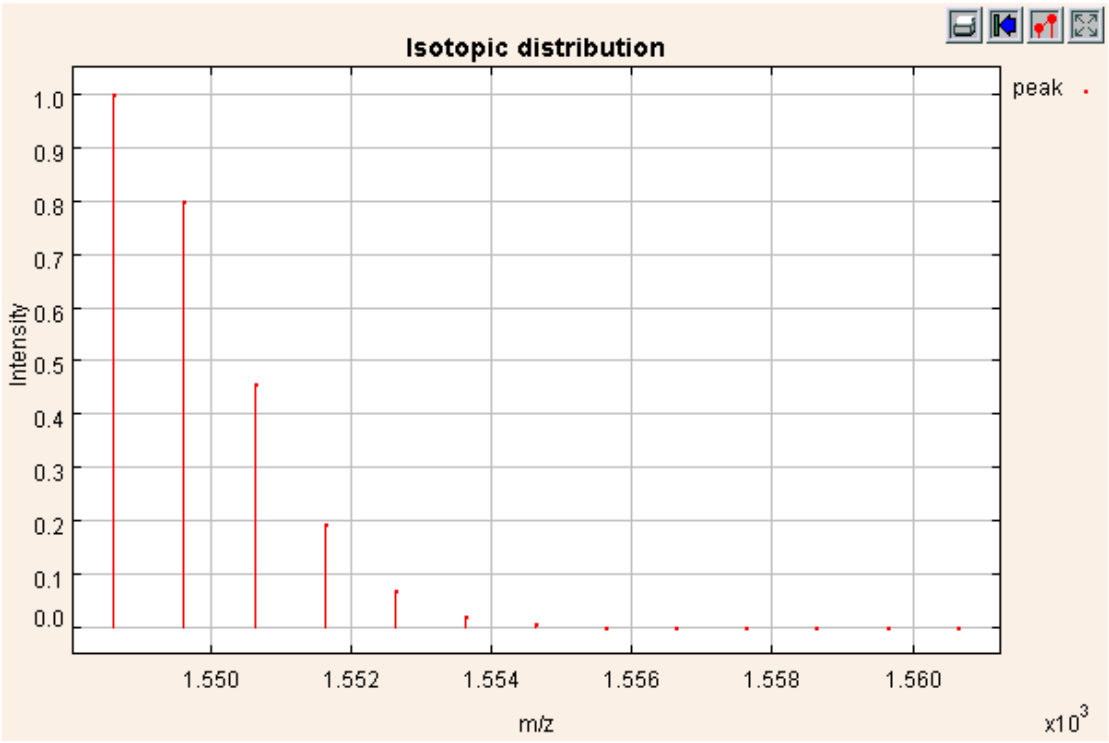


Online Calculator

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:
$^{12}C_{63}$
$^{1}H_{100}$
$^{14}N_{14}$
$^{16}O_{27}$
$^{32}S_2$

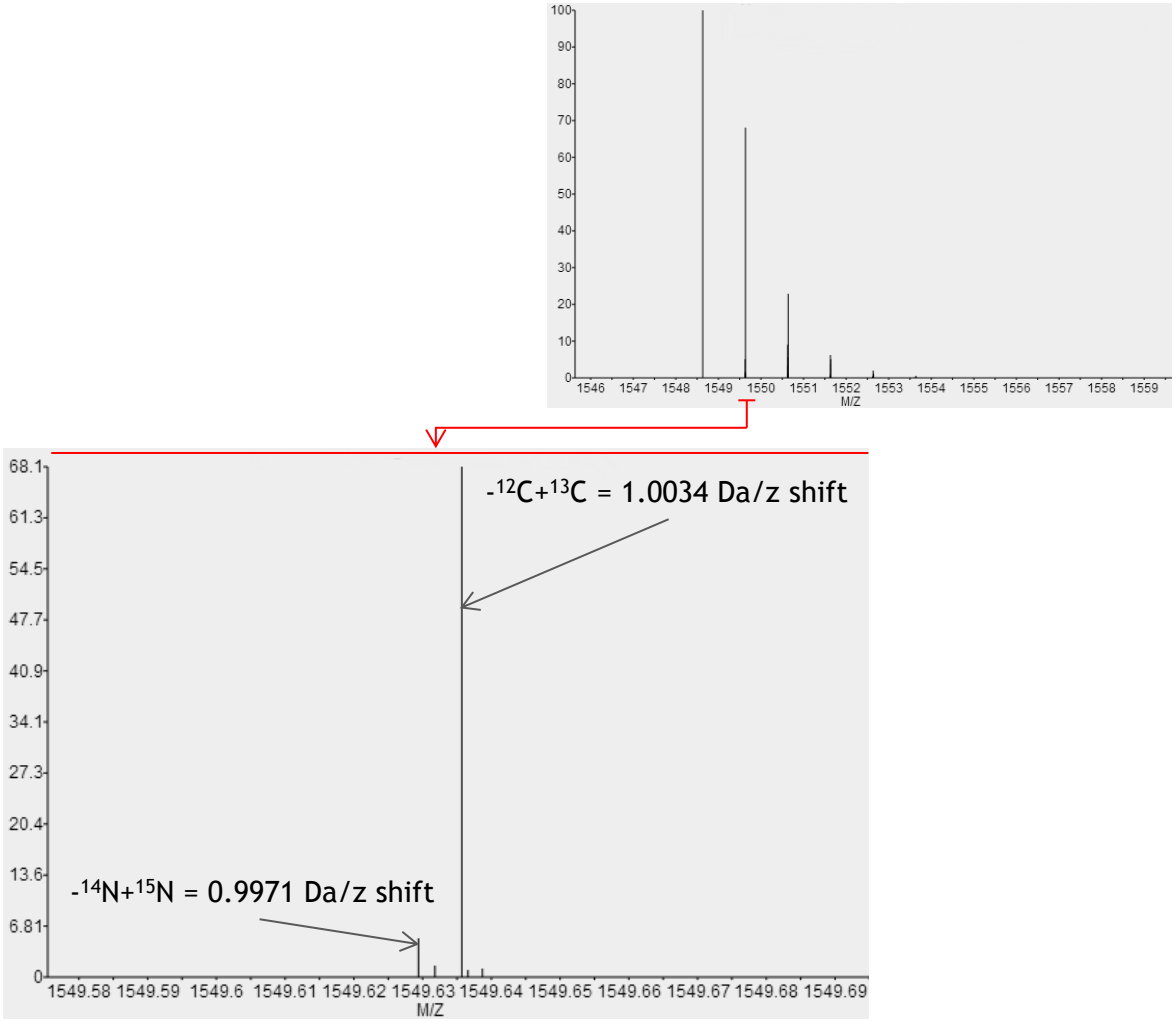


Most likely isotope combination:
$^{12}C_{63}$
$^{1}H_{100}$
$^{14}N_{14}$
$^{16}O_{27}$
$^{32}S_2$
Exact mass is 1548.632
Probability of combination is 39.255%
The most likely combination is 100.00% of those masses rounding to 1548 amu.

Download the data

Isotopic Fine Structure

- High-resolution MS reveals isotopic fine structure



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

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

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

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

OPENMS AND KNIME

- Workflows - definition
- Conceptual ideas behind OpenMS and TOPP
- Installation of KNIME and OpenMS extensions
- Overview of KNIME
- Simple workflows in KNIME
  - Loading tabular data, manipulating rows, columns
  - Visualization of data
  - Preparing simple reports
  - Embedding R scripts
  - Simple OpenMS ID workflow: finding all proteins in a sample

Growth of Omics Data (EBI Repositories)

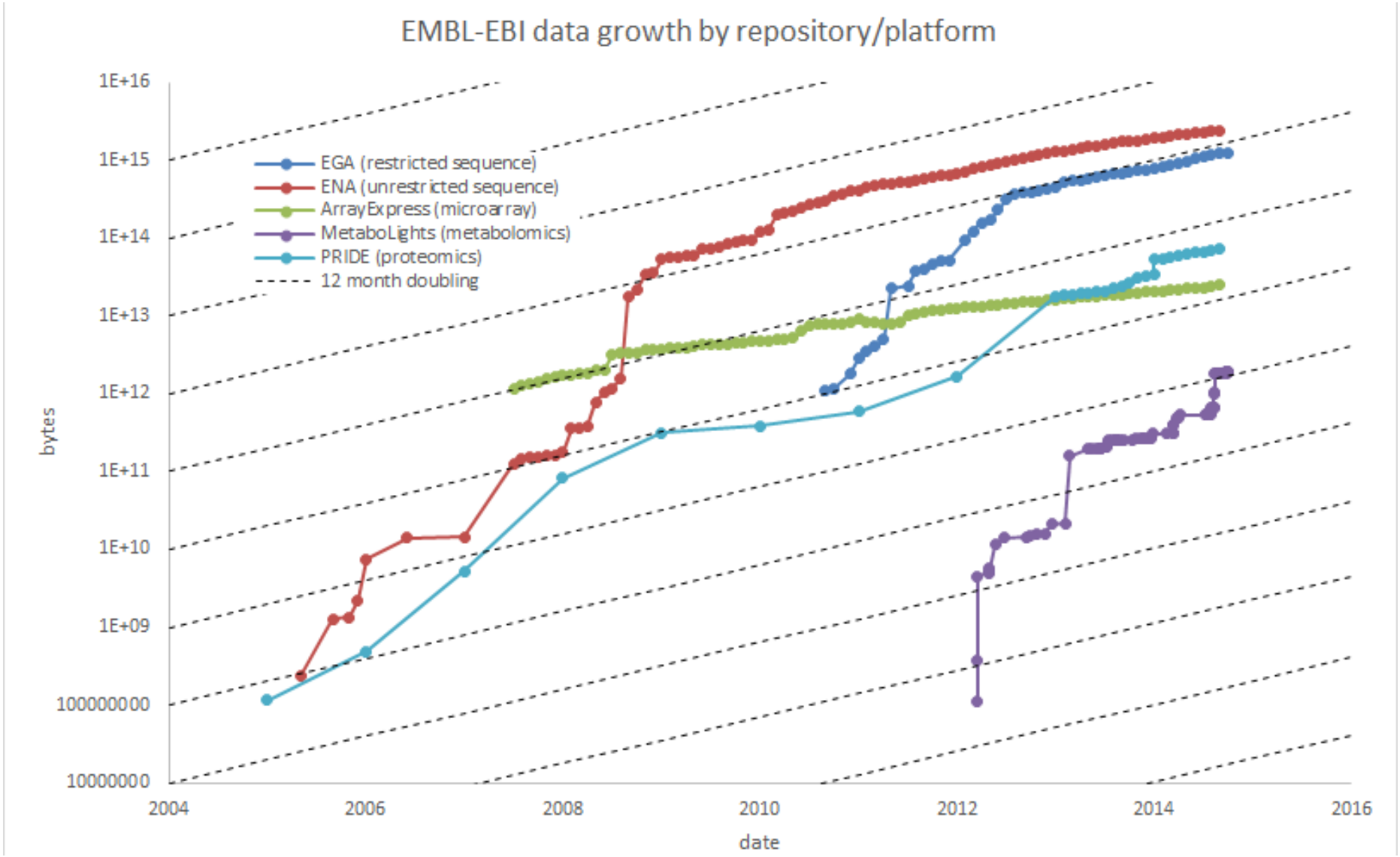
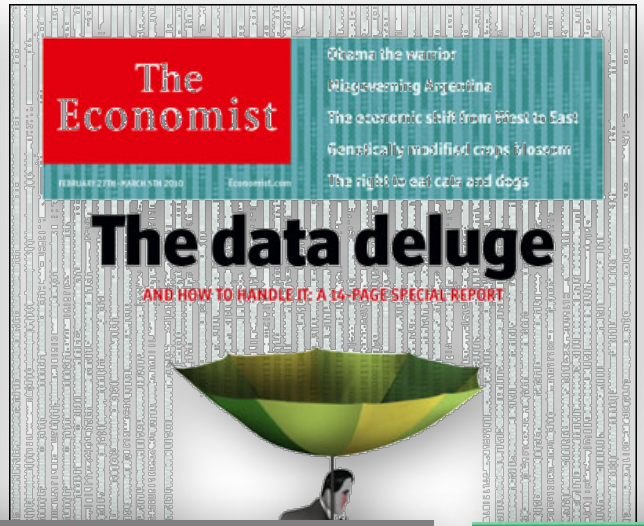


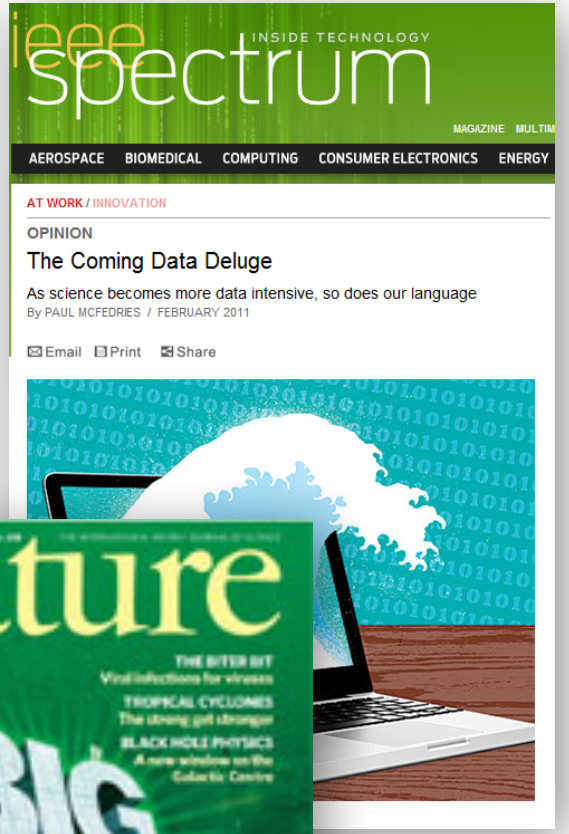
Illustration: Christoph Steinbeck, EBI

Multi-Omics/Polyomics

- Systems biology requires an integrative view spanning more than one omics level – this is called ‘multi-omics’ or ‘polyomics’
- Data sets are
  - Huge (often hundreds of GB)
  - Heterogeneous
  - Complex in their structure
- Integrative analysis is complex (usually takes longer than data generation)
- Complex analysis workflows are hard to reproduce

Big Data and Reproducible Science





Big Data and Reproducible Science

Error prone

Biologists must realize the pitfalls of work on massive amounts of data.

Genomics has the potential to revolutionize medical care, but it is becoming increasingly clear that the field is having to deal with growing pains.

In a Comment piece this week, Daniel MacArthur, a researcher at Massachusetts General Hospital in Boston, argues that the massive pools of data generated in even routine genome studies make it easy to misinterpret artefacts as biologically important results (see page 427). Such false positives, he says, can lead to embarrassing retractions, futile projects and stalled careers. More careful attention to methods and greater awareness of the potential pitfalls will help to cut down on the needless mistakes.

In a field as competitive as genomics, scientists will inevitably seek faster, more efficient ways to generate and analyse data. Just this week, the firm Ion Torrent in Guilford, Connecticut — part of Life Technologies in Carlsbad, California — announced that it will tackle a competition to accurately sequence 100 genomes in 30 days for less than US$1,000 per genome — and to win the US$10-million prize offered by the X Prize Foundation in Playa Vista, California (see page 417).

Genomics is not the only field of science to battle with quality-control issues. In March, *Nature* lamented the high number of corrections to research papers in the life sciences that arise from avoidable errors (see *Nature* **483**, 509; 2012). Scientists are making too many careless mistakes, and those mistakes are getting published.

Much of this sloppy science comes from the pressure to generate 'surprising' results and to publish them quickly, even though they are more likely to be driven by errors than are findings that more or less follow from previous work. A researcher who reveals something exciting is more likely to get a high-profile paper (and a permanent position) than is someone who spends years providing solid evidence for something that everyone in the field expected to be true.

This pressure extends throughout the careers of scientists, and is compounded by the preference of journals (including *Nature*) to publish significant findings — and of the media to report them. MacArthur asks scientists to weigh up the importance of avoiding being scooped against the embarrassment of a mistake, but to an ambitious scientist in a competitive field such as genomics, the risk of being out-published will often outweigh the potential damage of retraction.

Many areas of the life sciences now work with massive amounts of data, so technology-based artefacts are unlikely to be restricted to genomics. Any life scientist who works at a university or is affiliated with a hospital can now collect human samples and sequence them to create huge amounts of genomic data, with which they are perhaps not used to working. The problem goes beyond analysis — time and time again, biologists fail to design experiments properly, and so submit underpowered studies that have an insufficient sample size and trumpet chance observations as biological effects.

The problems are not hard to solve. Biologists must seek relevant training in experimental methods and collaborate with good statisticians. Principal investigators have a responsibility to their labs and to colleagues to ensure that any data they publish are robust. And the efforts of peer reviewers who thoroughly reanalyse data to double-check that submissions are solid deserve more formal acknowledgement, albeit in private.

Meanwhile, researchers who deal with large amounts of data must agree on standards that will protect against avoidable errors. Fields such as RNA sequencing have been slow to establish such guidelines (see *Nature* **484**, 428; 2012), but others have shown that it can be done. The human-genetics community, for instance, has established criteria for genome-wide association studies to ensure that findings are rigorous and comparable. Less-proactive genomics fields, and the rest of biology, should follow that lead. ■

NATURE.COM
To comment online, click on Editorials at: go.nature.com/xbuqqv

High-Throughput Proteomics

- Analyzing one sample is usually not a big deal
- Analyzing 20 can be tiresome
- Analyzing 100 is a really big deal

- ***High-throughput experiments require high-throughput analysis***

- ***Compute power scales much better than manpower***

Pipelines and Workflows

pipeline |ˈpīpˌlīn| *noun*

1. a long pipe, typically underground, for conveying oil, gas, etc., over long distances. […]

2. *Computing* a linear sequence of specialized modules used for pipelining.

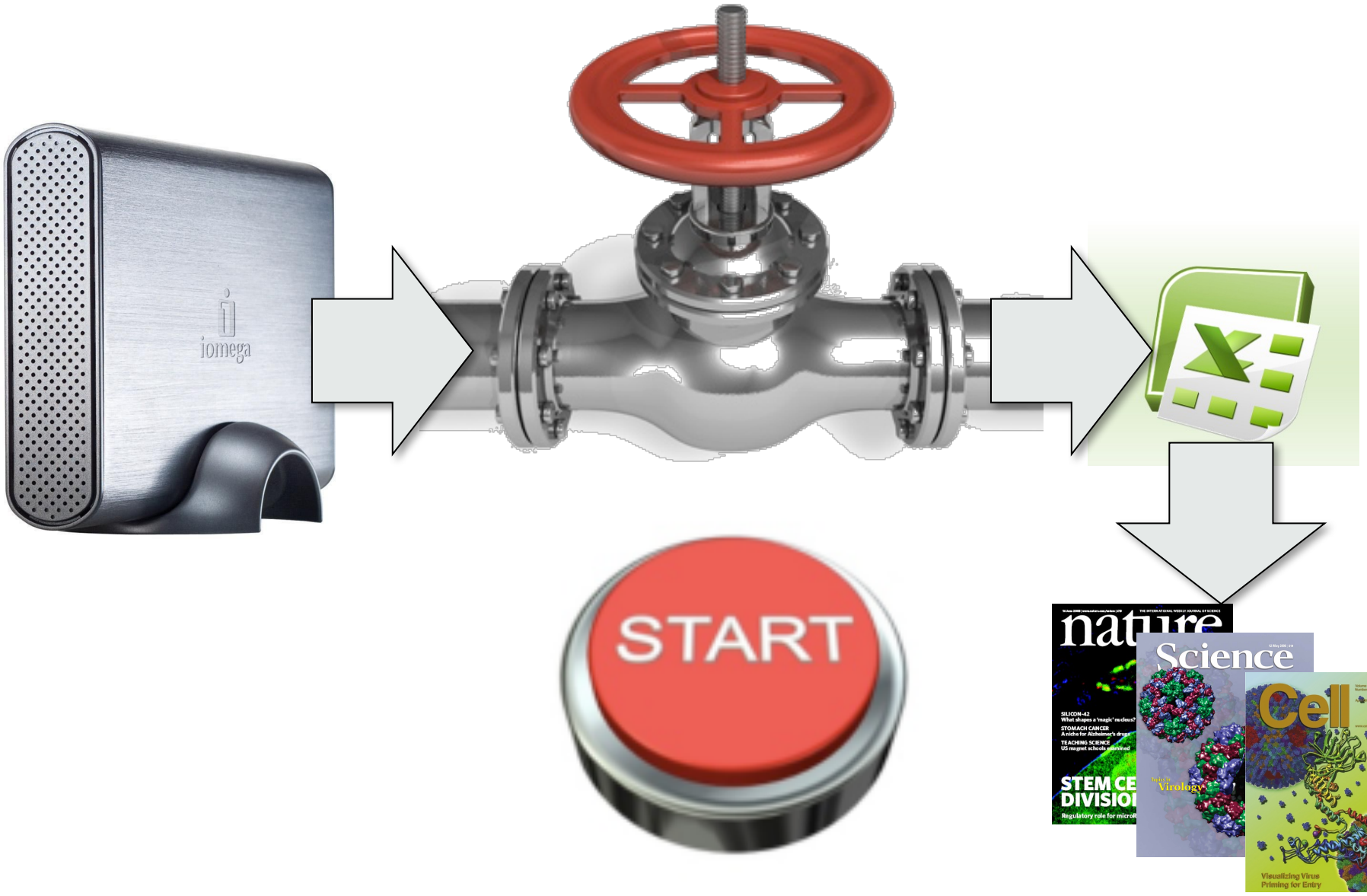
3. *(in surfing)* the hollow formed by the breaking of a large wave.

workflow |ˈwərkˌflō| *noun*

- the sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion.

http://oxforddictionaries.com/definition/american_english/pipeline
http://oxforddictionaries.com/definition/american_english/workflow

Bioinformatics – The Holy Grail



KNIME and OpenMS

- Constructing workflows requires
  - **Tools** – making up the nodes of the workflows
  - A **workflow engine** – executing the nodes in a predefined order
- In the context of this course, we will use **OpenMS** tools to analyze mass spectrometric data
- We will design the workflow engine and data mining tool **KNIME** to construct and execute these workflows in a convenient manner
- We will briefly intro both tools – they are open-source software and freely available on all major platforms

OpenMS/TOPP

- **OpenMS** – an open-source C++ framework for computational mass spectrometry
- Jointly developed at ETH Zürich, FU Berlin, University of Tübingen
- **Open source**: BSD 3-clause license
- **Portable**: available on Windows, OSX, Linux
- **Vendor-independent**: supports all standard formats and vendor-formats through proteowizard
- **TOPP – The OpenMS Proteomics Pipeline**
  - Building blocks: One application for each analysis step
  - All applications share **identical user interfaces**
  - Uses PSI **standard formats** and integrates seamlessly with other applications supporting these formats
- **TOPP tools** can be integrated in various **workflow systems**
  - TOPPAS – TOPP Pipeline Assistant
  - Galaxy
  - WS-PGRADE/gUSE
  - Proteome Discoverer/Compound Discoverer
  - **KNIME**

Kohlbacher et al., Bioinformatics (2007), 23:e191

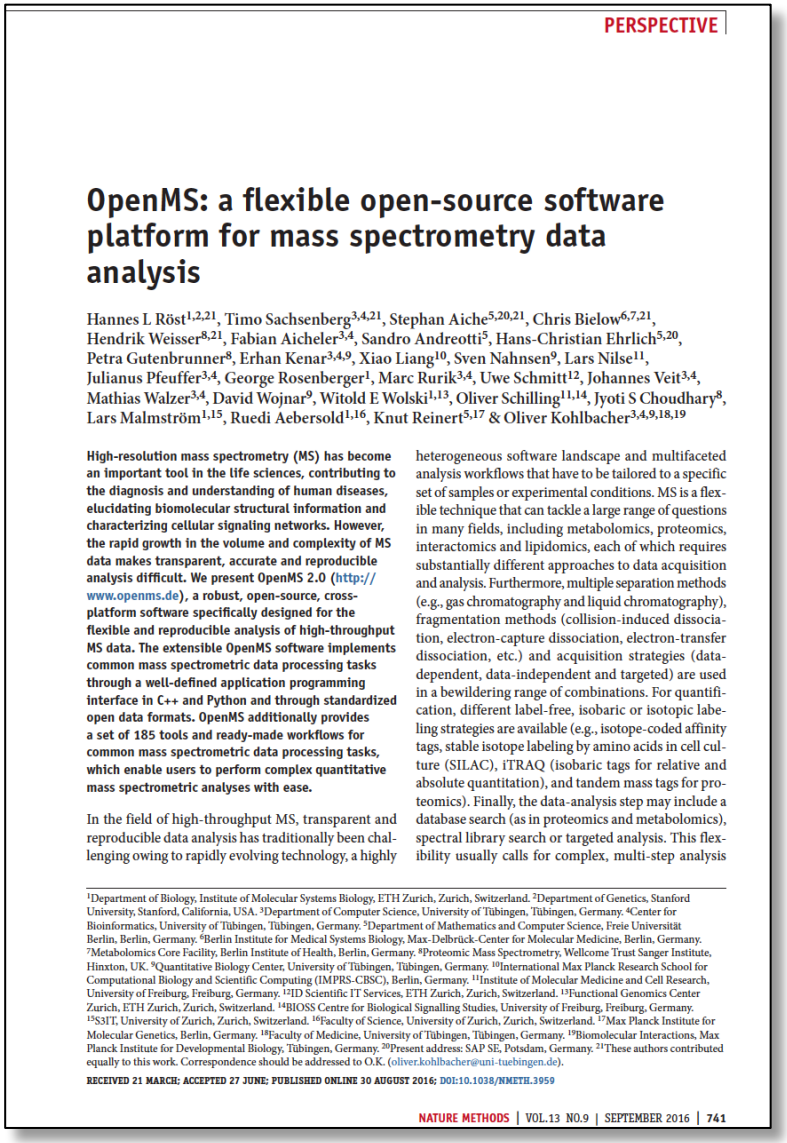
TOPP – Concepts

- **TOPP – The OpenMS Proteomics Pipeline**
- No programming skills required
- **Graphical User Interface**: TOPPView and TOPPAS
- Building blocks: One application for each analysis step
- All applications share **identical user interfaces**
- Uses PSI **standard formats** and integrates seamlessly with other applications supporting these formats

Kohlbacher et al., Bioinformatics (2007), 23:e191

OpenMS 2.0 - Features

- **Currently 185 distinct tools**
- **Utilities** – extract information from files, file conversion, visualization
- **PTX identification** – interface to DB search engines, de novo search, RNA-protein XL MS, protein inference, RT prediction, proteotypicity prediction
- **PTX quantification – label-free**, TMT, iTRAQ, SILAC, MRM, OpenSWATH (DIA), ProteinSIP (metaproteomics), RT alignment
- **MTX quantification – nontargeted metabolomics**, MRM
- **MTX identification** – accurate mass DB search, spectral matching, composition
- **Miscellaneous** – MRM scheduling, LC-MS simulator, …



PERSPECTIVE

OpenMS: a flexible open-source software platform for mass spectrometry data analysis

Hannes L Röst[1,2,21], Timo Sachsenberg[3,4,21], Stephan Aiche[5,20,21], Chris Bielow[6,7,21], Hendrik Weisser[8,21], Fabian Aicheler[3,4], Sandro Andreotti[5], Hans-Christian Ehrlich[5,20], Petra Gutenbrunner[8], Erhan Kenar[3,4,9], Xiao Liang[10], Sven Nahnsen[9], Lars Nilse[11], Julianus Pfeuffer[3,4], George Rosenberger[1], Marc Rurik[3,4], Uwe Schmitt[12], Johannes Veit[3,4], Mathias Walzer[3,4], David Wojnar[9], Witold E Wolski[1,13], Oliver Schilling[11,14], Jyoti S Choudhary[8], Lars Malmström[1,15], Ruedi Aebersold[1,16], Knut Reinert[5,17] & Oliver Kohlbacher[3,4,9,18,19]

High-resolution mass spectrometry (MS) has become an important tool in the life sciences, contributing to the diagnosis and understanding of human diseases, elucidating biomolecular structural information and characterizing cellular signaling networks. However, the rapid growth in the volume and complexity of MS data makes transparent, accurate and reproducible analysis difficult. We present OpenMS 2.0 (http://www.openms.de), a robust, open-source, cross-platform software specifically designed for the flexible and reproducible analysis of high-throughput MS data. The extensible OpenMS software implements common mass spectrometric data processing tasks through a well-defined application programming interface in C++ and Python and through standardized open data formats. OpenMS additionally provides a set of 185 tools and ready-made workflows for common mass spectrometric data processing tasks, which enable users to perform complex quantitative mass spectrometric analyses with ease.

In the field of high-throughput MS, transparent and reproducible data analysis has traditionally been challenging owing to rapidly evolving technology, a highly heterogeneous software landscape and multifaceted analysis workflows that have to be tailored to a specific set of samples or experimental conditions. MS is a flexible technique that can tackle a large range of questions in many fields, including metabolomics, proteomics, interactomics and lipidomics, each of which requires substantially different approaches to data acquisition and analysis. Furthermore, multiple separation methods (e.g., gas chromatography and liquid chromatography), fragmentation methods (collision-induced dissociation, electron-capture dissociation, electron-transfer dissociation, etc.) and acquisition strategies (data-dependent, data-independent and targeted) are used in a bewildering range of combinations. For quantification, different label-free, isobaric or isotopic labeling strategies are available (e.g., isotope-coded affinity tags, stable isotope labeling by amino acids in cell culture (SILAC), iTRAQ (isobaric tags for relative and absolute quantitation), and tandem mass tags for proteomics). Finally, the data-analysis step may include a database search (as in proteomics and metabolomics), spectral library search or targeted analysis. This flexibility usually calls for complex, multi-step analysis

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [2]Department of Genetics, Stanford University, Stanford, California, USA. [3]Department of Computer Science, University of Tübingen, Tübingen, Germany. [4]Center for Bioinformatics, University of Tübingen, Tübingen, Germany. [5]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [6]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Berlin, Germany. [7]Metabolomics Core Facility, Berlin Institute of Health, Berlin, Germany. [8]Proteomic Mass Spectrometry, Wellcome Trust Sanger Institute, Hinxton, UK. [9]Quantitative Biology Center, University of Tübingen, Tübingen, Germany. [10]International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC), Berlin, Germany. [11]Institute of Molecular Medicine and Cell Research, University of Freiburg, Freiburg, Germany. [12]ID Scientific IT Services, ETH Zurich, Zurich, Switzerland. [13]Functional Genomics Center Zurich, ETH Zurich, Zurich, Switzerland. [14]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. [15]S3IT, University of Zurich, Zurich, Switzerland. [16]Faculty of Science, University of Zurich, Zurich, Switzerland. [17]Max Planck Institute for Molecular Genetics, Berlin, Germany. [18]Faculty of Medicine, University of Tübingen, Tübingen, Germany. [19]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen, Germany. [20]Present address: SAP SE, Potsdam, Germany. [21]These authors contributed equally to this work. Correspondence should be addressed to O.K. (oliver.kohlbacher@uni-tuebingen.de).

RECEIVED 21 MARCH; ACCEPTED 27 JUNE; PUBLISHED ONLINE 30 AUGUST 2016; DOI:10.1038/NMETH.3959

Röst et al., Nat. Methods, 2016, 13:741

TOPP Tools – Implementation

- Very easy to implement thanks to the OpenMS framework
- Usually short (200 lines of code on average, mostly concerned with parameter handling)
- Use of the OpenMS core library

IDMapper.C:

```
[...]
vector<ProteinIdentification> protein_ids;
vector<PeptideIdentification> peptide_ids;
String document_id;
IdXMLFile().load(getStringOption_
   ("id"), protein_ids,peptide_ids, document_id);
IDMapper mapper;
[...]
ConsensusXMLFile file;
ConsensusMap map;
file.load(in, map);
mapper.annotate(map, peptide_ids, protein_ids, false);
file.store(out, map);
```

Interoperability

- Pipeline components (tools) have to be compatible
- Data formats have to be compatible
- Alternatives
  - **Glue code** to convert parameters, adapt settings
  - **Converters** translating one data format into another
- Issues
  - Portability
  - Loss of information

PSI Standard Formats

Numerous open and standardized **XML formats** have been proposed by the **HUPO Proteomics Standards Initiative (HUPO PSI)**:

- **mzML** (successor of mzData) for storing mass spectrometry data
- **mzIdentML** for storing peptide/protein identifications
- **traML** for storing transition and inclusion lists (Deutsch et al., MCP, 2012)
- **mzQuantML** for storing quantitation results (Walzer et al., MCP, 2013)
- **mzTab** for summary information of quantitative and qualitative results, Excel-compatible TSV format (Griss et al., MCP, 2014)
- **qcML** for storing and mining quality control information (Walzer et al., MCP, 2014)

Advantages

- Open, documented, no closed-source libraries required
- Will still be readable in 10 years from now
- Interoperable with different software packages

Disadvantages

- Initial raw data conversion required (and often awkward)
- File size
- Poor support by instrument software

Documentation

- Documentation for each tool is available as part of the OpenMS documentation (www.OpenMS.org)

FeatureFinder

The feature detection application for quantitation.

pot. predecessor tools		pot. successor tools
PeakPicker	→ FeatureFinder →	FeatureLinker
MapAligner		SeedListGenerator

This module identifies "features" in a LC/MS map. By feature, we understand a peptide in a MS sample that reveals a characteristic isotope distribution. The algorithm computes positions in rt and m/z dimension and a charge estimate of each peptide.

The algorithm identifies pronounced regions of the data around so-called seeds. In the next step, we iteratively fit a model of the isotope profile and the retention time to these data points. Data points with a low probability under this model are removed from the feature region. The intensity of the feature is then given by the sum of the data points included in its regions.

How to find suitable parameters and details of the different algorithms implemented are described in the **TOPP tutorial**.

Note:

that the wavelet transform is very slow on high-resolution spectra (i.e. FT, Orbitrap). We recommend to use a noise or intensity filter to remove spurious points first and to speed-up the feature detection process.

Specialized tools are available for some experimental techniques: **SILACAnalyzer**, **ITRAQAnalyzer**.

The command line parameters of this tool are:

```
FeatureFinder -- Detects two-dimensional features in LC-MS data.
Version: 1.7.0 Sep  3 2010, 15:13:04, Revision: 7349

Usage:
  FeatureFinder <options>
```

Documentation

- Documentation for each tool is available as part of the OpenMS documentation (www.OpenMS.org)

```
Common TOPP options:
  -ini <file>        Use the given TOPP INI file
  -threads <n>       Sets the number of threads allowed to be used by the TOPP tool (default: "1")
  -write_ini <file>  Writes the default configuration file
  --help             Shows options
  --helphelp         Shows all options (including advanced)

The following configuration subsections are valid:
 - algorithm   Algorithm section

You can write an example INI file using the '-write_ini' option.
Documentation of subsection parameters can be found in the
doxygen documentation or the INIFileEditor.
Have a look at OpenMS/doc/index.html for more information.
```

For the parameters of the algorithm section see the algorithms documentation:
centroided
isotope_wavelet
mrm

In the following table you can find example values of the most important parameters for different instrument types.
These parameters are not valid for all instruments of that type, but can be used as a starting point for finding suitable parameters.

'centroided' algorithm:

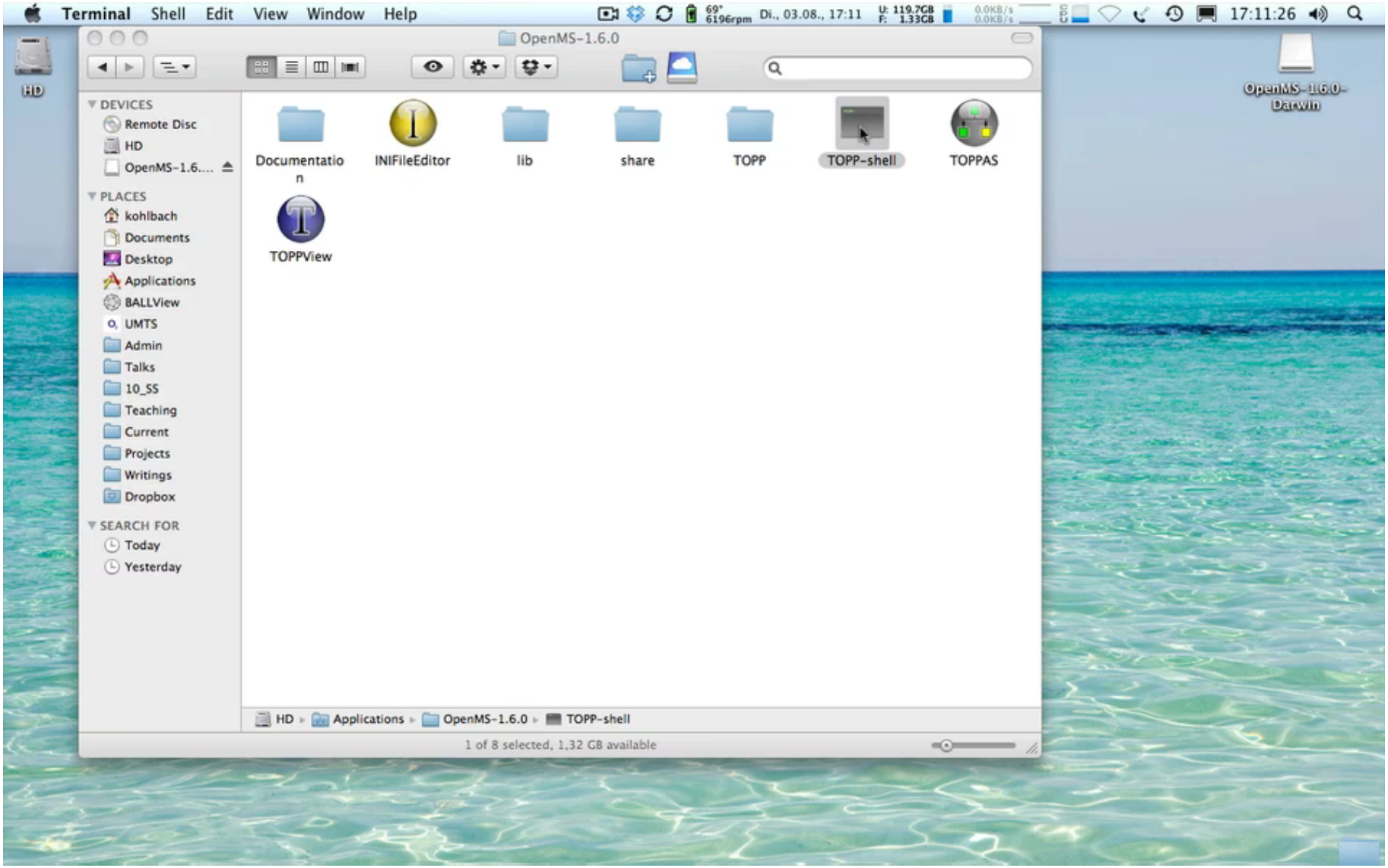
	Q-TOF	LTQ Orbitrap
intensity:bins	10	10
mass_trace:mz_tolerance	0.02	0.004
isotopic_pattern:mz_tolerance	0.04	0.005

For the *centroided* algorithm centroided data is needed. In order to create centroided data from profile data use the **PeakPicker**.

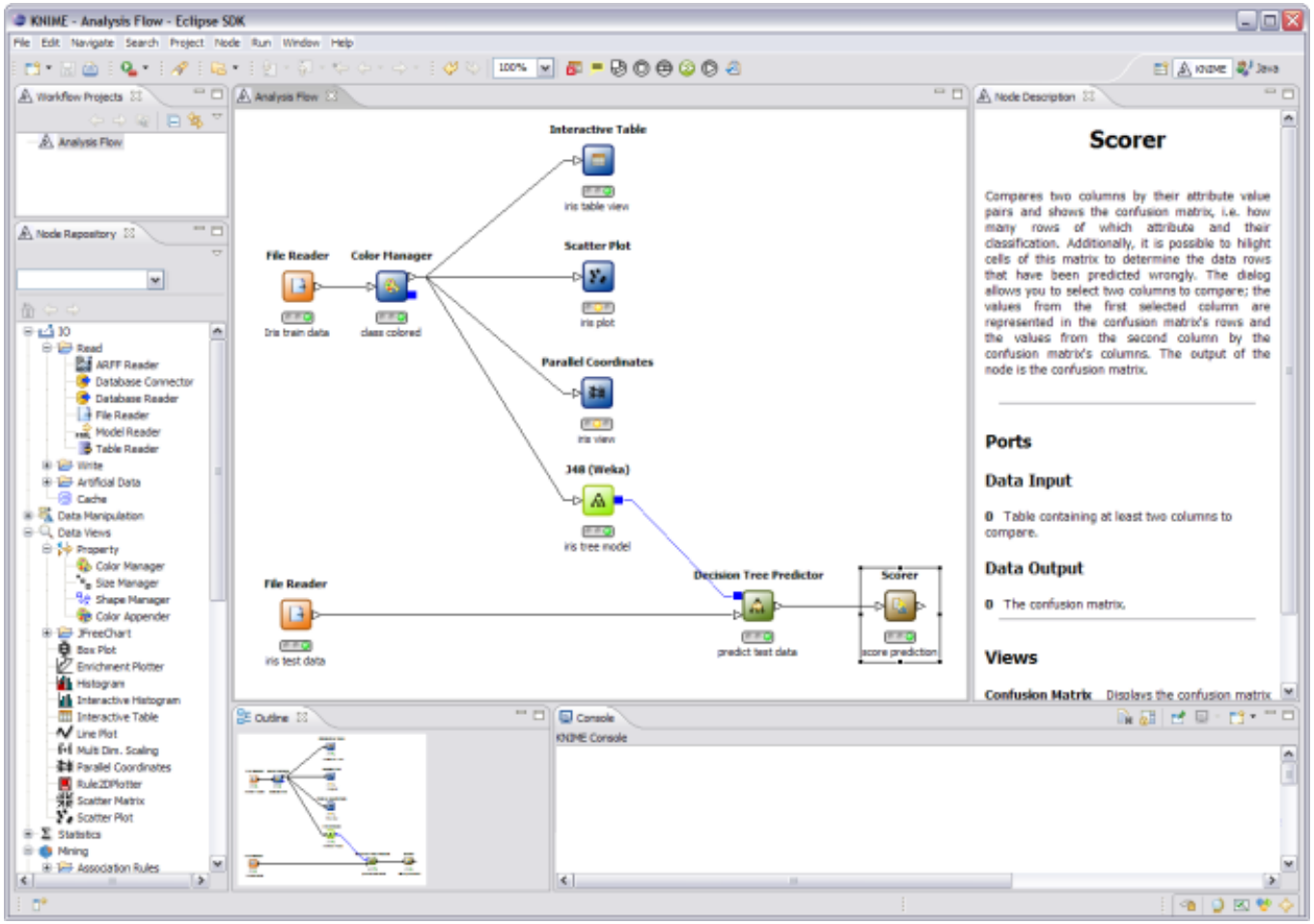
Installation of OpenMS

- Standalone version for command line and cluster environments
- Pre-built installers for Windows and Mac OS X
- Installer and installation instructions: http://open-ms.sourceforge.net/downloads/
- Bleeding edge development versions: http://ftp.mi.fu-berlin.de/OpenMS/nightly_binaries/
- Linux? Build your own OpenMS from git: https://github.com/OpenMS/OpenMS

Use on the Command Line



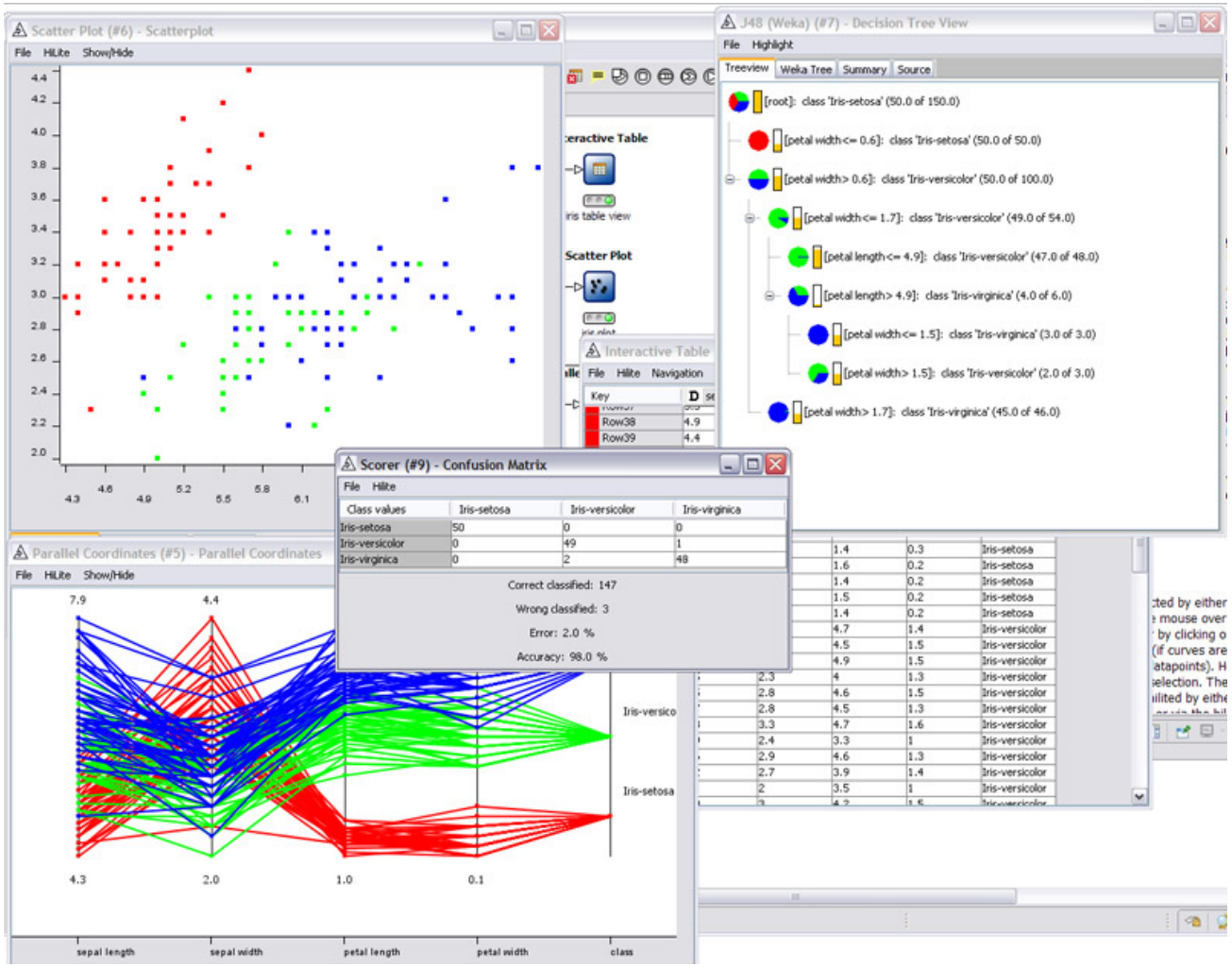
KNIME – KoNstanz Information MinEr



- Industrial-strength general-purpose workflow system
- Convenient and easy-to-use graphical user interface
- Available for Windows, OSX, Linux at http://KNIME.org

http://knime.org

KNIME – KoNstanz Information MinEr

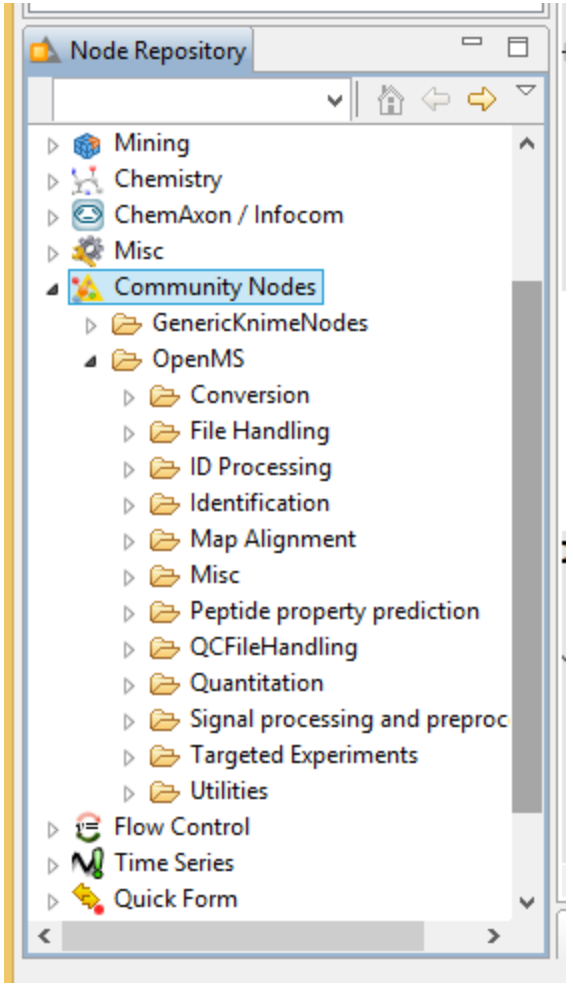


- Visualization capabilities
- Data mining & advanced statistical methods

http://knime.org

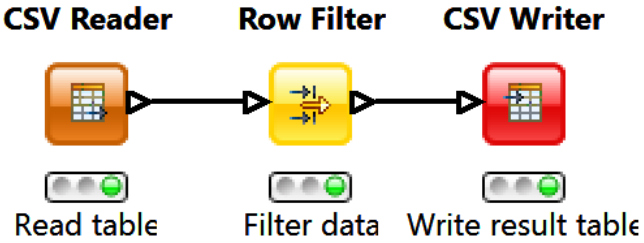
Installation of KNIME and OpenMS

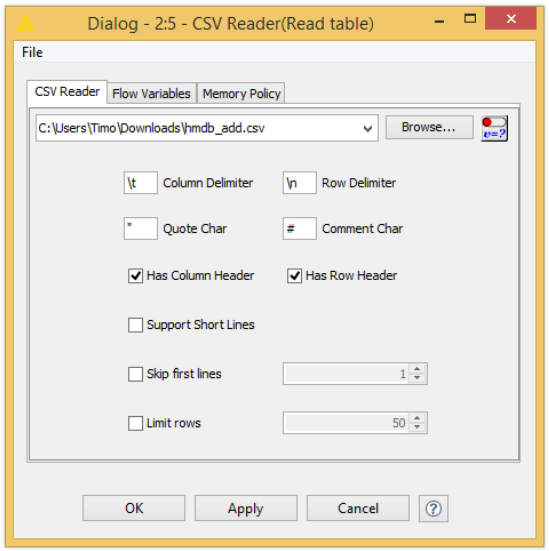
- KNIME installers available from: www.knime.org
- KNIME provides a sophisticated plugin system:
  - Many additional nodes can be installed as KNIME extensions
  - OpenMS installation in KNIME provides all TOPP tools as separate nodes
  - Nodes can be found in the folder ‘Community Nodes’
  - Detailed instructions on how to install OpenMS nodes in the additional materials

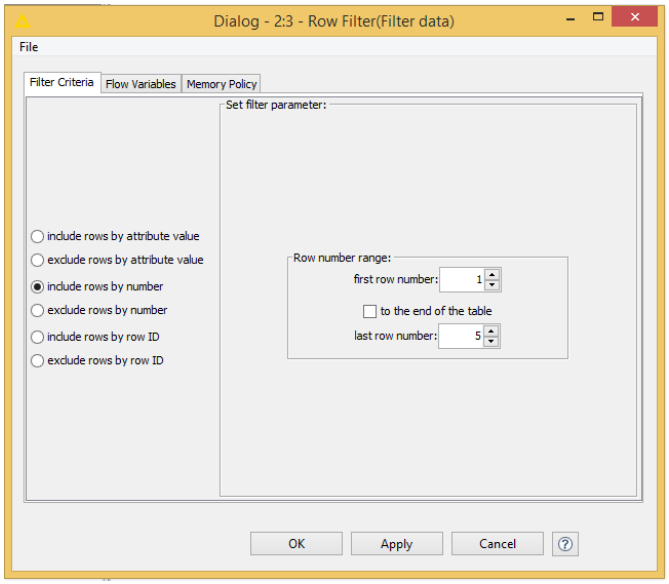


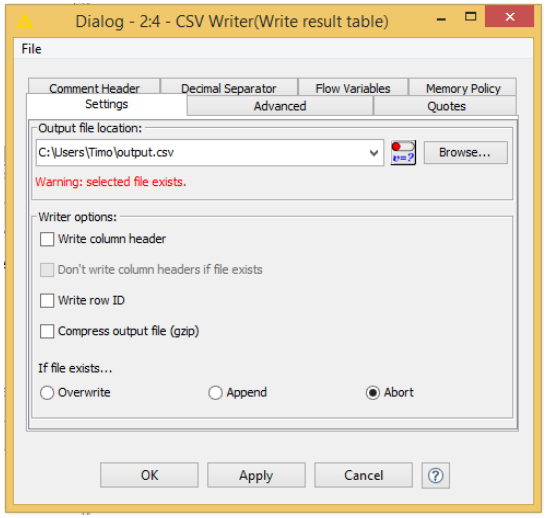
Simple Workflows in KNIME

- KNIME workflows consist of distinct nodes that can be assembled into workflows
- Workflow construction via drag and drop
- Either **tables** or **files** are exchanged between nodes along the edges of the workflow
- Files are marked by square ports, tables by triangular ports
- **Configuration dialogs** exist for all nodes

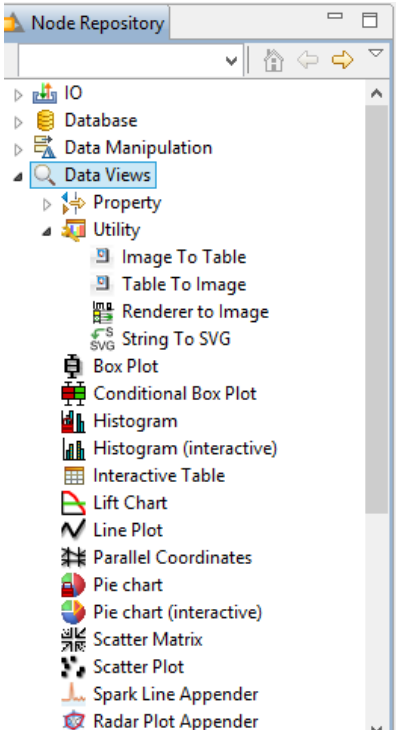








Simple Workflows in KNIME

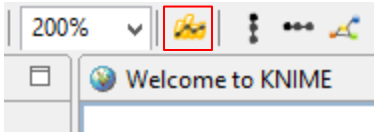


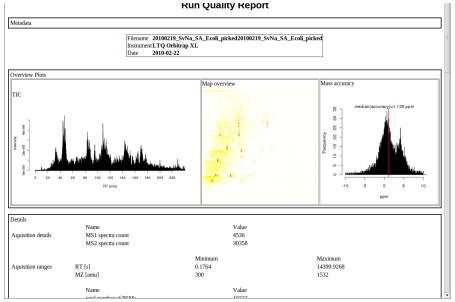
Plotting

- Data View nodes offer interactive visualization of tables
- Data can be explored interactively

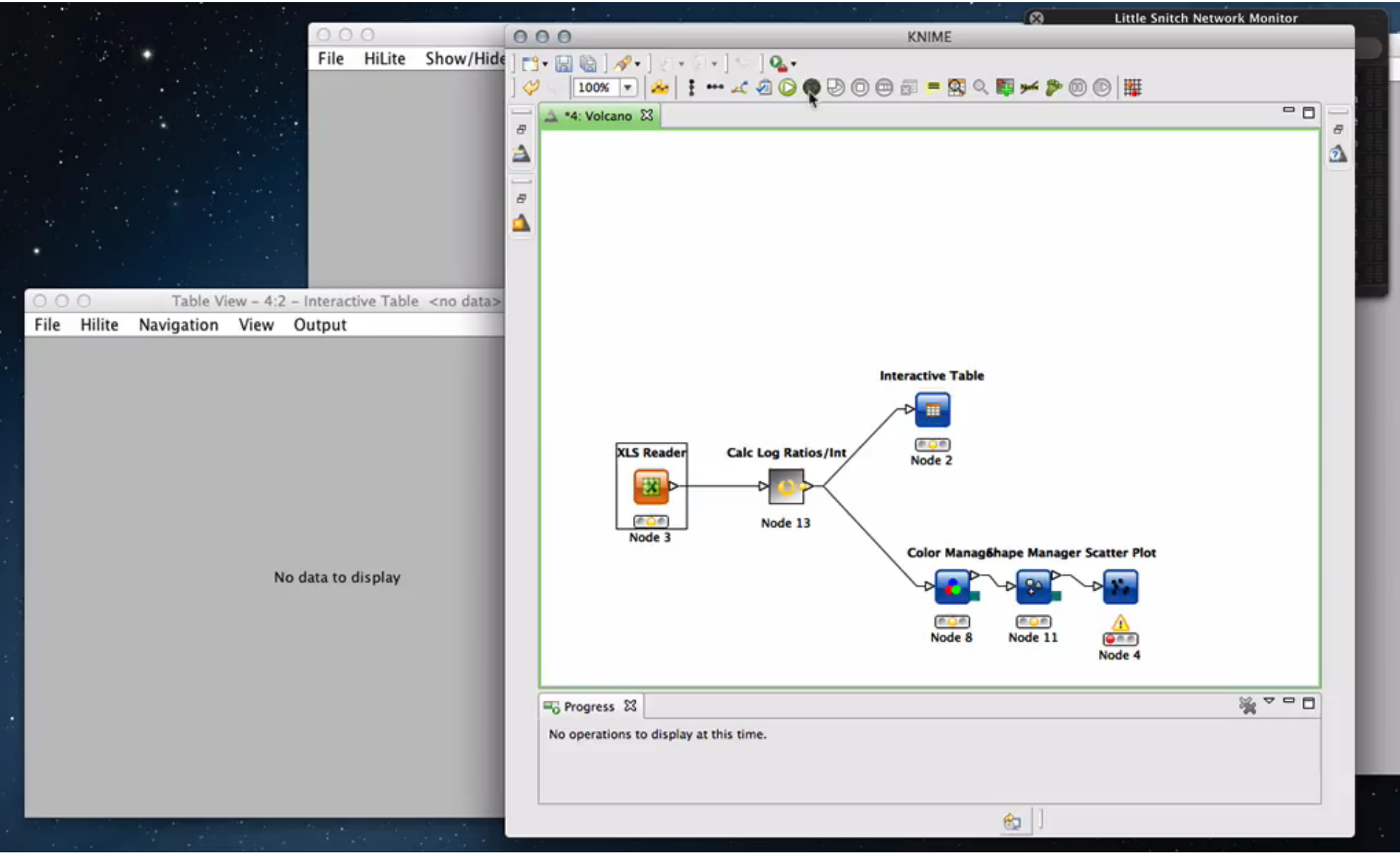
Generating **reports**

- Single file (e.g., pdf) from workflows
- "Data To Report" and "Image To Report" nodes specify what will be reported
- Visual construction and layout in **report perspective**



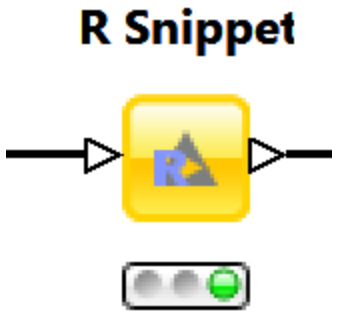


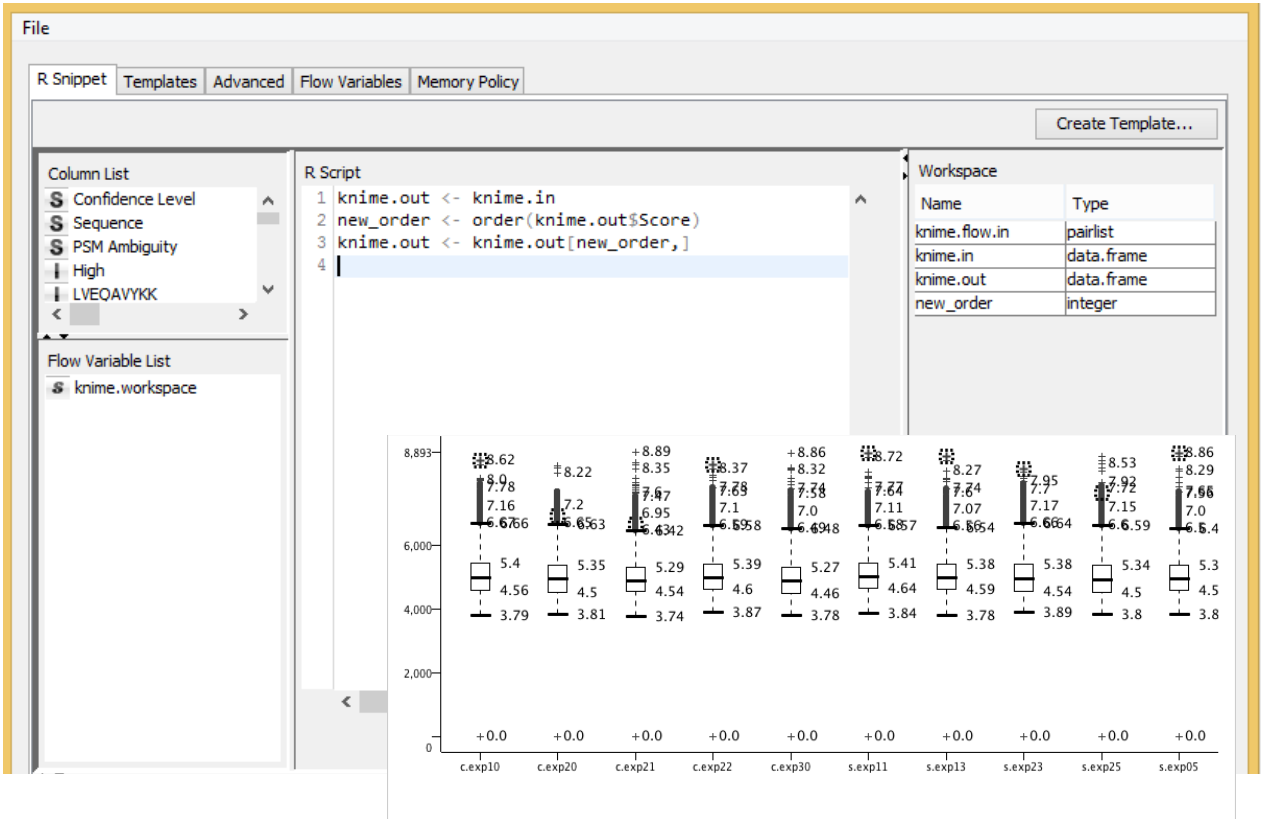
KNIME Interactive Analysis



Statistics Workflows in KNIME

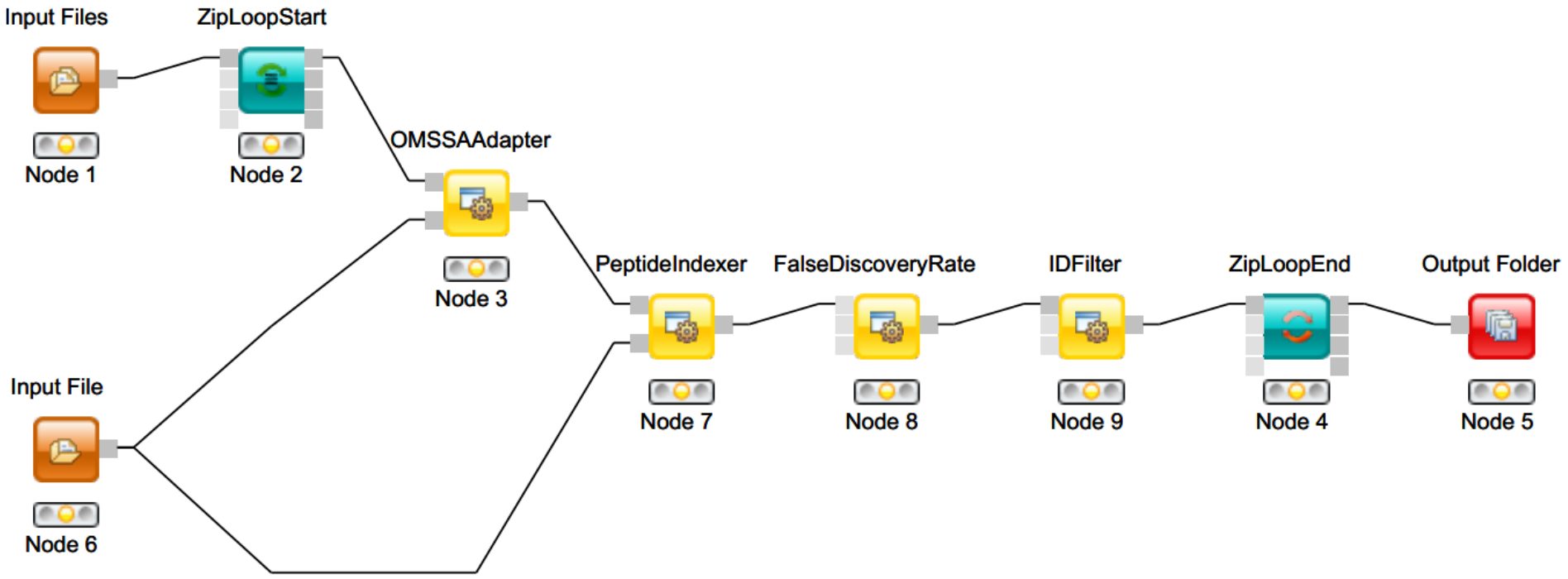
- KNIME permits the embedding of R code for advanced statistics
- Embedding of R scripts using the R Snippet node
- All plotting capabilities of R can be used as well



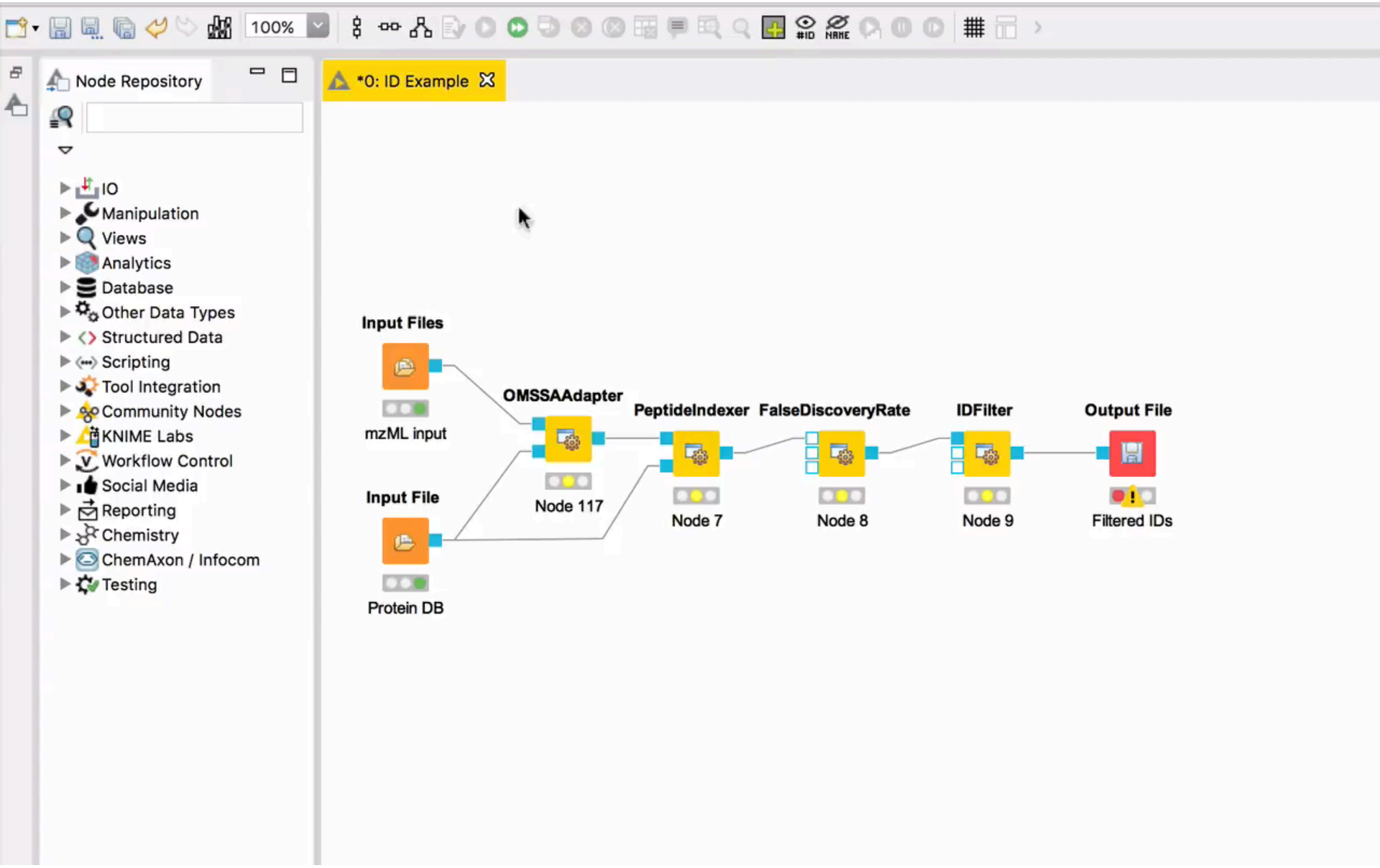


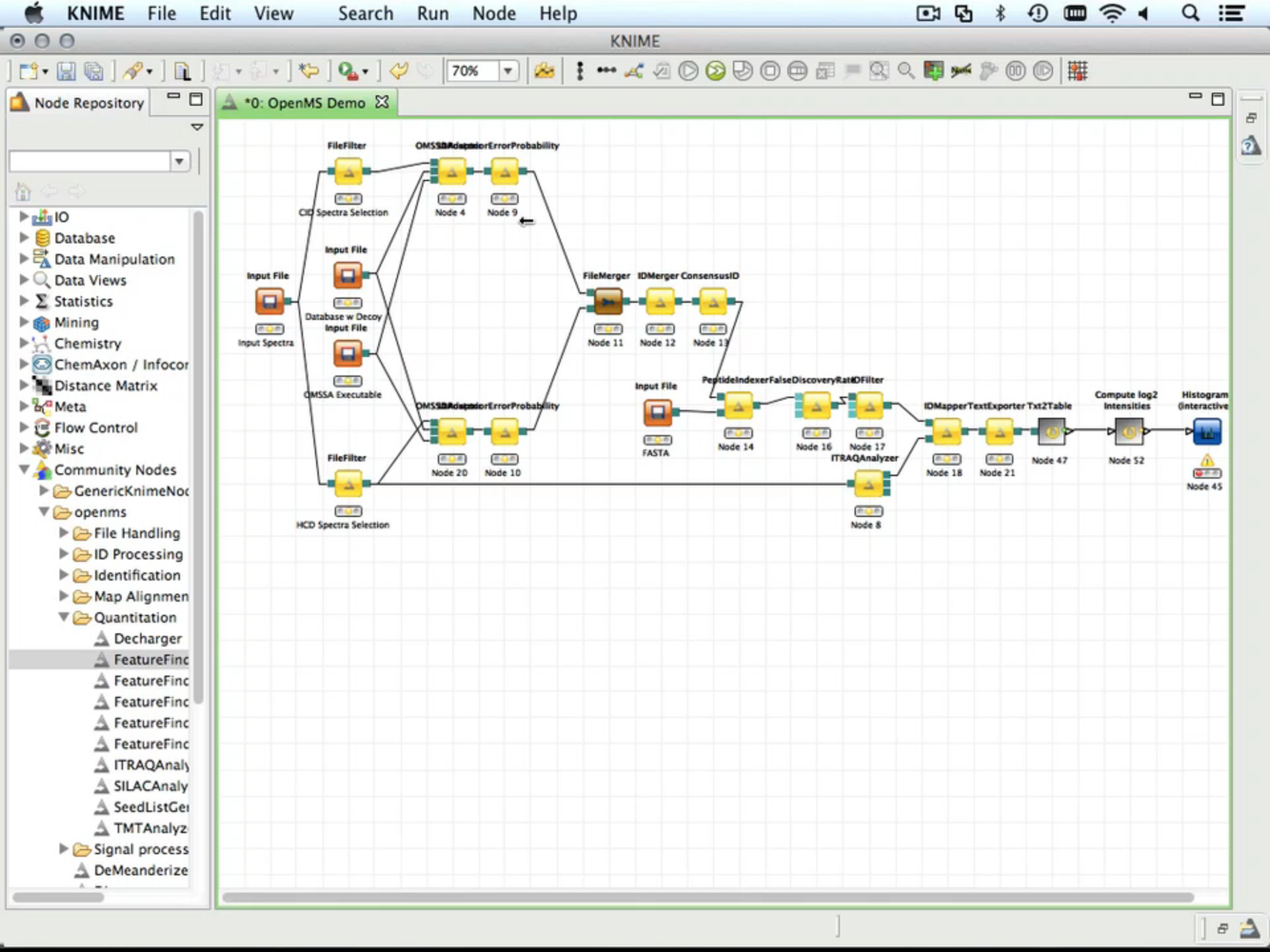
Protein Identification Workflow

- Finding all proteins in multiple samples
- Mass spectra enter workflow on the left
- Loop nodes permit execution of parts of the workflow
- Identified proteins end up in result files (right side)

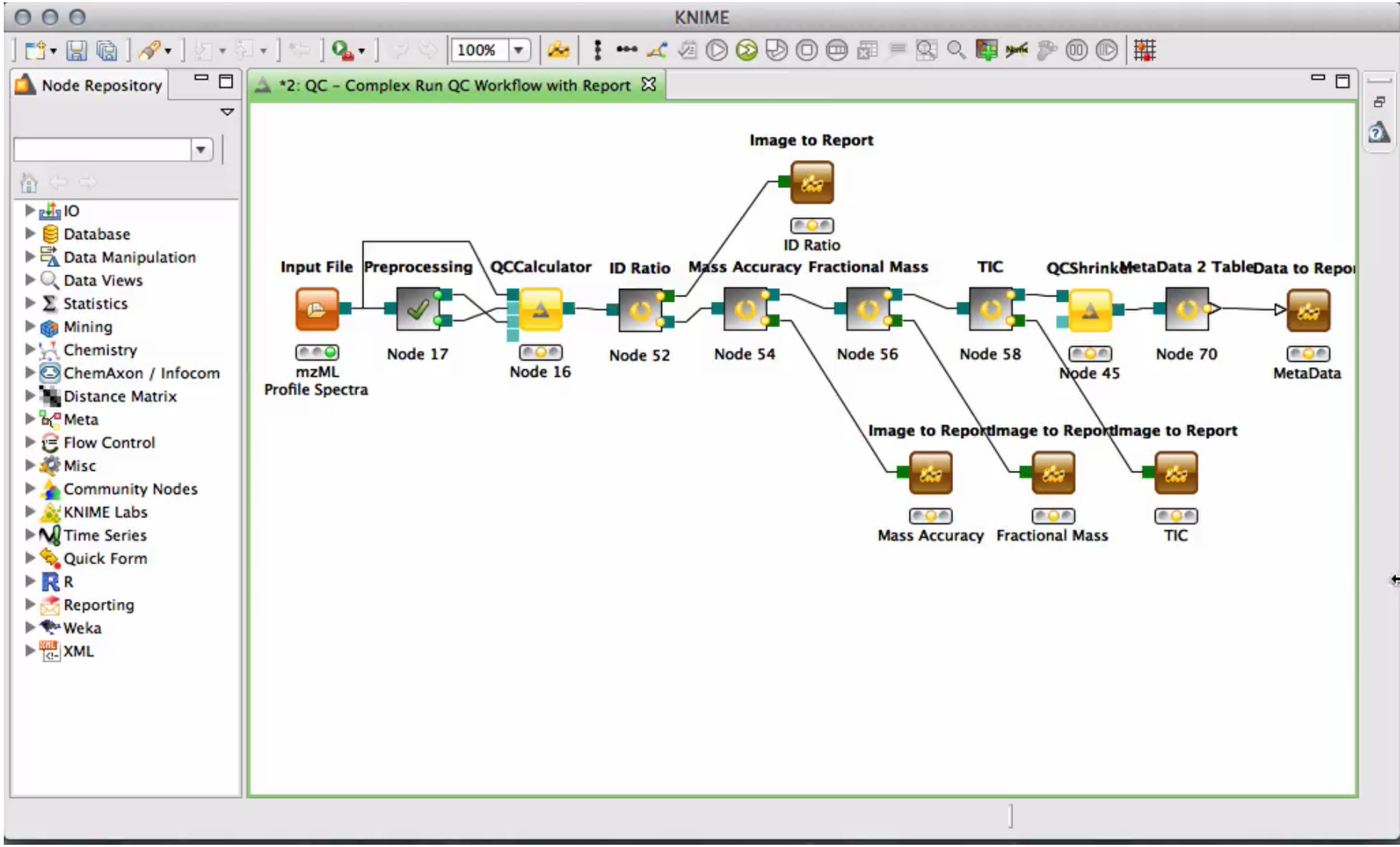


Workflows as an Abstraction

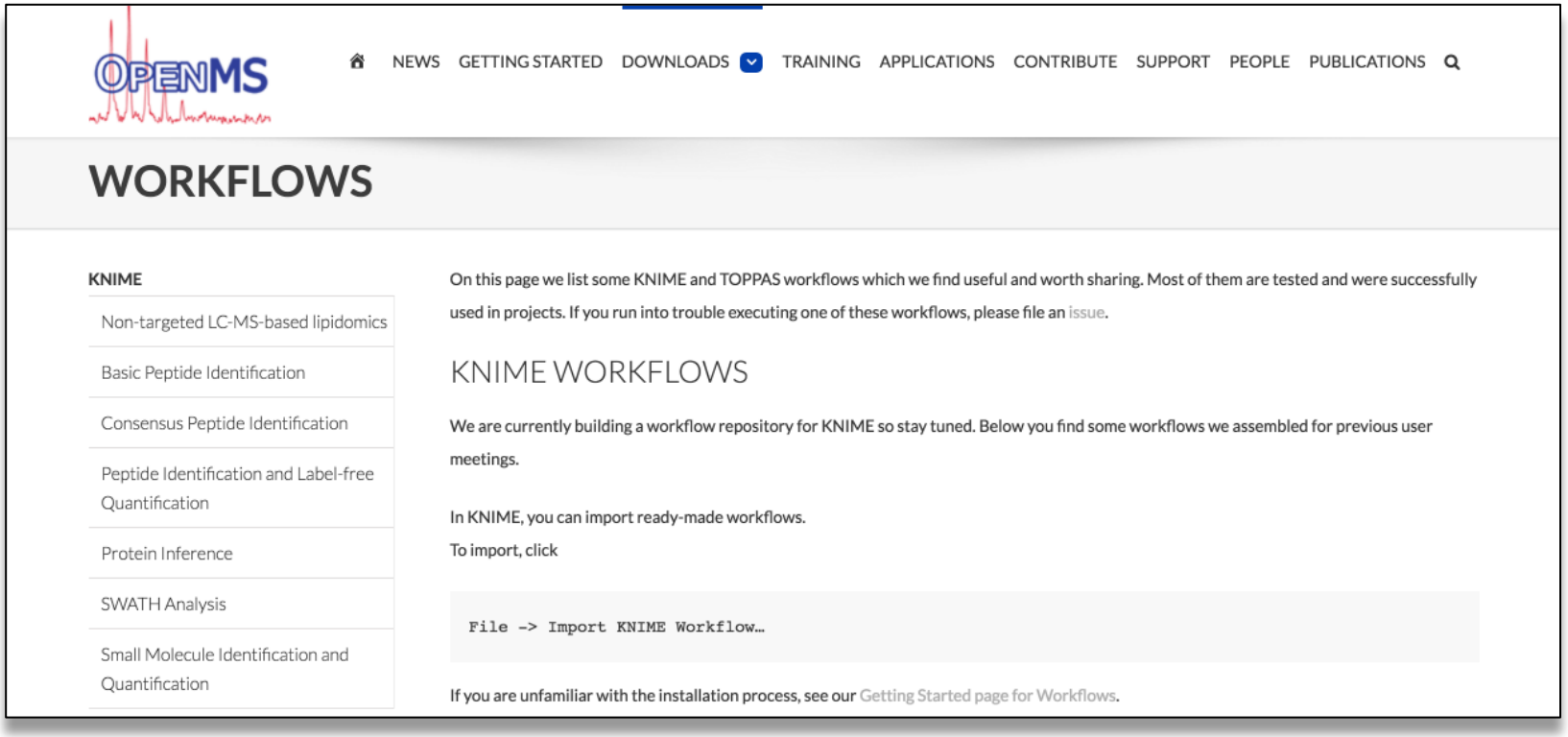



KNIME File Edit View Search Run Node Help
KNIME
70%
Node Repository
IO
Database
Data Manipulation
Data Views
Statistics
Mining
Chemistry
ChemAxon / Infocor
Distance Matrix
Meta
Flow Control
Misc
Community Nodes
GenericKnimeNoc
openms
File Handling
ID Processing
Identification
Map Alignmen
Quantitation
Decharger
FeatureFinc
FeatureFinc
FeatureFinc
FeatureFinc
FeatureFinc
ITRAQAnaly
SILACAnaly
SeedListGe
TMTAnalyz
Signal process
DeMeanderize
*0: OpenMS Demo
FileFilter
CID Spectra Selection
Node 4
Node 9
Input File
Input Spectra
Input File
Database w Decoy
Input File
OMSSA Executable
FileFilter
HCD Spectra Selection
Node 20
Node 10
FileMerger
Node 11
IDMerger
Node 12
ConsensusID
Node 13
Input File
FASTA
Node 14
Node 16
Node 17
Node 8
ITRAQAnalyzer
Node 18
Node 21
Node 47
Node 52
Node 45

Report Generation

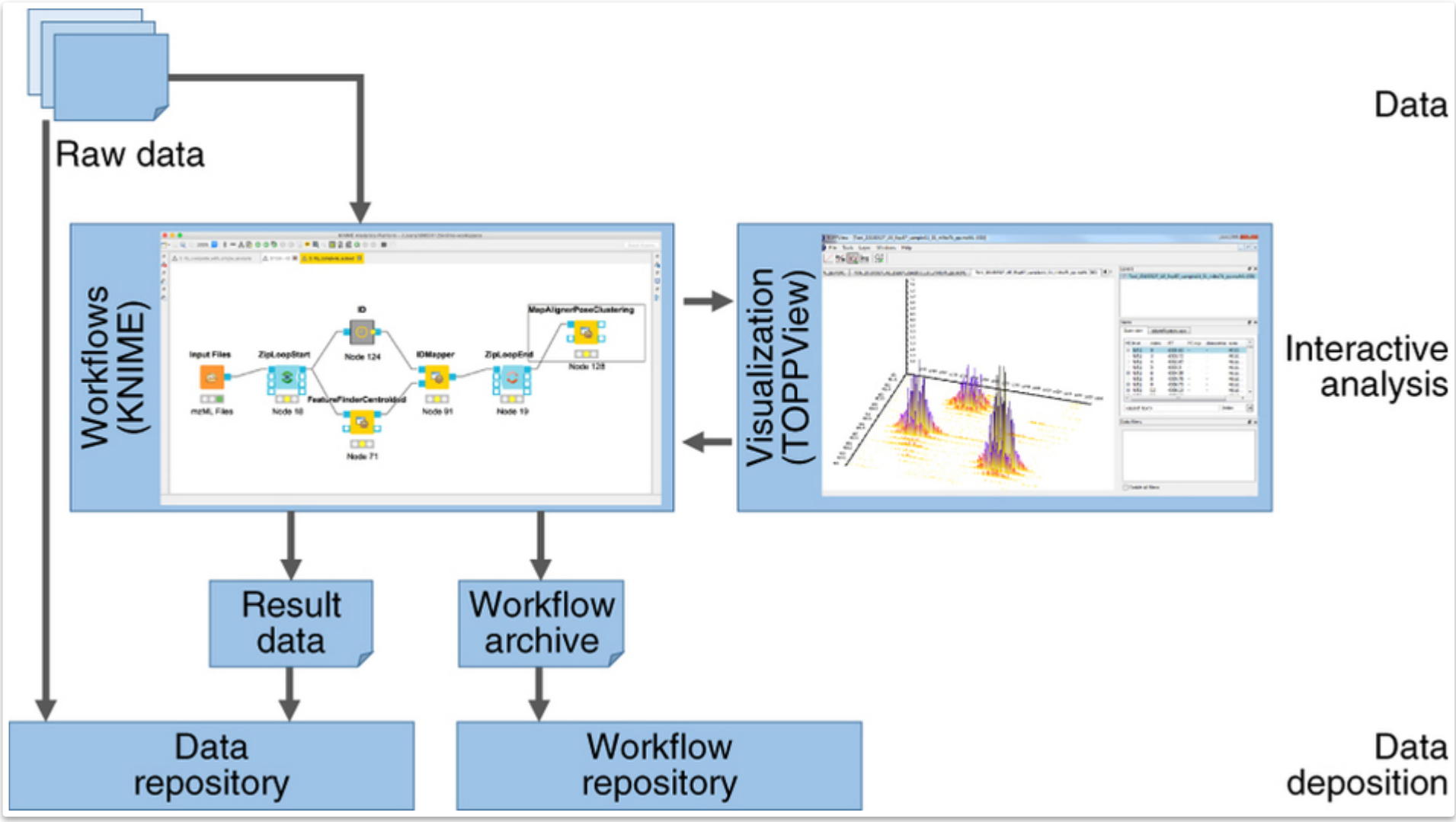


Workflow Repositories



- OpenMS website contains a workflow repository with selected example workflows (**www.OpenMS.org**)
- General-purpose workflow repository: **www.myexperiment.org**
  - Collects workflows from arbitrary workflow engines
  - Numerous applications, can be used to document data analysis

Open (Data | Source | Science)



Röst et al., Nat. Methods, 2016, 13:741

Materials

- **KNIME**
  - Quickstart guide http://tech.knime.org/files/KNIME_quickstart.pdf
  - Screencasts http://tech.knime.org/screencasts
  - Report generation http://tech.knime.org/getting-started-0
- **OpenMS**
  - Tutorials

    http://ftp.mi.fu-berlin.de/OpenMS/release-documentation/TOPP_tutorial.pdf

    https://sourceforge.net/p/open-ms/code/HEAD/tree/Tutorials/UM_2014/Handout/

Materials

- **Online Materials**
  - Online lecture ‘Computational Proteomics and Metabolomics’ (Kohlbacher, Reinert, Nahnsen)
    http://bit.ly/2d2kBSq
  - Comp MS Website @ http://CompMS.org

- **OpenMS website**: http://www.OpenMS.org
  - Documentation
  - Downloads
    - Binaries
    - Source code
    - Plugins for Proteome Discoverer
  - Access to mailing lists – this is where you can get help 24/7