COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

0. Introduction and Overview

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)

Integration vs. Reductionism

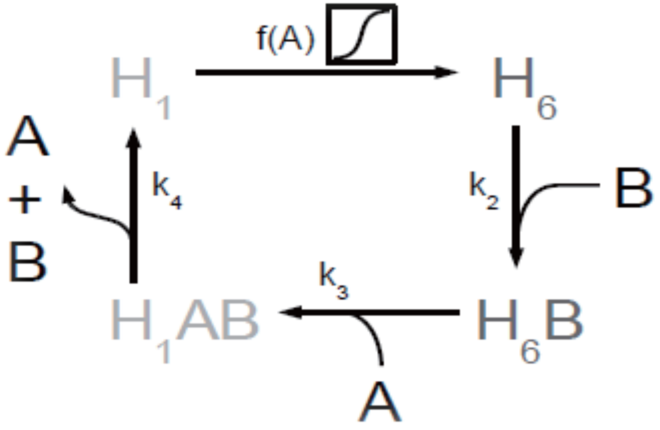
- Systems biology as an integrative approach takes the reductionist approach one step further
- Do not only understand the components, but understand 'emerging properties' of a system
- Key of this is the integration of different data, covering different aspects of the system
- Integrated modeling of the whole system can then reveal these emerging and dynamic properties

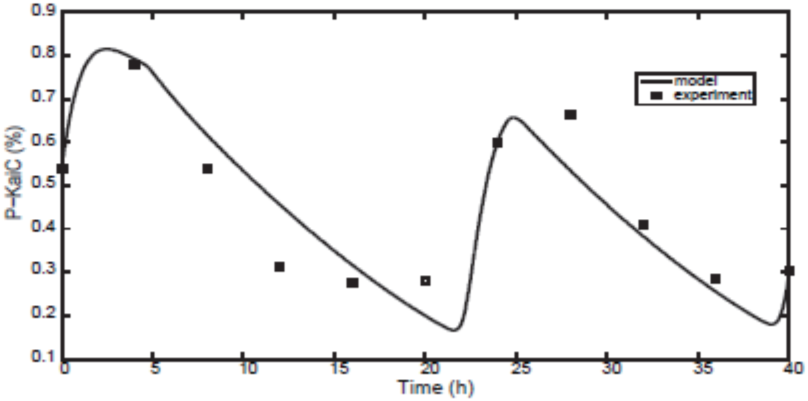
Example:

circadian clock – the temporal (dynamic) behavior is an emerging property of the rather simple interaction of a few key players.

Circadian Clock in Cyanobacteria

- Circadian clocks are internal oscillators implementing a 24 hour rhythm in most organisms
- The model shown on the right is a simple model for cyanobacteria including three genes (KaiA, KaiB, and KaiC – A/B/C)
- Their interaction, phosphorylation, hexamer formation (H_6), etc. are simple processes that can be described mathematically
- **Together** these simple processes give rise to the oscillation shown on the right, which agrees well with experimental data
- Looking at each of the processes in isolation will not reveal the osciallation

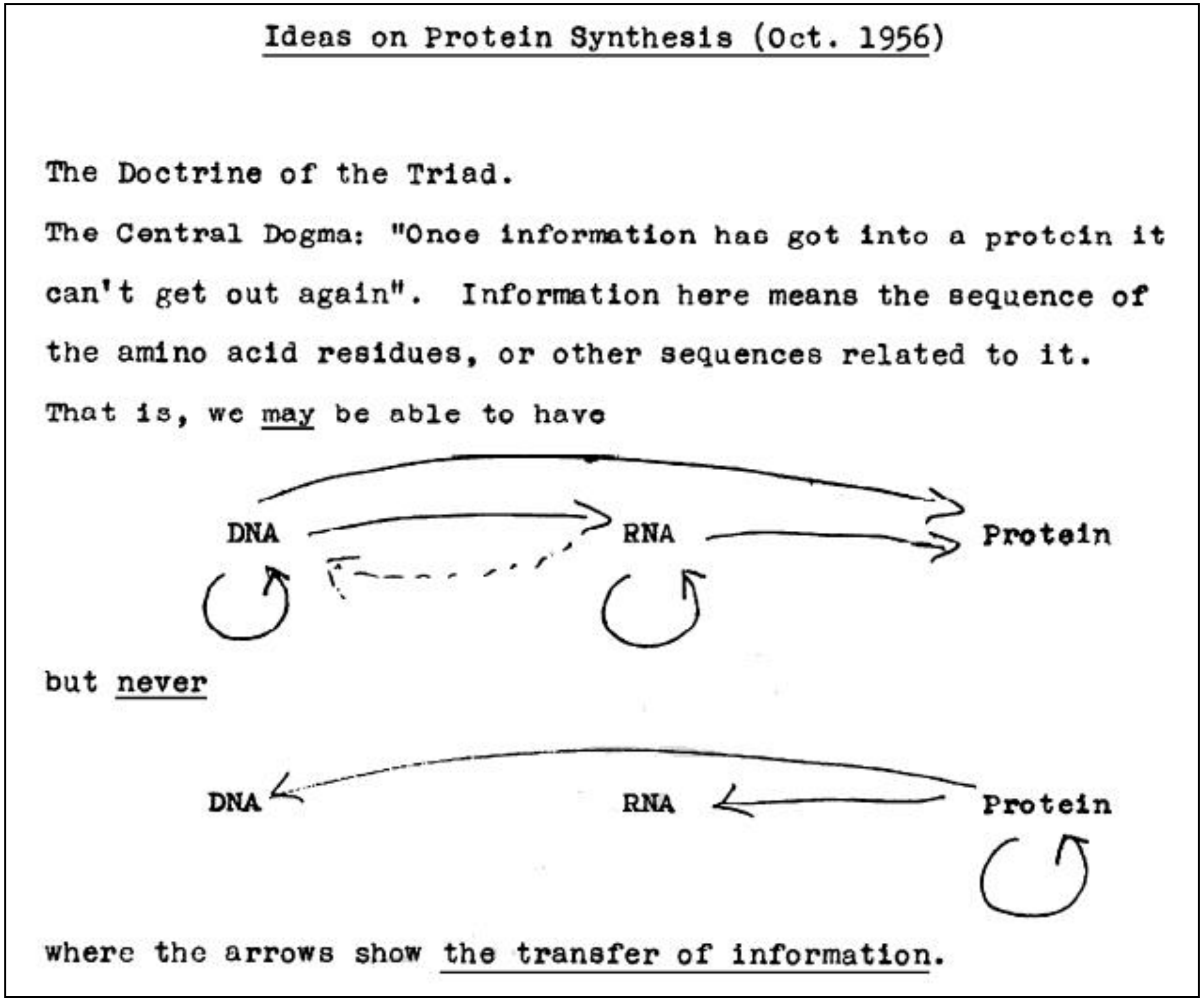




I. Axmann, S. Legewie, and H. Herzel (2007). A minimal circadian clock model. Genome Informatics. 18:54-64.

Central Dogma of Molecular Biology

- First articulation 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

DNA
RNA
Protein

but never

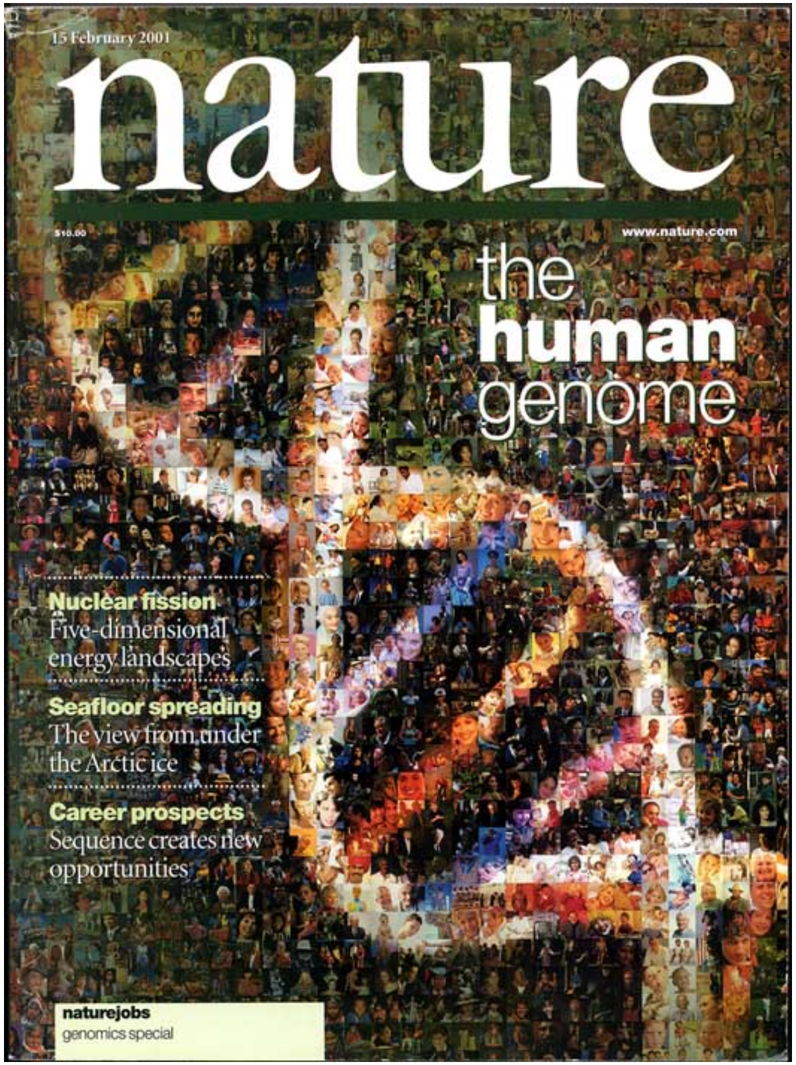
DNA
RNA
Protein

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*

[…]

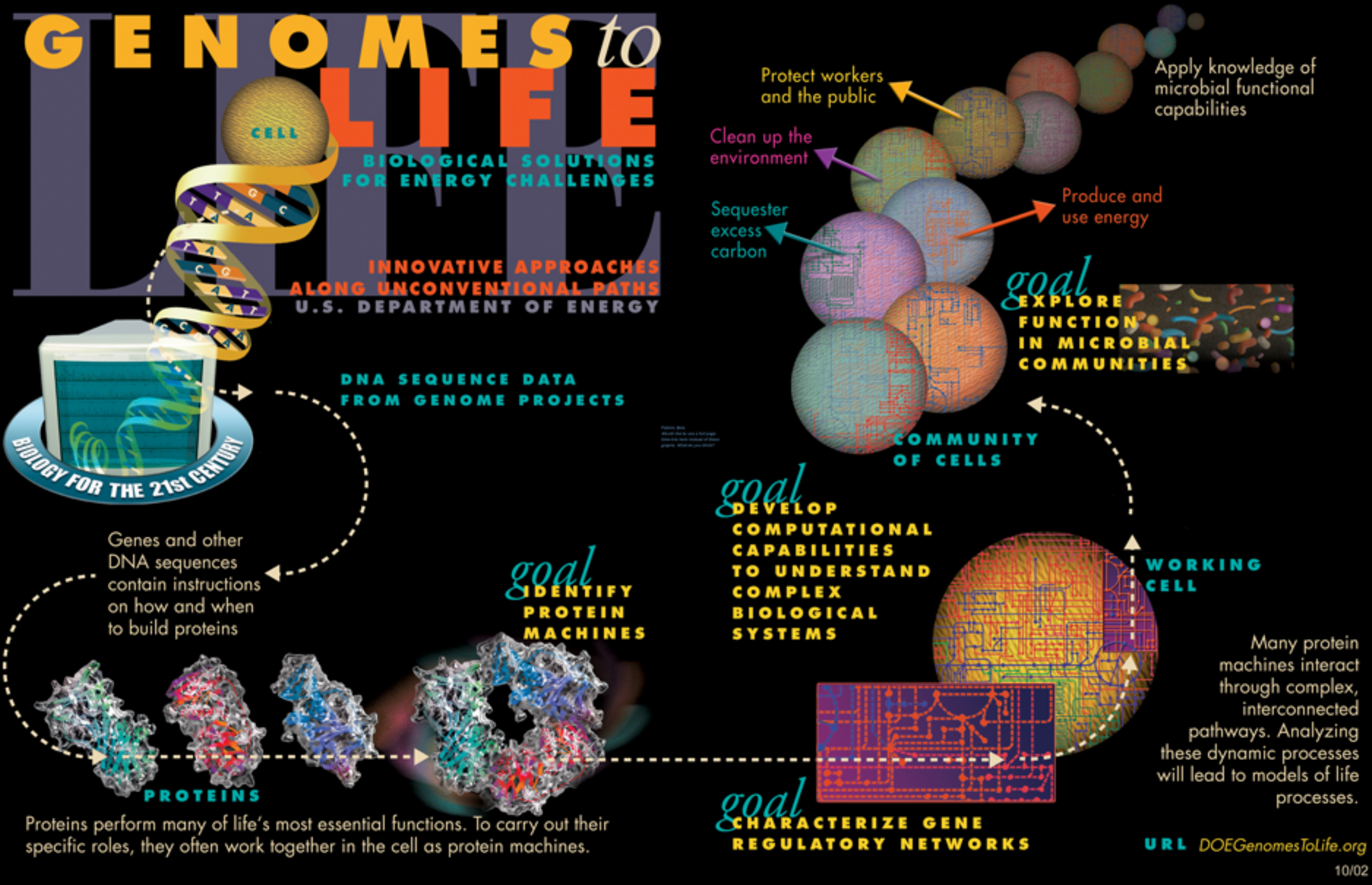
3. *Cell Biol. and Molecular Biol.* Forming nouns with the sense ‘all of the specified constituents of a cell, considered collectively or in total’, as plastidome n., plastome n., vacuome n.

(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



GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

U.S. DEPARTMENT OF ENERGY

DNA SEQUENCE DATA FROM GENOME PROJECTS

Genes and other DNA sequences contain instructions on how and when to build proteins

PROTEINS

Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

goal IDENTIFY PROTEIN MACHINES

goal CHARACTERIZE GENE REGULATORY NETWORKS

goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS

WORKING CELL

Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.

COMMUNITY OF CELLS

goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

Apply knowledge of microbial functional capabilities

Protect workers and the public

Clean up the environment

Sequester excess carbon

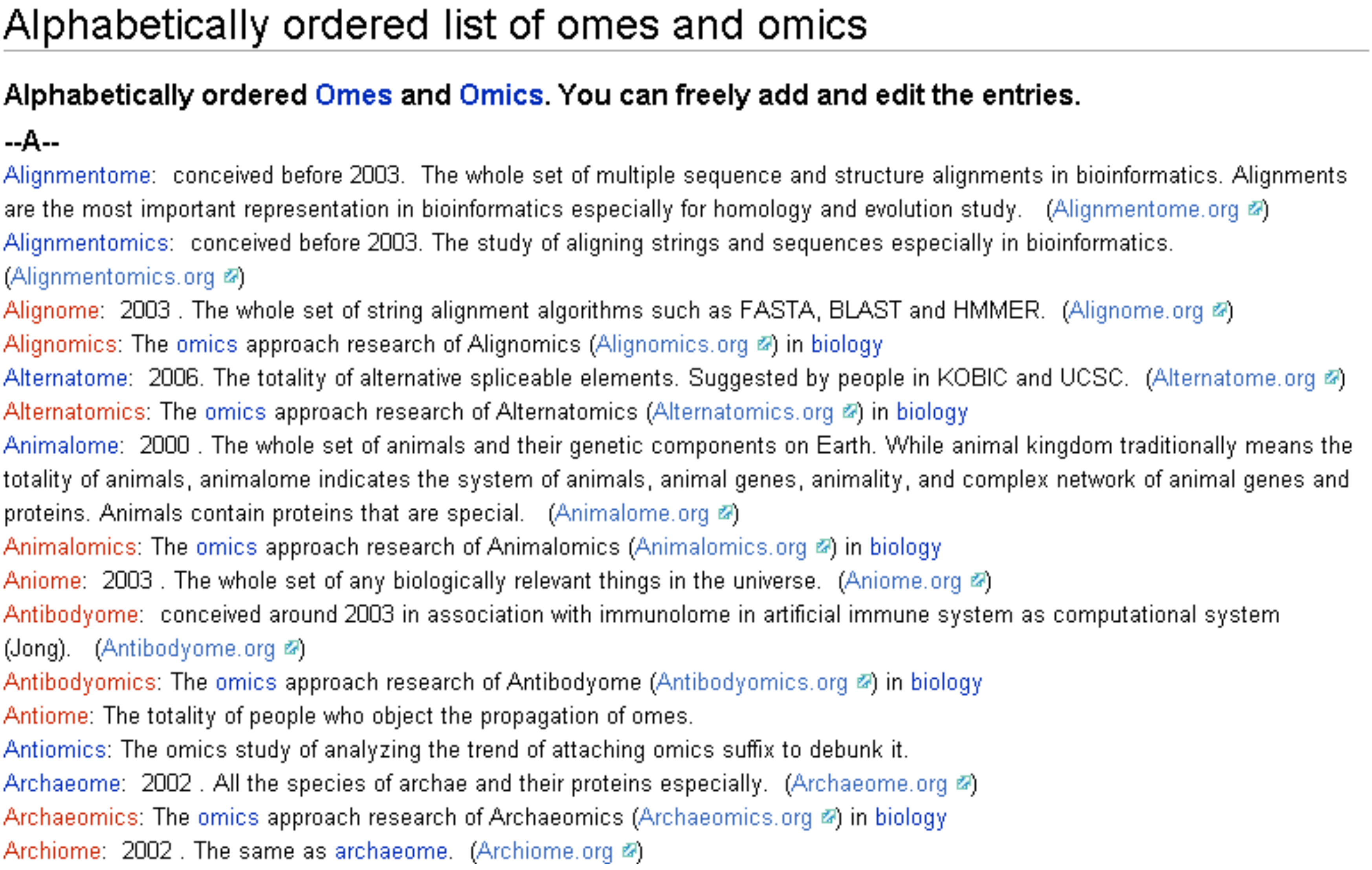
Produce and use energy

Patrick, Bob,
Would like to use a full page time-line here instead of these graphs. What do you think?

URL *DOEGenomesToLife.org*

10/02

OMICS Mania



Alphabetically ordered list of omes and omics

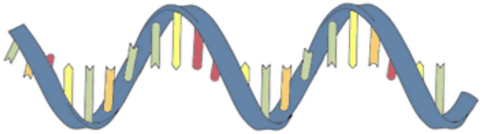
Alphabetically ordered Omes and Omics. You can freely add and edit the entries.

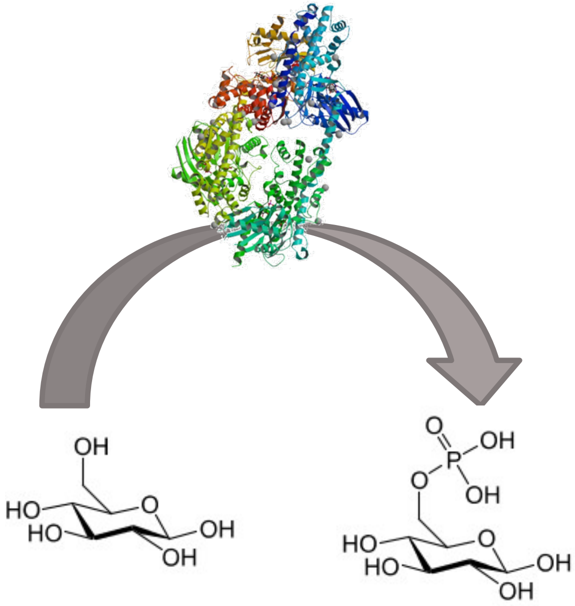
--A--

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)

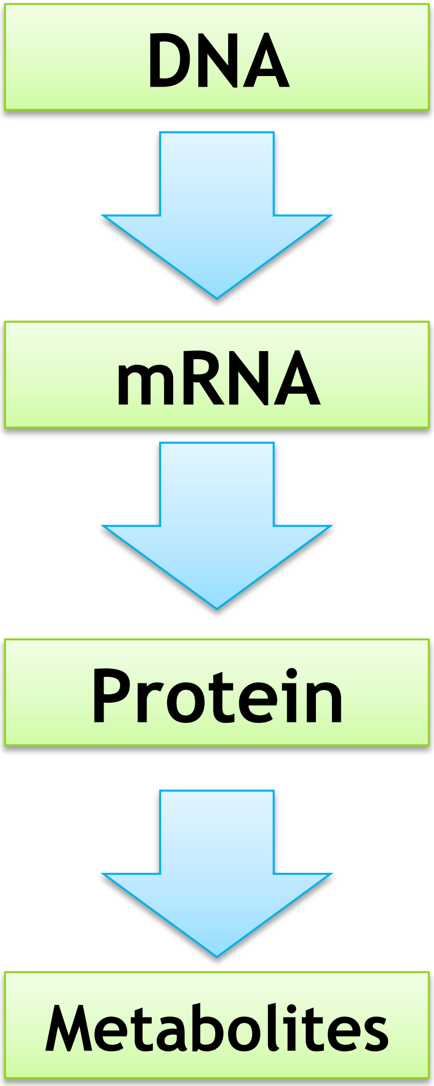
http://omics.org/index.php/Alphabetically_ordered_list_of_omes_and_omics

The World of Omes



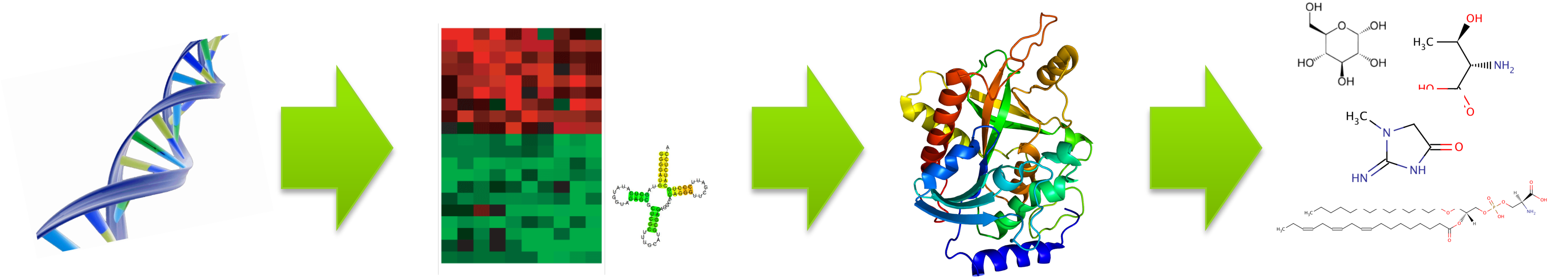


http://www.pdb.org/pdb/images/2nzt_bio_r_500.jpg



- Genome
- Transcriptome
- Proteome
- Metabolome

Technologies



Genome
Epigenome

Transcriptome
RNOme

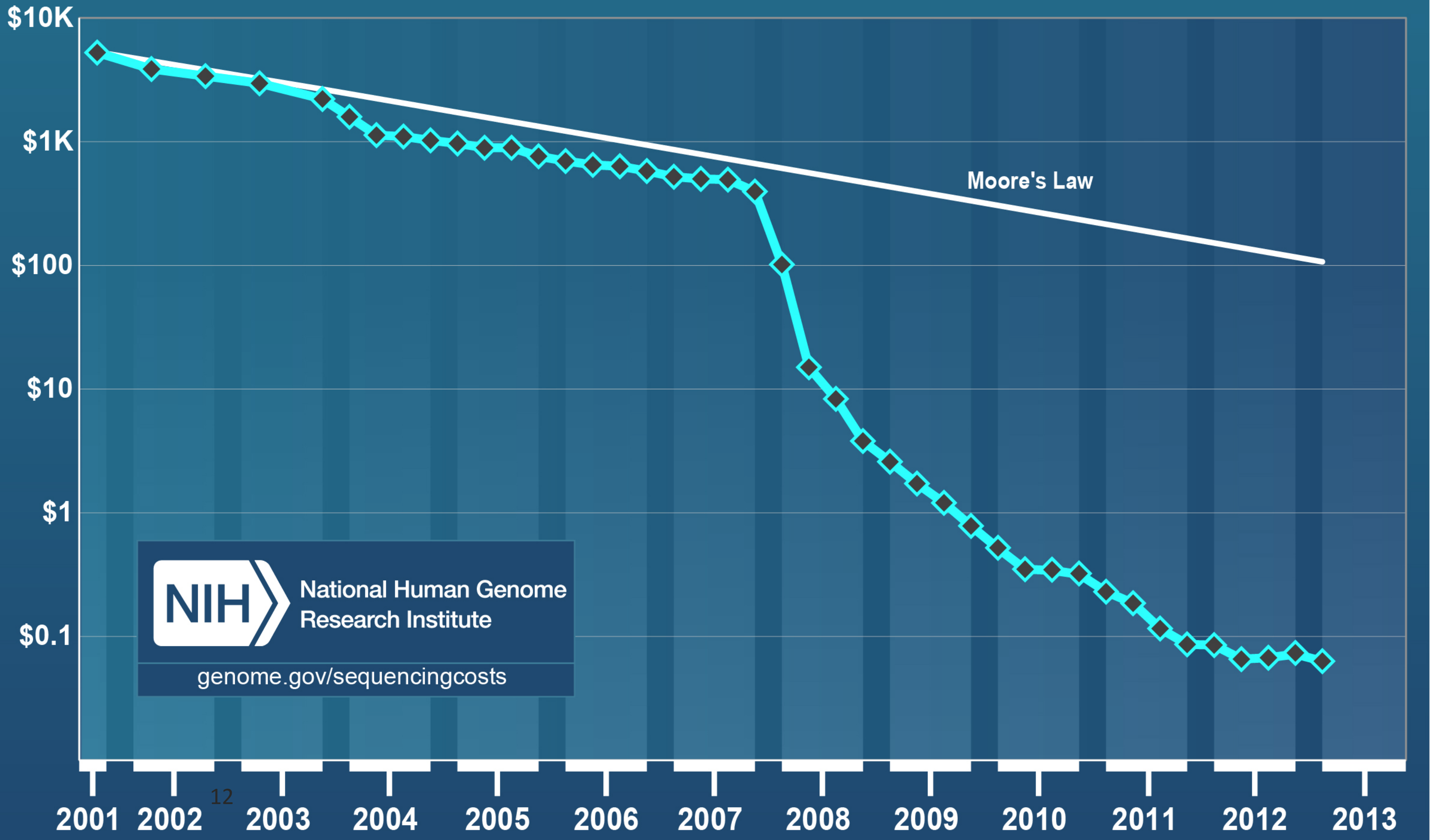
Proteome
Interactome

Metabolome
Lipidome

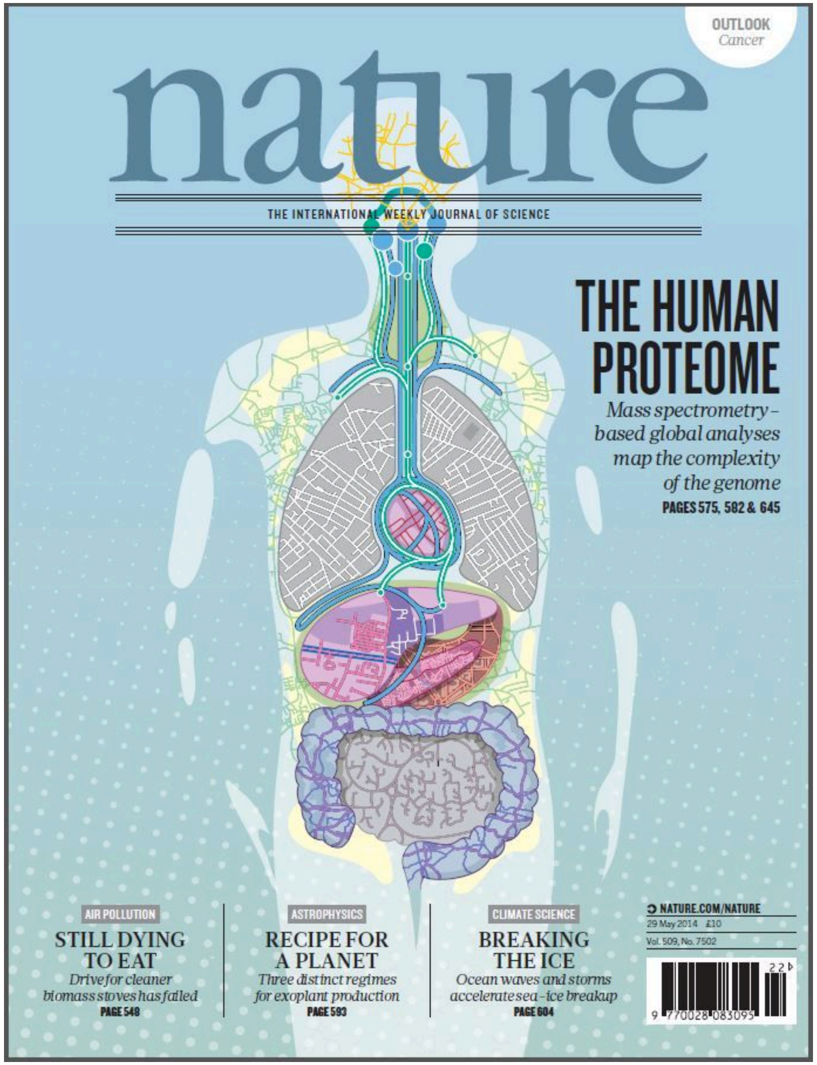
Next-Generation Sequencing

Mass Spectrometry

Cost per Raw Megabase of DNA Sequence



Human Proteome


OUTLOOK
Cancer
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE HUMAN PROTEOME
Mass spectrometry-based global analyses map the complexity of the genome
PAGES 575, 582 & 645
AIR POLLUTION
STILL DYING TO EAT
Drive for cleaner biomass stoves has failed
PAGE 548
ASTROPHYSICS
RECIPE FOR A PLANET
Three distinct regimes for exoplant production
PAGE 593
CLIMATE SCIENCE
BREAKING THE ICE
Ocean waves and storms accelerate sea-ice breakup
PAGE 604
NATURE.COM/NATURE
29 May 2014 £10
Vol. 509, No. 7502

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

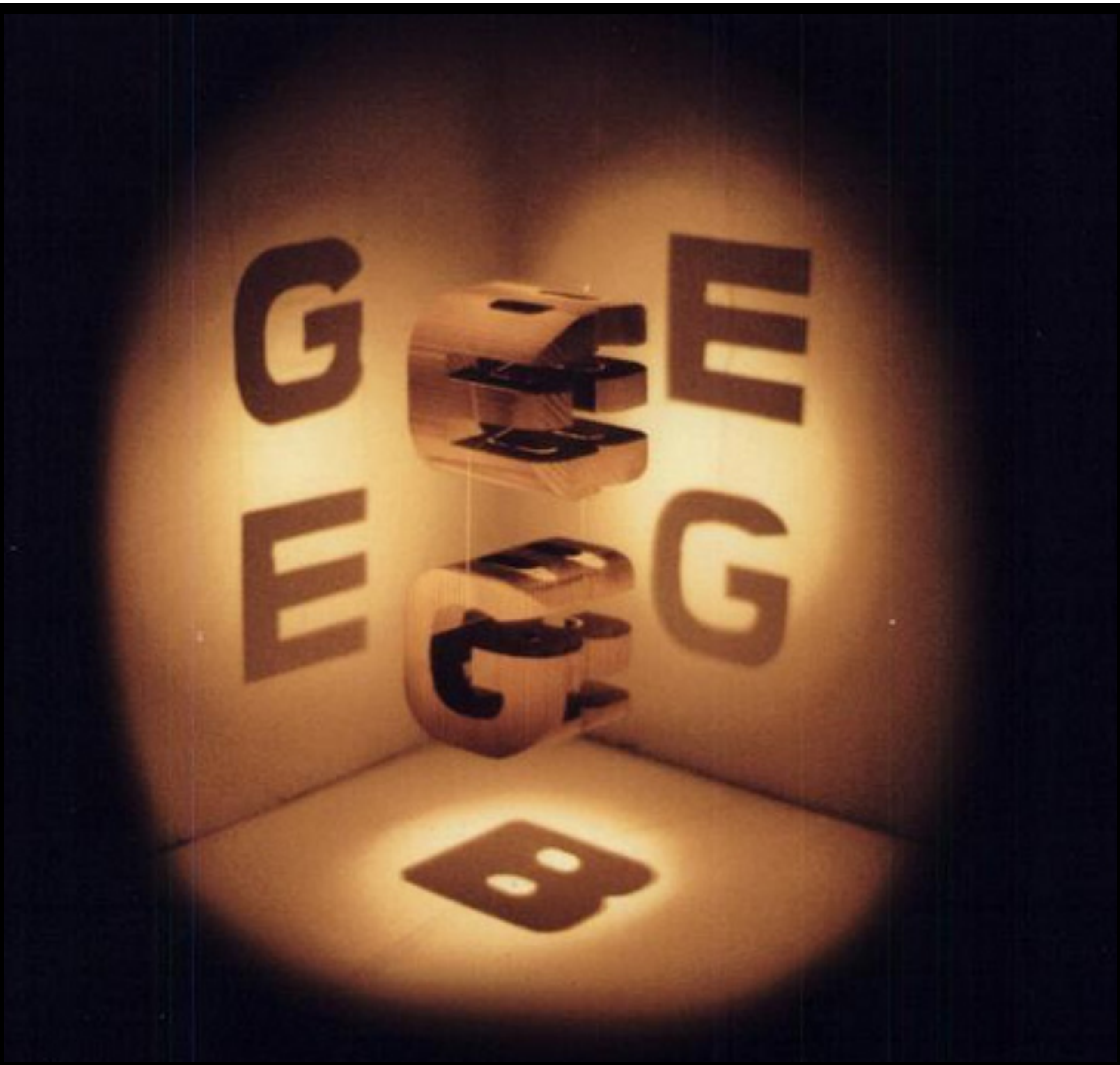
- **High-throughput techniques** provide data for one specific type of relationship
  - **Genomics**: DNA sequence data
  - **Transcriptomics**: mRNA concentration
  - **Proteomics**: protein concentrations/sequence
  - **Metabolomics**: metabolite concentrations
  - **Interactomics**: protein-protein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

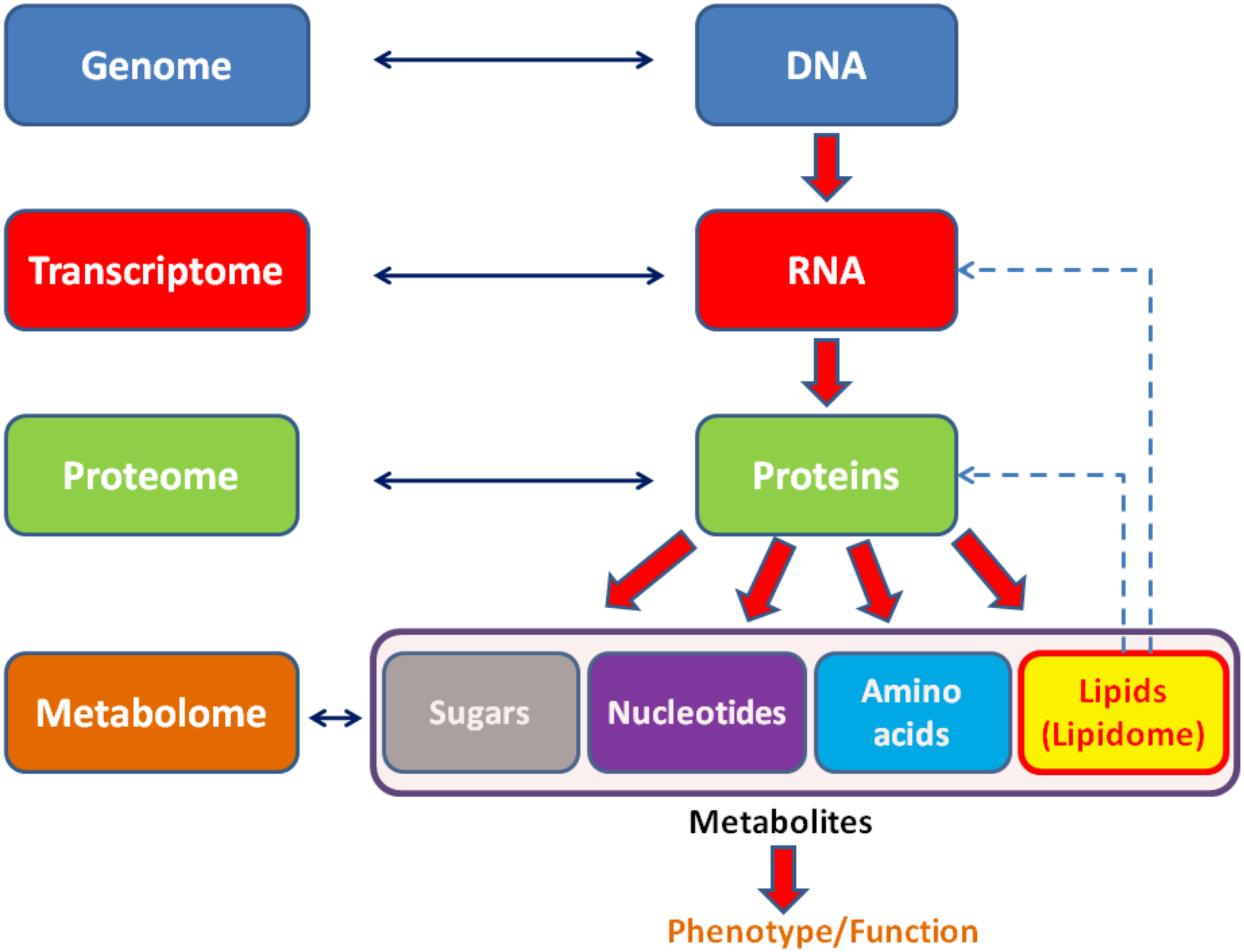
Omics is a Matter of Perspective!



Omics is a Matter of Perspective

- Each omics technology/level provides a cross-section of one particular type of biomolecules
- Different levels thus correlate (roughly) to distinct ??????
  - **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?
- Different levels thus provide a different functional perspective

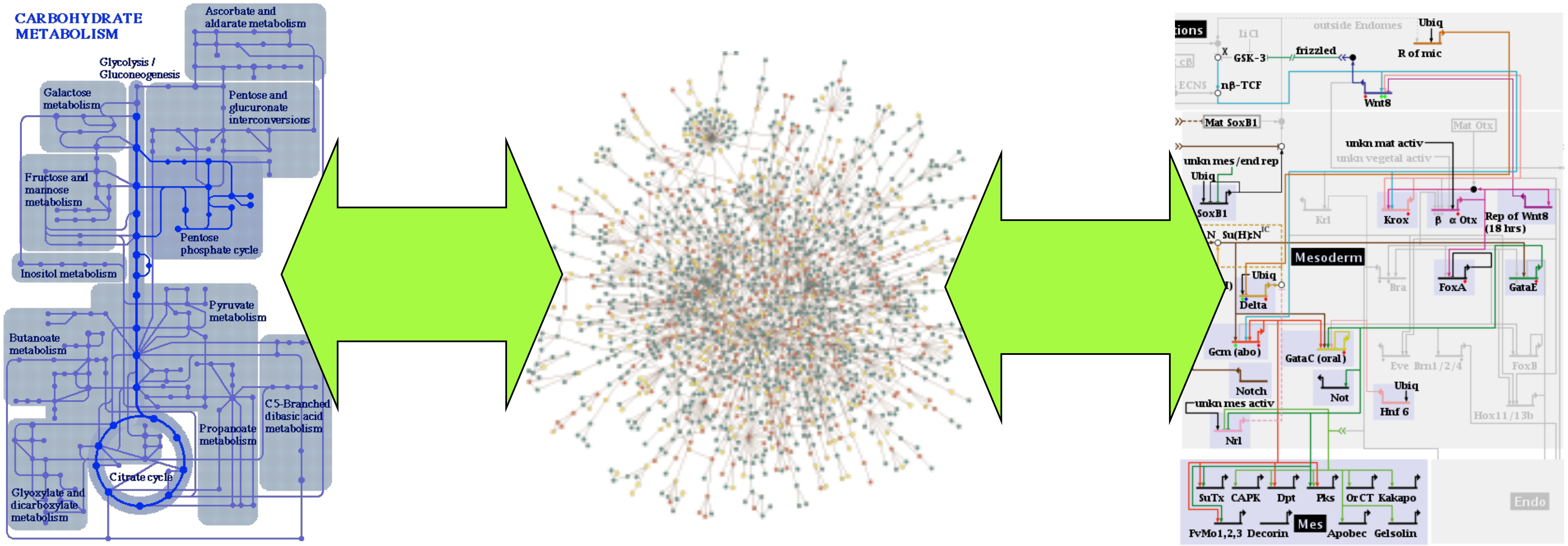
Omics Technologies



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!

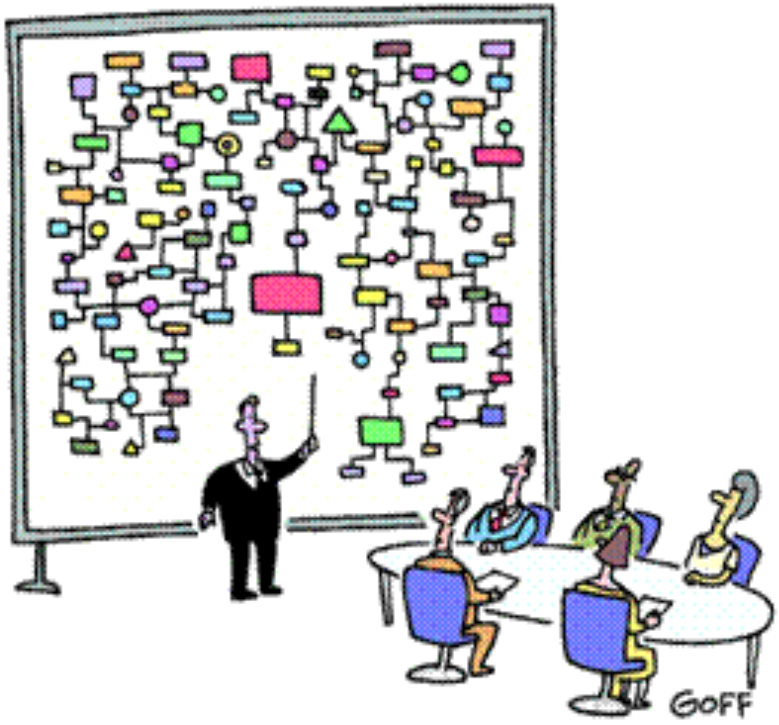


Computational Systems Biology

- 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."

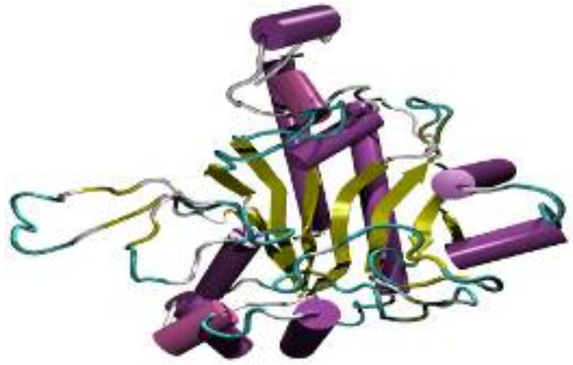
http://www.sys-bio.org/contentimages/WhyWeNeedComputer.png

Challenges in Data Integration

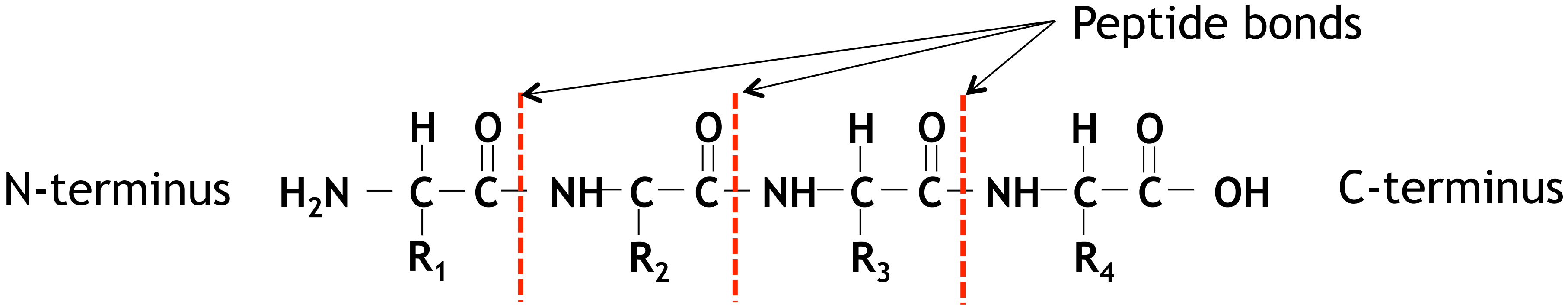
- **Semantic integration** of data from different sources
  - Different data formats
  - Ambiguities, nomenclature
- **Lack of data**
  - We do not know everything!
  - High-throughput methods show only a fraction of 'everything' (detection limits!)
- **Different scales**
  - Time scales different, length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

Protein

- A protein or polypeptide consists of a linear chain of amino acids that build 3-dimensional structures

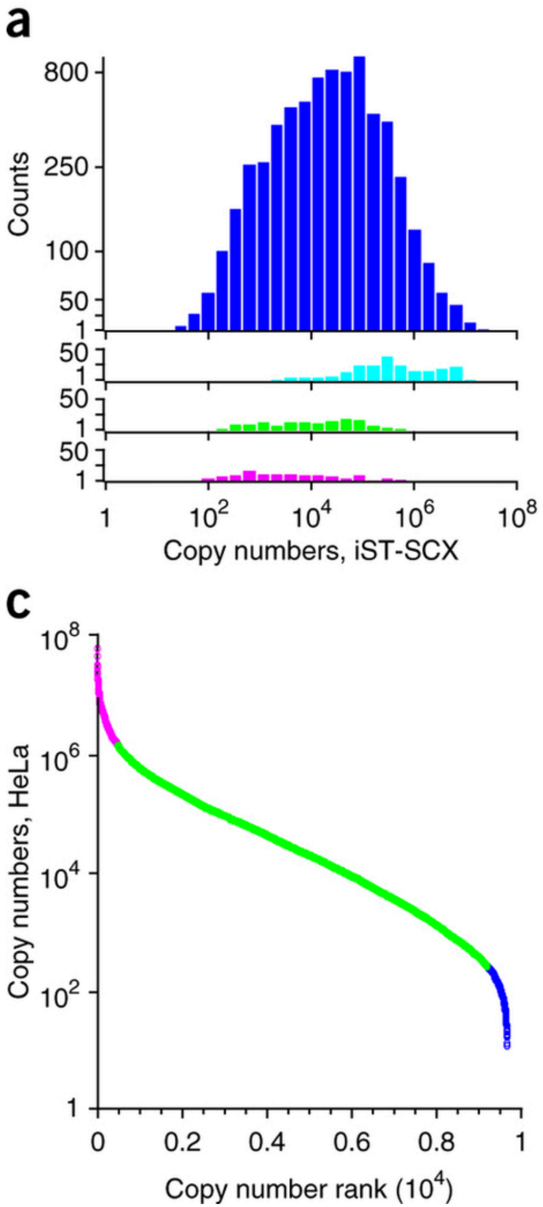


- Amino acids are connected via peptide bonds



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

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
  - Posttranslational modifications: PTMs are very hetero-geneous and significantly alter the function of the protein

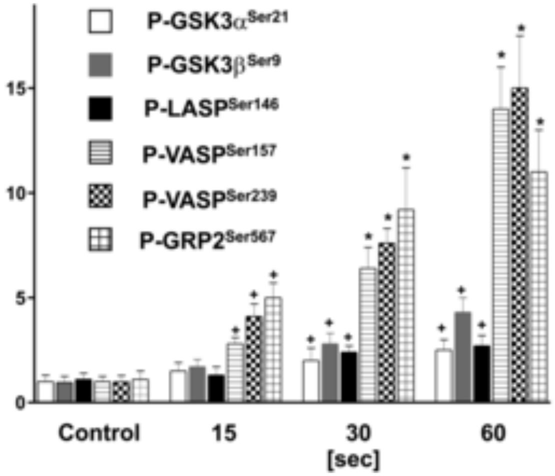
Proteomics - Examples

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

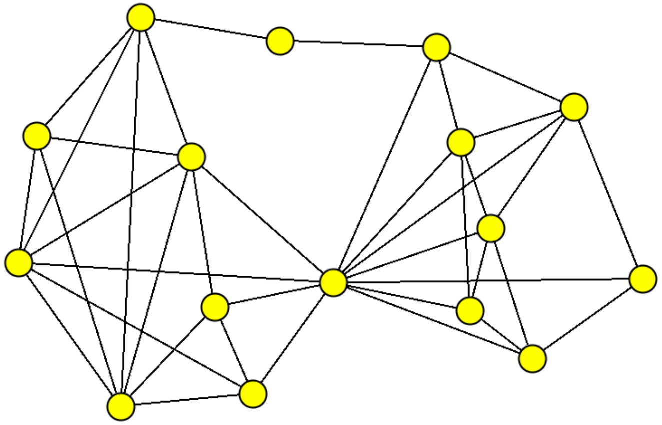


Time course of selected phosphopeptides (Beck et al., 2014)

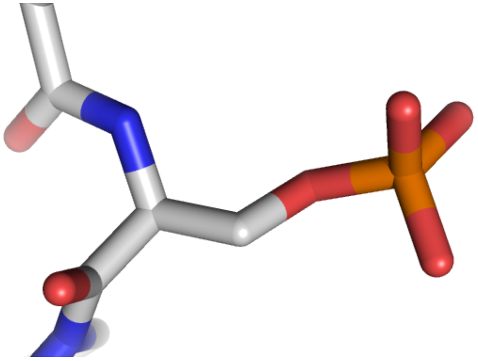
http://www.iamashcash.com/wp-content/uploads/2011/03/caterpillar-to-butterfly1.jpg, accessed: 14/10/2013 6 PM
http://www.ufrgs.br/laprotox/en/what-we-do/research-lines/ureases-non-enzymatic-properties/ureases-induce-platelet-secretion-and-aggregation
Beck et al., Blood (2014), 123(5):e1-e10. doi: 10.1182/blood-2013-07-512384

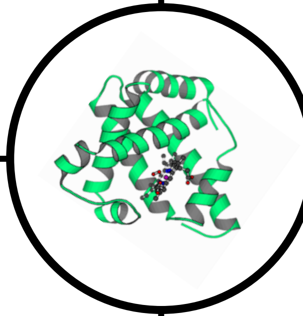
Main fields of proteomics

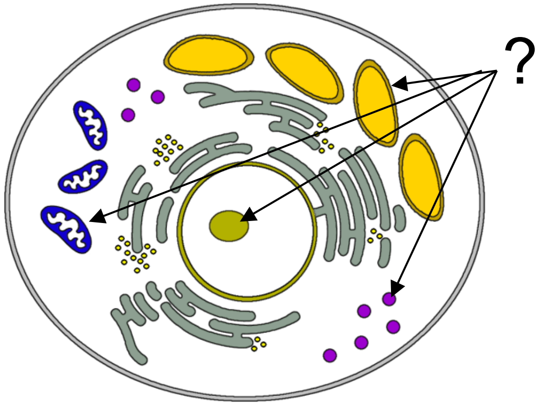
protein interaction



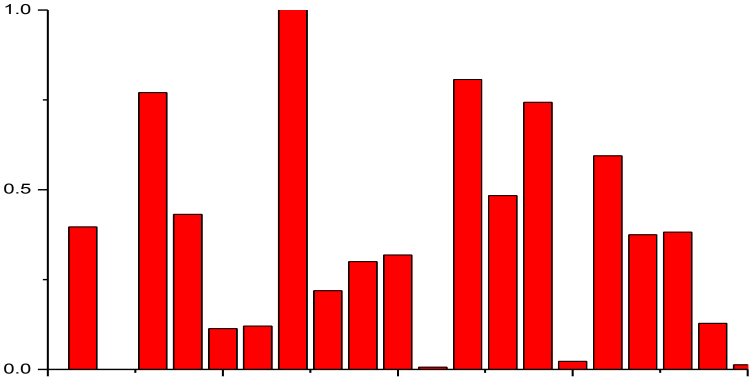
protein characterization (identification + PTMs)







protein localization



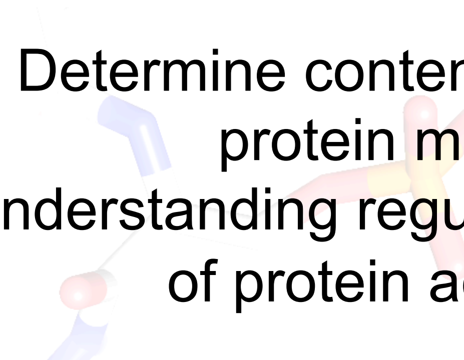
protein expression

Applications of proteomics

protein interaction

- Drug target identification

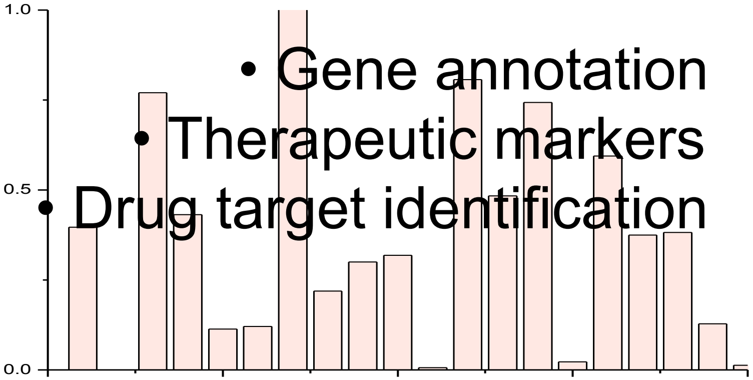
protein characterization
(identification + PTMs)



- Determine content of a protein mixture
- Understanding regulation of protein activity

- Functional annotation (compartment and function)
- Drug target identification

protein localization



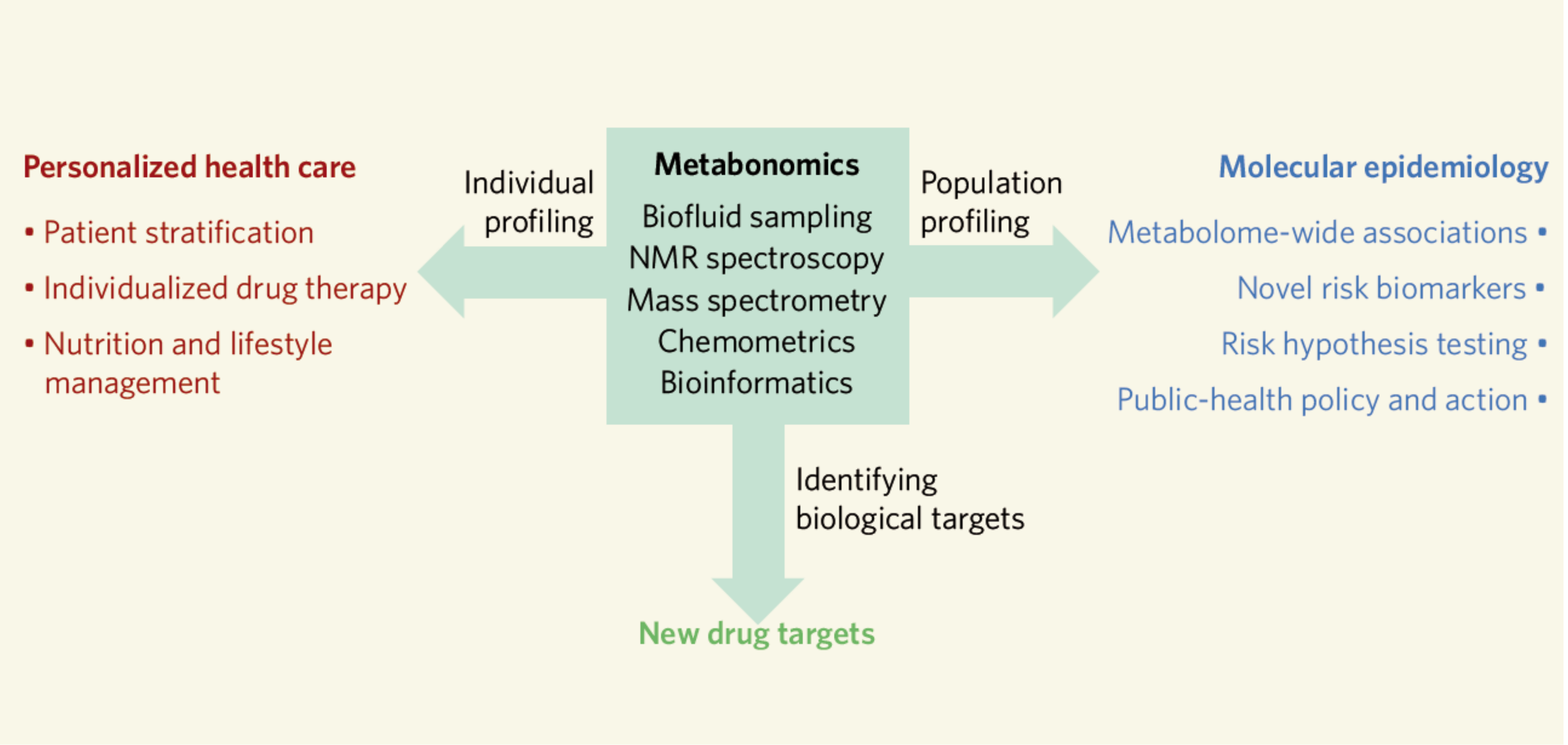
- Gene annotation
- Therapeutic markers
- Drug target identification

protein expression

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
  - Alcohols
  - Lipids
  - Sugars
  - …
- Chemically much more diverse than proteome!

Metabolomics – The Big Picture



Personalized health care

- Patient stratification
- Individualized drug therapy
- Nutrition and lifestyle management

Individual profiling

Metabonomics

Biofluid sampling
NMR spectroscopy
Mass spectrometry
Chemometrics
Bioinformatics

Population profiling

Molecular epidemiology

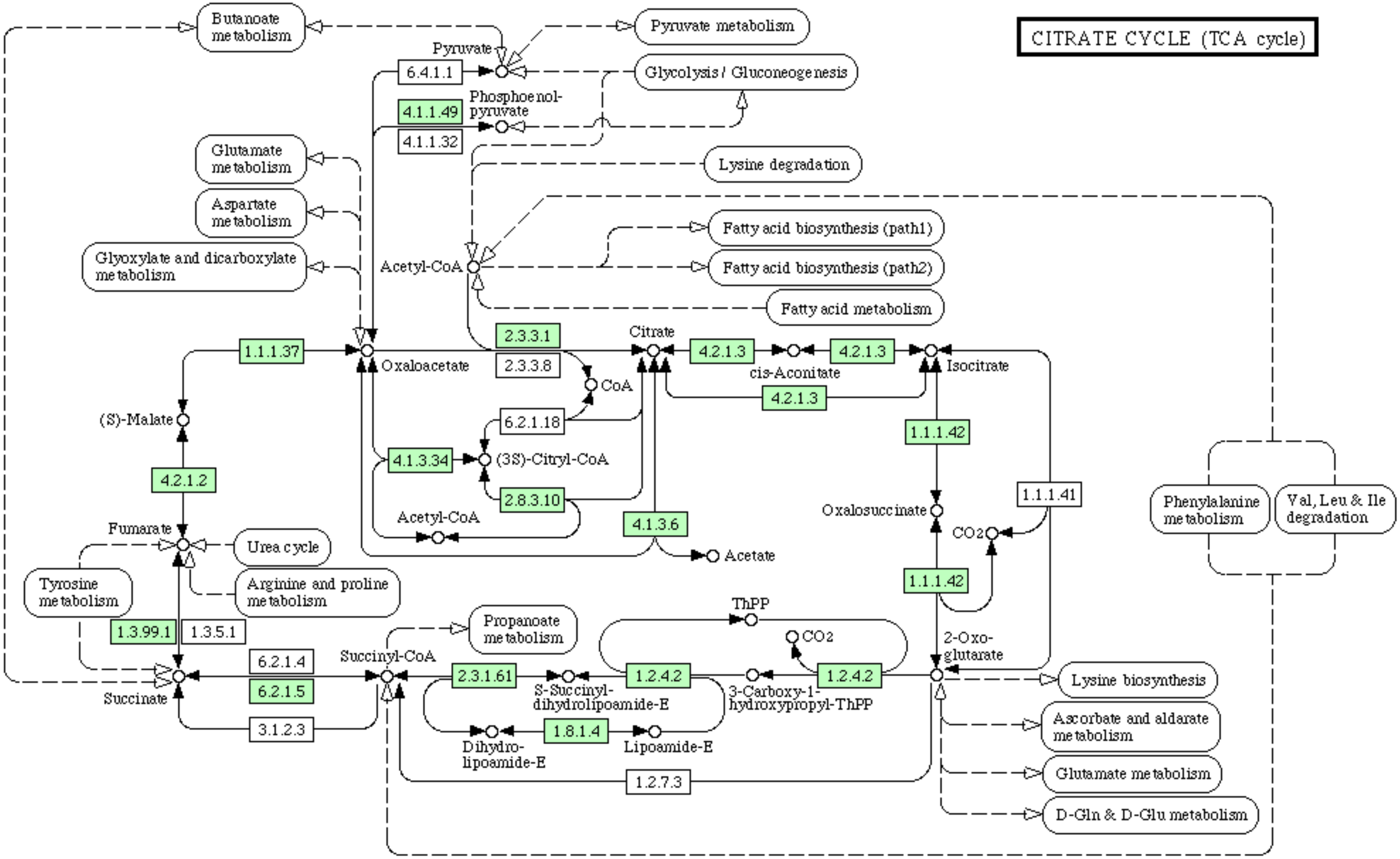
- Metabolome-wide associations
- Novel risk biomarkers
- Risk hypothesis testing
- Public-health policy and action

Identifying biological targets

New drug targets

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 are based on

Chromatography coupled to Mass spectrometry (MS)

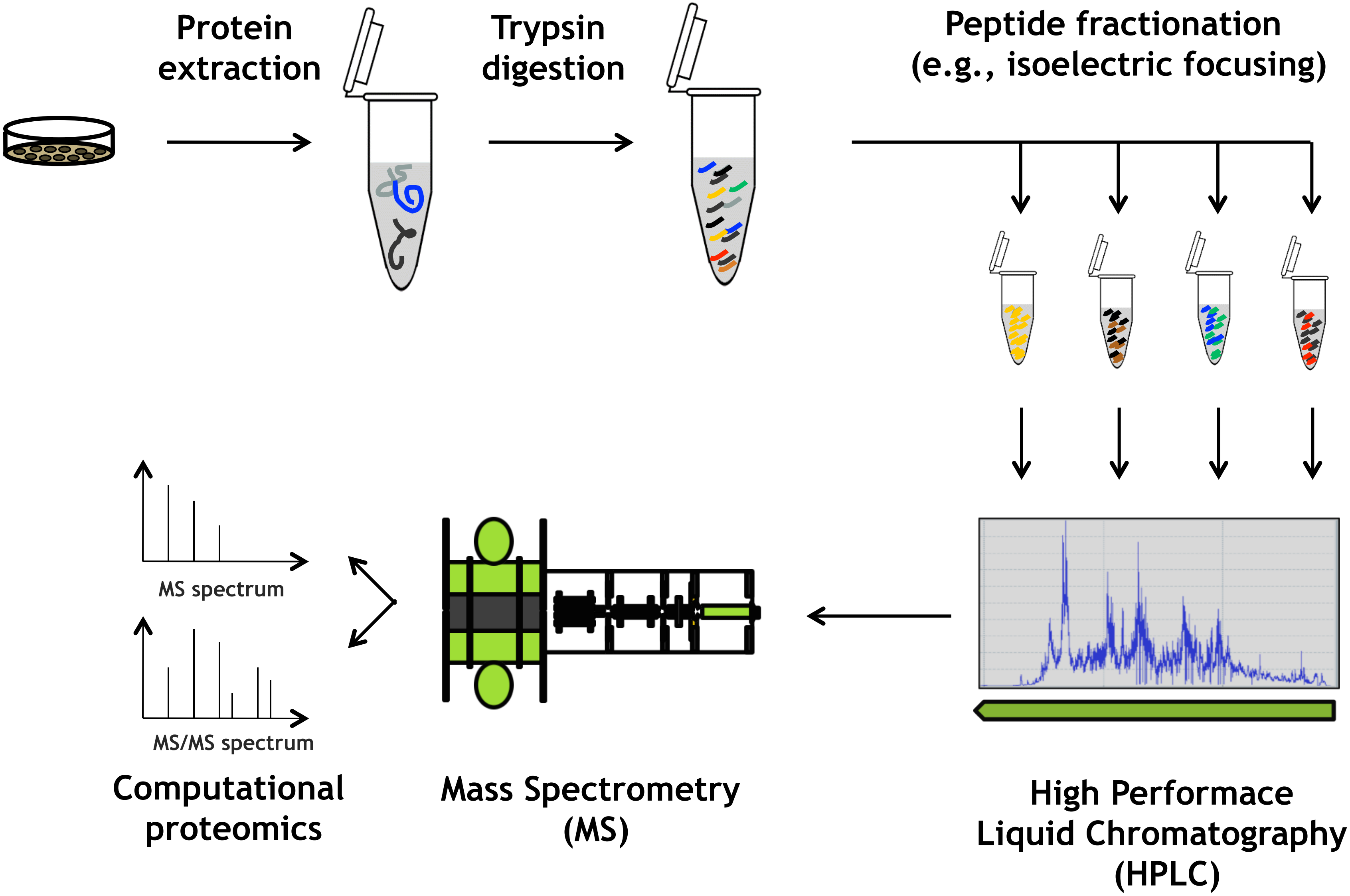
http://en.wikipedia.org/wiki/High-performance_liquid_chromatography.
Access 14/10/2013, 5 PM

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



Protein extraction

Trypsin digestion

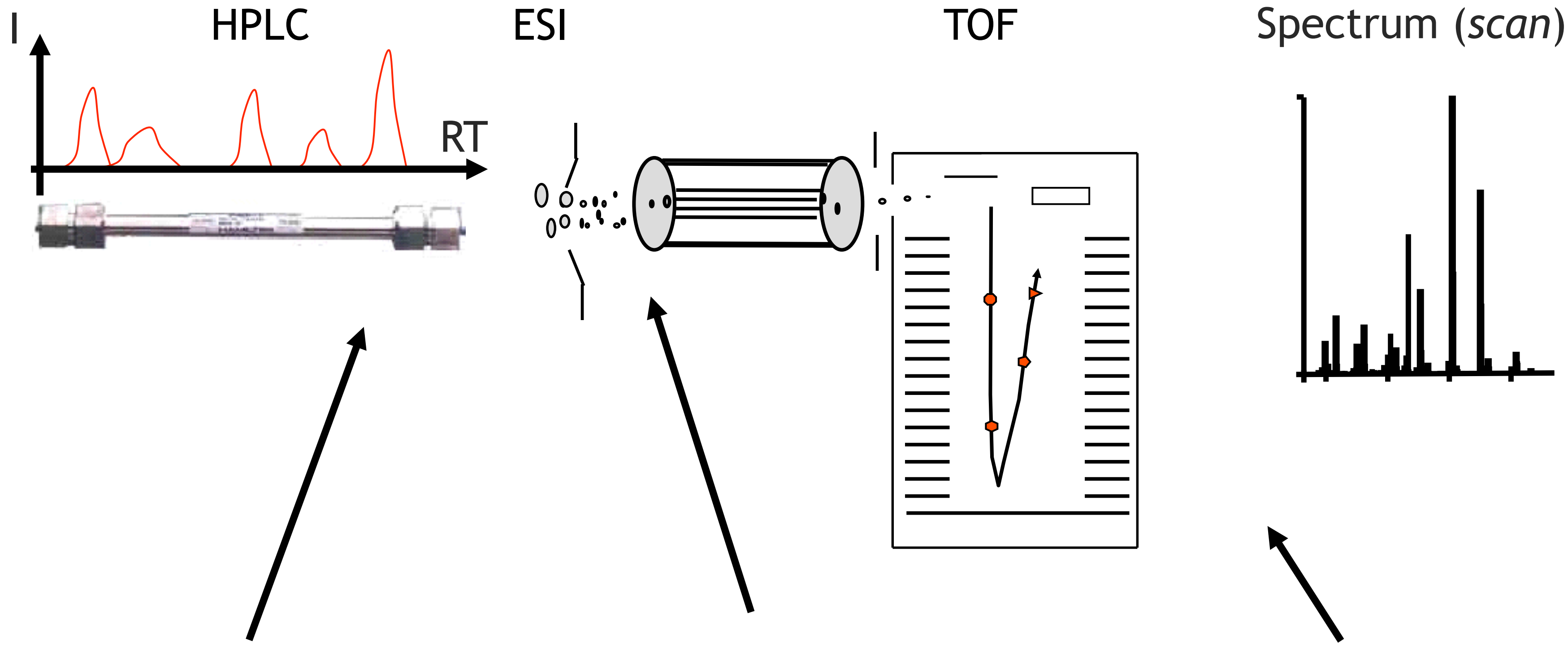
Peptide fractionation (e.g., isoelectric focusing)

Computational proteomics

Mass Spectrometry (MS)

High Performace Liquid Chromatography (HPLC)

At its core: HPLC-MS

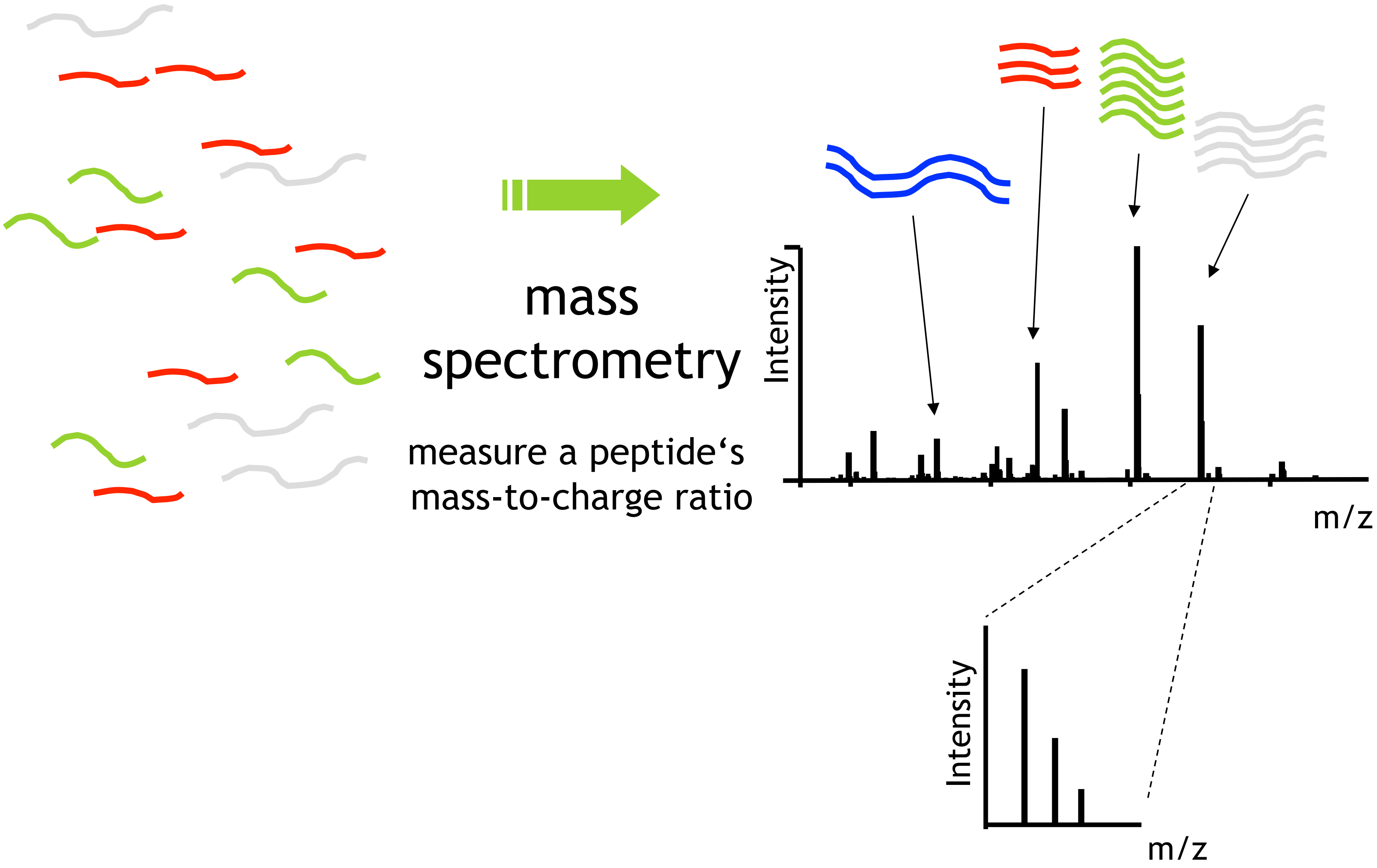


Separation 1
separate peptides by their retention time on column

Ionization
electrospray, transfers charge to the peptides

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

Mass Spectrometry

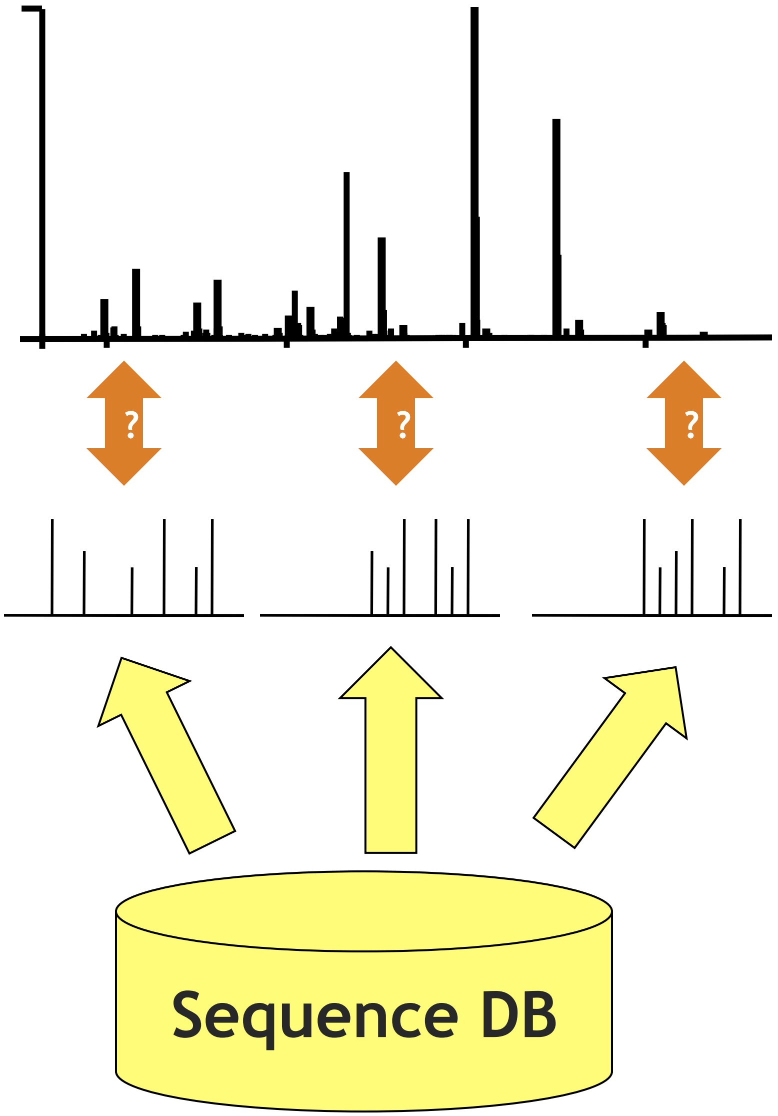


mass spectrometry

measure a peptide‘s mass-to-charge ratio

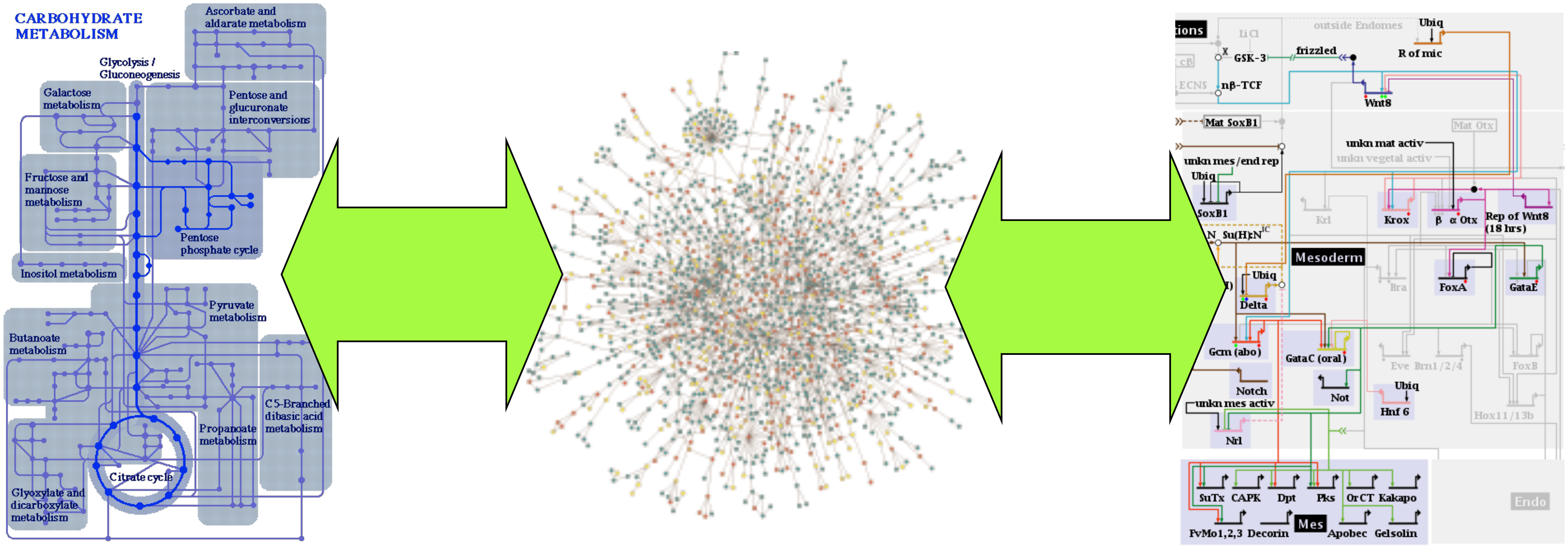
Proteomics: Database Search

- Identification of mass spectra is easily done through database search
  - Search all peptides of matching mass from a database
  - Construct a theoretical mass spectrum for these peptide candidates
  - Score against the experimental spectrum
- **Post-genomics**: database search is possible because we have a genome sequence



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!



Growth of Omics Data (EBI Repositories)

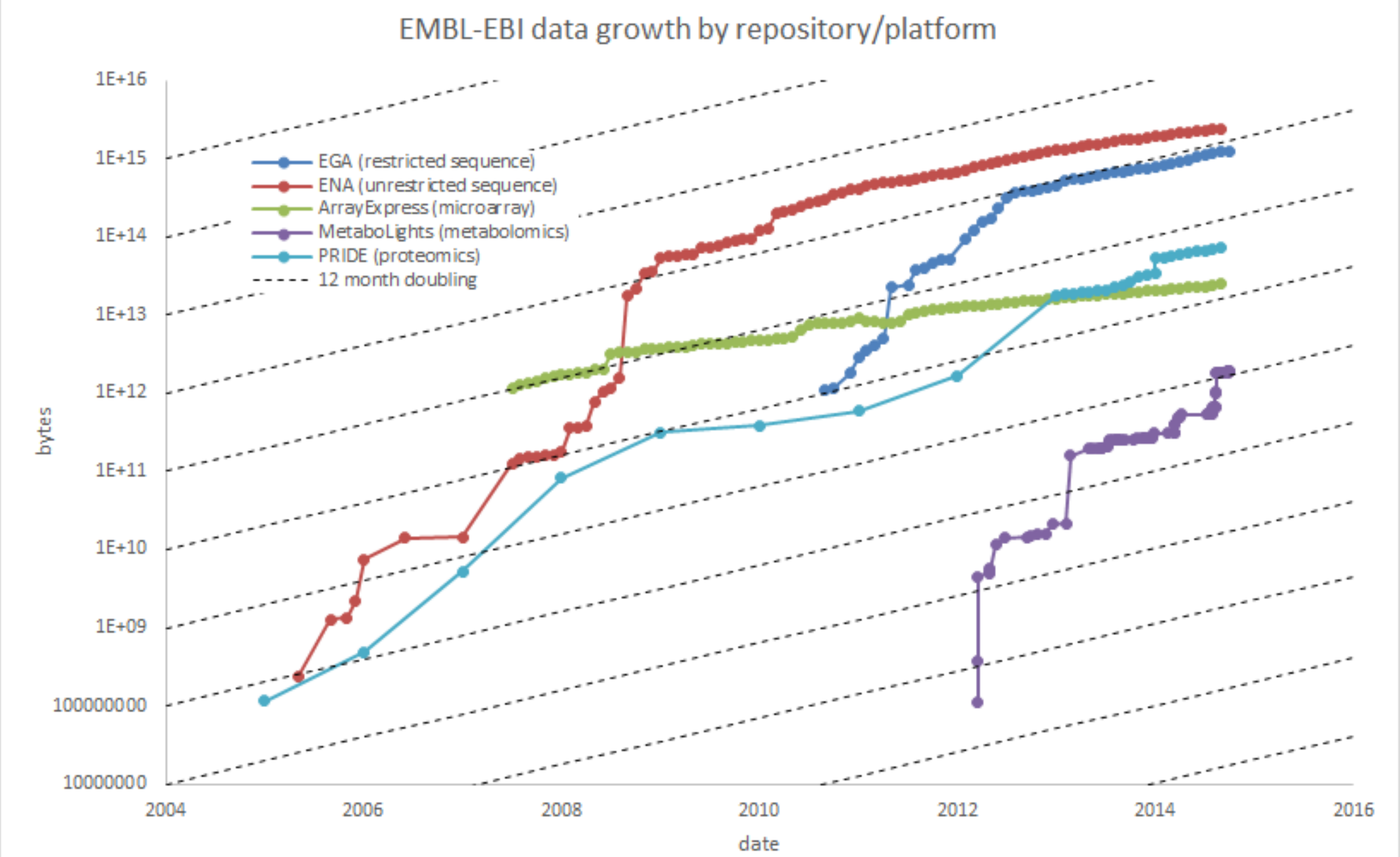


Illustration: Christoph Steinbeck, EBI

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

COMMUNICATIONS OF THE ACM
CACM.ACM.ORG
12/08 VOL.51 NO.12
Surviving the Data Deluge
Open Information Extraction from the Web
CTOs on Virtualization
Living Machines
High-Performance Web Sites
The Economist
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
SCIENCE IN THE PETABYTE ERA

ieee spectrum
INSIDE TECHNOLOGY
MAGAZINE MULTIM
AEROSPACE BIOMEDICAL COMPUTING CONSUMER ELECTRONICS ENERGY

AT WORK / INNOVATION

OPINION

The Coming Data Deluge

As science becomes more data intensive, so does our language

By PAUL MCFEDRIES / FEBRUARY 2011

Email Print Share

WIRED
SUBSCRIBE » SECTIONS » BLOGS » REVIEWS » VIDEO »
Sign In | RSS

WIRED MAGAZINE: 16.07

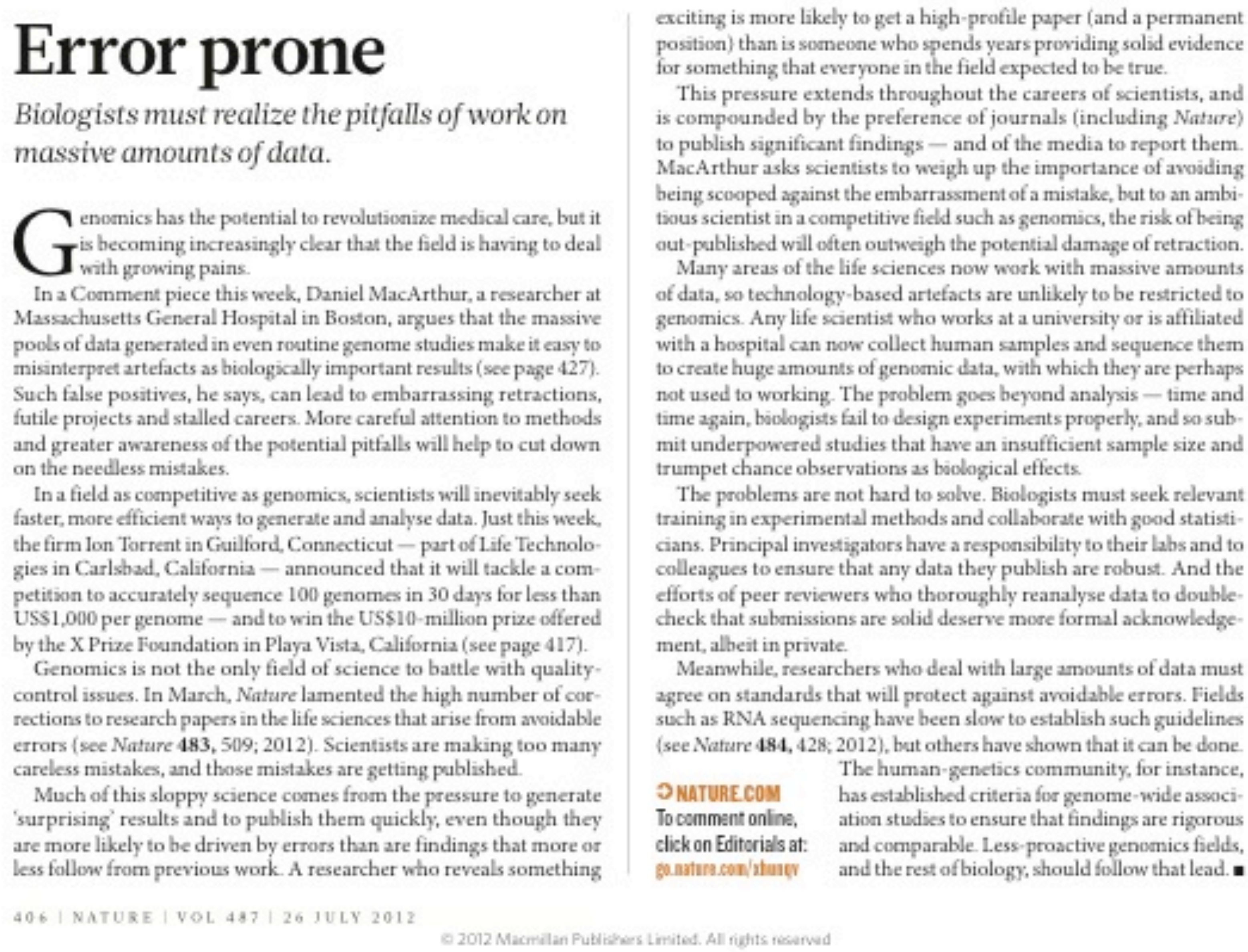
SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

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/xbunqv